Lecture Notes in Computer Science 3039

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA
Josef Kittler
University of Surrey, Guildford, UK
Jon M. Kleinberg
Cornell University, Ithaca, NY, USA
Friedemann Mattern
ETH Zurich, Switzerland
John C. Mitchell
Stanford University, CA, USA
Moni Naor
Weizmann Institute of Science, Rehovot, Israel
Oscar Nierstrasz
University of Bern, Switzerland
C. Pandu Rangan
Indian Institute of Technology, Madras, India
Bernhard Steffen
University of Dortmund, Germany
Madhu Sudan
Massachusetts Institute of Technology, MA, USA
Demetri Terzopoulos
New York University, NY, USA
Doug Tygar
University of California, Berkeley, CA, USA
Moshe Y. Vardi
Rice University, Houston, TX, USA
Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Marian Bubak Geert Dick van Albada
Peter M.A. Sloot Jack J. Dongarra (Eds.)

Computational Science - ICCS 2004

4th International Conference
Kraków, Poland, June 6-9, 2004
Proceedings, Part IV

Springer

Volume Editors

Marian Bubak
AGH University of Science and Technology
Institute of Computer Science and Academic Computer Center CYFRONET
Mickiewicza 30, 30-059 Kraków, Poland
E-mail: bubak@uci.agh.edu.pl

Geert Dick van Albada
Peter M.A. Sloot
University of Amsterdam, Informatics Institute, Section Computational Science
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
E-mail: {dick,sloot}@science.uva.nl

Jack J. Dongarra
University of Tennessee, Computer Science Departement
Suite 413, 1122 Volunteer Blvd., Knoxville, TN-37996, USA
E-mail: dongarra@cs.utk.edu

Library of Congress Control Number: Applied for

CR Subject Classification (1998): D, F, G, H, I, J, C.2-3

ISSN 0302-9743
ISBN 3-540-22129-8 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable to prosecution under the German Copyright Law.

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004
Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 11009597 06/3142 5 4 3 2 1 0

Preface

The International Conference on Computational Science (ICCS 2004) held in Kraków, Poland, June 6–9, 2004, was a follow-up to the highly successful ICCS 2003 held at two locations, in Melbourne, Australia and St. Petersburg, Russia; ICCS 2002 in Amsterdam, The Netherlands; and ICCS 2001 in San Francisco, USA.

As computational science is still evolving in its quest for subjects of investigation and efficient methods, ICCS 2004 was devised as a forum for scientists from mathematics and computer science, as the basic computing disciplines and application areas, interested in advanced computational methods for physics, chemistry, life sciences, engineering, arts and humanities, as well as computer system vendors and software developers. The main objective of this conference was to discuss problems and solutions in all areas, to identify new issues, to shape future directions of research, and to help users apply various advanced computational techniques. The event harvested recent developments in computational grids and next generation computing systems, tools, advanced numerical methods, data-driven systems, and novel application fields, such as complex systems, finance, econo-physics and population evolution.

Keynote lectures were delivered by David Abramson and Alexander V. Bogdanov, *From ICCS 2003 to ICCS 2004 – Personal Overview of Recent Advances in Computational Science*; Iain Duff, *Combining Direct and Iterative Methods for the Solution of Large Sparse Systems in Different Application Areas*; Chris Johnson, *Computational Multi-field Visualization*; John G. Michopoulos, *On the Pathology of High Performance Computing*; David De Roure, *Semantic Grid*; and Vaidy Sunderam, *True Grid: What Makes a Grid Special and Different?* In addition, three invited lectures were delivered by representatives of leading computer system vendors, namely: Frank Baetke from Hewlett Packard, Eng Lim Goh from SGI, and David Harper from the Intel Corporation.

Four tutorials extended the program of the conference: Paweł Płaszczak and Krzysztof Wilk, *Practical Introduction to Grid and Grid Services*; Grzegorz Młynarczyk, *Software Engineering Methods for Computational Science*; the *CrossGrid Tutorial* by the CYFRONET CG team; and the Intel tutorial.

We would like to thank all keynote, invited and tutorial speakers for their interesting and inspiring talks.

Aside of plenary lectures, the conference included 12 parallel oral sessions and 3 poster sessions. Ever since the first meeting in San Francisco, ICCS has attracted an increasing number of more researchers involved in the challenging field of computational science. For ICCS 2004, we received 489 contributions for the main track and 534 contributions for 41 originally-proposed workshops. Of these submissions, 117 were accepted for oral presentations and 117 for posters in the main track, while 328 papers were accepted for presentations at 30 workshops. This selection was possible thanks to the hard work of the Program

Committee members and 477 reviewers. The author index contains 1395 names, and almost 560 persons from 44 countries and all continents attended the conference: 337 participants from Europe, 129 from Asia, 62 from North America, 13 from South America, 11 from Australia, and 2 from Africa.

The ICCS 2004 proceedings consists of four volumes, the first two volumes, LNCS 3036 and 3037 contain the contributions presented in the main track, while volumes 3038 and 3039 contain the papers accepted for the workshops. Parts I and III are mostly related to pure computer science, while Parts II and IV are related to various computational research areas. For the first time, the ICCS proceedings are also available on CD. We would like to thank Springer-Verlag for their fruitful collaboration. During the conference the best papers from the main track and workshops as well as the best posters were nominated and presented on the ICCS 2004 Website. We hope that the ICCS 2004 proceedings will serve as a major intellectual resource for computational science researchers, pushing back the boundaries of this field. A number of papers will also be published as special issues of selected journals.

We owe thanks to all workshop organizers and members of the Program Committee for their diligent work, which ensured the very high quality of the event. We also wish to specifically acknowledge the collaboration of the following colleagues who organized their workshops for the third time: Nicoletta Del Buono (New Numerical Methods) Andres Iglesias (Computer Graphics), Dieter Kranzlmueller (Tools for Program Development and Analysis), Youngsong Mun (Modeling and Simulation in Supercomputing and Telecommunications).

We would like to express our gratitude to Prof. Ryszard Tadeusiewicz, Rector of the AGH University of Science and Technology, as well as to Prof. Marian Noga, Prof. Kazimierz Jeleń, Dr. Jan Kulka and Prof. Krzysztof Zieliński, for their personal involvement. We are indebted to all the members of the Local Organizing Committee for their enthusiastic work towards the success of ICCS 2004, and to numerous colleagues from ACC CYFRONET AGH and the Institute of Computer Science for their help in editing the proceedings and organizing the event. We very much appreciate the help of the Computer Science and Computational Physics students during the conference. We owe thanks to the ICCS 2004 sponsors: Hewlett-Packard, Intel, IBM, SGI and ATM, SUN Microsystems, Polish Airlines LOT, ACC CYFRONET AGH, the Institute of Computer Science AGH, the Polish Ministry for Scientific Research and Information Technology, and Springer-Verlag for their generous support.

We wholeheartedly invite you to once again visit the ICCS 2004 Website (http://www.cyfronet.krakow.pl/iccs2004/), to recall the atmosphere of those June days in Kraków.

June 2004

Marian Bubak, Scientific Chair 2004
on behalf of the co-editors:
G. Dick van Albada
Peter M.A. Sloot
Jack J. Dongarra

Organization

ICCS 2004 was organized by the Academic Computer Centre CYFRONET AGH University of Science and Technology (Kraków, Poland) in cooperation with the Institute of Computer Science AGH, the University of Amsterdam (The Netherlands) and the University of Tennessee (USA).

All the members of the Local Organizing Committee are the staff members of CYFRONET and/or ICS. The conference took place at the premises of the Faculty of Physics and Nuclear Techniques AGH and at the Institute of Computer Science AGH.

Conference Chairs

Scientific Chair – Marian Bubak (Institute of Computer Science and ACC CYFRONET AGH, Poland)
Workshop Chair – Dick van Albada (University of Amsterdam, The Netherlands)
Overall Chair – Peter M.A. Sloot (University of Amsterdam, The Netherlands)
Overall Co-chair – Jack Dongarra (University of Tennessee, USA)

Local Organizing Committee

Marian Noga
Marian Bubak
Zofia Mosurska
Maria Stawiarska
Milena Zając
Mietek Pilipczuk
Karol Frańczak
Aleksander Kusznir

Program Committee

Jemal Abawajy (Carleton University, Canada)
David Abramson (Monash University, Australia)
Dick van Albada (University of Amsterdam, The Netherlands)
Vassil Alexandrov (University of Reading, UK)
Srinivas Aluru (Iowa State University, USA)
David A. Bader (University of New Mexico, USA)

J.A. Rod Blais (University of Calgary, Canada)
Alexander Bogdanov (Institute for High Performance Computing and Information Systems, Russia)
Peter Brezany (University of Vienna, Austria)
Marian Bubak (Institute of Computer Science and CYFRONET AGH, Poland)
Rajkumar Buyya (University of Melbourne, Australia)
Bastien Chopard (University of Geneva, Switzerland)
Paul Coddington (University of Adelaide, Australia)
Toni Cortes (Universitat Politècnica de Catalunya, Spain)
Yiannis Cotronis (University of Athens, Greece)
Jose C. Cunha (New University of Lisbon, Portugal)
Brian D'Auriol (University of Texas at El Paso, USA)
Federic Desprez (INRIA, France)
Tom Dhaene (University of Antwerp, Belgium)
Hassan Diab (American University of Beirut, Lebanon)
Beniamino Di Martino (Second University of Naples, Italy)
Jack Dongarra (University of Tennessee, USA)
Robert A. Evarestov (SPbSU, Russia)
Marina Gavrilova (University of Calgary, Canada)
Michael Gerndt (Technical University of Munich, Germany)
Yuriy Gorbachev (Institute for High Performance Computing and Information Systems, Russia)
Andrzej Goscinski (Deakin University, Australia)
Ladislav Hluchy (Slovak Academy of Sciences, Slovakia)
Alfons Hoekstra (University of Amsterdam, The Netherlands)
Hai Jin (Huazhong University of Science and Technology, ROC)
Peter Kacsuk (MTA SZTAKI Research Institute, Hungary)
Jacek Kitowski (AGH University of Science and Technology, Poland)
Dieter Kranzlmüller (Johannes Kepler University Linz, Austria)
Domenico Laforenza (Italian National Research Council, Italy)
Antonio Lagana (Università di Perugia, Italy)
Francis Lau (University of Hong Kong, ROC)
Bogdan Lesyng (ICM Warszawa, Poland)
Thomas Ludwig (Ruprecht-Karls-Universität Heidelberg, Germany)
Emilio Luque (Universitat Autònoma de Barcelona, Spain)
Michael Mascagni (Florida State University, USA)
Edward Moreno (Euripides Foundation of Marilia, Brazil)
Jiri Nedoma (Institute of Computer Science AS CR, Czech Republic)
Genri Norman (Russian Academy of Sciences, Russia)
Stephan Olariu (Old Dominion University, USA)
Salvatore Orlando (University of Venice, Italy)
Marcin Paprzycki (Oklahoma State University, USA)
Ron Perrott (Queen's University of Belfast, UK)
Richard Ramaroson (ONERA, France)
Rosemary Renaut (Arizona State University, USA)

Alistair Rendell (Australian National University, Australia)
Paul Roe (Queensland University of Technology, Australia)
Hong Shen (Japan Advanced Institute of Science and Technology, Japan)
Dale Shires (U.S. Army Research Laboratory, USA)
Peter M.A. Sloot (University of Amsterdam, The Netherlands)
Gunther Stuer (University of Antwerp, Belgium)
Vaidy Sunderam (Emory University, USA)
Boleslaw Szymanski (Rensselaer Polytechnic Institute, USA)
Ryszard Tadeusiewicz (AGH University of Science and Technology, Poland)
Pavel Tvrdik (Czech Technical University, Czech Republic)
Putchong Uthayopas (Kasetsart University, Thailand)
Jesus Vigo-Aguiar (University of Salamanca, Spain)
Jens Volkert (University of Linz, Austria)
Koichi Wada (University of Tsukuba, Japan)
Jerzy Wasniewski (Technical University of Denmark, Denmark)
Greg Watson (Los Alamos National Laboratory, USA)
Jan Węglarz (Poznań University of Technology, Poland)
Roland Wismüller (LRR-TUM, Germany)
Roman Wyrzykowski (Technical University of Częstochowa, Poland)
Jinchao Xu (Pennsylvania State University, USA)
Yong Xue (Chinese Academy of Sciences, ROC)
Xiaodong Zhang (College of William and Mary, USA)
Alexander Zhmakin (Soft-Impact Ltd, Russia)
Krzysztof Zieliński (Institute of Computer Science and CYFRONET AGH, Poland)
Zahari Zlatev (National Environmental Research Institute, Denmark)
Albert Zomaya (University of Sydney, Australia)
Elena Zudilova (University of Amsterdam, The Netherlands)

Reviewers

Abawajy, J.H.
Abe, S.
Abramson, D.
Adali, S.
Adcock, M.
Adriaansen, T.
Ahn, G.
Ahn, S.J.
Albada, G.D. van
Albuquerque, P.
Alda, W.
Alexandrov, V.
Alt, M.
Aluru, S.
Anglano, C.
Archibald, R.
Arenas, A.
Astalos, J.
Ayani, R.
Ayyub, S.
Babik, M.
Bader, D.A.
Bajaj, C.
Baker, M.
Baliś, B.
Balk, I.
Balogh, Z.
Bang, Y.C.
Baraglia, R.
Barron, J.
Baumgartner, F.
Becakaert, P.
Belleman, R.G.
Bentes, C.
Bernardo Filho, O.
Beyls, K.
Blais, J.A.R.
Boada, I.
Bode, A.

Bogdanov, A.
Bollapragada, R.
Boukhanovsky, A.
Brandes, T.
Brezany, P.
Britanak, V.
Bronsvoort, W.
Brunst, H.
Bubak, M.
Budinska, I.
Buono, N. Del
Buyya, R.
Cai, W.
Cai, Y.
Cannataro, M.
Carbonell, N.
Carle, G.
Caron, E.
Carothers, C.
Castiello, C.
Chan, P.
Chassin-de-Kergommeaux, J.
Chaudet, C.
Chaves, J.C.
Chen, L.
Chen, Z.
Cheng, B.
Cheng, X.
Cheung, B.W.L.
Chin, S.
Cho, H.
Choi, Y.S.
Choo, H.S.
Chopard, B.
Chuang, J.H.
Chung, R.
Chung, S.T.
Coddington, P.
Coeurjolly, D.
Congiusta, A.
Coppola, M.
Corral, A.
Cortes, T.
Cotronis, Y.
Cramer, H.S.M.
Cunha, J.C.
Danilowicz, C.
D'Auriol, B.
Degtyarev, A.
Denazis, S.
Derntl, M.
Desprez, F.
Devendeville, L.
Dew, R.
Dhaene, T.
Dhoedt, B.
D'Hollander, E.
Diab, H.
Dokken, T.
Dongarra, J.
Donnelly, D.
Donnelly, W.
Dorogovtsev, S.
Duda, J.
Dudek-Dyduch, E.
Dufourd, J.F.
Dumitriu, L.
Duplaga, M.
Dupuis, A.
Dzwinel, W.
Embrechts, M.J.
Emiris, I.
Emrich, S.J.
Enticott, C.
Evangelos, F.
Evarestov, R.A.
Fagni, T.
Faik, J.
Fang, W.J.
Farin, G.
Fernandez, M.
Filho, B.O.
Fisher-Gewirtzman, D.
Floros, E.
Fogel, J.
Foukia, N.
Frankovic, B.
Fuehrlinger, K.
Funika, W.
Gabriel, E.
Gagliardi, F.
Galis, A.
Galvez, A.
Gao, X.S.
Garstecki, L.
Gatial, E.
Gava, F.
Gavidia, D.P.
Gavras, A.
Gavrilova, M.
Gelb, A.
Gerasimov, V.
Gerndt, M.
Getov, V.
Geusebroek, J.M.
Giang, T.
Gilbert, M.
Glasner, C.
Gobbert, M.K.
Gonzalez-Vega, L.
Gorbachev, Y.E.
Goscinski, A.M.
Goscinski, W.
Gourhant, Y.
Gualandris, A.
Guo, H.
Ha, R.
Habala, O.
Habib, A.
Halada, L.
Hawick, K.
He, K.
Heinzlreiter, P.
Heyfitch, V.
Hisley, D.M.
Hluchy, L.
Ho, R.S.C.
Ho, T.
Hobbs, M.
Hoekstra, A.
Hoffmann, C.
Holena, M.
Hong, C.S.
Hong, I.

Hong, S.
Horan, P.
Hu, S.M.
Huh, E.N.
Hutchins, M.
Huynh, J.
Hwang, I.S.
Hwang, J.
Iacono, M.
Iglesias, A.
Ingram, D.
Jakulin, A.
Janciak, I.
Janecek, J.
Janglova, D.
Janicki, A.
Jin, H.
Jost, G.
Juhola, A.
Kacsuk, P.
Kalousis, A.
Kalyanaraman, A.
Kang, M.G.
Karagiorgos, G.
Karaivanova, A.
Karl, W.
Karypis, G.
Katarzyniak, R.
Kelley, T.
Kelly, W.
Kennedy, E.
Kereku, E.
Kergommeaux, J.C. De
Kim, B.
Kim, C.H.
Kim, D.S.
Kim, D.Y.
Kim, M.
Kim, M.J.
Kim, T.W.
Kitowski, J.
Klein, C.
Ko, P.
Kokoszka, P.
Kolingerova, I.
Kommineni, J.
Korczak, J.J.
Korkhov, V.
Kou, G.
Kouniakis, C.
Kranzlmüller, D.
Krzhizhianovskaya, V.V.
Kuo, T.W.
Kurka, G.
Kurniawan, D.
Kurzyniec, D.
Laclavik, M.
Laforenza, D.
Lagan, A.
Lagana, A.
Lamehamedi, H.
Larrabeiti, D.
Latt, J.
Lau, F.
Lee, H.G.
Lee, M.
Lee, S.
Lee, S.S.
Lee, S.Y.
Lefevre, L.
Leone, P.
Lesyng, B.
Leszczynski, J.
Leymann, F.
Li, T.
Lindner, P.
Logan, B.
Lopes, G.P.
Lorencz, R.
Low, M.Y.H.
Ludwig, T.
Luethi, J.
Lukac, R.
Luksch, P.
Luque, E.
Mairandres, M.
Malawski, M.
Malony, A.
Malyshkin, V.E.
Maniatty, W.A.
Marconi, S.
Mareev, V.
Margalef, T.
Marrone, S.
Martino, B. Di
Marzolla, M.
Mascagni, M.
Mayer, M.
Medeiros, P.
Meer, H. De
Meyer, N.
Miller, B.
Miyaji, C.
Modave, F.
Mohr, B.
Monterde, J.
Moore, S.
Moreno, E.
Moscato, F.
Mourelle, L.M.
Mueller, M.S.
Mun, Y.
Na, W.S.
Nagel, W.E.
Nanni, M.
Narayanan, M.
Nasri, A.
Nau, B.
Nedjah, N.
Nedoma, J.
Negoita, C.
Neumann, L.
Nguyen, G.T.
Nguyen, N.T.
Norman, G.
Olariu, S.
Orlando, S.
Orley, S.
Otero, C.
Owen, J.
Palus, H.
Paprzycki, M.
Park, N.J.
Patten, C.
Peachey, T.C.

Peluso, R.
Peng, Y.
Perales, F.
Perrott, R.
Petit, F.
Petit, G.H.
Pfluger, P.
Philippe, L.
Platen, E.
Plemenos, D.
Pllana, S.
Polak, M.
Polak, N.
Politi, T.
Pooley, D.
Popov, E.V.
Puppin, D.
Qut, P.R.
Rachev, S.
Rajko, S.
Rak, M.
Ramaroson, R.
Ras, I.
Rathmayer, S.
Raz, D.
Recio, T.
Reichel, L.
Renaut, R.
Rendell, A.
Richta, K.
Robert, Y.
Rodgers, G.
Rodionov, A.S.
Roe, P.
Ronsse, M.
Ruder, K.S.
Ruede, U.
Rycerz, K.
Sanchez-Reyes, J.
Sarfraz, M.
Sbert, M.
Scarpa, M.
Schabanel, N.
Scharf, E.
Scharinger, J.
Schaubschlaeger, C.
Schmidt, A.
Scholz, S.B.
Schreiber, A.
Seal, S.K.
Seinstra, F.J.
Seron, F.
Serrat, J.
Shamonin, D.P.
Sheldon, F.
Shen, H.
Shende, S.
Shentu, Z.
Shi, Y.
Shin, H.Y.
Shires, D.
Shoshmina, I.
Shrikhande, N.
Silvestri, C.
Silvestri, F.
Simeoni, M.
Simo, B.
Simonov, N.
Siu, P.
Slizik, P.
Slominski, L.
Sloot, P.M.A.
Slota, R.
Smetek, M.
Smith, G.
Smolka, B.
Sneeuw, N.
Snoek, C.
Sobaniec, C.
Sobecki, J.
Sofroniou, M.
Sole, R.
Soofi, M.
Sosnov, A.
Sourin, A.
Spaletta, G.
Spiegl, E.
Stapor, K.
Stuer, G.
Suarez Rivero, J.P.
Sunderam, V.
Suzuki, H.
Szatzschneider, W.
Szczepanski, M.
Szirmay-Kalos, L.
Szymanski, B.
Tadeusiewicz, R.
Tadic, B.
Talia, D.
Tan, G.
Taylor, S.J.E.
Teixeira, J.C.
Telelis, O.A.
Teo, Y.M
Teresco, J.
Teyssiere, G.
Thalmann, D.
Theodoropoulos, G.
Theoharis, T.
Thurner, S.
Tirado-Ramos, A.
Tisserand, A.
Toda, K.
Tonellotto, N.
Torelli, L.
Torenvliet, L.
Tran, V.D.
Truong, H.L.
Tsang, K.
Tse, K.L.
Tvrdik, P.
Tzevelekas, L.
Uthayopas, P.
Valencia, P.
Vassilakis, C.
Vaughan, F.
Vazquez, P.P.
Venticinque, S.
Vigo-Aguiar, J.
Vivien, F.
Volkert, J.
Wada, K.
Walter, M.
Wasniewski, J.
Wasserbauer, A.

Watson, G.
Wawrzyniak, D.
Weglarz, J.
Weidendorfer, J.
Weispfenning, W.
Wendelborn, A.L.
Weron, R.
Wismüller, R.
Wojciechowski, K.
Wolf, F.
Worring, M.
Wyrzykowski, R.
Xiao, Y.
Xu, J.
Xue, Y.
Yahyapour, R.
Yan, N.
Yang, K.
Yener, B.
Yoo, S.M.
Yu, J.H.
Yu, Z.C.H.
Zara, J.
Zatevakhin, M.A.
Zhang, J.W.
Zhang, N.X.L.
Zhang, X.
Zhao, L.
Zhmakin, A.I.
Zhu, W.Z.
Zieliński, K.
Zlatev, Z.
Zomaya, A.
Zudilova, E.V.

Workshops Organizers

Programming Grids and Metasystems

V. Sunderam (Emory University, USA)
D. Kurzyniec (Emory University, USA)
V. Getov (University of Westminster, UK)
M. Malawski (Institute of Computer Science and CYFRONET AGH, Poland)

Active and Programmable Grids Architectures and Components

C. Anglano (Università del Piemonte Orientale, Italy)
F. Baumgartner (University of Bern, Switzerland)
G. Carle (Tubingen University, Germany)
X. Cheng (Institute of Computing Technology, Chinese Academy of Science, ROC)
K. Chen (Institut Galilée, Université Paris 13, France)
S. Denazis (Hitachi Europe, France)
B. Dhoedt (University of Gent, Belgium)
W. Donnelly (Waterford Institute of Technology, Ireland)
A. Galis (University College London, UK)
A. Gavras (Eurescom, Germany)
F. Gagliardi (CERN, Switzerland)
Y. Gourhant (France Telecom, France)
M. Gilbert (European Microsoft Innovation Center, Microsoft Corporation, Germany)
A. Juhola (VTT, Finland)
C. Klein (Siemens, Germany)
D. Larrabeiti (University Carlos III, Spain)
L. Lefevre (INRIA, France)
F. Leymann (IBM, Germany)
H. de Meer (University of Passau, Germany)
G. H. Petit (Alcatel, Belgium)

J. Serrat (Universitat Politècnica de Catalunya, Spain)
E. Scharf (QMUL, UK)
K. Skala (Ruder Boskoviç Institute, Croatia)
N. Shrikhande (European Microsoft Innovation Center, Microsoft Corporation, Germany)
M. Solarski (FhG FOKUS, Germany)
D. Raz (Technion Institute of Technology, Israel)
K. Zieliński (AGH University of Science and Technology, Poland)
R. Yahyapour (University Dortmund, Germany)
K. Yang (University of Essex, UK)

Next Generation Computing

E.-N. John Huh (Seoul Women's University, Korea)

Practical Aspects of High-Level Parallel Programming (PAPP 2004)

F. Loulergue (Laboratory of Algorithms, Complexity and Logic, University of Paris Val de Marne, France)

Parallel Input/Output Management Techniques (PIOMT 2004)

J. H. Abawajy (Carleton University, School of Computer Science, Canada)

OpenMP for Large Scale Applications

B. Chapman (University of Houston, USA)

Tools for Program Development and Analysis in Computational Science

D. Kranzlmüller (Johannes Kepler University Linz, Austria)
R. Wismüller (TU München, Germany)
A. Bode (Technische Universität München, Germany)
J. Volkert (Johannes Kepler University Linz, Austria)

Modern Technologies for Web-Based Adaptive Systems

N. Thanh Nguyen (Wrocław University of Technology, Poland)
J. Sobecki (Wrocław University of Technology, Poland)

Agent Day 2004 – Intelligent Agents in Computing Systems

E. Nawarecki (AGH University of Science and Technology, Poland)
K. Cetnarowicz (AGH University of Science and Technology, Poland)
G. Dobrowolski (AGH University of Science and Technology, Poland)
R. Schaefer (Jagiellonian University, Poland)
S. Ambroszkiewicz (Polish Academy of Sciences, Warsaw, Poland)
A. Koukam (Université de Belfort-Montbeliard, France)
V. Srovnal (VSB Technical University of Ostrava, Czech Republic)
C. Cotta (Universidad de Málaga, Spain)
S. Raczynski (Universidad Panamericana, Mexico)

Dynamic Data Driven Application Systems

F. Darema (NSF/CISE, USA)

HLA-Based Distributed Simulation on the Grid

S. J. Turner (Nanyang Technological University, Singapore)

Interactive Visualisation and Interaction Technologies

E. Zudilova (University of Amsterdam, The Netherlands)
T. Adriaansen (CSIRO, ICT Centre, Australia)

Computational Modeling of Transport on Networks

B. Tadic (Jozef Stefan Institute, Slovenia)
S. Thurner (Universität Wien, Austria)

Modeling and Simulation in Supercomputing and Telecommunications

Y. Mun (Soongsil University, Korea)

QoS Routing

H. Choo (Sungkyunkwan University, Korea)

Evolvable Hardware

N. Nedjah (State University of Rio de Janeiro, Brazil)
L. de Macedo Mourelle (State University of Rio de Janeiro, Brazil)

Advanced Methods of Digital Image Processing

B. Smolka (Silesian University of Technology, Laboratory of Multimedia Communication, Poland)

Computer Graphics and Geometric Modelling (CGGM 2004)

A. Iglesias Prieto (University of Cantabria, Spain)

Computer Algebra Systems and Applications (CASA 2004)

A. Iglesias Prieto (University of Cantabria, Spain)
A. Galvez (University of Cantabria, Spain)

New Numerical Methods for DEs: Applications to Linear Algebra, Control and Engineering

N. Del Buono (University of Bari, Italy)
L. Lopez (University of Bari, Italy)

Parallel Monte Carlo Algorithms for Diverse Applications in a Distributed Setting

V. N. Alexandrov (University of Reading, UK)
A. Karaivanova (Bulgarian Academy of Sciences, Bulgaria)
I. Dimov (Bulgarian Academy of Sciences, Bulgaria)

Modelling and Simulation of Multi-physics Multi-scale Systems

V. Krzhizhanovskaya (University of Amsterdam, The Netherlands)
B. Chopard (University of Geneva, CUI, Switzerland)
Y. Gorbachev (St. Petersburg State Polytechnical University, Russia)

Gene, Genome and Population Evolution

S. Cebrat (University of Wrocław, Poland)
D. Stauffer (Cologne University, Germany)
A. Maksymowicz (AGH University of Science and Technology, Poland)

Computational Methods in Finance and Insurance

A. Janicki (University of Wrocław, Poland)
J.J. Korczak (University Louis Pasteur, Strasbourg, France)

Computational Economics and Finance

X. Deng (City University of Hong Kong, Hong Kong)
S. Wang (Chinese Academy of Sciences, ROC)
Y. Shi (University of Nebraska at Omaha, USA)

GeoComputation

Y. Xue (Chinese Academy of Sciences, ROC)
C. Yarotsos (University of Athens, Greece)

Simulation and Modeling of 3D Integrated Circuits

I. Balk (R3Logic Inc., USA)

Computational Modeling and Simulation on Biomechanical Engineering

Y.H. Kim (Kyung Hee University, Korea)

Information Technologies Enhancing Health Care Delivery

M. Duplaga (Jagiellonian University Medical College, Poland)
D. Ingram (University College London, UK)
K. Zieliński (AGH University of Science and Technology, Poland)

Computing in Science and Engineering Academic Programs

D. Donnelly (Siena College, USA)

Sponsoring Institutions

Hewlett-Packard
Intel
SGI
ATM
SUN Microsystems
IBM
Polish Airlines LOT
ACC CYFRONET AGH
Institute of Computer Science AGH
Polish Ministry of Scientific Research and Information Technology
Springer-Verlag

Table of Contents – Part IV

Workshop on Advanced Methods of Digital Image Processing

The New Graphic Description of the Haar Wavelet Transform 1
P. Porwik, A. Lisowska

On New Radon-Based Translation, Rotation, and Scaling Invariant Transform for Face Recognition 9
T. Arodź

On Bit-Level Systolic Arrays for Least-Squares Digital Contour Smoothing 18
J. Glasa

Bayer Pattern Demosaicking Using Local-Correlation Approach 26
R. Lukac, K.N. Plataniotis, A.N. Venetsanopoulos

Edge Preserving Filters on Color Images 34
V. Hong, H. Palus, D. Paulus

Segmentation of Fundus Eye Images Using Methods of Mathematical Morphology for Glaucoma Diagnosis 41
K. Stąpor, A. Świtonski, R. Chrastek, G. Michelson

Automatic Detection of Glaucomatous Changes Using Adaptive Thresholding and Neural Networks 49
K. Stąpor, L. Pawlaczyk, R. Chrastek, G. Michelson

Analytical Design of 2-D Narrow Bandstop FIR Filters 56
P. Zahradnik, M. Vlček

Analytical Design of Arbitrary Oriented Asteroidal 2-D FIR Filters 64
P. Zahradnik, M. Vlček

A $\{k, n\}$-Secret Sharing Scheme for Color Images 72
R. Lukac, K.N. Plataniotis, A.N. Venetsanopoulos

Workshop on Computer Graphics and Geometric Modelling (CGGM 2004)

Declarative Modelling in Computer Graphics: Current Results and Future Issues 80
P.-F. Bonnefoi, D. Plemenos, W. Ruchaud

Geometric Snapping for 3D Meshes 90
K.-H. Yoo, J.S. Ha

Multiresolution Approximations of Generalized Voronoi Diagrams 98
I. Boada, N. Coll, J.A. Sellarès

LodStrips: Level of Detail Strips 107
J.F. Ramos, M. Chover

Declarative Specification of Ambiance in VRML Landscapes 115
V. Jolivet, D. Plemenos, P. Poulingeas

Using Constraints in Delaunay and Greedy Triangulation for Contour Lines Improvement 123
I. Kolingerová, V. Strych, V. Čada

An Effective Modeling of Single Cores Prostheses Using Geometric Techniques 131
K.-H. Yoo, J.S. Ha

GA and CHC. Two Evolutionary Algorithms to Solve the Root Identification Problem in Geometric Constraint Solving 139
M.V. Luzón, E. Barreiro, E. Yeguas, R. Joan-Arinyo

Manifold Extraction in Surface Reconstruction 147
M. Varnuška, I. Kolingerová

Expression of a Set of Points' Structure within a Specific Geometrical Model 156
J.-L. Mari, J. Sequeira

Effective Use of Procedural Shaders in Animated Scenes 164
P. Kondratieva, V. Havran, H.-P. Seidel

Real-Time Tree Rendering 173
I. Remolar, C. Rebollo, M. Chover, J. Ribelles

A Brush Stroke Generation Using Magnetic Field Model for Painterly Rendering 181
L.S. Yeon, Y.H. Soon, Y.K. Hyun

Reuse of Paths in Final Gathering Step with Moving Light Sources 189
M. Sbert, F. Castro

Real Time Tree Sketching 197
C. Campos, R. Quirós, J. Huerta, E. Camahort, R. Vivó, J. Lluch

Facial Expression Recognition Based on Dimension Model Using Sparse Coding 205
Y.-s. Shin

An Application to the Treatment of Geophysical Images through Orthogonal Projections 213
S. Romero, F. Moreno

A Derivative-Free Tracking Algorithm for Implicit Curves with Singularities 221
J.F.M. Morgado, A.J.P. Gomes

Framework for Simulating the Human Behavior for Intelligent Virtual Agents. Part I: Framework Architecture 229
F. Luengo, A. Iglesias

Framework for Simulating the Human Behavior for Intelligent Virtual Agents. Part II: Behavioral System 237
F. Luengo, A. Iglesias

Point-Based Modeling from a Single Image 245
P.-P. Vázquez, J. Marco, M. Sbert

Introducing Physical Boundaries in Virtual Environments 252
P. Herrero, A. de Antonio

Thin Client Access to a Visualization Environment 258
I. Fudos, I. Kyriazis

Interactive Visualization of Relativistic Effects with the Hardware Acceleration 264
R. Mantiuk, K. Murawko-Wisniewska, D. Zdrojewska

Workshop on Computer Algebra Systems and Applications (CASA 2004)

Design of Interactive Environment for Numerically Intensive Parallel Linear Algebra Calculations 270
P. Luszczek, J. Dongarra

Computer Algebra for Real-Time Dynamics of Robots with Large Numbers of Joints 278
R. Bansevicius, A. Cepulkauskas, R. Kulvietiene, G. Kulvietis

Development of SyNRAC—Formula Description and New Functions 286
H. Yanami, H. Anai

DisCAS: A Distributed-Parallel Computer Algebra System 295
Y. Wu, G. Yang, W. Zheng, D. Lin

A Mathematica Package for Solving and Displaying Inequalities 303
R. Ipanaqué, A. Iglesias

Choleski-Banachiewicz Approach to Systems with Non-positive Definite Matrices with MATHEMATICA® 311
R.A. Walentyński

A Closed Form Solution of the Run-Time of a Sliding Bead along a Freely Hanging Slinky 319
H. Sarafian

Analytical Theory of Motion of a Mars Orbiter 327
J.F. San Juan, S. Serrano, A. Abad

Computing Theta-Stable Parabolic Subalgebras Using LiE 335
A.G. Noël

Graphical and Computational Representation of Groups 343
A. Bretto, L. Gillibert

First Order ODEs: Mathematica and Symbolic-Numerical Methods 351
C. D'Apice, G. Gargiulo, M. Rosanna

Evaluation of the Fundamental Physical Constants in *Mathematica* 358
A.S. Siver

Symbolic Polynomial Interpolation Using Mathematica 364
A. Yazici, I. Altas, T. Ergenc

Constant Weight Codes with Package CodingTheory.m in Mathematica 370
I. Gashkov

Graph Coloring with web*Mathematica* 376
Ü. Ufuktepe, G. Bacak, T. Beseri

Construction of Solutions for Nonintegrable Systems with the Help of the Painlevé Test 382
S.Y. Vernov

Computer Algebra Manipulations in Astronomy 388
T. Ivanova

Workshop on New Numerical Methods for DEs: Applications to Linear Algebra, Control and Engineering

Higher Order Quadrature on Sparse Grids 394
H.-J. Bungartz, S. Dirnstorfer

Application of Extrapolation Methods to Numerical Solution of Fredholm Integral Equations Related to Boundary Value Problems 402
A. Sidi

Extrapolation Techniques for Computing Accurate Solutions of Elliptic Problems with Singular Solutions 410
H. Koestler, U. Ruede

Vandermonde–Type Matrices in Two Step Collocation Methods for Special Second Order Ordinary Differential Equations 418
S. Martucci, B. Paternoster

Direct Optimization Using Gaussian Quadrature and Continuous Runge-Kutta Methods: Application to an Innovation Diffusion Model 426
F. Diele, C. Marangi, S. Ragni

The ReLPM Exponential Integrator for FE Discretizations of Advection-Diffusion Equations 434
L. Bergamaschi, M. Caliari, M. Vianello

Function Fitting Two–Step BDF Algorithms for ODEs 443
L.G. Ixaru, B. Paternoster

Pseudospectral Iterated Method for Differential Equations with Delay Terms 451
J. Mead, B. Zubik-Kowal

A Hybrid Numerical Technique for the Solution of a Class of Implicit Matrix Differential Equation 459
N. Del Buono, L. Lopez

A Continuous Approach for the Computation of the Hyperbolic Singular Value Decomposition 467
T. Politi

Workshop on Parallel Monte Carlo Algorithms for Diverse Applications in a Distributed Setting

Using P-GRADE for Monte Carlo Computations in a Distributed Environment 475
V.N. Alexandrov, A. Thandavan, P. Kacsuk

Calculating Activation Energies in Diffusion Processes Using a Monte Carlo Approach in a Grid Environment 483
M. Calleja, M.T. Dove

Using Parallel Monte Carlo Methods in Large-Scale Air Pollution Modelling 491
V.N. Alexandrov, Z. Zlatev

Parallel Importance Separation for Multiple Integrals and Integral Equations 499
S. Ivanovska, A. Karaivanova

Investigation of the Sensitivity of the Monte Carlo Solution for the Barker-Ferry Equation with Sequential and Parallel Pseudo-Random Number Generators 507
T.V. Gurov, P.A. Whitlock

Design and Distributed Computer Simulation of Thin p^+–i–n^+ Avalanche Photodiodes Using Monte Carlo Model 515
M. Yakutovich

Convergence Proof for a Monte Carlo Method for Combinatorial Optimization Problems 523
S. Fidanova

Monte Carlo Algorithm for Maneuvering Target Tracking and Classification 531
D. Angelova, L. Mihaylova, T. Semerdjiev

Workshop on Modelling and Simulation of Multi-physics Multi-scale Systems

Coupling a Lattice Boltzmann and a Finite Difference Scheme 540
P. Albuquerque, D. Alemani, B. Chopard, P. Leone

Accuracy versus Performance in Lattice Boltzmann BGK Simulations of Systolic Flows 548
A.M. Artoli, L. Abrahamyan, A.G. Hoekstra

Mesoscopic Modelling of Droplets on Topologically Patterned Substrates 556
A. Dupuis, J.M. Yeomans

Soot Particle Deposition within Porous Structures Using a Method of Moments – Lattice Boltzmann Approach 564
B.F.W. Gschaider, C.C. Honeger, C.E.P. Redl

Numerical Bifurcation Analysis of Lattice Boltzmann Models: A Reaction-Diffusion Example 572
P. Van Leemput, K. Lust

Particle Models of Discharge Plasmas in Molecular Gases 580
S. Longo, M. Capitelli, P. Diomede

Fully Kinetic Particle-in-Cell Simulation of a Hall Thruster 588
F. Taccogna, S. Longo, M. Capitelli, R. Schneider

Standard of Molecular Dynamics Modeling and Simulation of Relaxation in Dense Media 596
A.Y. Kuksin, I.V. Morozov, G.E. Norman, V.V. Stegailov

Implicit and Explicit Higher Order Time Integration Schemes for Fluid-Structure Interaction Computations 604
A. van Zuijlen, H. Bijl

Accounting for Nonlinear Aspects in Multiphysics Problems: Application to Poroelasticity 612
D. Néron, P. Ladevèze, D. Dureisseix, B.A. Schrefler

Computational Modelling of Multi-field Ionic Continuum Systems 621
J. Michopoulos

Formation of Dwarf Galaxies in Reionized Universe with Heterogeneous Multi-computer System 629
T. Boku, H. Susa, K. Onuma, M. Umemura, M. Sato, D. Takahashi

A Multi-scale Numerical Study of the Flow, Heat, and Mass Transfer in Protective Clothing 637
M.P. Sobera, C.R. Kleijn, P. Brasser, H.E.A. Van den Akker

Thermomechanical Waves in SMA Patches under Small Mechanical Loadings 645
L. Wang, R.V.N. Melnik

Direct and Homogeneous Numerical Approaches to Multiphase Flows and Applications 653
R. Samulyak, T. Lu, Y. Prykarpatskyy

Molecular Dynamics and Monte Carlo Simulations for Heat Transfer in Micro and Nano-channels 661
A.J.H. Frijns, S.V. Nedea, A.J. Markvoort, A.A. van Steenhoven, P.A.J. Hilbers

Improved Semi-Lagrangian Stabilizing Correction Scheme for Shallow Water Equations 667
A. Bourchtein, L. Bourchtein

Bose-Einstein Condensation Studied by the Real-Time Monte Carlo Simulation in the Frame of Java Applet 673
M. Gall, R. Kutner, A. Majerowski, D. Żebrowski

Workshop on Gene, Genome, and Population Evolution

Life History Traits and Genome Structure: Aerobiosis and G+C Content in Bacteria 679
J.R. Lobry

Differential Gene Survival under Asymmetric Directional Mutational Pressure 687
P. Mackiewicz, M. Dudkiewicz, M. Kowalczuk, D. Mackiewicz, J. Banaszak, N. Polak, K. Smolarczyk, A. Nowicka, M.R. Dudek, S. Cebrat

How Gene Survival Depends on Their Length 694
N. Polak, J. Banaszak, P. Mackiewicz, M. Dudkiewicz, M. Kowalczuk, D. Mackiewicz, K. Smolarczyk, A. Nowicka, M.R. Dudek, S. Cebrat

Super-tree Approach for Studying the Phylogeny of Prokaryotes: New Results on Completely Sequenced Genomes 700
A. Calteau, V. Daubin, G. Perrieère

Genetic Paralog Analysis and Simulations 709
S. Cebrat, J.P. Radomski, D. Stauffer

Evolutionary Perspectives on Protein Thermodynamics 718
R.A. Goldstein

The Partition Function Variant of Sankoff's Algorithm 728
I.L. Hofacker, P.F. Stadler

Simulation of Molecular Evolution Using Population Dynamics Modelling 736
S.V. Semovski

Lotka-Volterra Model of Macro-Evolution on Dynamical Networks 742
F. Coppex, M. Droz, A. Lipowski

Simulation of a Horizontal and Vertical Disease Spread in Population 750
M. Magdoń-Maksymowicz

Evolution of Population with Interaction between Neighbours 758
M. Magdoń-Maksymowicz, A.Z. Maksymowicz

The Role of Dominant Mutations in the Population Expansion 765
S. Cebrat, A. Pękalski

Workshop on Computational Methods in Finance and Insurance

On the Efficiency of Simplified Weak Taylor Schemes for Monte Carlo Simulation in Finance 771
N. Bruti Liberati, E. Platen

Time-Scale Transformations: Effects on VaR Models 779
F. Lamantia, S. Ortobelli, S. Rachev

Environment and Financial Markets 787
W. Szatzschneider, M. Jeanblanc, T. Kwiatkowska

Pricing of Some Exotic Options with *NIG*-Lévy Input 795
S. Rasmus, S. Asmussen, M. Wiktorsson

Construction of Quasi Optimal Portfolio for Stochastic Models of Financial Market 803
A. Janicki, J. Zwierz

Euler Scheme for One-Dimensional SDEs with Time Dependent Reflecting Barriers 811
L. Słomiński, T. Wojciechowski

On Approximation of Average Expectation Prices for Path Dependent Options in Fractional Models 819
B. Ziemkiewicz

Confidence Intervals for the Autocorrelations of the Squares of GARCH Sequences 827
P. Kokoszka, G. Teyssière, A. Zhang

Performance Measures in an Evolutionary Stock Trading Expert System 835
P. Lipinski, J.J. Korczak

Stocks' Trading System Based on the Particle Swarm Optimization Algorithm 843
J. Nenortaite, R. Simutis

Parisian Options – The Implied Barrier Concept 851
J. Anderluh, H. van der Weide

Modeling Electricity Prices with Regime Switching Models 859
M. Bierbrauer, S. Trück, R. Weron

Modeling the Risk Process in the XploRe Computing Environment 868
K. Burnecki, R. Weron

Workshop on Computational Economics and Finance

A Dynamic Stochastic Programming Model for Bond Portfolio Management 876
L. Yu, S. Wang, Y. Wu, K.K. Lai

Communication Leading to Nash Equilibrium without Acyclic Condition (– **S4**-Knowledge Model Case –) 884
T. Matsuhisa

Support Vector Machines Approach to Credit Assessment 892
J. Li, J. Liu, W. Xu, Y. Shi

Measuring Scorecard Performance 900
Z. Yang, Y. Wang, Y. Bai, X. Zhang

Parallelism of Association Rules Mining and Its Application in Insurance Operations 907
J. Tian, L. Zhu, S. Zhang, G. Huang

No Speculation under Expectations in Awareness 915
K. Horie, T. Matsuhisa

A Method on Solving Multiobjective Conditional Value-at-Risk 923
M. Jiang, Q. Hu, Z. Meng

Cross-Validation and Ensemble Analyses on Multiple-Criteria Linear Programming Classification for Credit Cardholder Behavior 931
Y. Peng, G. Kou, Z. Chen, Y. Shi

Workshop on GeoComputation

A Cache Mechanism for Component-Based WebGIS 940
Y. Luo, X. Wang, Z. Xu

A Data Structure for Efficient Transmission of Generalised Vector Maps 948
M. Zhou, M. Bertolotto

Feasibility Study of Geo-spatial Analysis Using Grid Computing 956
Y. Hu, Y. Xue, J. Wang, X. Sun, G. Cai, J. Tang, Y. Luo, S. Zhong, Y. Wang, A. Zhang

An Optimum Vehicular Path Solution with Multi-heuristics 964
F. Lu, Y. Guan

An Extended Locking Method for Geographical Database with Spatial Rules 972
C. Cheng, P. Shen, M. Zhang, F. Lu

Preliminary Study on Unsupervised Classification of Remotely Sensed Images on the Grid 981
J. Wang, X. Sun, Y. Xue, Y. Hu, Y. Luo, Y. Wang, S. Zhong, A. Zhang, J. Tang, G. Cai

Experience of Remote Sensing Information Modelling with Grid Computing 989
G. Cai, Y. Xue, J. Tang, J. Wang, Y. Wang, Y. Luo, Y. Hu, S. Zhong, X. Sun

Load Analysis and Load Control in Geo-agents 997
Y. Luo, X. Wang, Z. Xu

Workshop on Simulation and Modeling of 3D Integrated Circuits

Challenges in Transmission Line Modeling at Multi-gigabit Data Rates 1004
V. Heyfitch

MPI-Based Parallelized Model Order Reduction Algorithm 1012
I. Balk, S. Zorin

3D-VLSI Design Tool 1017
R. Bollapragada

Analytical Solutions of the Diffusive Heat Equation as the Application for Multi-cellular Device Modeling – A Numerical Aspect . . . 1021
Z. Lisik, J. Wozny, M. Langer, N. Rinaldi

Layout Based 3D Thermal Simulations of Integrated Circuits Components 1029
K. Slusarczyk, M. Kaminski, A. Napieralski

Simulation of Electrical and Optical Interconnections for Future VLSI ICs 1037
G. Tosik, Z. Lisik, M. Langer, F. Gaffiot, I. O'Conor

Balanced Binary Search Trees Based Approach for Sparse Matrix Representation 1045
I. Balk, I. Pavlovsky, A. Ushakov, I. Landman

Principles of Rectangular Mesh Generation in Computational Physics . . . 1049
V. Ermolaev, E. Odintsov, A. Sobachkin, A. Kharitonovich, M. Bevzushenko, S. Zorin

Workshop on Computational Modeling and Simulation on Biomechanical Engineering

Inter-finger Connection Matrices 1056
V.M. Zatsiorsky, M.L. Latash, F. Danion, F. Gao, Z.-M. Li, R.W. Gregory, S. Li

Biomechanics of Bone Cement Augmentation with Compression Hip Screw System for the Treatment of Intertrochanteric Fractures 1065
S.J. Lee, B.J. Kim, S.Y. Kwon, G.R. Tack

Comparison of Knee Cruciate Ligaments Models Using Kinematics from a Living Subject during Chair Rising-Sitting 1073
R. Stagni, S. Fantozzi, M. Davinelli, M. Lannocca

Computer and Robotic Model of External Fixation System for Fracture Treatment 1081
Y.H. Kim, S.-G. Lee

Robust Path Design of Biomechanical Systems Using the Concept of Allowable Load Set 1088
J.H. Chang, J.H. Kim, B.M. Kwak

A New Modeling Method for Objects with Branching Problem Using Non-uniform B-Spline 1095
H.S. Kim, Y.H. Kim, Y.H. Choe, S.-M. Kim, T.-S. Cho, J.H. Mun

Motion Design of Two-Legged Locomotion Process of a Man 1103
S. Novikava, K. Miatliuk, K. Jaworek

Adaptive Microcalcification Detection in Computer Aided Diagnosis 1110
H.-K. Kang, S.-M. Kim, N.N. Thanh, Y.M. Ro, W.-H. Kim

Workshop on Information Technologies Enhancing Health Care Delivery

The Impact of Information Technology on Quality of Healthcare Services 1118
M. Duplaga

Computer Generated Patient Plans Based on Patterns of Care 1126
O.M. Winnem

On Direct Comparing of Medical Guidelines with Electronic Health Record 1133
J. Zvárová, A. Veselý, P. Hanzliček, J. Špidlen, D. Buchtela

Managing Information Models for E-health via Planned Evolutionary Pathways 1140
H. Duwe

An Attributable Role-Based Access Control for Healthcare 1148
D. Schwartmann

Aspects of a Massively Distributed Stable Component Space 1156
K. Schmaranz, D. Schwartmann

Demonstrating Wireless IPv6 Access to a Federated Health Record Server 1165
D. Kalra, D. Ingram, A. Austin, V. Griffith, D. Lloyd, D. Patterson, P. Kirstein, P. Conversin, W. Fritsche

Collaborative Teleradiology 1172
K. Zieliński, J. Cała, Ł. Czekierda, S. Zieliński

Workshop on Computing in Science and Engineering Academic Programs

Some Remarks on CSE Education in Germany 1180
H.-J. Bungartz

The Computational Science and Engineering (CS&E) Program at Purdue University 1188
T. Downar, T. Kozlowski

Adapting the CSE Program at ETH Zurich to the Bologna Process 1196
R. Jeltsch, K. Nipp

Computational Engineering and Science Program at the University of Utah 1202
C. DeTar, A.L. Fogelson, C.R. Johnson, C.A. Sikorski, T. Truong

A Comparison of C, MATLAB, and Python as Teaching Languages in Engineering 1210
H. Fangohr

Teaching Computational Science Using VPython and Virtual Reality 1218
S. Roberts, H. Gardner, S. Press, L. Stals

Student Exercises on Fossil Fuels, Global Warming, and Gaia 1226
B.W. Rust

Teaching Scientific Computing 1234
B.A. Shadwick

Creating a Sustainable High-Performance Scientific Computing Course . . . 1242
E.R. Jessup, H.M. Tufo

CSE without Math? A First Course in Modeling and Simulation 1249
W. Wiechert

Author Index 1257

Table of Contents – Part I

Track on Parallel and Distributed Computing

Optimization of Collective Reduction Operations 1
R. Rabenseifner

Predicting MPI Buffer Addresses 10
F. Freitag, M. Farreras, T. Cortes, J. Labarta

An Efficient Load-Sharing and Fault-Tolerance Algorithm in Internet-Based Clustering Systems 18
I.-B. Choi, J.-D. Lee

Dynamic Parallel Job Scheduling in Multi-cluster Computing Systems 27
J.H. Abawajy

Hunting for Bindings in Distributed Object-Oriented Systems 35
M. Sławińska

Design and Implementation of the Cooperative Cache for PVFS 43
I.-C. Hwang, H. Kim, H. Jung, D.-H. Kim, H. Ghim, S.-R. Maeng, J.-W. Cho

Track on Grid Computing

Towards OGSA Compatibility in Alternative Metacomputing Frameworks 51
G. Stuer, V. Sunderam, J. Broeckhove

DartGrid: Semantic-Based Database Grid 59
Z. Wu, H. Chen, Changhuang, G. Zheng, J. Xu

A 3-tier Grid Architecture and Interactive Applications Framework for Community Grids 67
O. Ardaiz, K. Sanjeevan, R. Sanguesa

Incorporation of Middleware and Grid Technologies to Enhance Usability in Computational Chemistry Applications 75
J.P. Greenberg, S. Mock, M. Katz, G. Bruno, F. Sacerdoti, P. Papadopoulos, K.K. Baldridge

An Open Grid Service Environment for Large-Scale Computational Finance Modeling Systems 83
C. Wiesinger, D. Giczi, R. Hochreiter

The Migrating Desktop as a GUI Framework for the "Applications on Demand" Concept 91
M. Kupczyk, R. Lichwała, N. Meyer, B. Palak, M. Płóciennik, M. Stroiński, P. Wolniewicz

Interactive Visualization for the UNICORE Grid Environment 99
P. Bała, K. Benedyczak, A. Nowiński, K.S. Nowiński, J. Wypychowski

Efficiency of the GSI Secured Network Transmission 107
B. Baliś, M. Bubak, W. Rząsa, T. Szepieniec

An Idle Compute Cycle Prediction Service for Computational Grids 116
S. Hwang, E.-J. Im, K. Jeong, H. Park

Infrastructure for Grid-Based Virtual Organizations 124
L. Hluchy, O. Habala, V.D. Tran, B. Simo, J. Astalos, M. Dobrucky

Air Pollution Modeling in the CrossGrid Project 132
J.C. Mouriño, M.J. Martín, P. González, R. Doallo

The Genetic Algorithms Population Pluglet for the H2O Metacomputing System 140
T. Ampuła, D. Kurzyniec, V. Sunderam, H. Witek

Applying Grid Computing to the Parameter Sweep of a Group Difference Pseudopotential 148
W. Sudholt, K.K. Baldridge, D. Abramson, C. Enticott, S. Garic

A Grid Enabled Parallel Hybrid Genetic Algorithm for SPN 156
G.L. Presti, G.L. Re, P. Storniolo, A. Urso

An Atmospheric Sciences Workflow and Its Implementation with Web Services 164
D. Abramson, J. Kommineni, J.L. McGregor, J. Katzfey

Twins: 2-hop Structured Overlay with High Scalability 174
J. Hu, H. Dong, W. Zheng, D. Wang, M. Li

Dispatching Mechanism of an Agent-Based Distributed Event System 184
O.K. Sahingoz, N. Erdogan

An Adaptive Communication Mechanism for Highly Mobile Agents 192
J. Ahn

Track on Models and Algorithms

Knapsack Model and Algorithm for HW/SW Partitioning Problem 200
A. Ray, W. Jigang, S. Thambipillai

A Simulated Annealing Algorithm for the Circles Packing Problem 206
D. Zhang, W. Huang

Parallel Genetic Algorithm for Graph Coloring Problem 215
Z. Kokosiński, M. Kołodziej, K. Kwarciany

Characterization of Efficiently Parallel Solvable Problems on a Class of Decomposable Graphs 223
S.-Y. Hsieh

The Computational Complexity of Orientation Search in Cryo-Electron Microscopy 231
T. Mielikäinen, J. Ravantti, E. Ukkonen

Track on Data Mining and Data Bases

Advanced High Performance Algorithms for Data Processing 239
A.V. Bogdanov, A.V. Boukhanovsky

Ontology-Based Partitioning of Data Steam for Web Mining: A Case Study of Web Logs 247
J.J. Jung

Single Trial Discrimination between Right and Left Hand Movement-Related EEG Activity 255
S. Cho, J.A. Kim, D.-U. Hwang, S.K. Han

WINGS: A Parallel Indexer for Web Contents 263
F. Silvestri, S. Orlando, R. Perego

A Database Server for Predicting Protein-Protein Interactions 271
K. Han, B. Park

PairAnalyzer: Extracting and Visualizing RNA Structure Elements Formed by Base Pairing 279
D. Lim, K. Han

A Parallel Crawling Schema Using Dynamic Partition 287
S. Dong, X. Lu, L. Zhang

Hybrid Collaborative Filtering and Content-Based Filtering for Improved Recommender System 295
K.-Y. Jung, D.-H. Park, J.-H. Lee

Object-Oriented Database Mining: Use of Object Oriented Concepts for Improving Data Classification Technique 303
K. Waiyamai, C. Songsiri, T. Rakthanmanon

Data-Mining Based Skin-Color Modeling Using the ECL Skin-Color Images Database 310
M. Hammami, D. Tsishkou, L. Chen

Maximum Likelihood Based Quantum Set Separation 318
S. Imre, F. Balázs

Chunking-Coordinated-Synthetic Approaches to Large-Scale Kernel Machines 326
F.J. González-Castaño, R.R. Meyer

Computational Identification of -1 Frameshift Signals 334
S. Moon, Y. Byun, K. Han

Track on Networking

Mobility Management Scheme for Reducing Location Traffic Cost in Mobile Networks 342
B.-M. Min, J.-G. Jee, H.S. Oh

Performance Analysis of Active Queue Management Schemes for IP Network 349
J. Koo, S. Ahn, J. Chung

A Real-Time Total Order Multicast Protocol 357
K. Erciyes, A. Şahan

A Rule-Based Intrusion Alert Correlation System for Integrated Security Management 365
S.-H. Lee, H.-H. Lee, B.-N. Noh

Stable Neighbor Based Adaptive Replica Allocation in Mobile Ad Hoc Networks 373
Z. Jing, S. Jinshu, Y. Kan, W. Yijie

Mobile-Based Synchronization Model for Presentation of Multimedia Objects 381
K.-W. Lee, H.-S. Cho, K.-H. Lee

Synchronization Scheme of Multimedia Streams
in Mobile Handoff Control 389
G.-S. Lee

Poster Papers

The Development of a Language for Specifying Structure
of a Distributed and Parallel Application 397
R. Dew, P. Horan, A. Goscinski

Communication Primitives for Minimally Synchronous Parallel ML 401
F. Loulergue

Dependence Analysis of Concurrent Programs
Based on Reachability Graph and Its Applications 405
X. Qi, B. Xu

Applying Loop Tiling and Unrolling to a Sparse Kernel Code 409
E. Herruzo, G. Bandera, O. Plata

A Combined Method for Texture Analysis and Its Application 413
Y. Zhang, R. Wang

Reliability of Cluster System with a Lot of Software Instances 417
M. Szymczyk, P. Szymczyk

A Structural Complexity Measure for UML Class Diagrams 421
B. Xu, D. Kang, J. Lu

Parallelizing Flood Models with MPI:
Approaches and Experiences 425
V.D. Tran, L. Hluchy

Using Parallelism in Experimenting and Fine Tuning of Parameters
for Metaheuristics 429
M. Blesa, F. Xhafa

DEVMA: Developing Virtual Environments with Awareness Models 433
P. Herrero, A. de Antonio

A Two-Leveled Mobile Agent System for E-commerce
with Constraint-Based Filtering 437
O.K. Sahingoz, N. Erdogan

ABSDM: Agent Based Service Discovery Mechanism
in Internet 441
S. Li, C. Xu, Z. Wu, Y. Pan, X. Li

Meta Scheduling Framework for Workflow Service on the Grids 445
S. Hwang, J. Choi, H. Park

Resources Virtualization in Fault-Tolerance and Migration Issues 449
G. Jankowski, R. Mikolajczak, R. Januszewski, N. Meyer, M. Stroiński

On the Availability of Information Dispersal Scheme for Distributed Storage Systems 453
S.K. Song, H.Y. Youn, G.-L. Park, K.S. Tae

Virtual Storage System for the Grid Environment 458
D. Nikolow, R. Słota, J. Kitowski, Ł. Skitał

Performance Measurement Model in the G-PM Tool 462
R. Wismüller, M. Bubak, W. Funika, T. Arodź, M. Kurdziel

Paramedir: A Tool for Programmable Performance Analysis 466
G. Jost, J. Labarta, J. Gimenez

Semantic Browser: an Intelligent Client for Dart-Grid 470
Y. Mao, Z. Wu, H. Chen

On Identity-Based Cryptography and GRID Computing 474
H.W. Lim, M.J.B. Robshaw

The Cambridge CFD Grid Portal for Large-Scale Distributed CFD Applications 478
X. Yang, M. Hayes, K. Jenkins, S. Cant

Grid Computing Based Simulations of the Electrical Activity of the Heart 482
J.M. Alonso, V. Hernández, G. Moltó

Artificial Neural Networks and the Grid 486
E. Schikuta, T. Weishäupl

Towards a Grid-Aware Computer Algebra System 490
D. Petcu, D. Dubu, M. Paprzycki

Grid Computing and Component-Based Software Engineering in Computer Supported Collaborative Learning 495
M.L. Bote-Lorenzo, J.I. Asensio-Pérez, G. Vega-Gorgojo, L.M. Vaquero-González, E. Gómez-Sánchez, Y.A. Dimitriadis

An NAT-Based Communication Relay Scheme for Private-IP-Enabled MPI over Grid Environments 499
S. Choi, K. Park, S. Han, S. Park, O. Kwon, Y. Kim, H. Park

A Knowledge Fusion Framework in the Grid Environment 503
J. Gou, J. Yang, H. Qi

A Research of Grid Manufacturing and Its Application in Custom Artificial Joint 507
L. Chen, H. Deng, Q. Deng, Z. Wu

Toward a Virtual Grid Service of High Availability 511
X. Zhi, W. Tong

The Measurement Architecture of the Virtual Traffic Laboratory 515
A. Visser, J. Zoetebier, H. Yakali, B. Hertzberger

Adaptive QoS Framework for Multiview 3D Streaming 519
J.R. Kim, Y. Won, Y. Iwadate

CORBA-Based Open Platform for Processes Monitoring. An Application to a Complex Electromechanical Process 523
K. Cantillo, R.E. Haber, J.E. Jiménez, Á. Alique, R. Galán

An Approach to Web-Oriented Discrete Event Simulation Modeling 527
E. Ochmańska

Query Execution Algorithm in Web Environment with Limited Availability of Statistics 532
J. Jezierski, T. Morzy

Using Adaptive Priority Controls for Service Differentiation in QoS-Enabled Web Servers 537
M.M. Teixeira, M.J. Santana, R.H. Carlucci Santana

On the Evaluation of x86 Web Servers Using Simics: Limitations and Trade-Offs 541
F.J. Villa, M.E. Acacio, J.M. García

MADEW: Modelling a Constraint Awareness Model to Web-Based Learning Environments 545
P. Herrero, A. de Antonio

An EC Services System Using Evolutionary Algorithm 549
W.D. Lin

A Fast and Efficient Method for Processing Web Documents 553
D. Szegő

Online Internet Monitoring System of Sea Regions 557
M. Piotrowski, H. Krawczyk

Modeling a 3G Power Control Algorithm in the MAC Layer for Multimedia Support 561
U. Pineda, C. Vargas, J. Acosta-Elías, J.M. Luna, G. Pérez, E. Stevens

Network Probabilistic Connectivity: Exact Calculation with Use of Chains 565
O.K. Rodionova, A.S. Rodionov, H. Choo

A Study of Anycast Application for Efficiency Improvement of Multicast Trees 569
K.-J. Lee, W.-H. Choi, J.-S. Kim

Performance Analysis of IP-Based Multimedia Communication Networks to Support Video Traffic 573
A.F. Yaroslavtsev, T.-J. Lee, M.Y. Chung, H. Choo

Limited Deflection Routing with QoS-Support 577
H. Kim, S. Lee, J. Song

Advanced Multicasting for DVBMT Solution 582
M. Kim, Y.-C. Bang, H. Choo

Server Mobility Using Domain Name System in Mobile IPv6 Networks 586
H. Sung, S. Han

Resource Reservation and Allocation Method for Next Generation Mobile Communication Systems 590
J. Lee, S.-P. Cho, C. Kang

Improved Location Scheme Using Circle Location Register in Mobile Networks 594
D.C. Lee, H. Kim, I.-S. Hwang

An Energy Efficient Broadcasting for Mobile Devices Using a Cache Scheme 598
K.-H. Han, J.-H. Kim, Y.-B. Ko, W.-S. Yoon

On Balancing Delay and Cost for Routing Paths 602
M. Kim, Y.-C. Bang, H. Choo

Performance of Optical Burst Switching in Time Division Multiplexed Wavelength-Routing Networks 607
T.-W. Um, Y. Kwon, J.K. Choi

On Algorithm for All-Pairs Most Reliable Quickest Paths 611
Y.-C. Bang, I. Hong, H. Choo

Performance Evaluation of the Fast Consistency Algorithms in Large Decentralized Systems 615
J. Acosta-Elías, L. Navarro-Moldes

Building a Formal Framework for Mobile Ad Hoc Computing 619
L. Yan, J. Ni

Efficient Immunization Algorithm for Peer-to-Peer Networks 623
H. Chen, H. Jin, J. Sun, Z. Han

A Secure Process-Service Model 627
S. Deng, Z. Wu, Z. Yu, L. Huang

Multi-level Protection Building for Virus Protection Infrastructure 631
S.-C. Noh, D.C. Lee, K.J. Kim

Parallelization of the IDEA Algorithm 635
V. Beletskyy, D. Burak

A New Authorization Model for Workflow Management System Using the RPI-RBAC Model 639
S. Lee, Y. Kim, B. Noh, H. Lee

Reducing the State Space of RC4 Stream Cipher 644
V. Tomašević, S. Bojanić

A Pair-Wise Key Agreement Scheme in Ad Hoc Networks 648
W. Cha, G. Wang, G. Cho

Visual Analysis of the Multidimensional Meteorological Data 652
G. Dzemyda

Using Branch-Grafted R-trees for Spatial Data Mining 657
P. Dubey, Z. Chen, Y. Shi

Using Runtime Measurements and Historical Traces for Acquiring Knowledge in Parallel Applications 661
L.J. Senger, M.J. Santana, R.H.C. Santana

Words as Rules: Feature Selection in Text Categorization 666
E. Montañés, E.F. Combarro, I. Díaz, J. Ranilla, J.R. Quevedo

Proper Noun Learning from Unannotated Corpora for Information Extraction 670
S.-S. Kang

Proposition of Boosting Algorithm for Probabilistic Decision Support System 675
M. Wozniak

Efficient Algorithm for Linear Pattern Separation 679
C. Tadonki, J.-P. Vial

Improved Face Detection Algorithm in Mobile Environment 683
S.-B. Rhee, Y.-H. Lee

Real-Time Face Recognition by the PCA
(Principal Component Analysis) with Color Images 687
J.O. Kim, S.J. Seo, C.H. Chung

Consistency of Global Checkpoints Based on Characteristics
of Communication Events in Multimedia Applications 691
M. Ono, H. Higaki

Combining the Radon, Markov, and Stieltjes Transforms
for Object Reconstruction 695
A. Cuyt, B. Verdonk

Author Index 699

Table of Contents – Part II

Track on Numerical Algorithms

Hierarchical Matrix-Matrix Multiplication Based on Multiprocessor Tasks 1
S. Hunold, T. Rauber, G. Rünger

Improving Geographical Locality of Data for Shared Memory Implementations of PDE Solvers 9
H. Löf, M. Nordén, S. Holmgren

Cache Oblivious Matrix Transposition: Simulation and Experiment 17
D. Tsifakis, A.P. Rendell, P.E. Strazdins

An Intelligent Hybrid Algorithm for Solving Non-linear Polynomial Systems 26
J. Xue, Y. Li, Y. Feng, L. Yang, Z. Liu

A Jacobi–Davidson Method for Nonlinear Eigenproblems 34
H. Voss

Numerical Continuation of Branch Points of Limit Cycles in MATCONT 42
A. Dhooge, W. Govaerts, Y.A. Kuznetsov

Online Algorithm for Time Series Prediction Based on Support Vector Machine Philosophy 50
J.M. Górriz, C.G. Puntonet, M. Salmerón

Improved A-P Iterative Algorithm in Spline Subspaces 58
J. Xian, S.P. Luo, W. Lin

Solving Differential Equations in Developmental Models of Multicellular Structures Expressed Using L-systems 65
P. Federl, P. Prusinkiewicz

On a Family of A-stable Collocation Methods with High Derivatives 73
G.Y. Kulikov, A.I. Merkulov, E.Y. Khrustaleva

Local Sampling Problems 81
S.-Y. Yang, W. Lin

Recent Advances in Semi-Lagrangian Modelling of Flow through the Strait of Gibraltar 89
M. Seaïd, M. El-Amrani, A. Machmoum

Efficiency Study of the "Black-Box" Component Decomposition Preconditioning for Discrete Stress Analysis Problems 97
M.D. Mihajlović, S. Mijalković

Direct Solver Based on FFT and SEL for Diffraction Problems with Distribution 105
H. Koshigoe

Non-negative Matrix Factorization for Filtering Chinese Document 113
J. Lu, B. Xu, J. Jiang, D. Kang

On Highly Secure and Available Data Storage Systems 121
S.J. Choi, H.Y. Youn, H.S. Lee

Track on Finite Element Method

A Numerical Adaptive Algorithm for the Obstacle Problem 130
F.A. Pérez, J.M. Cascón, L. Ferragut

Finite Element Model of Fracture Formation on Growing Surfaces 138
P. Federl, P. Prusinkiewicz

An Adaptive, 3-Dimensional, Hexahedral Finite Element Implementation for Distributed Memory 146
J. Hippold, A. Meyer, G. Rünger

A Modular Design for Parallel Adaptive Finite Element Computational Kernels 155
K. Banaś

Load Balancing Issues for a Multiple Front Method 163
C. Denis, J.P. Boufflet, P. Breitkopf, M. Vayssade, B. Glut

Multiresolutional Techniques in Finite Element Method Solution of Eigenvalue Problem 171
M. Kamiński

Track on Neural Networks

Self-Organizing Multi-layer Fuzzy Polynomial Neural Networks Based on Genetic Optimization 179
S.-K. Oh, W. Pedrycz, H.-K. Kim, J.-B. Lee

Information Granulation-Based Multi-layer Hybrid Fuzzy Neural Networks: Analysis and Design 188
B.-J. Park, S.-K. Oh, W. Pedrycz, T.-C. Ahn

Efficient Learning of Contextual Mappings by Context-Dependent Neural Nets 196
P. Ciskowski

An Unsupervised Neural Model to Analyse Thermal Properties of Construction Materials 204
E. Corchado, P. Burgos, M. Rodríguez, V. Tricio

Intrusion Detection Based on Feature Transform Using Neural Network 212
W. Kim, S.-C. Oh, K. Yoon

Track on Applications

Accelerating Wildland Fire Prediction on Cluster Systems 220
B. Abdalhaq, A. Cortés, T. Margalef, E. Luque

High Precision Simulation of Near Earth Satellite Orbits for SAR-Applications 228
M. Kalkuhl, K. Nöh, O. Loffeld, W. Wiechert

Hybrid Approach to Reliability and Functional Analysis of Discrete Transport System 236
T. Walkowiak, J. Mazurkiewicz

Mathematical Model of Gas Transport in Anisotropic Porous Electrode of the PEM Fuel Cell 244
E. Kurgan, P. Schmidt

Numerical Simulation of Anisotropic Shielding of Weak Magnetic Fields 252
E. Kurgan

Functionalization of Single-Wall Carbon Nanotubes: An Assessment of Computational Methods 260
B. Akdim, T. Kar, X. Duan, R. Pachter

Improved Sampling for Biological Molecules Using Shadow Hybrid Monte Carlo 268
S.S. Hampton, J.A. Izaguirre

A New Monte Carlo Approach for Conservation Laws and Relaxation Systems 276
L. Pareschi, M. Seaïd

A Parallel Implementation of Gillespie's Direct Method 284
A.M. Ridwan, A. Krishnan, P. Dhar

Simulation of Deformable Objects Using Sliding Mode Control with Application to Cloth Animation 292
F. Rum, B.W. Gordon

Constraint-Based Contact Analysis between Deformable Objects 300
M. Hong, M.-H. Choi, C. Lee

Prediction of Binding Sites in Protein-Nucleic Acid Complexes 309
N. Han, K. Han

Prediction of Protein Functions Using Protein Interaction Data 317
H. Jung, K. Han

Interactions of Magainin-2 Amide with Membrane Lipids 325
K. Murzyn, T. Róg, M. Pasenkiewicz-Gierula

Dynamics of Granular Heaplets: A Phenomenological Model 332
Y.K. Goh, R.L. Jacobs

Modelling of Shear Zones in Granular Materials within Hypoplasticity 340
J. Tejchman

Effective Algorithm for Detection of a Collision between Spherical Particles 348
J.S. Leszczynski, M. Ciesielski

Vorticity Particle Method for Simulation of 3D Flow 356
H. Kudela, P. Regucki

Crack Analysis in Single Plate Stressing of Particle Compounds 364
M. Khanal, W. Schubert, J. Tomas

A Uniform and Reduced Mathematical Model for Sucker Rod Pumping 372
L. Liu, C. Tong, J. Wang, R. Liu

Distributed Computation of Optical Flow 380
A.G. Dopico, M.V. Correia, J.A. Santos, L.M. Nunes

Analytical Test on Effectiveness of MCDF Operations 388
J. Kong, B. Zhang, W. Guo

An Efficient Perspective Projection Using VolumePro^{TM} 396
S. Lim, B.-S. Shin

Reconstruction of 3D Curvilinear Wireframe Model from 2D Orthographic Views 404
A. Zhang, Y. Xue, X. Sun, Y. Hu, Y. Luo, Y. Wang, S. Zhong, J. Wang, J. Tang, G. Cai

Surface Curvature Estimation for Edge Spinning Algorithm 412
M. Cermak, V. Skala

Visualization of Very Large Oceanography Time-Varying Volume Datasets 419
S. Park, C. Bajaj, I. Ihm

Sphere-Spin-Image: A Viewpoint-Invariant Surface Representation for 3D Face Recognition 427
Y. Wang, G. Pan, Z. Wu, S. Han

Design and Implementation of Integrated Assembly Object Model for Intelligent Virtual Assembly Planning 435
J. Fan, Y. Ye, J.-M. Cai

Adaptive Model Based Parameter Estimation, Based on Sparse Data and Frequency Derivatives 443
D. Deschrijver, T. Dhaene, J. Broeckhove

Towards Efficient Parallel Image Processing on Cluster Grids Using GIMP 451
P. Czarnul, A. Ciereszko, M. Frączak

Benchmarking Parallel Three Dimensional FFT Kernels with ZENTURIO 459
R. Prodan, A. Bonelli, A. Adelmann, T. Fahringer, C. Überhuber

The Proof and Illustration of the Central Limit Theorem by Brownian Numerical Experiments in Real Time within the Java Applet 467
M. Gall, R. Kutner, W. Wesela

An Extended Coherence Protocol for Recoverable DSM Systems with Causal Consistency 475
J. Brzezinski, M. Szychowiak

2D and 3D Representations of Solution Spaces for CO Problems 483
E. Nowicki, C. Smutnicki

Effective Detector Set Generation and Evolution for Artificial Immune System 491
C. Kim, W. Kim, M. Hong

Artificial Immune System against Viral Attack 499
H. Lee, W. Kim, M. Hong

Proposal of the Programming Rules for VHDL Designs 507
J. Borgosz, B. Cyganek

A Weight Adaptation Method for Fuzzy Cognitive Maps to a Process Control Problem 515
E. Papageorgiou, P. Groumpos

A Method Based on Fuzzy Logic Technique for Smoothing in 2D 523
A. Çinar

Proportional-Integral-Derivative Controllers Tuning for Unstable and Integral Processes Using Genetic Algorithms 532
M.A. Paz-Ramos, J. Torres-Jimenez, E. Quintero-Marmol-Marquez

Enabling Systems Biology: A Scientific Problem-Solving Environment . . . 540
M. Singhal, E.G. Stephan, K.R. Klicker, L.L. Trease, G. Chin Jr., D.K. Gracio, D.A. Payne

Poster Papers

Depth Recovery with an Area Based Version of the Stereo Matching Method with Scale-Space Tensor Representation of Local Neighborhoods 548
B. Cyganek

Symbolic Calculation for Frölicher-Nijenhuis $\mathbb{R}$-Algebra for Exploring in Electromagnetic Field Theory 552
J. de Cruz Guzmán, Z. Oziewicz

Spherical Orthogonal Polynomials and Symbolic-Numeric Gaussian Cubature Formulas 557
A. Cuyt, B. Benouahmane, B. Verdonk

The Berlekamp-Massey Algorithm. A Sight from Theory of Pade Approximants and Orthogonal Polynomials 561
S.B. Gashkov, I.B. Gashkov

An Advanced Version of the Local-Global Step Size Control for Runge-Kutta Methods Applied to Index 1 Differential-Algebraic Systems 565
G.Y. Kulikov

INTEGRATOR: A Computational Tool to Solve Ordinary Differential Equations with Global Error Control 570
G.Y. Kulikov, S.K. Shindin

Reconstruction of Signal from Samples of Its Integral in Spline Subspaces 574
J. Xian, Y. Li, W. Lin

The Vectorized and Parallelized Solving of Markovian Models for Optical Networks 578
B. Bylina, J. Bylina

A Parallel Splitting up Algorithm for the Determination of an Unknown Coefficient in Multi Dimensional Parabolic Problem 582
D.S. Daoud, D. Subasi

A-Posteriori Error Analysis of a Mixed Method for Linear Parabolic Problem 586
M.I. Asensio, J.M. Cascón, L. Ferragut

Analysis of Parallel Numerical Libraries to Solve the 3D Electron Continuity Equation 590
N. Seoane, A.J. García-Loureiro

Parallel Solution of Cascaded ODE Systems Applied to ^{13}C-Labeling Experiments 594
K. Nöh, W. Wiechert

A k-way Graph Partitioning Algorithm Based on Clustering by Eigenvector 598
T.-Y. Choe, C.-I. Park

Network of Networks 602
J. de Cruz Guzmán, Z. Oziewicz

MSL: An Efficient Adaptive In-Place Radix Sort Algorithm 606
F. El-Aker, A. Al-Badarneh

Parallel Chip Firing Game Associated with *n-cube* Edges Orientations 610
R. Ndoundam, C. Tadonki, M. Tchuente

A Fast Multifrontal Solver for Non-linear Multi-physics Problems 614
A. Bertoldo, M. Bianco, G. Pucci

Modelling of Interaction between Surface Waves and Mud Layer 618
L. Balas

Computational Modelling of Pulsating Biological Flow 622
X.S. Yang, R.W. Lewis, H. Zhang

Effect of Heterogeneity on Formation of Shear Zones in Granular Bodies 626
J. Tejchman

Effect of Structural Disorder on the Electronic Density of States in One-Dimensional Chain of Atoms 630
M. Wołoszyn, B.J. Spisak

The Estimation of the Mathematical Exactness of System Dynamics Method on the Base of Some Economic System 634
E. Kasperska, D. Słota

Size of the Stable Population in the Penna Bit-String Model of Biological Aging 638
K. Malarz, M. Sitarz, P. Gronek, A. Dydejczyk

Velocity Field Modelling for Pollutant Plume Using 3-D Adaptive Finite Element Method 642
G. Montero, R. Montenegro, J.M. Escobar, E. Rodríguez, J.M. González-Yuste

Organization of the Mesh Structure 646
T. Jurczyk, B. Głut

Kernel Maximum Likelihood Hebbian Learning 650
J. Koetsier, E. Corchado, D. MacDonald, J. Corchado, C. Fyfe

Discovery of Chemical Transformations with the Use of Machine Learning 654
G. Fic, G. Nowak

Extraction of Document Descriptive Terms with a Linguistic-Based Machine Learning Approach 658
J. Fernández, E. Montañés, I. Díaz, J. Ranilla, E.F. Combarro

Application of Brain Emotional Learning Based Intelligent Controller (BELBIC) to Active Queue Management 662
M. Jalili-Kharaajoo

A Hybrid Algorithm Based on PSO and SA and Its Application for Two-Dimensional Non-guillotine Cutting Stock Problem 666
J.Q. Jiang, Y.C. Liang, X.H. Shi, H.P. Lee

Evolving TSP Heuristics Using Multi Expression Programming 670
M. Oltean, D. Dumitrescu

Improving the Performance of Evolutionary Algorithms for the Multiobjective 0/1 Knapsack Problem Using ε-Dominance 674
C. Groşan, M. Oltean

Genetic Evolution Approach for Target Movement Prediction 678
S. Baik, J. Bala, A. Hadjarian, P. Pachowicz

Adaptive Transfer Functions
in Radial Basis Function (RBF) Networks 682
G.A. Hoffmann

Disturbance Rejection Control of Thermal Power Plant
Using Immune Algorithm 687
D.H. Kim, J.H. Cho

The Design Methodology of Fuzzy Controller Based on Information
Granulation (IG)-Based Optimization Approach 691
S.-K. Oh, S.-B. Roh, D.-Y. Lee

PID Controller Tuning of a Boiler Control System
Using Immune Algorithm Typed Neural Network 695
D.H. Kim

A Framework to Investigate and Evaluate Genetic Clustering
Algorithms for Automatic Modularization of Software Systems 699
S. Parsa, O. Bushehrian

An Artificial Immune Algorithms Apply to Pre-processing Signals 703
M. Święcicki, W. Wajs, P. Wais

Identification and Control Using Direction Basis
Function Neural Network 708
M. Jalili-Kharaajoo

A New Dynamic Structure Neural Network for Control
of Nonlinear Systems 713
M. Jalili-Kharaajoo

Proposing a New Learning Algorithm to Improve Fault Tolerance
of Neural Networks 717
M. Jalili-Kharaajoo

Nonlinear Parametric Model Identification and Model Based Control
of *S. cerevisiae* Production 722
B. Akay

The Notion of Community in United States
Computational Science Education Initiatives 726
M.E. Searcy, J.T. Richie

Author Index 731

Table of Contents – Part III

Workshop on Programming Grids and Metasystems

High-Performance Parallel and Distributed Scientific Computing with the Common Component Architecture 1
D.E. Bernholdt

Multiparadigm Model Oriented to Development of Grid Systems 2
J.L.V. Barbosa, C.A. da Costa, A.C. Yamin, C.F.R. Geyer

The Effect of the 2^{nd} Generation Clusters: Changes in the Parallel Programming Paradigms 10
J. Porras, P. Huttunen, J. Ikonen

JavaSymphony, a Programming Model for the Grid 18
A. Jugravu, T. Fahringer

Adaptation of Legacy Software to Grid Services 26
B. Baliś, M. Bubak, M. Węgiel

Grid Service Registry for Workflow Composition Framework 34
M. Bubak, T. Gubała, M. Kapałka, M. Malawski, K. Rycerz

A-GWL: Abstract Grid Workflow Language 42
T. Fahringer, S. Pllana, A. Villazon

Design of Departmental Metacomputing ML 50
F. Gava

A Grid-Enabled Scene Rendering Application 54
M. Caballer, V. Hernández, J.E. Román

Rule-Based Visualization in a Computational Steering Collaboratory 58
L. Jiang, H. Liu, M. Parashar, D. Silver

Placement of File Replicas in Data Grid Environments 66
J.H. Abawajy

Generating Reliable Conformance Test Suites for Parallel and Distributed Languages, Libraries, and APIs 74
Ł. Garstecki

A Concept of Replicated Remote Method Invocation 82
J. Brzezinski, C. Sobaniec

Workshop on First International Workshop on Active and Programmable Grids Architectures and Components

Discovery of Web Services with a P2P Network 90
F. Forster, H. De Meer

Achieving Load Balancing in Structured Peer-to-Peer Grids 98
C. Pairot, P. García, A.F.G. Skarmeta, R. Mondéjar

A Conceptual Model for Grid-Adaptivity of HPC Applications and Its Logical Implementation Using Components Technology 106
A. Machì, S. Lombardo

Global Discovery Service for JMX Architecture 114
J. Midura, K. Balos, K. Zielinski

Towards a Grid Applicable Parallel Architecture Machine 119
K. Skala, Z. Sojat

A XKMS-Based Security Framework for Mobile Grid into the XML Web Services 124
N. Park, K. Moon, J. Jang, S. Sohn

A Proposal of Policy-Based System Architecture for Grid Services Management 133
E. Magaña, E. Salamanca, J. Serrat

Self-Management GRID Services – A Programmable Network Approach 141
L. Cheng, A. Galis, A. Savanović, B.J. Blažič, J. Bešter

Application-Specific Hints in Reconfigurable Grid Scheduling Algorithms 149
B. Volckaert, P. Thysebaert, F. De Turck, B. Dhoedt, P. Demeester

Self-Configuration of Grid Nodes Using a Policy-Based Management Architecture 158
F.J. García, Ó. Cánovas, G. Martínez, A.F.G. Skarmeta

Context-Aware GRID Services: Issues and Approaches 166
K. Jean, A. Galis, A. Tan

Security Issues in Virtual Grid Environments 174
J.L. Muñoz, J. Pegueroles, J. Forné, O. Esparza, M. Soriano

Implementation and Evaluation of Integrity Protection Facilities for Active Grids 179
A. Savanović, D. Gabrijelčič, B.J. Blažič, J. Bešter

A Convergence Architecture for GRID Computing and Programmable Networks 187
C. Bachmeir, P. Tabery, D. Marinov, G. Nachev, J. Eberspächer

Programmable Grids Framework Enabling QoS in an OGSA Context 195
J. Soldatos, L. Polymenakos, G. Kormentzas

Active and Logistical Networking for Grid Computing: The E-toile Architecture 202
A. Bassi, M. Beck, F. Chanussot, J.-P. Gelas, R. Harakaly, L. Lefèvre, T. Moore, J. Plank, P. Primet

Distributed Resource Discovery in Wide Area Grid Environments 210
T.N. Ellahi, M.T. Kechadi

Trusted Group Membership Service for JXTA 218
L. Kawulok, K. Zielinski, M. Jaeschke

Workshop on Next Generation Computing

An Implementation of Budget-Based Resource Reservation for Real-Time Linux 226
C.S. Liu, N.C. Perng, T.W. Kuo

Similarity Retrieval Based on SOM-Based R*-Tree 234
K.H. Choi, M.H. Shin, S.H. Bae, C.H. Kwon, I.H. Ra

Extending the Power of Server Based Computing 242
H.L. Yu, W.M. Zhen, M.M. Shen

Specifying Policies for Service Negotiations of Response Time 250
T.K. Kim, O.H. Byeon, K.J. Chun, T.M. Chung

Determination and Combination of Quantitative Weight Value from Multiple Preference Information 258
J.H. Yoo, B.G. Lee, H.S. Han

Forwarding Based Data Parallel Handoff for Real-Time QoS in Mobile IPv6 Networks 266
H.Y. Jeong, J. Lim, J.D. Park, H. Choo

Mobile Agent-Based Load Monitoring System for the Safety Web Server Environment 274
H.J. Park, K.J. Jyung, S.S. Kim

A Study on TCP Buffer Management Algorithm for Improvement of Network Performance in Grid Environment 281
Y. Jeong, M. Noh, H.K. Lee, Y. Mun

Workshop on Practical Aspects of High-Level Parallel Programming (PAPP 2004)

Evaluating the Performance of Skeleton-Based High Level Parallel Programs 289
A. Benoit, M. Cole, S. Gilmore, J. Hillston

Towards a Generalised Runtime Environment for Parallel Haskells 297
J. Berthold

Extending Camelot with Mutable State and Concurrency 306
S. Gilmore

EVE, an Object Oriented SIMD Library 314
J. Falcou, J. Sérot

Petri Nets as Executable Specifications of High-Level Timed Parallel Systems 322
F. Pommereau

Parallel I/O in Bulk-Synchronous Parallel ML 331
F. Gava

Workshop on Parallel Input/Output Management Techniques (PIOMT04)

File Replacement Algorithm for Storage Resource Managers in Data Grids 339
J.H. Abawajy

Optimizations Based on Hints in a Parallel File System 347
M.S. Pérez, A. Sánchez, V. Robles, J.M. Peña, F. Pérez

Using DMA Aligned Buffer to Improve Software RAID Performance 355
Z. Shi, J. Zhang, X. Zhou

mNFS: Multicast-Based NFS Cluster 363
W.-G. Lee, C.-I. Park, D.-W. Kim

Balanced RM2: An Improved Data Placement Scheme for Tolerating Double Disk Failures in Disk Arrays 371
D.-W. Kim, S.-H. Lee, C.-I. Park

Diagonal Replication on Grid for Efficient Access of Data in Distributed Database Systems 379
M. Mat Deris, N. Bakar, M. Rabiei, H.M. Suzuri

Workshop on OpenMP for Large Scale Applications

Performance Comparison between OpenMP and MPI on IA64 Architecture 388
L. Qi, M. Shen, Y. Chen, J. Li

Defining Synthesizable OpenMP Directives and Clauses 398
P. Dziurzanski, V. Beletskyy

Efficient Translation of OpenMP to Distributed Memory 408
L. Huang, B. Chapman, Z. Liu, R. Kendall

ORC-OpenMP: An OpenMP Compiler Based on ORC 414
Y. Chen, J. Li, S. Wang, D. Wang

Workshop on Tools for Program Development and Analysis in Computational Science

Performance Analysis, Data Sharing, and Tools Integration in Grids: New Approach Based on Ontology 424
H.-L. Truong, T. Fahringer

Accurate Cache and TLB Characterization Using Hardware Counters 432
J. Dongarra, S. Moore, P. Mucci, K. Seymour, H. You

A Tool Suite for Simulation Based Analysis of Memory Access Behavior 440
J. Weidendorfer, M. Kowarschik, C. Trinitis

Platform-Independent Cache Optimization by Pinpointing Low-Locality Reuse 448
K. Beyls, E.H. D'Hollander

Teuta: Tool Support for Performance Modeling of Distributed and Parallel Applications 456
T. Fahringer, S. Pllana, J. Testori

MPI Application Development Using the Analysis Tool MARMOT 464
B. Krammer, M.S. Müller, M.M. Resch

Monitoring System for Distributed Java Applications 472
W. Funika, M. Bubak, M. Smętek

Automatic Parallel-Discrete Event Simulation 480
M. Marín

Workshop on Modern Technologies for Web-Based Adaptive Systems

Creation of Information Profiles in Distributed Databases as a n-Person Game 488
J.L. Kulikowski

Domain Knowledge Modelling for Intelligent Instructional Systems 497
E. Pecheanu, L. Dumitriu, C. Segal

Hybrid Adaptation of Web-Based Systems User Interfaces 505
J. Sobecki

Collaborative Web Browsing Based on Ontology Learning from Bookmarks 513
J.J. Jung, Y.-H. Yu, G.-S. Jo

Information Retrieval Using Bayesian Networks 521
L. Neuman, J. Kozlowski, A. Zgrzywa

An Application of the DEDS Control Synthesis Method 529
F. Čapkovič

Using Consistency Measures and Attribute Dependencies for Solving Conflicts in Adaptive Systems 537
M. Malowiecki, N.T. Nguyen, M. Zgrzywa

Logical Methods for Representing Meaning of Natural Language Texts . . . 545
T. Batura, F. Murzin

Software Self-Adaptability by Means of Artificial Evolution 552
M. Nowostawski, M. Purvis, A. Gecow

Professor:e – An IMS Standard Based Adaptive E-learning Platform 560
C. Segal, L. Dumitriu

Workshop on Agent Day 2004 – Intelligent Agents in Computing Systems

Towards Measure of Semantic Correlation between Messages in Multiagent System 567
A. Pieczyńska-Kuchtiak, R. Katarzyniak

Modelling Intelligent Virtual Agent Skills with Human-Like Senses 575
P. Herrero, A. de Antonio

Reuse of Organisational Experience Harnessing Software Agents 583
K. Krawczyk, M. Majewska, M. Dziewierz, R. Słota, Z. Balogh, J. Kitowski, S. Lambert

The Construction and Analysis of Agent Fault-Tolerance Model Based on π-Calculus 591
Y. Jiang, Z. Xia, Y. Zhong, S. Zhang

REMARK – Reusable Agent-Based Experience Management and Recommender Framework 599
Z. Balogh, M. Laclavik, L. Hluchy, I. Budinska, K. Krawczyk

Behavior Based Detection of Unfavorable Resources 607
K. Cetnarowicz, G. Rojek

Policy Modeling in Four Agent Economy 615
A. Woźniak

Multi-agent System for Irregular Parallel Genetic Computations 623
J. Momot, K. Kosacki, M. Grochowski, P. Uhruski, R. Schaefer

Strategy Extraction for Mobile Embedded Control Systems Apply the Multi-agent Technology 631
V. Srovnal, B. Horák, R. Bernatík, V. Snášel

Multi-agent Environment for Dynamic Transport Planning and Scheduling 638
J. Kozlak, J.-C. Créput, V. Hilaire, A. Koukam

Agent-Based Models and Platforms for Parallel Evolutionary Algorithms 646
M. Kisiel-Dorohinicki

A Co-evolutionary Multi-agent System for Multi-modal Function Optimization 654
R. Dreżewski

Workshop on Dynamic Data Driven Applications Systems

Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements 662
F. Darema

Distributed Collaborative Adaptive Sensing for Hazardous Weather Detection, Tracking, and Predicting 670
J. Brotzge, V. Chandresakar, K. Droegemeier, J. Kurose, D. McLaughlin, B. Philips, M. Preston, S. Sekelsky

Rule-Based Support Vector Machine Classifiers Applied to Tornado Prediction 678
T.B. Trafalis, B. Santosa, M.B. Richman

Adaptive Coupled Physical and Biogeochemical Ocean Predictions: A Conceptual Basis 685
P.F.J. Lermusiaux, C. Evangelinos, R. Tian, P.J. Haley, J.J. McCarthy, N.M. Patrikalakis, A.R. Robinson, H. Schmidt

Dynamic-Data-Driven Real-Time Computational Mechanics Environment 693
J. Michopoulos, C. Farhat, E. Houstis

A Note on Data-Driven Contaminant Simulation 701
C.C. Douglas, C.E. Shannon, Y. Efendiev, R. Ewing, V. Ginting, R. Lazarov, M.J. Cole, G. Jones, C.R. Johnson, J. Simpson

Computational Aspects of Data Assimilation for Aerosol Dynamics 709
A. Sandu, W. Liao, G.R. Carmichael, D. Henze, J.H. Seinfeld, T. Chai, D. Daescu

A Framework for Online Inversion-Based 3D Site Characterization 717
V. Akçelik, J. Bielak, G. Biros, I. Epanomeritakis, O. Ghattas, L.F. Kallivokas, E.J. Kim

A Note on Dynamic Data Driven Wildfire Modeling 725
J. Mandel, M. Chen, L.P. Franca, C. Johns, A. Puhalskii, J.L. Coen, C.C. Douglas, R. Kremens, A. Vodacek, W. Zhao

Agent-Based Simulation of Data-Driven Fire Propagation Dynamics 732
J. Michopoulos, P. Tsompanopoulou, E. Houstis, A. Joshi

Model Reduction of Large-Scale Dynamical Systems 740
A. Antoulas, D. Sorensen, K.A. Gallivan, P. Van Dooren, A. Grama, C. Hoffmann, A. Sameh

Data Driven Design Optimization Methodology Development and Application 748
H. Zhao, D. Knight, E. Taskinoglu, V. Jovanovic

A Dynamic Data Driven Computational Infrastructure for Reliable Computer Simulations 756
J.T. Oden, J.C. Browne, I. Babuška, C. Bajaj, L.F. Demkowicz, L. Gray, J. Bass, Y. Feng, S. Prudhomme, F. Nobile, R. Tempone

Improvements to Response-Surface Based Vehicle Design Using a Feature-Centric Approach 764
D. Thompson, S. Parthasarathy, R. Machiraju, S. Lawrence

An Experiment for the Virtual Traffic Laboratory: Calibrating Speed Dependency on Heavy Traffic (A Demonstration of a Study in a Data Driven Trafic Analysis) 771
A. Visser, J. Zoetebier, H. Yakali, B. Hertzberger

SAMAS: Scalable Architecture for Multi-resolution Agent-Based Simulation 779
A. Chaturvedi, J. Chi, S. Mehta, D. Dolk

Simulation Coercion Applied to Multiagent DDDAS 789
Y. Loitière, D. Brogan, P. Reynolds

O'SOAP – A Web Services Framework for DDDAS Applications 797
K. Pingali, P. Stodghill

Application of Grid-Enabled Technologies for Solving Optimization Problems in Data-Driven Reservoir Studies 805
M. Parashar, H. Klie, U. Catalyurek, T. Kurc, V. Matossian, J. Saltz, M.F. Wheeler

Image-Based Stress Recognition Using a Model-Based Dynamic Face Tracking System 813
D. Metaxas, S. Venkataraman, C. Vogler

Developing a Data Driven System for Computational Neuroscience 822
R. Snider, Y. Zhu

Karhunen–Loeve Representation of Periodic Second-Order Autoregressive Processes 827
D. Lucor, C.-H. Su, G.E. Karniadakis

Workshop on HLA-Based Distributed Simulation on the Grid

Using Web Services to Integrate Heterogeneous Simulations in a Grid Environment 835
J.M. Pullen, R. Brunton, D. Brutzman, D. Drake, M. Hieb, K.L. Morse, A. Tolk

Support for Effective and Fault Tolerant Execution of HLA-Based Applications in the OGSA Framework 848
K. Rycerz, M. Bubak, M. Malawski, P.M.A. Sloot

Federate Migration in HLA-Based Simulation 856
Z. Yuan, W. Cai, M.Y.H. Low, S.J. Turner

FT-RSS: A Flexible Framework for Fault Tolerant HLA Federations 865
J. Lüthi, S. Großmann

Design and Implementation of GPDS 873
T.-D. Lee, S.-H. Yoo, C.-S. Jeong

HLA_AGENT: Distributed Simulation of Agent-Based Systems with HLA 881
M. Lees, B. Logan, T. Oguara, G. Theodoropoulos

FedGrid: An HLA Approach to Federating Grids 889
S. Vuong, X. Cai, J. Li, S. Pramanik, D. Suttles, R. Chen

Workshop on Interactive Visualisation and Interaction Technologies

Do Colors Affect Our Recognition Memory for Haptic Rough Surfaces? . . 897
Z. Luo, A. Imamiya

Enhancing Human Computer Interaction in Networked Hapto-Acoustic Virtual Reality Environments on the CeNTIE Network 905
T. Adriaansen, A. Krumm-Heller, C. Gunn

Collaborative Integration of Speech and 3D Gesture for Map-Based Applications 913
A. Corradini

Mobile Augmented Reality Support for Architects Based on Feature Tracking Techniques 921
M. Bang Nielsen, G. Kramp, K. Grønbæk

User Interface Design for a Navigation and Communication System in the Automotive World 929
O. Preißner

Target Selection in Augmented Reality Worlds 936
J. Sands, S.W. Lawson, D. Benyon

Towards Believable Behavior Generation for Embodied Conversational Agents 946
A. Corradini, M. Fredriksson, M. Mehta, J. Königsmann, N.O. Bernsen, L. Johannesson

A Performance Analysis of Movement Patterns 954
C. Sas, G. O'Hare, R. Reilly

On the Motivation and Attractiveness Scope of the Virtual Reality User Interface of an Educational Game 962
M. Virvou, G. Katsionis, K. Manos

A Client-Server Engine for Parallel Computation of High-Resolution Planes 970
D.P. Gavidia, E.V. Zudilova, P.M.A. Sloot

A Framework for 3D Polysensometric Comparative Visualization 978
J.I. Khan, X. Xu, Y. Ma

An Incremental Editor for Dynamic Hierarchical Drawing of Trees 986
D. Workman, M. Bernard, S. Pothoven

Using Indexed-Sequential Geometric Glyphs to Explore Visual Patterns 996
J. Morey, K. Sedig

Studying the Acceptance or Rejection of Newcomers in Virtual Environments 1004
P. Herrero, A. de Antonio, J. Segovia

Open Standard Based Visualization of Complex Internet Computing Systems 1008
S.S. Yang, J.I. Khan

General Conception of the Virtual Laboratory 1013
M. Lawenda, N. Meyer, T. Rajtar, M. Okoń, D. Stokłosa, M. Stroiński, Ł. Popenda, Z. Gdaniec, R.W. Adamiak

Individual Differences in Virtual Environments 1017
C. Sas

Ecological Strategies and Knowledge Mapping 1025
J. Bidarra, A. Dias

Need for a Prescriptive Taxonomy of Interaction for Mathematical Cognitive Tools 1030
K. Sedig

Workshop on Computational Modeling of Transport on Networks

Evolution of the Internet Map and Load Distribution 1038
K.-I. Goh, B. Kahng, D. Kim

Complex Network of Earthquakes 1046
S. Abe, N. Suzuki

Universal Features of Network Topology 1054
K. Austin, G.J. Rodgers

Network Brownian Motion: A New Method to Measure Vertex-Vertex Proximity and to Identify Communities and Subcommunities 1062
H. Zhou, R. Lipowsky

Contagion Flow through Banking Networks 1070
M. Boss, M. Summer, S. Thurner

Local Search with Congestion in Complex Communication Networks 1078
A. Arenas, L. Danon, A. Díaz-Guilera, R. Guimerà

Guided Search and Distribution of Information Flow on Complex Graphs 1086
B. Tadić

Network Topology in Immune System Shape Space 1094
J. Burns, H.J. Ruskin

An Evolutionary Approach to Pickup and Delivery Problem with Time Windows 1102
J.-C. Créput, A. Koukam, J. Kozlak, J. Lukasik

Automatic Extraction of Hierarchical Urban Networks: A Micro-Spatial Approach 1109
R. Carvalho, M. Batty

Workshop on Modeling and Simulation in Supercomputing and Telecommunications

Design and Implementation of the Web-Based PSE *GridGate* 1117
K. Kang, Y. Kang, K. Cho

Performance Evaluation of ENUM Directory Service Design 1124
H.K. Lee, Y. Mun

A Single Thread Discrete Event Simulation Toolkit for Java: STSimJ 1131
W. Chen, D. Wang, W. Zheng

Routing and Wavelength Assignments in Optical WDM Networks with Maximum Quantity of Edge Disjoint Paths 1138
H. Choo, V.V. Shakhov

Parallelism for Nested Loops with Non-uniform and Flow Dependences 1146
S.-J. Jeong

Comparison Based Diagnostics as a Probabilistic Deduction Problem 1153
B. Polgár

Dynamic Threshold for Monitor Systems on Grid Service Environments 1162
E.N. Huh

Multiuser CDMA Parameters Estimation by Particle Filter with Resampling Schemes 1170
J.-S. Kim, D.-R. Shin, W.-G. Chung

Workshop on QoS Routing

Routing, Wavelength Assignment in Optical Networks Using an Efficient and Fair EDP Algorithm 1178
P. Manohar, V. Sridhar

Route Optimization Technique to Support Multicast in Mobile Networks 1185
K. Park, S. Han, B.-g. Joo, K. Kim, J. Hong

PRED: Prediction-Enabled RED 1193
M.G. Chung, E.N. Huh

An Efficient Aggregation and Routing Algorithm Using Multi-hop Clustering in Sensor Networks 1201
B.-H. Lee, H.-W. Yoon, T.-J. Lee, M.Y. Chung

Explicit Routing for Traffic Engineering in Labeled Optical Burst-Switched WDM Networks 1209
J. Zhang, H.-J. Lee, S. Wang, X. Qiu, K. Zhu, Y. Huang, D. Datta, Y.-C. Kim, B. Mukherjee

A Mutual Authentication and Route Optimization Method between MN and CN Using AAA in Mobile IPv6 1217
M. Kim, H.K. Lee, Y. Mun

Studies on a Class of AWG-Based Node Architectures for Optical Burst-Switched Networks 1224
Y. Huang, D. Datta, X. Qiu, J. Zhang, H.-K. Park, Y.-C. Kim, J.P. Heritage, B. Mukherjee

Self-Organizing Sensor Networks 1233
D. Bein, A.K. Datta

Workshop on Evolvable Hardware

The Application of GLS Algorithm to 2 Dimension Irregular-Shape Cutting Problem 1241
L. Budzyńska, P. Kominek

Biologically-Inspired:
A Rule-Based Self-Reconfiguration of a Virtex Chip 1249
G. Tufte, P.C. Haddow

Designing Digital Circuits for the Knapsack Problem 1257
M. Oltean, C. Groşan, M. Oltean

Improvements in FSM Evolutions from Partial Input/Output Sequences 1265
S.G. Araújo, A. Mesquita, A.C.P. Pedroza

Intrinsic Evolution of Analog Circuits on a Programmable Analog Multiplexer Array 1273
J.F.M. Amaral, J.L.M. Amaral, C.C. Santini, M.A.C. Pacheco, R. Tanscheit, M.H. Szwarcman

Encoding Multiple Solutions in a Linear Genetic Programming Chromosome 1281
M. Oltean, C. Groşan, M. Oltean

Evolutionary State Assignment for Synchronous Finite State Machines . . . 1289
N. Nedjah, L. de Macedo Mourelle

Author Index 1297

The New Graphic Description of the Haar Wavelet Transform

Piotr Porwik[1] and Agnieszka Lisowska [2]

[1] Institute of Informatics, Silesian University, ul.B dzi ska 39,
41-200 Sosnowiec, Poland
porwik@us.edu.pl
[2] Institute of Mathematics, Silesian University, ul.Bankowa 14,
40-007 Katowice, Poland
alisow@ux2.math.us.edu.pl

Abstract. The image processing and analysis based on the continuous or discrete image transforms are the classic processing technique. The image transforms are widely used in image filtering, data description, etc. The image transform theory is a well known area, but in many cases some transforms have particular properties which are not still investigated. This paper for the first time presents graphic dependences between parts of Haar and wavelets images. The extraction of image features immediately from spectral coefficients distribution has been shown. In this paper it has been presented that two-dimensional both, the Haar and wavelets functions products, can be treated as extractors of particular image features.

1 Introduction

The computer and video-media applications have developed rapidly the field of multimedia, which requires the high performance, speedy digital video and audio capabilities. The digital signal processing is widely used in many areas of electronics, communication and information techniques [1,2,3,6,12]. In the signals compression, filtration, systems identification, the commonly used transforms are based on sinusoidal basic functions such as: Discrete Fourier, Sine or Cosine Transform or rectangular basic functions: Discrete Walsh and Wavelet Transform, (Haar, Daubechies, etc.) [2,3,7]. All these functions are orthogonal, and their transforms require only additions and subtractions. It makes that it is easy to implement them on the computer. It not only simplifies computations but also permits to use different (linear and nonlinear) filters [3,4,9] to get the spectrum. One should remember that researches in this topic are still in progress and new improvements have been found [5,8,9].

Fourier methods are not always good tools to recapture the non-smooth signal [2]; too much information is needed to reconstruct the signal locally. In these cases the wavelet analysis is often very effective because it provides a simple approach for dealing with the local aspects of signal, therefore particular properties of the Haar or wavelet transforms allow analyzing original image on spectral domain effectively.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1–8, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 The Discrete Haar and Wavelet Transforms

Alfred Haar in [7] has defined a complete orthogonal system of functions in $L^p([0,1])$, $p \in [1,\infty]$. Nowadays, in the literature there are some other definitions of the Haar functions [3,12]. Discrete Haar functions can be defined as functions determined by sampling the Haar functions at 2^n points. These functions can be conveniently represented by means of matrix form. Each row of the matrix $\mathbf{H}(n)$ includes the discrete Haar sequence $haar(w,t)$ (or otherwise the discrete Haar function). In this notation, index w identifies the number of the Haar function and index t discrete point of the function determination interval. In this case, the Haar matrix of any dimension can be obtained by the following recurrence relation:

$$\mathbf{H}(n) = \begin{bmatrix} \mathbf{H}(n-1) & \otimes[1 \quad 1] \\ 2^{(n-1)/2}\mathbf{I}(n-1) & \otimes[1 \quad -1] \end{bmatrix}, \qquad \mathbf{H}(0) = 1 \tag{1}$$

and: $\mathbf{H}(n) \neq \mathbf{H}(n)^T$ for $n > 1$ and $[\mathbf{H}(n)]^{-1} = 2^{-n} \cdot \mathbf{H}(n)^T$,

where: $\mathbf{H}(n)-$ matrix of the discrete Haar functions of degree 2^n, $\mathbf{I}(n)-$ identity matrix of degree 2^n, $\otimes-$ the Kronecker (tensor) product.

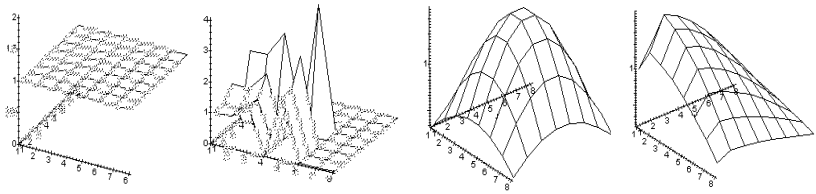
Definition 1. Two-dimensional $N \times N = 2^n \times 2^n$ forward and inverse Discrete Haar Transform can be defined in matrix notation as:

$$\mathbf{S} = a \cdot \mathbf{H}(n) \cdot \mathbf{F} \cdot a \cdot \mathbf{H}(n)^T, \qquad \mathbf{F} = b \cdot \mathbf{H}(n)^T \cdot \mathbf{S} \cdot b \cdot \mathbf{H}(n), \tag{2}$$

where: $\mathbf{F}$ – the image in matrix form. The matrix has dimension $N \times N$ pixels. $\mathbf{S}$ – the spectrum matrix, and $a \cdot b = 1/N$. Hence a or b parameters can be defined as values: 1/N, $1/\sqrt{N}$ or 1, $n = \log_2 N$.

Fig. 1 presents some known transforms of a test image. The test image contains a simple test impulse represented as 8×8 matrix, which has 0 values everywhere, except the upper left element, which has the value of 8. From Fig. 1 we can observe that all N^2 elements of these transforms are nonzero except the Haar transform, which has only 2N nonzero entries. These features are very important in image processing and convenient from image compression point of view. The energy distribution informs us where there are situated the important features of image [2,10,12].

It is easy to observe from Fig.1 that the Walsh transform gives the worst results here: distribution of spectral energy is uniformable. In c) and d) cases distribution of spectral energy has sharply outlined maximum, outside of which, one can observe the decrease of energy. The distribution of the Haar spectrum is not proper too, but we can treat this transform differently. Presented discrete transforms, enable us to observe where energy concentrations occur but from this representation, it is not possible to find more precisely information about real image. For example, it is difficult to point places, which describe horizontal, vertical, etc. details of real image. These troubles can be overcome by well known multiresolution analysis [3,5].



a) b) Note: $a = 1/\sqrt{N}$ c) d)

Fig. 1. The **S** transform of image containing the test impulse: a) Walsh-Hadamard; b) Haar; c) DST (Discrete Sine Transform); d) DCT (Discrete Cosine Transform)

The motivation for usage of the wavelet transform is to obtain information that is more discriminating by providing a different resolution at different parts of the time-frequency plane. The wavelet transforms allow partitioning of the time-frequency domain into non-uniform tiles in connection with the time-spectral contents of the signal. The wavelet methods are connected with classical basis of the Haar functions – scaling and dilation of a basic wavelet can generate the basis Haar functions.

Any Haar function basis (1) can be generated as: $\psi_i^j(t) = \sqrt{2^j}\psi(2^j t - i)$, $i = 0,1,...,2^j - 1$, $j = 0,1,...,\log_2 N - 1$, or generally $\psi_i^j(t) = haar(2^j + i, t)$. From this example follows that functions $\psi_i^j(t)$ are orthogonal to one another. Hence, we obtain linear span of vector space $W^j = spn\{\psi_i^j\}_{i=0,...,2^j-1}$. A collection of linearly independent functions $\{\psi_i^j(t)\}_{i=0,...,2^j-1}$ spanning W^j we called wavelets. The Haar scaling function is defined by the formula: $\phi_i^j(t) = \sqrt{2^j}\phi(2^j t - i)$, $i = 0,1,...,2^j - 1$, $j = 0,1,...,\log_2 N - 1$. The index j refers to dilation and index i refers to translation [3,11]. Hence, we obtain linear span of vector space $V^j = spn\{\phi_i^j\}_{i=0,...,2^j-1}$. The basic functions from the space V^j are called scaling functions. In multiresolution analysis the Haar basis has important property of orthogonality: $V^j = V^{j-1} \oplus W^{j-1}$. The space W^j can be treated as the orthogonal complement of V^j in V^{j+1}. So, the basis functions of W^j together with the basis functions of V^j form a basis for V^{j+1}.

3 The Haar and Wavelet Basic Images

Due to its low computing requirements, the Haar transform has been mainly used for image processing and pattern recognition. From this reason two dimensional signal processing is an area of efficient applications of Haar transforms due to their wavelet-like structure. Because $\mathbf{H}(n)$ and $\mathbf{H}(n)^T$ are the square matrices, their product is commutative, therefore equations (2) can be rewritten and expressed as:

$$s(k,m)=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\times haar(k,x)\times haar(m,y) \tag{3}$$

where: $\mathbf{S}=[s_{km}]$, $\mathbf{F}=[f_{xy}]$, $x,y,k,m\in\{0,1,...,N-1\}$.

Basing on equation of analysis (2) we can conclude that in 2D spectral domain the values of coefficients s_{ij} depend on appropriate product of the two Haar functions. Fig. 2 presents an example of product of the arbitrary selected Haar functions.

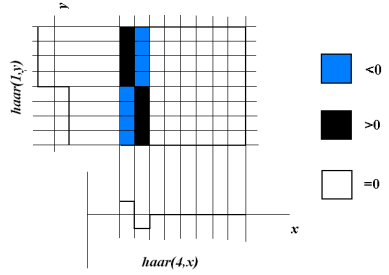


Fig. 2. The example of product of two discrete Haar functions

Because this product is multiplied by image matrix, the result of such multiplication can be treated as a particular extractor – it can be used to locate the specific edges hidden in image. By looking for all coefficients in the spectral space, one can find all important edge directions in the image. In this case, we must find decomposition matrices of matrix $\mathbf{H}(n)$. For last decomposition level, it can be noticed that $\mathbf{M}_n=\mathbf{H}(n)$. If each orthogonal matrix $\mathbf{M}_i$, $i=1,2,3$ one multiplies by $1/\sqrt{2}$ factor, then procedure of calculations will be according to the classical Mallat algorithm [11]. The product of the decomposition levels for all 2D Haar functions (for case $N=8$) is shown in Fig. 3 – the pictures have been composed on the basis of $\mathbf{M}_i$ matrices and the method shown in Fig. 2. From Fig. 3 we can conclude that the classical Haar transform gives different spectral coefficients on different decomposition levels. The construction of decomposition matrices can be as follows:

Step 1. According to the formula $V^n=V^{n-1}\oplus W^{n-1}$, the matrix $\mathbf{M}_1$ has a form $\mathbf{M}_1=[\phi^{n-1}_{j=0,...,2^{n-1}-1}\subset V^{n-1},\psi^{n-1}_{j=0,...,2^{n-1}-1}\subset W^{n-1}]^T$.

Step 2. Because $V^{n-1}=V^{n-2}\oplus W^{n-2}\oplus W^{n-1}$, the matrix $\mathbf{M}_2$ can be constructed as follows $\mathbf{M}_2=[\phi^{n-2}_{j=0,...,2^{n-2}-1}\subset V^{n-2},\psi^{n-2}_{j=0,...,2^{n-2}-1}\subset W^{n-2},\psi^{n-1}_{j=0...,2^{n-1}-1}\subset W^{n-1}]^T$.

Step n. Finally, after n steps of calculations, we can construct the formula $V^1=V^0\oplus W^0\oplus W^1\oplus W^2\oplus...\oplus W^{n-1}$, hence the matrix $\mathbf{M}_n$ has a structure $\mathbf{M}_n=[\phi^0_0\subset V^0,\psi^0_0\subset W^0,\psi^1_{j=0,1}\subset W^1,\psi^2_{j=0,...,3}\subset W^2,...,\psi^{n-1}_{j=0,...,2^{n-1}-1}\subset W^{n-1}]^T$.

Example 1. Let $n = 3$ then:

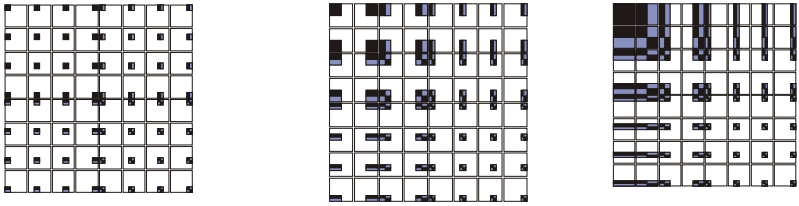
$$V^3 = V^2 \oplus W^2, \qquad \mathbf{M}_1 = [\phi_0^2, \phi_1^2, \phi_2^2, \phi_3^2, \psi_0^2, \psi_1^2, \psi_2^2, \psi_3^2]^T,$$

$$V^2 = V^1 \oplus W^1 \oplus W^2, \qquad \mathbf{M}_2 = [\phi_0^1, \phi_1^1, \psi_0^1, \psi_1^1, \psi_{j=0...,3}^2 \subset W^2]^T,$$

$$V^1 = V^0 \oplus W^0 \oplus W^1 \oplus W^2, \qquad \mathbf{M}_3 = [\phi_0^0, \psi_0^0, \psi_{j=0,1}^1 \subset W^1, \psi_{j=0,...3}^2 \subset W^2]^T,$$

$$\mathbf{M}_1 = \begin{bmatrix} 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} & \sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix},$$

$$\mathbf{M}_3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix}.$$



a) b) c)

Fig. 3. The 2D Haar functions product treated as extractors. Decomposition levels: a) first, b) second, c) third

One advantage of the method presented above is that often a large number of the detail coefficients turn out to be very small in magnitude, as in the example of Fig. 1. Truncating, or removing, these small coefficients introduce only small errors in the reconstructed image. Additionally, we can control which coefficients will be removed, because its distribution is known (Fig. 3).

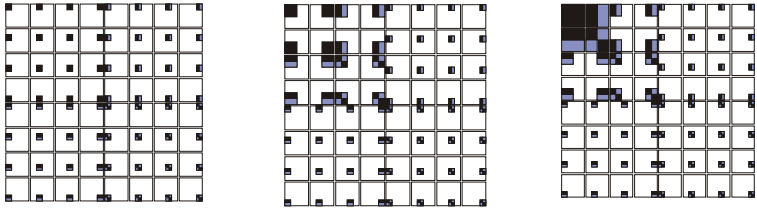
Basing on the facts that $W^j = spn\{\varphi_i^j\}_{i=0,...,2^j-1}$, $V^j = spn\{\phi_i^j\}_{i=0,...,2^j-1}$ we can express functions ϕ and ψ as a linear combination of the basis functions from V and W spaces. Let us denote **F** as an image in matrix form and define the operators:

$$\mathbf{A}(i) = 1/\sqrt{2} \cdot [\mathbf{F}(2i) + \mathbf{F}(2i+1)], \qquad \mathbf{D}(i) = 1/\sqrt{2} \cdot [\mathbf{F}(2i) - \mathbf{F}(2i+1)], \qquad (4)$$

where: $\mathbf{F}(i)$ – vector of size N, containing row or column of matrix **F**, $i \in \{0,1,...,N/2-1\}$, $\mathbf{A}(i)$ – vector of size $N/2$, containing approximation coefficients, $\mathbf{D}(i)$ – vector of size $N/2$, containing detail coefficients.

To get wavelet decomposition on the first level of an image **F** (the spectrum matrix called $\mathbf{S}_1$) we first apply the operators (4) to all columns of the matrix and then to all rows [3,8,11]. To get the second level of wavelet decomposition (matrix $\mathbf{S}_2$) one can apply similar analysis to upper left sub-matrix of size $\frac{N}{2}\times\frac{N}{2}$ of matrix $\mathbf{S}_1$. And generally, to get k-th level – matrix $\mathbf{S}_k$, one can apply this analysis to upper left sub-matrix of size $\frac{N}{2^{k-1}}\times\frac{N}{2^{k-1}}$ of matrix $\mathbf{S}_{k-1}$, where $k \in \{1,...,\log_2 N\}$.

Note, that applying filters (4) to an image, give the same results as multiplying matrices $\mathbf{S}_1 = \frac{1}{8}\mathbf{M}_1 \cdot \mathbf{F} \cdot \mathbf{M}_1^T$, where matrix $\mathbf{M}_1$ is taken from Example 1. Therefore, $\mathbf{S}_1$ may be treated as extractor of image features on the first level of wavelet decomposition, similar as above in the Haar decomposition case. Because on the second and next levels only the part of a matrix is transformed (opposite to Haar decomposition) these extractors on these levels are different. For example, for N=8 the products of the non-standard wavelet decomposition levels are shown in Fig. 4.



a) b) c)

Fig. 4. The 2D wavelet functions product treated as extractors. Decomposition levels: a) first, b) second, c) third

All considerations, which have been presented until now for the classical of Haar functions, have applications in that case too, with the exception of extractors' distribution (Fig. 4). The Haar decomposition can be simply implemented as matrix multiplication. The wavelet algorithm is a little more efficient.

4 Experimental Results

To test our method the well known benchmarks have been used. Each of these images was of size $a \times a \times 8$ bits, where $a \in \{32, 64, 128, 256\}$. By analysing the Figs. 3-4 we can divide areas of a figure into 4 equal pieces. Each piece has dimension $(N/2)\times(N/2)$ and is called *A*, *H*, *V* and *D*. Location of these areas presents Fig. 5.

Each piece (A, H, V or D) for $N = 8$ includes sixteen appropriate sub-squares from Fig. 3-4. According to presented arguments, mentioned areas possess different features: A (Approximation), H (Horizontal), V (Vertical), D (Diagonal). Fig. 5 presents "Baboon" – one of the grey-level test images and its wavelet and Haar spectra. The spectra images are different what directly follows from Figs. 3-4. Taking into account mentioned features of areas some differences between spectres can be shown.

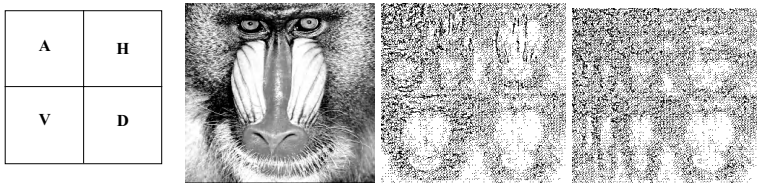


Fig. 5. Principle of spectra partitioning; original image and its wavelet and Haar spectra respectively

In Fig. 6 are shown differences between origin image and compressed one for wavelet and Haar method of analysis, respectively after removing some coefficients.

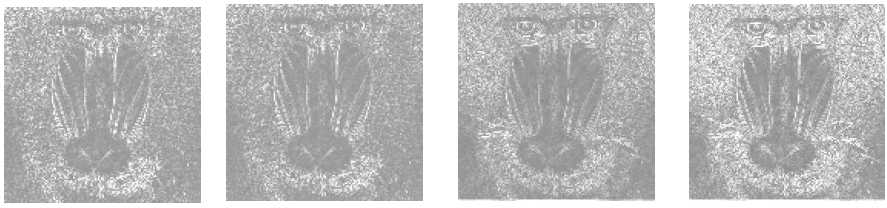
The exact information about distribution of spectral coefficients allows us to match easily up the compression ratio with the type of image. Obtained results for Haar matrix-based method and wavelet method were compared by means of PSNR coefficients. These results of investigations are collected in Tab. 1. From this table one can see that Haar reconstructed images have slightly better quality. From Tab. 1 (last column) follows, that after removing all horizontal and vertical details on the first level of decomposition we get exactly the same PSNR of both methods reconstructed images because of the proportionality of the diagonal detail coefficients.

The different cases of removing the spectral coefficients can be applied as well. These entire processes are based on the fact, that appropriate selection and modification of the spectral coefficients may preserve the contents of the image.

Between Haar matrix-based method and the wavelet one can be observed quantitative and graphic relationship. Let d_H and d_W stand for diagonal coefficients from Haar and wavelet spectrum matrix respectively, both of degree 2^n. Then $d_H = 2^n d_W$.

Table 1. The PSNR of reconstructed images after appropriate details elimination

Method \ Details	Horizontal (H)	Vertical (V)	Diagonal (D)	Horizotal +Vertical (H+V)
Wavelet decomposition	29,7254	27,3697	31,4822	25,3813
Haar decomposition	29,7269	27,3702	31,4827	25,3813



a) b)

Fig. 6. Horizontal – (a) and horizontal with vertical – (b) details elimination and loosed information after applied wavelet and Haar matrix-based method, respectively

5 Concluding Remarks

In the paper it has been shown the new graphic way of presentation of decomposition levels for both the Haar matrix-based method and wavelets. As it has been shown both methods can be modelled on the basis of the wavelets theorem.

The 2D Haar matrix method of calculations like the 2D Fast Fourier Transform has complexity $O(4N^2 \log_2 N)$ [4], classical 2×1 D fast wavelet method of calculations has complexity $O(16/3N^2)$ only [3,11,12]. This complexity can be decreased to $O(14/3N^2)$ by suitable organization of calculations [10]. Described complexity factors are determined as number of additions and multiplications in computation process. The graphic distribution of the Haar-wavelet spectral coefficients also has been presented. Additionally, knowledge about spectra distribution allows us to point appropriate selection or modification (reduction) of the Haar-wavelet coefficients.

References

1. Addison P. S., Watson J. N., Feng T.: Low-Oscillation Complex Wavelets, Journal of Sound and Vibration, Vol. 254 (4), 733-762. (2002)
2. Ahmed N., Rao K. R.: Orthogonal Transforms for Digital Signals Processing. Springer-Verlag. Berlin, Heidelberg. (1975)
3. Castleman K.R.: Digital Image Processing. Prentice-Hall. New Jersey. (1996)
4. Daubechies I.: Recent results in wavelet applications, Journal of Electronic Imaging, Vol. 7 No. 4, pp. 719-724. (1998)
5. Drori I., Lischinski D.: Fast Multiresolution Image Operations in the Wavelet Domain, IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 3, July-September, pp. 395-411. (2003)
6. Harmuth H.F.: Sequence Theory. Foundations and applications. Academic Press. New York. (1977)
7. Haar A.: Zur Theorie der orthogonalen Functionsysteme. Math Annal. No 69, pp. 331-371. (1910)
8. Jorgensen P.: Matrix Factorizations, Algorithms, Wavelets, Notices of the American Mathematical Society, Vol. 50, No 8, September, pp. 880-894. (2003)
9. Lisowska A.: Nonlinear Weighted Median Filters in Dyadic Decomposition of Images, Annales UMCS Informatica AI, Vol. 1, pp.157-164. (2003)
10. Lisowska A., Porwik P.: New Extended Wavelet Method of 2D Signal Decomposition Based on Haar Transform. Mathematics and Computers in Simulation. Elsevier Journal. (to appear)
11. Mallat S. A.: Theory for Multiresolution Signal Decomposition: The Wavelet Representation, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 11,
12. Walker J. S.: Fourier Analysis and Wavelet Analysis. Notices of the American Mathematical Society. Vol. 44, No 6, pp. 658-670. (1997)

On New Radon-Based Translation, Rotation, and Scaling Invariant Transform for Face Recognition

Tomasz Arodź[1,2]

[1] Institute of Computer Science, AGH, al. Mickiewicza 30, 30-059 Kraków, Poland
[2] Academic Computer Centre – CYFRONET, Nawojki 11, 30-950 Kraków, Poland
gefarodz@cyf-kr.edu.pl

Abstract. The Radon transform has some interesting properties concerning the scaling, rotation-in-plane and translation of the input image. In the paper, these properties are a basis for deriving a transformation invariant to the aforementioned spatial image variations, a transformation that uses direct translation, angle representation and 1-D Fourier transform. As the face images often differ in pose and scale of the face, such a transformation can ease the recognition task. Experimental results show that the proposed method can achieve 96% and 89% recognition accuracy for, respectively, uniformly and non-uniformly illuminated images.

Keywords: Face recognition; Radon transform; Invariant recognition

1 Introduction

In face recognition, one often encounters the problem that pictures of the same person are taken in different conditions. These include variations in facial expression, illumination as well as spatial location and pose of the face in the picture. The latter ones consist of translation and scaling of the face, rotation on-the-plane and rotation in depth. In this paper, a method for compensating translation, scaling and rotation-on-the-plane will be shown. The goal is to present a transformation which will give identical results if applied to a pair of images that are in the similarity relation with each other. The result of such a transformation form a new, invariant set of features for recognition.

Several approaches to the problem of invariant object recognition have been proposed up to date. These include group of methods based on the Fourier and log-polar or the Fourier-Mellin transform [6,8] or Taylor and Hessian invariants [2]. Methods based on algebraic moments, e.g. the Zernike [11] or Hu [3] moments are also proposed. Another approach, using only object boundary has been proposed, e.g. a method utilizing wavelets [4] or multi-vector eigenvector shape descriptors [5]. Other methods, using e.g. a group of random lines through the object are also used [10]. The Radon transform has been used as part of the invariant transform [7], albeit not in the face recognition. In this paper the Radon transform will be used as a preliminary step for deriving the invariance.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 9–17, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The paper is arranged as follows. In Sect. 2 the Radon transform is studied and subsequently the full invariance is derived. Section 3 discusses implementation choices and the experimental results for Yale [1] dataset. Finally, Sect. 4 concludes the paper.

2 Method Details

The proposed method consists of two main steps. First, the Radon transform of the image is calculated. Then, the result of the transform is processed to achieve full similarity invariance, using direct translation, angle representation and 1-D Fourier transform.

2.1 Radon Transform

The behaviour of the Radon transform for translated, rotated and scaled images is presented in this section. The result of the Radon transform RAD of the image $f : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is a function $g : \mathbb{R} \times [0, 2\pi) \to \mathbb{R}_+$ defined as:

$$g(s, \theta) = RAD(f(x, y)) = \int_{-\infty}^{\infty} f(s\cos\theta - u\sin\theta, s\sin\theta + u\cos\theta)\,\mathrm{d}u \quad , \tag{1}$$

where:

$$\begin{bmatrix} s \\ u \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad . \tag{2}$$

Given the original image f in Cartesian $(f(x, y))$ and polar $(f_{\text{polar}}(r, \phi))$ coordinates, the following images and their Radon transforms can be defined [7]:

- translated image: $RAD(f(x - x_0, y - y_0)) = g(s - x_0\cos\theta - y_0\sin\theta, \theta)$,
- rotated image: $RAD(f_{\text{polar}}(r, \phi + \theta_0)) = g(s, (\theta + \theta_0) \bmod 2\pi)$,
- scaled image: $RAD(f(\alpha x, \alpha y)) = \frac{1}{|\alpha|} g(\alpha s, \theta)$.

The translation in any direction in spatial domain leads to translation in the s direction in the Radon domain, with the translation value varying with the θ dimension. The rotation in spatial domain leads to circular translation along the θ axis in the Radon domain. Finally, the scaling along both axes in the spatial domain results in the scaling along the s axis in the Radon domain and scaling of the value of the transform. These properties are depicted in the Fig. 1.

2.2 Achieving Full Similarity-Transform Invariance

The result of the Radon transform is further processed to obtain translation, rotation and scale invariance. To simplify the notation, the $g(s)$ will denote the values of $g(s, \theta)$ for any specified θ in places where it will not lead to confusion.

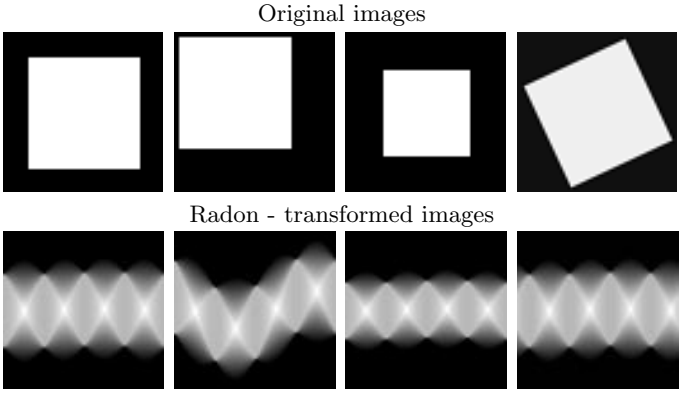


Fig. 1. Effects of the Radon transform for different image transformations

Translation Invariance. To achieve the translation invariance, the result g of the Radon transform is directly translated along the s axis. The value of the translation $s_{\mathrm{g}}(\theta)$ depends on the θ dimension. The resulting function g_{t} is defined as follows:

$$g_{\mathrm{t}}(s) = g(s + s_{\mathrm{g}}), \quad \text{where} \quad s_{\mathrm{g}}(\theta) = \min\{s : g(s, \theta) > 0\} \quad . \tag{3}$$

The function g_{t} is translation invariant and preserves the scale variance.

Theorem 1. *Let $g(s)$ and $G(\alpha s + S)$ be two functions that are translated and scaled version of each other, i.e. $\alpha g(s) = G(\alpha s + S)$. The functions g_{t} and G_{t}, as defined in (3), are scaled version of each other.*

The proof of the theorem is straightforward and is omitted for brevity.

Scale Invariance. Scale invariance is based on the method that can be called angle-representation. The process of achieving translation and scale invariance is depicted in Fig. 2.

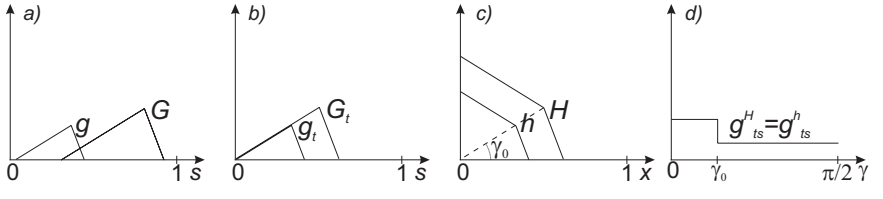


Fig. 2. Achieving invariance to translation: a, b and scale: b, c, d

In the method, a group of functions defined on the basis of the function g_{t} is used. Let:

$$g_{\mathrm{rev}}(s) = g_{\mathrm{t}}(1-s), \quad \text{where } s \in [0,1] \ ; \tag{4}$$

$$h_{\mathrm{rev}}(x) = \int_0^x \left| \frac{\mathrm{d}g_{\mathrm{rev}}}{\mathrm{d}s} \right| \mathrm{d}s, \quad \text{where } x \in [0,1] \ ; \tag{5}$$

$$h(x) = h_{\mathrm{rev}}(1-x) \quad . \tag{6}$$

It can be shown that these functions have the following properties.

Theorem 2. *If one assumes that function $g_{\mathrm{t}}(s)$ has continuous first derivative in $[0,1]$ then the function $h(x)$ is well defined, differentiable, nonincreasing and nonnegative in $[0,1]$. Also, $h(1) = 0$ and $h(0) > 0$.*

The function h can be used to derive scale invariant transforms. First, let h_{P} be a variant of the function h narrowed to the domain $(0, X]$, where $X = \min(\{x : h(x) = 0\})$. Let $g_{\mathrm{ts}}^{\mathrm{h}}$ be a function $g_{\mathrm{ts}}^{\mathrm{h}} : \left[0, \frac{\pi}{2}\right) \to \mathbb{R}_+$ defined as:

$$g_{\mathrm{ts}}^{\mathrm{h}}(\gamma) = |h'_{\mathrm{P}}(x)|, \quad \text{where } \gamma = \arctan\left(\frac{h_{\mathrm{P}}(x)}{x}\right) \quad . \tag{7}$$

It can be shown the function $g_{\mathrm{ts}}^{\mathrm{h}}$ is scale-invariant.

Theorem 3. *Let $G_{\mathrm{t}}(y)$ and $g_{\mathrm{t}}(x)$ be two functions with the translation removed and meeting the constraints of Theorem 2, such that $G_{\mathrm{t}}(y) = \alpha g_{\mathrm{t}}(x)$ and $y = \alpha x$, i.e two functions that are scaled versions of each other. The function $g_{\mathrm{ts}}^{\mathrm{h}}$ is scale invariant, i.e. $g_{\mathrm{ts}}^{\mathrm{H}}(\gamma) = g_{\mathrm{ts}}^{\mathrm{h}}(\gamma)$, where the functions h and H are defined on the basis of g_{t} and G_{t} according to (6).*

Proof. Since the function h_{P} is derived from g_{t} with only integration, differentiation, reorientation of the axis and narrowing of the domain, the functions h_{P} and H_{P} are also scaled versions of each other, i.e. $H_{\mathrm{P}}(y) = \alpha h_{\mathrm{P}}(x)$. Thus:

$$\gamma_{\mathrm{H}} = \arctan\left(\frac{H_{\mathrm{P}}(y)}{y}\right) = \arctan\left(\frac{\alpha h_{\mathrm{P}}(x)}{\alpha x}\right) = \arctan\left(\frac{h_{\mathrm{P}}(x)}{x}\right) = \gamma_{\mathrm{h}} \ , \tag{8}$$

$$g_{\mathrm{ts}}^{\mathrm{H}}(\gamma) = \frac{\mathrm{d}H}{\mathrm{d}y} = \frac{\mathrm{d}H(\alpha x)}{\mathrm{d}\alpha x} = \frac{1}{\alpha}\frac{\mathrm{d}\alpha h(x)}{\mathrm{d}x} = \frac{\mathrm{d}h}{\mathrm{d}x} = g_{\mathrm{ts}}^{\mathrm{h}}(\gamma) \quad . \tag{9}$$

Therefore, $g_{\mathrm{ts}}^{\mathrm{h}}$ is indeed scale invariant. □

For clarity, the $g_{\mathrm{ts}}^{\mathrm{h}}(\gamma)$, or simply g_{ts}, have been derived for 1-D function $g_{\mathrm{t}}(s) = g_{\mathrm{t}}(s, \theta = \mathrm{const})$, but in fact it is defined in 2-D: $g_{\mathrm{ts}}(\gamma, \theta)$ as g_{t} is.

Rotation Invariance. In order to eliminate the rotation variance, modulus of the one-dimensional discrete Fourier transform is used along the θ axis.

It can be shown [11], that for the discrete function $f : [0, X] \to R$ the magnitude of the Fourier transform is invariant with respect to circular translation:

$$|DFT\left(f\left(x\right)\right)| = |DFT\left(f\left(\left(x + x_0\right) \bmod X\right)\right)| \quad . \tag{10}$$

The rotation in the image is reduced to the circular translation in the θ direction by the Radon transform and is preserved in this form by translation and scale variance elimination, both operating along the s axis. Therefore, the function:

$$G_{\mathrm{tsr}}\left(\gamma, \Theta\right) = |DFT_{\gamma}\left(G_{\mathrm{ts}}\left(\gamma = \mathrm{const}, \theta\right)\right)| \quad . \tag{11}$$

where G_{ts} is a discrete approximation of g_{ts}, is translation, scale and rotation invariant.

3 Experimental Results and Discussion

The developed invariant transform can be applied to face recognition, either as a pre-processing for methods that are not invariant or as a stand-alone method.

In the latter case, the transform G_{tsr} of input image has to be compared with a set of labelled, transformed sample images. This can be done using the nearest-neighbour decision rule, i.e. the person minimizing the distance is chosen. Several metrics for transformed images has been evaluated, i.e. the Euclidean and Manhattan distance and the Tanimoto dissimilarity measure, defined as the inverse of the Tanimoto similarity measure [9].

It should be noted that apart from spatial variance, other variances, such as illumination, exist in face images. The issue of eliminating the effects of different lighting in images is beyond the scope of this paper, but simple histogram equalization is used as a pre-processing step to compensate for minor lighting variances. Also, simple wavelet-based method is used in some experiments.

The experiments were conducted on the Yale faces dataset [1], consisting of 11 pictures for each of 15 individuals. These images have relatively high resolution and consist of the whole head and neck of a person (see Fig. 3).

Fig. 3. Faces from the Yale dataset [1]

Table 1. Results for Yale dataset

	3-NN, leave-one-out, 9 imgs/person			8 test and 1 sample imgs/person		
	Tanimoto	Euclidean	Manhattan	Tanimoto	Euclidean	Manhattan
256×256	0.941	0.933	0.919	0.909	0.892	0.892
64×64	0.963	0.956	0.963	0.900	0.867	0.850
32×32	0.919	0.933	0.933	0.775	0.758	0.775

3.1 Experimental Setup

Two pictures for each person from the Yale dataset, containing faces with side illumination has been eliminated from the tests. Two experimental configurations were used. In the first one the "leave-one-out" cross-validation was used to obtain the recognition accuracy. In the second configuration the faces were split into two groups, the sample images database containing 1 image per person and a test set containing remaining 8 images of each person. The picture labelled "normal" in the Yale set, was chosen as a sample image. In both configurations, the nearest-neighbour decision rule was used.

The images consist of a person's head on black background. All heads in the test and sample sets were randomly rotated in the range of $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, scaled in the range of $[66\%, 100\%]$ and placed randomly in the picture.

The images, after histogram equalization, were transformed using the Radon transform to obtain $256\sqrt{2} \times 256$ images. Since the data points in the angle representation after the application of scale-invariance transform are not regularly spaced along the γ axis, piecewise cubic Hermite interpolation was used to obtain regular grid of the size 256×256 pixels. Finally, modulus of the FFT along the Θ dimension was calculated.

3.2 Discussion of Results

The results of the tests for the Yale images are summarized in the Table 1. Since the calculation of the Radon transform of the 256×256 image is time consuming, the method has been applied to images downsized to the size 64×64 and 32×32. While in the nearest-neighbour scenario the reduction of size slightly increases the performance of the recognition, in the scenario with single sample image per person a decrease in accuracy can be observed, especially for 32×32 images. This decrease is caused by the difference in the scale among the images. The relation of the decrease in accuracy to the variations in scale between the faces can be observed with results for 32×32 images randomly rotated and translated, but with no random scaling introduced. For such pictures, the decrease of recognition accuracy is not present, as the recognition accuracy ranges from 0.88 to 0.9 depending on the metric used in the test.

Unlike the translation and rotation, the method for achieving scale invariance utilizes a transformation from the spatial domain to the angle domain. As the size of the image becomes small, the discrete nature of the image becomes more

evident. First, the Radon transform becomes less accurate, which affects also the rotation invariance. Next, the transformation from the Radon spatial domain to the angle domain becomes less stable, e.g. the γ parameter in (7) cannot be determined with good accuracy. Finally, the quality of the interpolation used to change the irregular points along the γ axis into a regular grid deteriorates. While the scale invariance method is used also during the tests with no random scaling, since the scaling in all images is the same, the inaccuracies introduced by the transformation become systematic and do not influence the results significantly. A similar effect can be observed in the nearest-neighbour scenario. Since the number of images that an image is compared with is large, there always exist an image with similar scale of face, for which the errors of transformation are of the same nature as for the tested image. Thus, the recognition rate is not decreased.

Finally, to allow for comparison with other methods, the results for the full Yale set, including two side-illuminated images per person were computed. To deal with the large illumination variance, a wavelet-based method for compensation of illumination effects was used in addition to histogram equalization. This method is labelled HEQ-WV, while the normal histogram equalization is labelled HEQ-64. The results for the images using the nearest-neighbour method are presented in Table 2.

Table 2. Results for 3-NN,leave-one-out, 11 images per person, Yale dataset

	Tanimoto	Euclidean	Manhattan
HEQ-64	0.830	0.824	0.818
HEQ-WV	0.891	0.885	0.879

The results of the tests summarized above allow for some insight into the optimal configuration of the method. Usually the Tanimoto and Euclidean metrics allow for better recognition accuracy than the Manhattan metric. In case of large illumination variance, the wavelet based method for illumination compensation outperforms the histogram equalization. In case the method is used with a large set of sample images, as in the "leave-one-out" method, the optimal input image size is 64×64. In case of small samples set size, larger input images yield better results.

3.3 Comparison with Other Methods

The proposed method operating on randomly translated, rotated and scaled faces has been compared with results for other methods cited after [6]. Nearest-neighbour method with "leave-one-out" cross-validation technique has been used for the measurement of the recognition accuracy. Two representative methods were used: Eigenface (PCA) and Fisherface. The results are summarized in Table 3. It should be noted that, since the proposed method does not claim to be illumination invariant, the tests of the method were conducted on a test set

with the two images per person, containing right- and left-side illumination, eliminated. Results of the best performing variant of the method, i.e. images of size 64×64 and Tanimoto metric were used. Additionally, for ease of comparison, the results for the whole set, including the two aforementioned images are also presented for the images of size 256×256 and the Tanimoto metric. The results for the benchmark methods are for full Yale dataset, but contrary to the tests of the presented method, the images are not translated, rotated nor scaled.

Table 3. Comparison with other methods using Yale dataset

Method	Recognition accuracy
Proposed method (no side-illuminated images)	96%
Proposed method	89%
Eigenface [6]	81%
Eigenface w/o 1st three components[6]	89%
Fisherface [6]	94%

4 Conclusions

The new Radon-based face recognition method has been proposed. It has the property of invariance with respect to spatial image translation, on-the-plane rotation and scaling. While allowing for recognition of faces pictured in different positions, rotation and scale, the proposed method gives results comparable or event better than some existing, established non-invariant methods. The proposed transformation can be used as a stand-alone method for face recognition or as a preliminary step providing an invariant feature set for some other non-invariant methods.

Acknowledgements. The author would like to thank prof. Witold Dzwinel for guidance. The author is also grateful to Mr. Marcin Kurdziel for his remarks.

References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 19(7) pp. 711-720
2. Brandt, R.D., Lin, F., 1996. Representations that uniquely characterize images modulo translation, rotation and scaling. Pattern Recognition Letters 17 pp. 1001-1015
3. Hu, M.K., 1962. Visual pattern recognition by moment invariants. IEEE Trans. Inform. Theory, vol. IT-8, pp. 179-187

4. Khalil, M.I., Bayoumi, M.M., 2002. Affine invariants for object recognition using the wavelet transform. Pattern Recognition Letters 23 pp. 57-72
5. Kim, H-K., Kim, J-D., 2000. Region-based shape descriptor invariant to rotation, scale and translation. Signal Processing: Image Communication 16 pp. 87-93
6. Lai, J.H., Yuen, P.C., Feng, G.C., 2001. Face recognition using holistic Fourier invariant features. Pattern Recognition 34 pp. 95-109
7. Shao, Y., Celenk, M., 2001. Higher-order spectra (HOS) invariants for shape recognition. Pattern Recognition 34 pp. 2097-2113
8. Sujan, V.A., Mulqueen, M.P., 2002. Fingerprint identification using space invariant transforms. Pattern Recognition Letters 23 pp. 609-919
9. Theodoridis, S., Koutroumbas, K., 1999. Pattern Recognition. Academic Press,
10. de Vel, O., Aeberhard, S., 2000. Object recognition using random image-lines. Image and Vision Computing 18 pp. 193-198
11. Wood, J., 1996, Invariant pattern recognition: A review. Pattern Recognition, Vol 29. No. 1, pp. 1-17

On Bit-Level Systolic Arrays for Least-Squares Digital Contour Smoothing

Ján Glasa

Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, 845 07 Bratislava, Slovak Republic
utrrglas@savba.sk

Abstract. In this paper bit-level systolic arrays for the least-squares digital contour smoothing are described. The smoothing is represented by linear operators defined by sparse circulant Toeplitz matrices with integer coefficients. A suitable decomposition of such matrices allows to analyze different bit-level pipelined strategies of the computation and to suggest corresponding bit-level systolic arrays which have a simple regular structure and achieve very high throughputs.

1 Introduction

Development of efficient techniques for the real-time analysis of digital arcs and curves plays an important role in image processing [2,7,8,16,18]. The methods for digital curves analysis can have a principal impact for analysis, representation and characterization of significant regions in digital images and for analysis of shapes of real objects, so that they belong to basic tools of commercial computer vision systems. Serious attention is paid to the introduction of new approaches and methods to improve the estimation accuracy of shape characteristics and invariants. The high speed computation, which is required in many concrete image processing applications [1,4,10,13,14,19], influences the development of VLSI systems utilizing parallelism, such as pipelined architectures, systolic arrays, string processors and wavefront arrays.

In this paper[1] we focus on efficient pipelined calculation of the constrained least-squares digital contour smoothing [15] which is based on the least-squares approximation to functions on equidistant subsets of points by orthogonal polynomials. Such a smoothing is represented by linear operators defined by circulant Toeplitz matrices with integer coefficients which can be efficiently realized. A suitable decomposition of these matrices allows to investigate different pipelined bit-level computation strategies and to suggest bit-level systolic arrays with a simple regular structure and very high throughputs.

In the following, let a digital picture be a finite rectangular array represented by a finite square grid Ω where a distance between neighbouring grid points of Ω is equal to 1.

[1] This work was partially supported by NSGA, grant No. 2/4149/24.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 18–25, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Let *a digital contour* be a planar simply closed digital curve $\Gamma \equiv \bigcup_{j=0}^{N} S_j$, where S_j is a linear segment of the length 1 whose end points (x_j, y_j) and $(x_{(j+1)modN}, y_{(j+1)modN})$ belong to grid points of Ω, and for each j there are exactly two points $(x_{(j-1)modN}, y_{(j-1)modN})$, $(x_{(j+1)modN}, y_{(j+1)modN})$ for which it holds $|x_j\text{-}x_{(j-1)modN}|+|y_j\text{-}y_{(j-1)modN}|=1$, $|x_{(j+1)modN}\text{-}x_j|+|y_{(j+1)modN}\text{-}y_j|=1$.

Let us denote by

$$X=\begin{pmatrix} x_0 & x_1 & \dots & x_N \\ y_0 & y_1 & \dots & y_N \end{pmatrix}^T,\quad C=\begin{pmatrix} c_0 & c_1 & c_2 & \dots & c_{N-1} & c_N \\ c_N & c_0 & c_1 & \dots & c_{N-2} & c_{N-1} \\ c_{N-1} & c_N & c_0 & \dots & c_{N-3} & c_{N-2} \\ & & & \dots & & \\ c_2 & c_3 & c_4 & \dots & c_0 & c_1 \\ c_1 & c_2 & c_3 & \dots & c_N & c_0 \end{pmatrix},\ \text{and}\ \ c=\sum_{j=0}^{N} c_j,$$

where C is an $(N+1)$x$(N+1)$ circulant Toeplitz matrix.

Digital contour smoothing in the least-squares sense is described by linear operator $\frac{1}{c}C$ which is applied on X [15],

$$\tfrac{1}{c}\ C\ X = X', \tag{1}$$

where the coefficients of C are obtained by the least-squares approximation to functions on equidistant subsets of points by orthogonal polynomials (for more details see [15,6]). Some examples of the operators $\frac{1}{c}\ C\ X$ are shown on Fig. 1.

The operators $\frac{1}{c}C$ defined by (1) are *position invariant* [15], i.e., it holds that

$$\tfrac{1}{N+1}\ J\ C\ (X - \tfrac{1}{N+1}\ J\ X\ ) = O,$$

where J is an $(N+1)$x$(N+1)$ matrix whose elements are 1s and O is an $(N+1)$x$(N+1)$ null matrix whose elements are 0s. This means that the smoothed contour has the same centroid as the original digital contour.

$$\tfrac{1}{3}\begin{pmatrix} 1 & 1 & & & 1 \\ 1 & 1 & 1 & & \\ & 1 & 1 & 1 & \\ & & & 1 & 1 & 1 \\ 1 & & & & 1 & 1 \end{pmatrix} \qquad \tfrac{1}{35}\begin{pmatrix} 17 & 12 & -3 & & & -3 & 12 \\ 12 & 17 & 12 & -3 & & & -3 \\ -3 & 12 & 17 & 12 & -3 & & \\ & & & & & & \\ & & -3 & 12 & 17 & 12 & -3 \\ -3 & & & -3 & 12 & 17 & 12 \\ 12 & -3 & & & -3 & 12 & 17 \end{pmatrix}$$

$$\tfrac{1}{21}\begin{pmatrix} 7 & 6 & 3 & -2 & -2 & 3 & 6 \\ 6 & 7 & 6 & 3 & -2 & -2 & 3 \\ 3 & 6 & 7 & 6 & 3 & -2 & -2 \\ & & & & & & \\ -2 & 3 & 6 & 7 & 6 & 3 & -2 \\ -2 & -2 & 3 & 6 & 7 & 6 & 3 \\ 3 & -2 & -2 & 3 & 6 & 7 & 6 \\ 6 & 3 & -2 & -2 & 3 & 6 & 7 \end{pmatrix}$$

(a) (b) (c)

Fig. 1. Examples of the linear operators $\frac{1}{c}C$ which correspond to the least-squares 3-, 5-, and 7-point approximation by polynomials of the 1-st, 3-rd, and 3-rd degree, respectively (zero values of the matrix coefficients are not registered)

A subset of operators $\frac{1}{c}C$ defined by (1) are *feasible* [15], i.e., they fulfil the condition

$$| x_j - x'_j \ | < \tfrac{1}{2}, \quad | \ y_j - y'_j \ | < \tfrac{1}{2}, \tag{2}$$

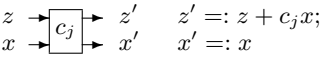
for all j, where x_j, y_j are elements of X and x'_j, y'_j are elements of X'. According to this definition, the feasible operator is defined by the constrained least-squares smoothing with the constraints (2) and it generates points which lie in the interior of the corridor $\bigcup_{j=0}^{N}\{(u,v) \in R^2 : | \ x_j - u \ | \leq \frac{1}{2}, | \ y_j - v \ | \leq \frac{1}{2}\}$. It has been shown that all operators on Fig. 1 are feasible [15]. The operator defined by polynomials of the third degree and by seven points allows to perform in some sense "maximal" feasible smoothing. These operators minimize the undersampling, digitizing and quantizing error and so they are able to improve the stability of calculation of invariants which correspond to invariants related to original pre-digitized contours investigated (for more details see [15,6]). Note that in this paper, the effect of noise by which digital images can be corrupted is not considered (it is assumed to be suppressed by a suitable pre-processing technique [3,9,11,12,17]).

2 Word-Level Pipelining of the Calculation

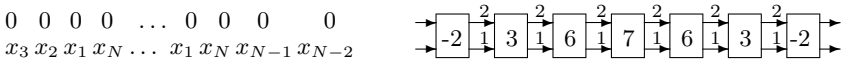
Let us consider the matrix C with coefficients $c_{-3} = c_3 = -2$, $c_{-2} = c_2 = 3$, $c_{-1} = c_1 = 6$, $c_0 = 7$ and let $c = 21$ (see Fig. 1*c*). The matrix-matrix multiplication CX can be represented by two circulant convolutions

$$cx'_i = \sum_{j=-3}^{3} c_j x_{(i-j)modN}, \ cy'_i = \sum_{j=-3}^{3} c_j y_{(i-j)modN}, \tag{3}$$

$i = 0, 1, 2, ..., N$. The word-level serial calculation as well as the word-level pipelining of (3) is straightforward.



(*a*)



(*b*)

Fig. 2. A word-level systolic array for the least-squares 7-point digital contour smoothing by polynomials of the third degree

A linear systolic array for parallel pipelined calculation of the first circulant convolution in (3) is shown on Fig. 2 (for more details see [15]). It has simple cells (Fig. 2*a*) separated by delay elements controlled by a common clock. The numbers of delay elements at z- and x-connections and the necessary input operations are shown on Fig. 2*b*. Since the primitive operations performed are operations on the word level, the array is qualified as a *word-level systolic array.* It has the throughputs (N+13)T, where T is the execution time of a single cell function.

3 Bit-Level Pipelining of the Calculation

The bit-level pipelining of the calculation of (1) is based on a suitable decomposition of the matrix C. Since the integer coefficients of the matrix C can be represented by sums of powers of 2, it is possible to find such decompositions of C by which the calculation of (1) can be represented by circulant convolutions in which neighbouring coefficients are represented by neighbouring powers of 2. It means that the corresponding multiplications can be carried out by shifting of the x-data.

More formally, the matrix C can be represented by a sum of n matrices $C^{(k)}$ with integer coefficients $c_j^{(k)}$,

$$C = \sum_{k=1}^{n} C^{(k)}, \; c_j = \sum_{k=1}^{n} c_j^{(k)}, \tag{4}$$

which fulfil the following conditions:

(i) Each non-zero coefficient $c_j^{(k)}$ has the form $c_j^{(k)} = 2^{\alpha}$, $\alpha \in Z_0^+$.

(ii) For each couple of neighbouring non-zero coefficients $c_r^{(k)} = 2^{\alpha}$, $c_s^{(k)} = 2^{\beta}$, $r < s$ of the matrix $C^{(k)}$, it holds that $\mid \alpha - \beta \mid \leq 1$.

(iii) For each two coefficients $c_p^{(k)} = 2^{\alpha}$, $c_q^{(k+1)} = 2^{\beta}$ of matrices $C^{(k)}$, $C^{(k+1)}$, where $c_p^{(k)}$ is the last non-zero coefficient from the set $\{c_j^{(k)}\}_{j=-3}^{3}$ and $c_q^{(k+1)}$ is the first non-zero coefficient from the set $\{c_j^{(k+1)}\}_{j=-3}^{3}$, it holds that $\mid \alpha - \beta \mid \leq 1$.

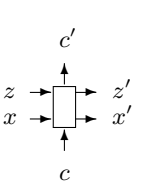
The decomposition (4) of the matrix C allows to represent the calculation of CX by two circulant convolutions, [6]

$$cx'_i = \sum_{k=1}^{n} \sum_{j=-3}^{3} c_j^{(k)} x_{(i-j) mod N} \;, \quad cy'_i = \sum_{k=1}^{n} \sum_{j=-3}^{3} c_j^{(k)} y_{(i-j) mod N} \;, \tag{5}$$

$i = 0, 1, ..., N$. Each particular decomposition of C mentioned above represents a concrete bit-level pipelined systolic implementation strategy of the calculation of (5), in which the neighbouring coefficients are represented by the neighbouring powers of 2. This corresponds to shifting by just one position in the neighbouring

cells of the corresponding array. The procedure how to obtain a particular systolic array completely pipelined on a bit-level for the first circulant convolution in (5) can be summerized as follows (for more details see [15,6]).

The array consists of a sequence of n word-level systolic arrays which have a structure similar to the array shown on Fig. 2. The array cells correspond to the coefficients $c_j^{(k)}$, $j = -3, -2, ..., 3$, $k = 1, 2, ..., n$. The cells which correspond to zero coefficients $c_j^{(k)}$ are omitted modifying the numbers of used delay elements. In such a word-level systolic array, each word-level cell can be realized as a linear vertical systolic array of full adders (see Fig. 3a) separated by delay elements. To achieve a systolic array completely pipelined on a bit-level, a horizontal pipelining by delay elements is involved and the input data are scewed to provide the proper data alignment (the corresponding changes of the numbers od delay elements on z- and x-connections are also required, for more details see [15,6]. In the case of 8-bit input data, the intermediate results can be represented by $(8 + \lceil \log_2 21 \rceil + 1) = 14$-bit numbers. The whole systolic array consists then of 2-dimensional regular matrix of full adders arranged into (at least) 12 columns



z	x	c	z'	x'	c'
0	0	0	0	0	0
0	0	1	1	0	0
0	1	0	1	1	0
0	1	1	0	1	1
1	0	0	1	0	0
1	0	1	0	0	1
1	1	0	0	1	1
1	1	1	1	1	1

column	1	2	3	4	5	6	7	8	9	10	11	12
coefficient	1	2	-1	2	1	-2	-2	2	4	8	4	2
NDE_z	1	1	3	1	3	1	1	2	1	1	3	3
NDE_x	0	1	1	1	1	1	7	1	1	1	1	1

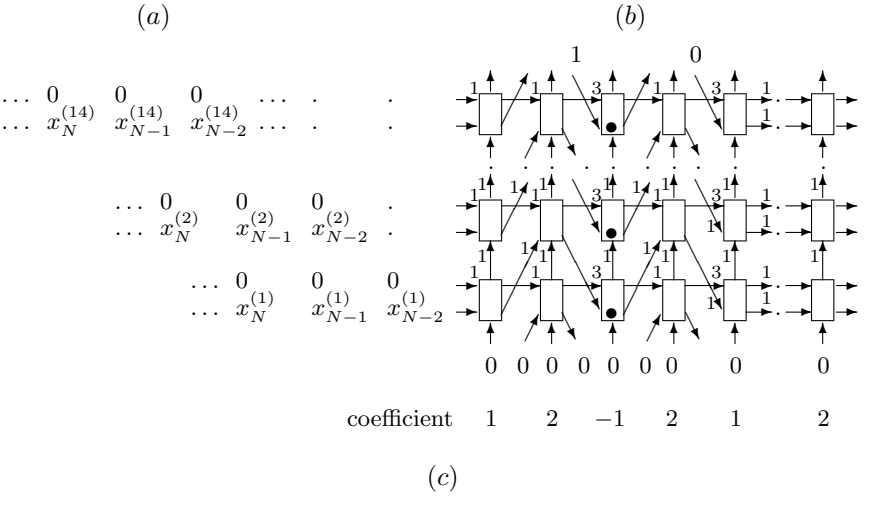


Fig. 3. A bit-level systolic array for the 7-point least-squares digital contour smoothing by polynomials of the third degree: NDE_* - number of delay elements on $_*$-connections

(corresponding to the non-zero coefficients of the matrices $C^{(k)}$) and 14 rows (corresponding to the bits of different significance) which are connected by delay elements. An example of such systolic arrays is shown on Fig. 3. The directions of z- and x-connections are constant for the whole array; the directions of x-connections are constant within one column and change from column to column. The shifting of x-data by just one bit position upwards, downwards, or horizontally, corresponds to the multiplication by 2^1, 2^{-1}, or 2^0, respectively. The multiplication by -1 is realized by converting the x-inputs into their 2's complements (inverters are denoted by small black circles, see Fig. 3b). The numbers of delay elements on x- and z-connections, constant within one column, change from column to column. The number of delay elements on c-connections is constant in the whole array. The clock period of such completely pipelined system is controlled by the delay t of a single full adder. Since the primitive operations performed are operations on the bit level, the array is qualified as a *bit-level systolic array.*

Table 1.

j	-3	-2	-1	0	1	2	3
c_j	-2	3	6	7	6	3	-2
$c_j^{(1)}$	0	2^0	2^1	-2^0	2^1	2^0	-2^1
$c_j^{(2)}$	-2^1	2^1	2^2	2^3	2^2	2^1	0

Table 2.

j	-3	-2	-1	0	1	2	3
c_j	-2	3	6	7	6	3	-2
$c_j^{(1)}$	-2^1	2^0	2^1	2^0	2^1	2^0	-2^1
$c_j^{(2)}$	0	2^1	2^2	2^1	2^2	2^1	0
$c_j^{(3)}$	0	0	0	2^2	0	0	0

The first bit-level systolic array for the calculation of (1) has been suggested in [15] (see Fig. 3). This array corresponds to the decomposition of C shown in Table 1. In this case, it holds that $n = 2$ and the number of non-zero coefficients $c_j^{(k)}$ equals to 12 (see Table 1).

Other different decompositions (4) have been used in [5] to achieve better throughputs (an example of such a decomposition is shown in Table 2). For these decompositions it holds that $n = 1$ and the corresponding bit-level systolic arrays have higher throughputs. In comparison with the array from [15], they have a little bit greather number of cells (one extra column of full adders), but in the same time there is a notable decrease of numbers of delay elements between array cells.

Finally, in [6] the complete analysis of all possible decompositions (4) of C has been performed to achieve the maximal throughputs of corresponding bit-level systolic arrays. In order to minimize also the required area of such arrays, the arrays with the minimum number of delay elements used (and with 12x14 cells) have been chosen. The complete set of bit-level systolic arrays for the calculation of (1) is shown on Fig. 4 [6]. The throughputs of these arrays is (N+36)t, where t is the delay of a single full adder and N is the number of points of digital contour.

column	1	2	3	4	5	6	7	8	9	10	11	12
coefficient	-1	2	1	-2	-1	2	4	2	4	8	4	-2
NDE_z	1	4	2	1	1	1	1	1	1	1	1	1
NDE_x	0	1	1	1	3	1	2	2	2	1	2	1
coefficient	-1	2	4	2	1	-2	-1	2	4	8	4	-2
NDE_z	1	1	2	3	2	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	1	3	1	4	1	2	1
coefficient	-1	2	4	2	-1	2	1	-2	4	8	4	-2
NDE_z	1	1	2	2	1	2	2	1	1	1	1	1
NDE_x	0	1	1	1	1	1	1	1	6	1	2	1
coefficient	-1	-2	1	2	-1	2	4	2	4	8	4	-2
NDE_z	1	5	1	1	1	1	1	1	1	1	1	1
NDE_x	0	1	1	2	2	1	2	2	2	1	2	1
coefficient	-1	2	4	-2	1	2	-1	2	4	8	4	-2
NDE_z	1	1	2	4	1	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	2	2	1	4	1	2	1
coefficient	-1	2	4	2	-1	-2	1	2	4	8	4	-2
NDE_z	1	1	2	2	1	3	1	1	1	1	1	1
NDE_x	0	1	1	1	1	1	1	2	5	1	2	1
coefficient	-1	2	-1	2	1	-2	4	2	4	8	4	-2
NDE_z	1	3	1	2	2	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	1	4	2	2	1	2	1
coefficient	-1	2	-1	-2	-1	-2	4	8	4	8	4	-2
NDE_z	1	1	3	1	3	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	1	3	3	2	1	2	1
coefficient	-1	2	-1	-2	-1	-2	4	8	4	8	4	-2
NDE_z	1	1	5	1	1	1	1	1	1	1	1	1
NDE_x	0	1	1	1	3	1	1	3	2	1	2	1
coefficient	-1	2	-1	-2	1	2	4	2	4	8	4	-2
NDE_z	1	3	1	3	1	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	2	3	2	2	1	2	1
coefficient	-1	2	-1	-2	-1	-2	4	8	4	8	4	-2
NDE_z	1	1	3	3	1	1	1	1	1	1	1	1
NDE_x	0	1	1	1	1	3	1	3	2	1	2	1

Fig. 4. Bit-level systolic arrays for the 7-point least-squares digital contour smoothing by polynomials of the third degree with the maximal throughputs in which the minimal number of the structural elements are used

4 Conclusion

In this paper different bit-level pipelined computation strategies for the real-time calculation of the constrained least-squares digital contour smoothing which represent different bit-level systolic arrays are summerized. They are based on

suitable decomposition of integer coefficients of the circulant Toeplitz matrix, by which the smoothing is represented. The suggested systolic arrays completely pipelined on the bit-level have simple regular structure and achieve very high throughputs. They can be implemented on single VLSI chips and can be used for the purposes of real-time digital contour analysis.

References

1. Bennamoun, M., Mamic, G.: Object Recognition. Springer-Verlag, Berlin (2002)
2. Dougherty, E.R., Laplante, P.A.: Introduction to Real-Time Imaging. IEEE Press, NY (2001)
3. Fisher, V., Drutarovsky, M., Lukac, R.: Implementation of 3D Adaptive LUM Smoother in Reconfigurable Hardware. Lect. Notes in Comp. Sci., Vol. 2438. Springer-Verlag, Berlin Heidelberg New York (2002) 720-729
4. Fu, K.S.: VLSI for Pattern Recognition and Image Processing. Springer-Verlag, Berlin (1984)
5. Glasa, J.: Bit-Level Systolic Arrays for Digital Contour Smoothing, Lect. Notes in Comp. Sci., Vol. 399. Springer-Verlag, Berlin Heidelberg New York (1989) 105-120
6. Glasa, J.: Circulant Toeplitz Matrices for Digital Contour Smoothing and Their Parallel Implementation on Systolic Arrays (in Slovak). PhD. theses. Slovak Academy of Sciences, Bratislava (1993)
7. Gonzales, R.C., Woods, R.E.: Digital Image Processing. 2nd edn. Prentice-Hall, Upper Saddle River, NJ (2002)
8. Jahne, B.: Digital Image Processing. 5th edn. Springer-Verlag, Berlin (2001)
9. Halada, L.: About Some Methods of Image Point Classification. Proc. of the Int. Conf. on Mechanical Engineering, Bratislava (1997) 26-29
10. Laplante, P.A., Stoyenko, A.D., Stoyenko, A.: Real-Time Imaging: Theory, Techniques and Applications. IEEE Press, NY (1996)
11. Lukac, R.: Binary LUM Smoothing. IEEE Signal Proc. Let. **9** (2002) 400-403
12. Lukac, R., Smolka, B., Plataniotis, K.N., Venetsanopoulos, A.N.: Entropy Vector Median Filter, Lect. Notes in Comp. Sci., Vol. 2653. Springer-Verlag, Berlin Heidelberg New York (2003) 1117-1125
13. Nishitani, T., Ang, P.H., Catthoor, F.: VLSI Video/Image Signal Processing. Kluwer Academic Publishers, Boston (1993)
14. Petkov, N.: Systolic Parallel Processing. North Holland, Elsevier Science Publ., Amsterdam (1993)
15. Petkov, N., Sloboda, F.: A Bit-Level Systolic Array for Digital Contour Smoothing. Parallel Computing **12** (1989) 301-313
16. Pitas, I.: Digital Image Processing: Algorithms and Applications. Wiley Interscience, NY (2000)
17. Pitas, I., Venetsanopoulos, A.N.: Non-Linear Digital Filters: Principles and Applications. Kluwer Academic Publishers, Boston (1990)
18. Pratt, W.K.: Digital Image Processing. 3rd edn. Wiley Interscience, NY (2001)
19. van der Heijden, F.: Image Based Measurement Systems: Object Recognition and Parameter Estimation. 1st edn. John Wiley and Sons, NY (1995)

Bayer Pattern Demosaicking Using Local-Correlation Approach

Rastislav Lukac, Konstantinos N. Plataniotis, and
Anastasios N. Venetsanopoulos

The Edward S. Rogers Sr. Dept. of Electrical and Computer Engineering,
University of Toronto, 10 King's College Road, Toronto, M5S 3G4, Canada
{lukacr,kostas,anv}@dsp.utoronto.ca

Abstract. A new Bayer pattern demosaicking scheme for single-sensor digital cameras is introduced. The raw output from a sensor, mostly a charge coupled device (CCD) or a complementary metal oxide semiconductor (CMOS) sensor, with a Bayer filter represents a mosaic of red, green and blue pixels of different intensity. To interpolate the two missing color components in each spatial location and constitute the full color, camera output, the proposed method utilizes edge-sensing interpolation and correction steps. Since the correction step is suitable only for the image regions with high spectral correlation, otherwise is counter productive, the scheme is adaptively controlled through the comparisons between the correlation coefficient and the pre-determined parameter. The proposed method yields excellent performance, in terms of subjective and objective image quality measures, and outperforms previously developed CFA interpolation solutions.

1 Introduction

Single-sensor digital color cameras use color filter array (CFA) to separate incoming light into a specific spatial arrangement of the color components. Most popular solutions utilize a Red-Green-Blue (RGB) Bayer CFA pattern shown in Fig. 1 [2]. The raw output from a sensor, mostly a charge coupled device (CCD) or a complementary metal oxide semiconductor (CMOS) sensor, with a Bayer filter represents a mosaic of red, green and blue pixels of different intensity. Since the two color components are missing in each spatial location (Fig. 2), they must be interpolated from the spatially adjacent CFA data. The process of interpolating missing spectral components is known as CFA interpolation or demosaicking.

The proposed demosaicking method produces the full color, camera output using edge-sensing interpolation and correction steps. After initial interpolation of the G components, the method utilizes color-difference model of [1] to generate R and B estimates using both R and G or B and G components, respectively. Thus, the procedure produces more accurate outputs compared to the schemes operating on R or B components, only. In the next step, the color appearance of the restored image is improved, first through the correction of the interpolated G

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 26–33, 2004.
© Springer-Verlag Berlin Heidelberg 2004

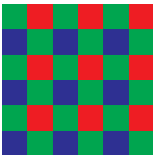
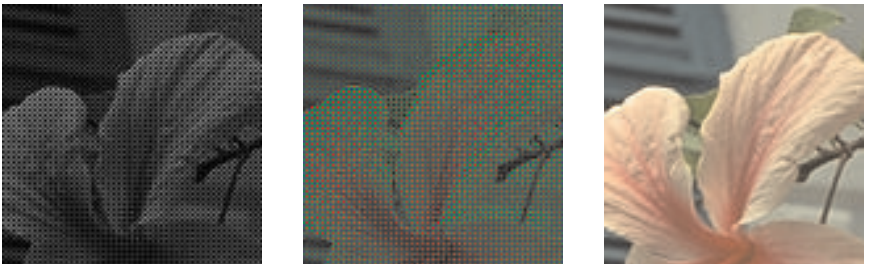


Fig. 1. RGB Bayer CFA pattern



(a) (b) (c)

Fig. 2. Enlarged parts of the image Window arranged as: (a) a gray-scale Bayer image and (b) a color Bayer image, both acquired by a single-sensor digital camera, (c) a full color image captured by a three-sensor digital camera

components and then increasing precision of the previously interpolated R and B components. To restrict the correction process only for the image regions with high correlation among the spectral components, the local correlation characteristics are compared to the pre-determined parameter. This preserves image quality in the regions with weak correlation, where the correction step may be counter productive.

2 Color Filter Array Basics

Let us consider, a $K_1 \times K_2$ gray-scale image $z(i) : Z^2 \to Z$ representing a two-dimensional matrix of integer samples. In the Bayer CFA pattern, half of the pixels $z_i \in Z^2$, for $i = 1, 2, ..., K_1K_2$, correspond to the G channel, whereas R,B channels are assigned the other half of the pixels. Assuming that $p = 1, 2, ..., K_1$ and $q = 1, 2, ..., K_2$ denote the spatial position of the pixels in vertical (image rows) and horizontal (image columns) directions, gray-scale pixels z_i can be transformed into the RGB vectors $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}) \in Z^2$, for $i = (p-1)K_2+q$, as follows:

$$\mathbf{x}_i = \begin{cases} (z_i, 0, 0) \text{ for } p \text{ odd and } q \text{ even} \\ (0, 0, z_i) \text{ for } p \text{ even and } q \text{ odd} \\ (0, z_i, 0) \text{ otherwise} \end{cases} \quad (1)$$

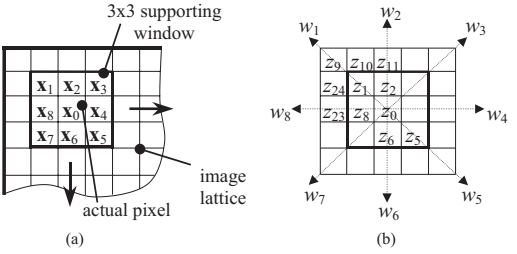


Fig. 3. Essential elements in the proposed scheme: (a) sliding supporting window determines the uncomplete color vectors $\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_{N-1}}$, (b) edge-sensing mechanism is expressed through the weighted coefficients $w_1, w_2, ..., w_8$

This transformation forms a $K_1 \times K_2$ RGB image $\mathbf{x}(i) : Z^2 \rightarrow Z^3$ representing a two-dimensional matrix of three-component samples. Note that the color vectors $\mathbf{x}_i$ relate to one true component varying in k from position to position, whereas other two components of $\mathbf{x}_i$ are set to zero. Estimating the missing color components of $\mathbf{x}(i)$ constitutes the interpolated RGB image $\mathbf{y}(i) : Z^2 \rightarrow Z^3$.

This process relates to a sliding supporting window $W = \{\mathbf{x}_i \in Z^2; i = 0, 1, ..., N-1\}$ of finite size N, considered with the sample under consideration, sample $\mathbf{x}_0$, placed in the center of the window (Fig. 3a). The procedure replaces the center $\mathbf{x}_0$ by some function of the local neighborhood area $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N-1}\}$ at a time. The rationale of this approach is to minimize the local distortion and ensure the stationarity of the processes generating the image.

3 Demosaicking Approach Based on the Local-Correlation Characteristics

To follow structural information and interpolate missing image components in sharp shapes, efficient interpolation algorithms incorporate edge information into the interpolation process. The proposed method utilizes the edge-sensing mechanism of [8]:

$$w_1 = \frac{1}{1 + \left(|z_0 - z_9| + |z_1 - z_5|\right)/(2\sqrt{2})} \tag{2}$$

$$w_2 = \frac{1}{1 + \left(|z_0 - z_{11}| + |z_2 - z_6|\right)/2} \tag{3}$$

where w_1 and w_2 denote weights in north-west and north directions. The original gray-scale values $\{z_0, z_1, ..., z_{24}\} \in z(i)$ have an spatial arrangement shown in (Fig.3b). The weighting coefficients w_3, w_5, w_7 and w_4, w_6, w_8 are calculated applying appropriately the concept of (2) and (3), respectively.

Using eight weighting coefficients, the G channel is interpolated as follows:

$$\bar{y}_{02} = \begin{cases} x_{02} & \text{if } z_0 \cong x_{02} \\ \sum_{i=1}^{N-1} w_i' x_{i2}' & \text{otherwise} \end{cases} \tag{4}$$

where $N = 9$ relates to a 3×3 sliding window, z_0 is the acquired gray-scale component positioned in the interpolated location, operator $\cong$ denotes a one to one relationship and $w'_i = w_i / \sum_{j=1}^{N-1} w_j$ is the normalized weighting coefficient corresponding to pre-determined G values x'_{i2}. For illustration purposes, quantities $x'_{(1)2}$ and $x'_{(2)2}$ are given by:

$$x'_{(1)2} = \left(x_{(2)2} + x_{(8)2}\right)/2 + ((z_1 - z_5)/(2\sqrt{2}) + (z_{11} - z_0 + z_{23} - z_0)/4)/2 \quad (5)$$

$$x'_{(2)2} = x_{(2)2} + (z_{11} - z_0 + z_2 - z_6)/4 \quad (6)$$

Using the color-difference model of [1] and the G values obtained in (4), the R and B channels are estimated as follows:

$$\bar{y}_{0k} = \begin{cases} x_{0k} & \text{if } z_0 \cong x_{0k} \\ \bar{y}_{02} + f''_{(2i)k} & \text{if } z_0 \cong x_{02} \\ \bar{y}_{02} + f'''_{(2i-1)k} & \text{if } z_0 \cong x_{0(k\pm2)} \end{cases} \quad (7)$$

where $k = 1$ and $k = 3$ characterize the R and B components, respectively, and the quantities $f''_{(2i)k}$ and $f'''_{(2i-1)k}$ are defined via

$$f''_{(2i)k} = \sum\nolimits_{i=1}^{(N-1)/2} w''_i (x_{(2i)k} - \bar{y}_{(2i)2}) \quad (8)$$

$$f'''_{(2i-1)k} = \sum\nolimits_{i=1}^{(N-1)/2} w'''_i (x_{(2i-1)k} - \bar{y}_{(2i-1)2}) \quad (9)$$

with the normalized weights $w''_i = w_{2i} \sum_{j=1}^{(N-1)/2} w_{2j}$ corresponding to edges in north, east, south and west directions. The weighting coefficients $w'''_i = w_{(2i-1)} / \sum_{j=1}^{(N-1)/2} w_{(2j-1)}$ correspond to diagonally positioned edges.

The utilization of the correction mechanism in the interpolation process improves contrast and accuracy of the initially interpolated G channel. Using the color-difference quantities similarly as in (7), the G values are corrected as follows:

$$y_{02} = \begin{cases} \bar{y}_{0k} + g''_{(2i)k} & \text{if } z_0 \cong x_{0k} \\ \bar{y}_{02} & \text{otherwise} \end{cases} \quad (10)$$

where $g''_{(2i)k} = \sum_{i=1}^{(N-1)/2} w''_i (\bar{y}_{(2i)2} - \bar{y}_{(2i)k})$ is defined using the weighting coefficients w''_i of (7).

Considering the corrected G values of (10) the update of R and B components is completed using the proposed approach as follows:

$$y_{0k} = \begin{cases} y_{0k} & \text{if } z_0 \cong x_{0k} \\ y_{02} + h''_{(2i)k} & \text{if } z_0 \cong x_{02} \\ y_{02} + h'''_{(2i-1)k} & \text{if } z_0 \cong x_{0(k\pm2)} \end{cases} \quad (11)$$

where

$$h''_{(2i)k} = \sum\nolimits_{i=1}^{(N-1)/2} w''_i (\bar{y}_{(2i)k} - y_{(2i)2}) \quad (12)$$

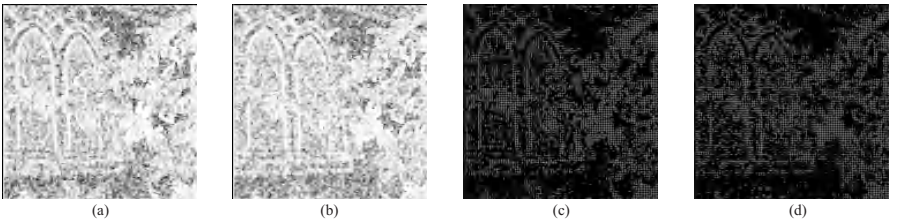


Fig. 4. Local correlation characteristics obtained using the image Window before (a,b) and after (c,d) thresholding: (a,c) RG correlation map and (b,d) GB correlation map

$$h'''_{(2i-1)k} = \sum_{i=1}^{(N-1)/2} w'''_i(\bar{y}_{(2i-1)k} - y_{(2i-1)2}) \tag{13}$$

are defined using the weighting coefficients of (8).

The correction procedure of (11) is performed only in image regions with the high spectral correlation. The method utilizes correlation characteristics (Fig.4) between the interpolated G channel of $\bar{\mathbf{y}}(i)$ and the original R,B channels of $\mathbf{x}(i)$. The correlation coefficient C_k is defined over a 3×3 neighborhood as follows:

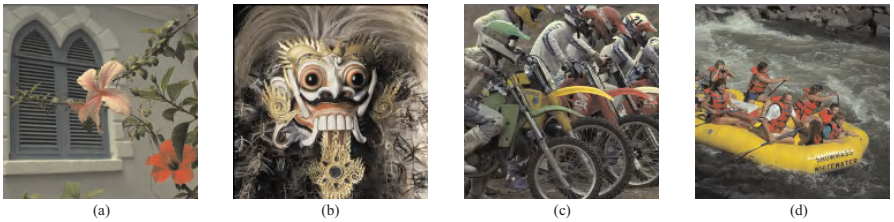
$$C_k = \frac{\sum (x_{ik} - \hat{x}_k)(y_{i2} - \hat{y}_2)}{\sqrt{\sum (x_{ik} - \hat{x}_k)^2} \sqrt{\sum (y_{i2} - \hat{y}_2)^2}} \tag{14}$$

where i, for $z_i \cong x_{ik}$, characterizes the spatial position of the samples corresponding to the original position of R,B values in the Bayer pattern. The mean values $\hat{x}_k$ and $\hat{y}_2$ denote sample mean values related to the k-th original CFA components and the interpolated G components, respectively. Experimentation with a wide set of the test images showed that the correction operations should be performed in the image areas, where the local correlation coefficient is larger than $\beta = 0.125$.

4 Experimental Results

A number of test color images have been used to evaluate the proposed demosaicking scheme. Examples are shown in Fig. 5. These images have been captured using professional three-sensor digital cameras. To facilitate the comparisons, the images have been normalized to a standard size of 512×512 pixels with a 8-bits per channel RGB representation. Since the original Bayer image is usually unavailable in order to obtain test Bayer image data used in CFA interpolation researchers sample original images with the Bayer CFA pattern [10],[11].

The restored outputs are obtained using demosaicking (CFA interpolation) solutions. Results obtained via the proposed method are compared with those obtained using the bilinear interpolation (BI) scheme [10], the high definition color interpolation (HDCI) scheme [7], the median filtering (MFI) scheme [5], the effective color interpolation (ECI) approach [11], the alternative projection (AP)



(a) (b) (c) (d)

Fig. 5. Test color images: (a) Window, (b) Mask, (c) Bikes, (d) Rafting

Table 1. Obtained objective results

Image	Window		Mask		Bikes		Rafting	
Method	MSE	NCD	MSE	NCD	MSE	NCD	MSE	NCD
BI	35.1	0.0417	212.9	0.1328	157.1	0.1209	92.2	0.0753
HDCI	11.7	0.0252	94.8	0.0825	58.8	0.0744	45.3	0.0492
MFI	7.7	0.0239	57.3	0.0822	27.4	0.0620	25.0	0.0434
ECI	6.8	0.0228	49.2	0.0819	20.5	0.581	21.0	0.0396
AP	5.6	0.0212	42.3	0.0754	16.9	0.0534	20.6	0.0358
C2D2	6.5	0.0193	56.3	0.0750	26.7	0.0545	22.5	0.0375
SAIG	11.5	0.0280	87.4	0.0854	53.1	0.0768	41.5	0.0507
SHT	21.3	0.0349	131.7	0.1070	94.8	0.1025	61.1	0.0626
KA	22.3	0.0324	60.2	0.0736	71.7	0.0792	55.7	0.0504
Proposed	4.4	0.0181	35.4	0.0648	12.1	0.0425	15.0	0.0308

approach [6], the color correlation directional derivative (C2D2) scheme [8], the smooth hue transition approach (SHT) [4], the saturation based adaptive inverse gradient (SAIG) [3], and the Kimmel's algorithm (KA) [9]. The efficiency of the all these methods is measured, objectively, via the mean square error (MSE) and the normalized color difference criterion (NCD) [7].

Table 1 summarizes the results corresponding to restoration of the test images shown in Fig. 5. As it can be observed the conventional BI scheme introduces significant inaccuracy into the restoration process. Other techniques such as HDCI, MFI, ECI, AP, C2D2, SAIG and KA which utilize more advanced interpolators compared to the BI scheme, provide better results in terms of both objective criteria. It has to be mentioned that some sophisticated algorithms such as HDCI, SAIG, and KA often fail in image scenarios with color corresponding to zero or very small portion of any additive primary. In such a case, the aforementioned schemes produce color artifacts.

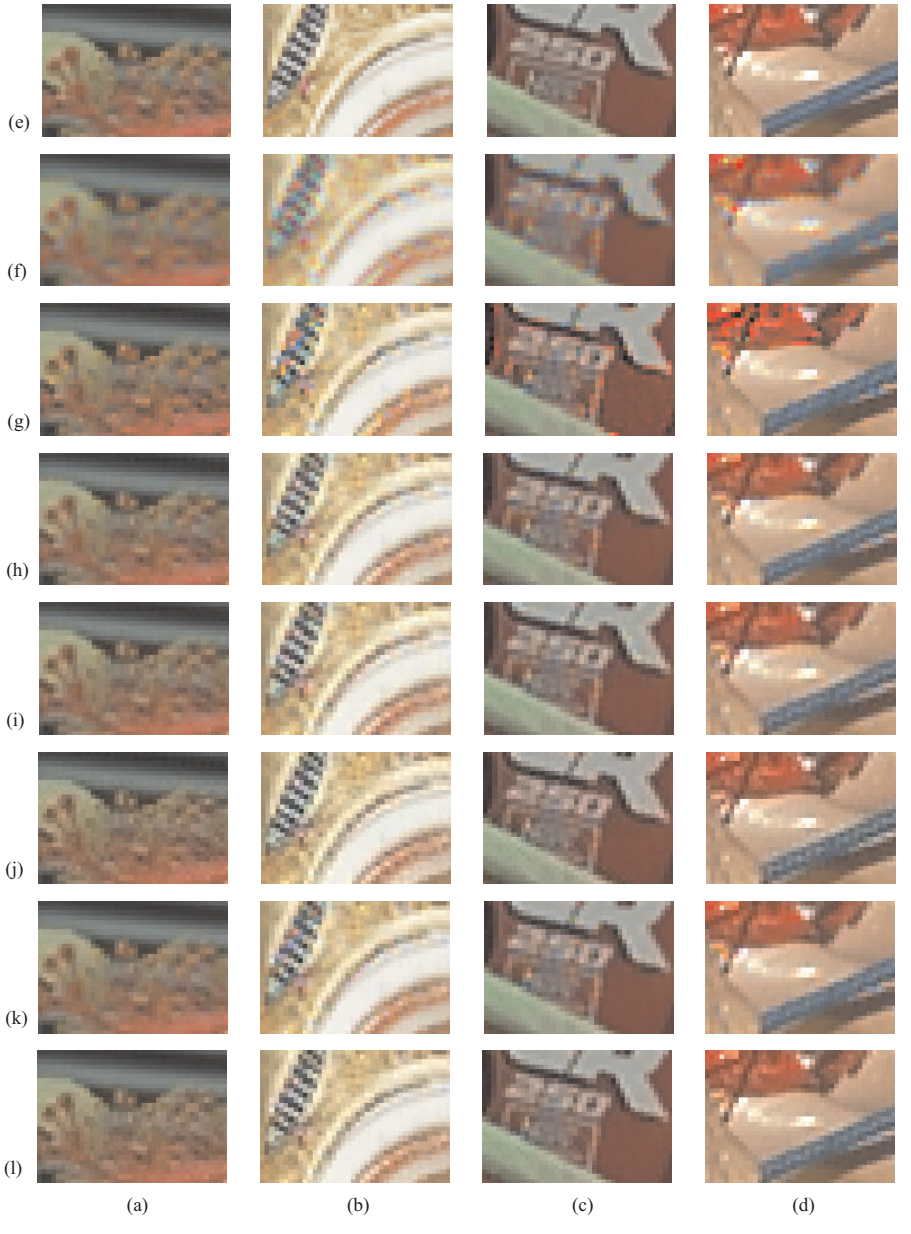


Fig. 6. Enlarged parts of the obtained results corresponding to the images: (a) Window, (b) Mask, (c) Bikes, (d) Rafting; and the methods: (e) original images, (f) BI output, (g) HDCI output, (h) MFI output, (i) ECI output, (j) AP output, (k) C2D2 output, (l) the proposed method output

Figure 6 facilitates the visual comparison of zoomed parts corresponding to the original images as well as the restored outputs obtained using the methods under consideration. Results indicate that the BI and HDCI schemes blur image edges and delete both structural content and fine details (Fig. 6f,g). The MFI, ECI, AP and C2D2 approaches exhibit improved detail-preserving characteristics (Fig. 6h-k). It can be seen that the proposed method preserves the image details and avoids color artifacts (Fig. 6l). Excellent design characteristics of the proposed method allow to restore the image with the highest fidelity compared to the original. At the same time, the proposed solution preserves the original structural content. This results in visually pleasing, naturally colored outputs.

5 Conclusion

A new, edge-sensing CFA interpolation approach for single-sensor digital cameras was presented. The method utilizes local correlation characteristics and a color-difference model to produce full color camera output. Excellent design characteristics of the proposed method result in sharp, naturally colored interpolated images. At the same time, this produces significant improvements and excellent results compared to the previously developed CFA interpolation methods.

References

1. Adams, J.: Design of practical color filter array interpolation algorithms for digital cameras. Proc. of the SPIE, **3028** (1997) 117–125
2. Bayer, B.E.: Color imaging array. U.S. Patent 3 971 065, (1976)
3. Cai, C., Yu, T.H., Mitra, S.K.: Saturation-based adaptive inverse gradient interpolation for Bayer pattern images. IEE Proceedings - Vision, Image, Signal Processing **148** (2001) 202–208
4. Cok, D.R.: Signal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal. U.S. Patent 4 642 678 (1987)
5. Freeman, W.T.: Median filter for reconstructing missing color samples. U.S. Patent 5 373 322,(1988)
6. Gunturk, B., Altunbasak, Y., Mersereau, R.: Color plane interpolation using alternating projections. IEEE Trans. Image Processing **11** (2002) 997–1013
7. Hur, B.S., Kang, M.G.: High definition color interpolation scheme for progressive scan CCD image sensor. IEEE Trans. Consumer Electronics **47** (2001) 179–186
8. Kehtarnavaz, N., Oh, H.J., Yoo, Y.: Color filter array interpolation using color correlation and directional derivatives. Journal of Electronic Imaging **12** (2003) 621–632
9. Kimmel, R.: Demosaicing: image reconstruction from color CCD samples. IEEE Trans. Image Processing **8** (1999) 1221–1228
10. Longere, P., Zhang, X., Delahunt, P.B., Brainard, D.H.: Perceptual assessment of demosaicing algorithm performance. Proceedings of the IEEE **90** (2002) 123–132
11. Pei, S.C., Tam, I.K., Effective color interpolation in CCD color filter arrays using signal correlation. IEEE Trans. Circuits and Systems for Video Technology **13** (2003) 503–513

Edge Preserving Filters on Color Images

Vinh Hong[1], Henryk Palus[2], and Dietrich Paulus[1]

[1] Institut für Computervisualistik, Universität Koblenz-Landau, Universitätsstr. 1, 56070 KOBLENZ – Germany, {hong,paulus}@uni-koblenz.de
[2] Institute of Automatic Control, Silesian University of Technology, Akademicka 16, 44-100 GLIWICE – Poland, hpalus@polsl.gliwice.pl

Abstract. In this contribution we present experiments on color image enhancement for several different non-linear filters which originally were defined for gray-level images. We disturb sample images by different types of noise and measure performance of the filters. We provide signal-to-noise measurements as well as perceived color difference in ΔE as defined by the CIE. All images and test programs are provided online on the internet so that experiments can be validated by arbitrary users on any image data.

Keywords: color image enhancement, edge-preserving filters, ΔE, performance measures.

1 Introduction

Color images as processed in various applications are recorded by different acquisition devices. Cameras as well as scanners have their specific noise characteristics. Image transmission may as well introduce noise into the image data. Typical models for noise are either Gaussian or salt-and-pepper noise; Gaussian noise is used as a model for sensor errors, drop-outs during transmission and errors on the CCD chip can be modelled by salt-and-pepper noise.

In our contribution we use images from a data base [3] and disturb them by artificial noise of varying degree and type. In Sect. 2.1 we describe some non-linear smoothing filters, such as edge preserving smoothing algorithm [8], and extend them to color images where they were defined for gray-level images originally. We apply these filters to the disturbed input images and compare each result with its corresponding original image to compute difference measures. Standard measures are the signal-to-noise ratio (SNR) and maximum differences for color vectors. As all disturbances are modelled in RGB, we compute these measures in RGB as well. More important for human perception than SNR is the so-called ΔE difference [12] which describes the perceived color difference (Sect. 3). In Sect. 4 we conclude our contribution with a summary of the evaluation and the prospective work.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 34–40, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Color Image Processing

Color image processing has become a central part of automatic image analysis as color can provide valuable cues for identification and localization of objects [11].

2.1 Color Filters

In research and literature there exist several filters that can be classified into linear and non-linear filters. Filters can either operate in the spatial or in the frequency domain [5]. In the following we compare non-linear filters in the spatial domain and additionally use an AMF (arithmetic mean filter).

Linear Filters. A huge number of filters for single band images has been proposed in the long history of image processing. As color images became affordable with respect to sensors, memory and processor speed, some of these filters have been extended to color.

If a linear filter, such as a Gaussian or mean filter, is applied to each channel of an RGB image separately, the resulting image will contain usually color triplets which are not present in the input image. Such artifacts yield perceptional differences which can be avoided by non-linear filtering. On the other hand, additive noise, such as additive Gaussian noise, can be removed by a low-pass filter which averages color vectors.

Non-linear Filters. Filters which are quoted to smooth homogeneous areas while preserving edges are the

- EPS (Edge preserving smoothing), presented for gray-level images in [8],
- SNN (Symmetric Nearest Neighbour Filter) described in [9,6],
- K-N (Kuwahara-Nagao Filter) proposed in [7,8] for gray-level images,
- VMF (Vector Median Filter, also known as CVF (Color Vector Median)), presented in [10,1].

EPS Filter. In the so-called *edge preserving smoothing* algorithm introduced in [8], the selection of gray-level pixels for averaging is done based on statistical principles. The algorithm uses nine different 5×5 masks for each pixel; three of them are shown in Fig. 1. The pixels marked in the neighborhood are used for the following computations. The symmetrical use of 1 (a) and (b) results in eight different masks. Each of these masks includes seven points for the calculation of the new gray-level. The contrast mask (c) includes nine elements for the following computations. For each mask we compute the variance. The mask with the lowest variance is selected. The central pixel gets the mean value of all points marked in this mask.

To extend this algorithm to color, we compute the color covariance matrix inside each mask. We decide for that mask for which the Frobenius norm of the covariance matrix is minimal and compute the mean color vector for the resulting pixel. This may, of course, introduce artifacts.

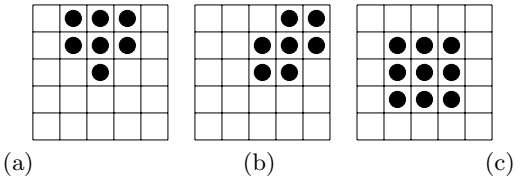


Fig. 1. Masks for edge preserving smoothing

SNN Filter. The SNN is a filter related to the mean and median filters but with better edge-preserving properties. The neighbours of the central pixel in a window are considered as four pairs of symmetric pixels (N-S, W-E, NW-SE and NE-SW). For each pair the pixel closest in color to the central pixel is selected. The colors of these four selected pixels are averaged and the mean color value is a new color for central pixel. The mask for the SNN filter is shown in Fig. 2.

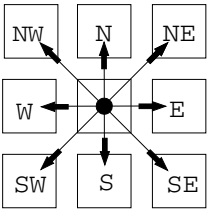


Fig. 2. Mask for SNN filter

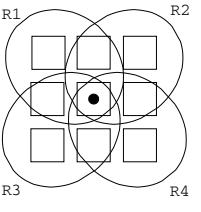


Fig. 3. Mask for Kuwahara-Nagao filter

K-N Filter. The 3×3 mask shown in Fig. 3 is split into four 2×2 slightly overlapping windows with the mask's central pixel as a common part. For each window in a gray-level image, the variance is calculated. The mean value of the window with minimal variance (maximal homogeneous region) is used as the output value of the central pixel.

As in the case of the edge-preserving smoothing, we extend this filter to color as we compute the color covariance matrix and use the Frobenius norm.

Recently the gray-level version of K-N filter has been generalized for round windows [2] and it has been shown that the filter is composition of linear diffusion and morphological sharpening [14].

Vector Median Filter. The definition of the vector median of a set of color vectors $\boldsymbol{f}_i$ in a window $\mathcal{W}$ is given in [10] as

$$\boldsymbol{f}_{\mathrm{v}} = \mathrm{argmin}_{\boldsymbol{f}_i \in \mathcal{W}} \sum_{\boldsymbol{f}_j \in \mathcal{W}} ||\boldsymbol{f}_i - \boldsymbol{f}_j|| \quad . \tag{1}$$

In our experiments we use $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_9$ in a square 3×3 window and apply the Euclidean norm on the difference color vector in (1). As this filter does not include averaging, it is the only filter described here which does not introduce color artifacts.

3 Experiments

Signal-to-noise ratio A measure for the accuracy of a filter is given by the *signal-to-noise ratio* (SNR). For color images pairs we define the SNR by a quotient of means:

$$\mathrm{SNR} = 10 \log_{10} \frac{E[\boldsymbol{f}_i{}^{\mathrm{T}} \boldsymbol{f}_i]}{E[\boldsymbol{n}_i{}^{\mathrm{T}} \boldsymbol{n}_i]} \quad , \tag{2}$$

where $\boldsymbol{f}_i$ is the color vector and $\boldsymbol{n}_i$ is the noise vector computed by the vector difference of the two pixels.

Color Metric To measure perceptual color distances between two color stimuli several metrics such in color spaces such as CIELUV, CIELAB, CIE94, etc. can be used [13,15]. In this paper we will prefer the CIE-recommended color metric CIE94 (see equation (6) below). That is a modification of the CIELAB color difference formula (3):

$$\Delta \mathrm{E}^*_{ab} = \sqrt{(\Delta \mathrm{L}^*_{ab})^2 + (\Delta \mathrm{a}^*_{ab})^2 + (\Delta \mathrm{b}^*_{ab})^2} \quad . \tag{3}$$

A value of $\Delta \mathrm{E}^*_{ab} = 1$ resp. $\Delta \mathrm{E}^*_{CH} = 1$ corresponds to the human's eye minimal perceivable difference between two colors. The greater the color difference between two stimuli is, the greater is its $\Delta \mathrm{E}^*$ value [4]. The CIELAB color metric from the CIE (International Commission on Illumination) describes the color difference between two color points in the uniform $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ space. The axes of this color space compound of the lightness-axis L^*, the red-green-axis a^* and the yellow-blue-axis b^* [13]. In this color space the Euclidean distance between two points corresponds to perceived difference[16]. The symbols $\Delta \mathrm{L}^*_{ab}$, $\Delta \mathrm{a}^*_{ab}$ and $\Delta \mathrm{b}^*_{ab}$ represent the componentwise differences (lightness, the red-green and the yellow-blue) between the two colors.

The color difference (3): has several drawbacks in practice [16]. So the CIE introduced an improved color metric called CIE94 which computes the weighted Euclidian distance between two points in the uniform $\mathrm{L}^*\mathrm{C}^*\mathrm{H}^*$ space. This color space uses the polar coordinates chroma C^* and hue H^* rather than the cartesian coordinates a^* and b^* of the $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ space. In comparison to the $\mathrm{L}^*\mathrm{a}^*\mathrm{b}^*$ space the $\mathrm{L}^*\mathrm{C}^*\mathrm{H}^*$ space is a more intuitive representation of color, because for example hue can be uniquely computed [16]. Chroma can be computed by

$$\mathrm{C}^*_{ab} = \sqrt{(a^*)^2 + (b^*)^2} \quad , \tag{4}$$

and hue can be caculated from

$$\mathrm{H}^*_{ab} = \arctan \frac{b^*}{a^*} \quad . \tag{5}$$

CIE94 computes the color difference between two colors in the $L^*C^*H^*$ space by

$$\Delta E^*_{CH} = \sqrt{\left(\frac{\Delta L^*_{ab}}{k_L S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2} \quad . \tag{6}$$

The symbols $\Delta L^*_{ab}, \Delta C^*_{ab}$ and ΔH^*_{ab} represent the differences between the two given colors corresponding to those lightness, chroma and hue attributes. S_L, S_C and S_H represent parameters calculated from the chroma coordinates of the two color stimuli. k_L, k_S and k_H are parameters those take specific experimental conditions into account[16]

We use the following parameter configuration [16]:

$$k_L = k_S = k_H = S_L = 1 \tag{7}$$

$$S_C = 1 + 0.045 \mu C^*_{ab} \tag{8}$$

$$S_H = 1 + 0.015 \mu C^*_{ab} \tag{9}$$

The values S_C and S_H are computed from the mean chroma value μC^*_{ab} of the two given color stimuli.

Noise model An ideal color image $\boldsymbol{f}$ consisting of color vectors $\boldsymbol{f}_i$ is disturbed by additive noise $\boldsymbol{\beta}_i$ and multiplicative noise $\boldsymbol{\gamma}_i$

$$\boldsymbol{g}_i = \boldsymbol{\gamma}_i \cdot \boldsymbol{f}_i + \boldsymbol{\beta}_i \tag{10}$$

to yield the observed image $\boldsymbol{g}$. We added zero-mean Gaussian noise $\boldsymbol{\beta}$ with varying σ to images in a test data base where the noise was statistically independent for the color channels. In another experiment we introduced impulsive noise which can be considered as a multiplicative noise $\boldsymbol{\gamma}$ with $\gamma_i = 0$ for drop-outs, $\gamma_i = 1$ for undisturbed image information, and $\gamma_i = 255$ to introduce white spots; with a given probability p white and black spots are created, each with probability 0.5. All test images can be found in a public image data base.[1] For each corrupted image we applied the filters described in Sect. 2.1. An example is shown in Fig. 4.

We then compared original and filtered image and computed SNR and mean ΔE^*_{CH}. Of course, if little noise is added to the image, the values for filtered images are worse than for the unfiltered noisy image, as can be seen from Fig. 5 and Fig. 6. The higher the corruption is, the higher the improvement can be by filtering.

4 Conclusion. Prospective Work

The vector median filter outperforms the other filter methods for impulsive noise, if we use ΔE^*_{CH} as a measure. This is as expected, as a measure for perceived color differences should be sensitive to color artifacts. The vector median filter

[1] `http://www.uni-koblenz.de/~puma`

Fig. 4. Example image "`peppers`" (left), corrupted image (center), filtered image (right)

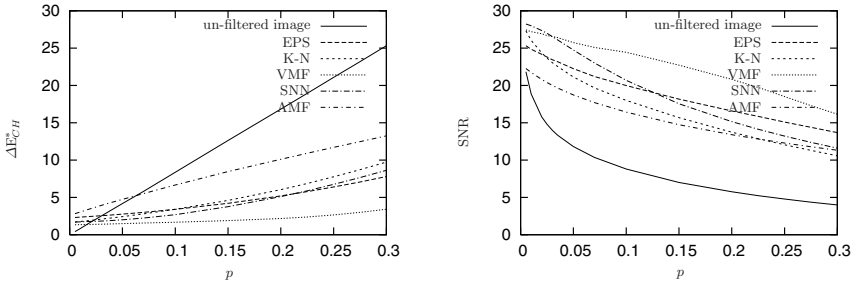


Fig. 5. Example image "`peppers`" corrupted by impulsive noise

also outperforms the other filters in the case of the SNR-measure, if the SNR of the input image is low.

Naturally, linear filtering reduces Gaussian noise better than rank-order filters. The Arithmetic Mean filter returns the best ΔE^*_{CH} for Gaussian noise. In general, the Vector Median filter outperforms the other filters for both distance measures.

Both measures SNR and ΔE^*_{CH} are consistent for extreme cases, i.e. very little or very large noise as they mostly have the same ordering for a qualitative judgement of the filters. Details in the medium range noise reveal differences of

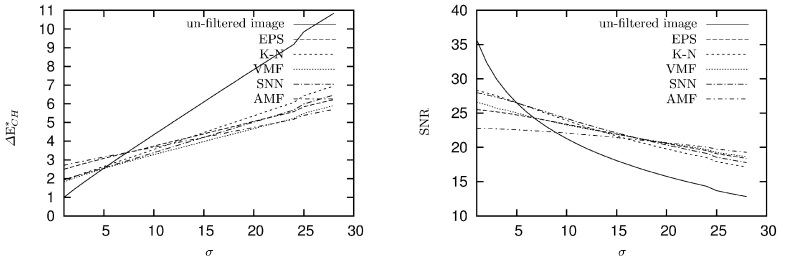


Fig. 6. Example image "`peppers`" corrupted by additive Gaussian noise

the measurements. For example, the arithmetic mean filter is judged differently for impulsive noise as it yields similar SNR but considerably different $\Delta \mathrm{E}^*_{CH}$.

As a conclusion we realize that better looking images (as quantitatively judged by $\Delta \mathrm{E}^*_{CH}$) will not always be best suited for further processing, as they may contain less information (as quantitatively judged by SNR) than images appearing worse visually.

References

1. Jaakko Astola, Pekka Haavisto, and Yrjo Neuvo. Vector median filters. *Proceedings of the IEEE*, 78:678–689, 1990.
2. Peter Bakker, L.J. van Fliet, and Piet W. Verbeek. Edge preserving orientation adaptive filtering. In *Proc. 5th Annual Conference of the Advanced School for Computing and Imaging*, pages 207–213, 1999.
3. Serge Chastel, Guido Schwab, and Dietrich Paulus. Web interface for image processing algorithms. In Simone Santini and Raimundo Schettini, editors, *Internet Imaging V* San Jose, 1 2004. Proc. of SPIE. volume 5304.
4. Rolf Gierling. *Farbmanagement*. Moderne Industrie Buch AG & Co. KG, Bonn 2001.
5. Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2001.
6. David Harwood, Murali Subbarao, H. Hakalahti, and L. Davis. A new class of edge-preserving smoothing filters. *Pattern Recognition Letters*, 5:155–162, 1987.
7. M. Kuwahara, K. Hachimura, S. Eiho, and M. Kinoshita. *Digital Processing of Biomedical Images*, chapter Processing of ri-angiocardiographic images, pages 187–202. Plenum Press, New York, USA, 1976.
8. M. Nagao and T. Matsuyama. Edge preserving smoothing. *Computer Graphics and Image Processing*, 9:394–407, 1979.
9. M. Pietikainen and David Harwood. *Advances in Image Processing and Pattern Recognition*, chapter Segmentation of color images using edge-preserving, pages 94–99. North Holland, Amsterdam, Netherlands, 1986.
10. Konstantinos N. Plataniotis and Anastasios N. Venetsanopoulos. *Color Image Processing and Applications*. Springer Verlag, 2000.
11. Volker Rehrmann, editor. *Erster Workshop Farbbildverarbeitung*, Universität Koblenz–Landau, 1995.
12. Manfred Richter. *Einführung in die Farbmetrik*. Walter de Gruyter, Berlin, New York, 2 edition, 1981.
13. Stephen J. Sangwine and R. E. N. Horne. *The Colour Image Processing Handbook*. Chapman Hall, London 1998.
14. Rein van den Boomgaard. Decomposition of the Kuwahara-Nagao operator in terms of linear smoothing and morphological sharpening. In *Proc. of the 6th International Symposium on Mathematical Morphology*, pages 283–292, 2002.
15. Günter Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulas*. John Wiley & Sons, second edition, 1982.
16. Xuemei Zhang and Brian A. Wandell. Color image fidelity metrics evaluated using image distortion maps. *Signal Processing*, 70(3):201–214, 11 1998.

Segmentation of Fundus Eye Images Using Methods of Mathematical Morphology for Glaucoma Diagnosis

Katarzyna Stąpor[1], Adam Świtonski[1], Radim Chrastek[2], and Georg Michelson[3]

[1] Institute of Computer Science, Silesian University of Technology, Akademicka 16, PL-44-100 Gliwice, Poland,
[2] Chair for Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg, Martenstrasse 3, D-91058 Erlangen, Germany,
[3] Department of Ophthalmology, Friedrich-Alexander-University Erlangen-Nuremberg Schwabachanlage 6, D-91054 Erlangen, Germany
delta@ivp.iinf.polsl.gliwice.pl

Abstract. In this paper the new method for automatic segmentation of cup and optic disc in fundus eye images taken from classical fundus camera is proposed. The proposed method is fully based on techniques from mathematical morphology. Detection of cup region makes use of watershed transformation with markers imposed, while optic disk is extracted based on geodesic reconstruction by dilation. The obtained results are encouraging.

1 Introduction

Glaucoma is a group of diseases characterized by the proceeding optic nerve neuropathy which leads to the rising diminution in vision field, ending with blindness. The correct optic disk (i.e. the exit of the optic nerve from the eye known as "blind spot") structure contains: neuroretinal rim of pink color and centrally placed yellowish cup [5]. The shape of the optic disc is more or less circular, interrupted by the outgoing vessels. Its size varies from patient to patient. Its diameter lies between 40 and 60 piksels on 640x480 color photographs. The cup is the area within the optic disc where no nerve fibers and blood vessels are present and in 3D image appears as an excavation. The neuroretinal rim is the area between optic disc border and cup border.

Glaucomatous changes in retina appearance embrace various changes in neuroretinal rim and cup, as the result of nerve fibers damages.

Optic disc structures evaluation is one of the most important examinations in glaucoma progress monitoring and diagnosis. Searching for glaucoma damages during routine examination is not an easy task and gives uncertain results even with the experienced ophthalmologist [5]. The existing methods of qualitative analysis are very subjective, while quantitative methods of optic disc morphology evaluation (cup to disc ratio, neuroretinal rim area) do not result in full

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 41–48, 2004.
© Springer-Verlag Berlin Heidelberg 2004

diagnosis. The new, quantitative methods based on scanning-laser-tomography are expensive and accessible only in specialized ophthalmic centers.

Thus, there is a need for cheaper and more objective methods that enable automatic classification of digital fundus eye images (fei) into normal and glaucomatous ones. The first, but most important step in all such methods is automatic segmentation of optic disc structures from fei.

In the existing approaches to automatic segmentation of fei for supporting glaucoma examinations researchers focused on the detection of the optic disk. These methods make use of Hough transform [8], active contours [6], tools from mathematical morphology [12].

In [4] important proofs that shape of the eye cup and its numerical characteristics correlate with progress of glaucoma disease were presented.

As far as we know, no automatic method for the segmentation of cup from fei has been reported.

This paper presents the new method for automatic segmentation of cup region as well as optic disc from fei taken from classical fundus camera. The proposed method is fully based on techniques from mathematical morphology: mainly watershed transformation and geodesic reconstruction.

2 Some Grey-Level Morphological Operators

In this section we briefly define the basic morphological operators used in this paper (for a comprehensive presentation see [1,11]). Let D_f and D_B be subsets of Z^2 and $T = \{t_{min}, ..., t_{max}\}$ be an ordered set of gray levels. A grey-level image f can be defined as a function:

$$f : D_f \subset Z^2 \to T, \tag{1}$$

Furthermore, we define another image known as a structuring element B:

$$B : D_B \subset Z^2 \to T, \tag{2}$$

We will restrict to flat, symmetric structuring elements B. We can now write the four basic morphological operators: erosion, dilation, opening and closing as:

$$E^B(f)(x,y) = \inf_{(j,k)\in D_B} \{f(x-j, y-k)\}, \tag{3}$$

$$D^B(f)(x,y) = \sup_{(j,k)\in D_B} \{f(x-j, y-k)\}, \tag{4}$$

$$O^B(f)(x,y) = D^B(E^B(f)(x,y)), \tag{5}$$

$$C^B(f)(x,y) = E^B(D^B(f)(x,y)), \tag{6}$$

Symmetric, morphological gradient of an image f can be defined as:

$$grad(f)(x,y) = D^B(f)(x,y) - E^B(f)(x,y), \tag{7}$$

Furthermore, we shall define geodesic transformation of an image f: geodesic erosion and dilation of size n:

$$E^{B,g}_{c(n)}(f)(x,y) = E^{B,g}_{c(1)}((E^{B,g}_{c(n-1)}(f))(x,y) \tag{8}$$

$$E^{B,g}_{c(1)}(f)(x,y) = sup\{E^{B}(f)(x,y), g(x,y)\} \tag{9}$$

$$D^{B,g}_{c(n)}(f)(x,y) = D^{B,g}_{c(1)}((D^{B,g}_{c(n-1)}(f))(x,y) \tag{10}$$

$$D^{B,g}_{c(1)}(f)(x,y) = inf\{D^{B}(f)(x,y), g(x,y)\} \tag{11}$$

reconstruction by dilation and reconstruction by erosion:

$$R^{B,f}(g)(x,y) = D^{B,g}_{c(\infty)}(f)(x,y) \tag{12}$$

$$R^{B,f}_{*}(g)(x,y) = E^{B,g}_{c(\infty)}(f)(x,y) \tag{13}$$

where $D^{B,g}_{c(\infty)}$ ($E^{B,g}_{c(\infty)}$) is the limit which is obtained by iterating unit geodesic erosion of f above (under) g until stability is reached, i.e.:

$$D^{B,g}_{c(i)}(f)(x,y) = D^{B,g}_{c(i+1)}(f)(x,y) \tag{14}$$

Furthermore, we shall make use of the watershed transformation, for which we do not give a mathematical definition here.

3 Detection of Cup Region Based on Watershed Transformation

3.1 The Color Space

Having compared several color spaces, we found the contours of the cup region to appear most continuous and most contrasted against the background in the a channel of the Lab color space [3] (image f_a- Fig. 1a).

3.2 Pre-processing

First, we apply a shade-correction operator in order to remove slow background variations. Thus, we calculate:

$$f^1_a = f_a - ASF(f_a) + k, \tag{15}$$

where k is a positive constant and

$$ASF(f_a) = C^{nB}(...(O^{2B}(O^{B}(f_a)))...), \tag{16}$$

is the result of alternating sequential filtering of f_a with n sufficiently large to remove the cup. This is the approximation of the slow variations of the background of image f_a.

Next, the image f_a^1 is filtered in order to eliminate large grey level variations within a cup region caused by the outgoing vessels. We "fill" the vessels by applying a simple closing operation:

$$f_a^2 = C^B(f_a^1), \tag{17}$$

with a structuring element B bigger than the maximal width of vessels - we use a circle with a radius 15. The result is shown in Fig. 1b.

3.3 Finding Markers for Watershed Transformation

To automatically find the internal marker, we first localize approximately the cup region. As we know approximately the size of the cup, and assuming that parts of the cup region belong to the darkest parts of the image f_a^2, we apply a simple global thresholding:

$$f_a^3 = T_{t1}(f_a^2), \tag{18}$$

to obtain a binary image f_a^3. The parameter $t1$ is chosen based on a cumulative histogram in a way that 2% of the darkest pixels will be considered as objects. The image f_a^3 contains cup region as well as other pathologies, like for example exudates. They are not very big, and they are far from reaching the size of the cup. Hence, we select the biggest particle of the image f_a^3 (giving image f_a^4 shown in Fig. 1c) which coincides with the candidate region containing the cup. The coordinates of the centroid c of the candidate region in the image f_a^4 are calculated as:

$$x_c = \frac{\sum_x \sum_y f_a^4(x,y) \cdot x}{\sum_x \sum_y f_a^4(x,y)}, y_c = \frac{\sum_x \sum_y f_a^4(x,y) \cdot y}{\sum_x \sum_y f_a^4(x,y)} \tag{19}$$

The calculated centroid $c(x_c, y_c)$ is taken as the internal marker m_{int} for watershed transformation. As external marker m_{ext} we use a circle $C_{cup}(c, r_{cup})$ with a center at the calculated centroid c and a radius r_{cup} bigger than the diameter of the biggest cup (see Fig. 1d).

3.4 Watershed Transformation

In order to detect exact contours of the cup region, we apply the classical watershed transformation:

$$f_a^6 = WT^{m_{int} \cup m_{ext}}(f_a^5), \tag{20}$$

to the morphological gradient of the filtered image f_a^2:

$$f_a^5 = grad(f_a^2) = D^B(f_a^2) - E^B(f_a^2) \tag{21}$$

with internal m_{int} and external m_{ext} markers imposed to avoid oversegmentation of the image. The watershed transformation assigns to each local minimum of f_a^5 one catchment basin (one connected region), in a way that all $x \in D_{f_a^5}$ belong to a basin except a one pixel strong line that delimits the basins (the watershed line). This watershed line defines the exact contour of the cup.

3.5 Post-processing

The shape irregularities in the segmentation result are due to the outgoing vessels or to low contrast. We can regularize the shape of the contour using standard morphological filtering techniques, i.e. smoothing by binary opening-closing operations of the resulted cup region. The final segmentation result is shown in Fig. 1f.

4 Detection of the Optic Disc Using Geodesic Reconstruction

Having compared several color spaces, we found that the contours of the optic disc are most contrasted against the background in the G channel of the RGB color space (Fig. 2a - image f_G). Finding the contours of the optic disc is based on the morphological reconstruction by dilation of the image f_G :

$$f_G^1 = R_D^{B,g}(f_G), \tag{22}$$

starting from a properly constructed marker image $g(x)$ (Fig. 2b)

$$g(x) = \begin{cases} 0 \quad if \quad x \in C_{disc}(c, r_{disc}) \\ f_G \; if \quad x \notin C_{disc}(c, r_{disc}) \end{cases} \tag{23}$$

where $C_{disc}(c, r_{disc})$ is a circle with a center, being the centroid $c(x_c, y_c)$ calculated during cup region detection (as described in the previous section), and a radius r_{disc} equal to a radius of a typical optic disc. This circle is a candidate region comprising the optic disc inside. All the pixels in a marker image $g(x)$ which are comprised in a circle $C_{disc}(c, r_{disc})$ are set to zero (means black). The reconstruction operator propagates the values of $f_G(x)$ of pixels x next to the circle into the circle by the successive geodesic dilation under the mask f_G. As optic disc is entirely comprised within the circle, it is completely removed, whereas regions that are not entirely comprised in the circle are nearly entirely reconstructed. Next, a simple thresholding operation is applied to the difference between the original image f_G and the reconstructed image f_G^1:

$$f_G^2 = T_{t2}(f_G - f_G^1), \tag{24}$$

The threshold T_{t2} has been experimentally set so as to differentiate between optic disc region and the rest of the image. To avoid influence of the vessels in the image f_G^2 (Fig. 2e), the binary closing operation is performed on the image f_G^2 which gives the result - the optic disc region. Its contour is shown in Fig. 2f imposed on the input image.

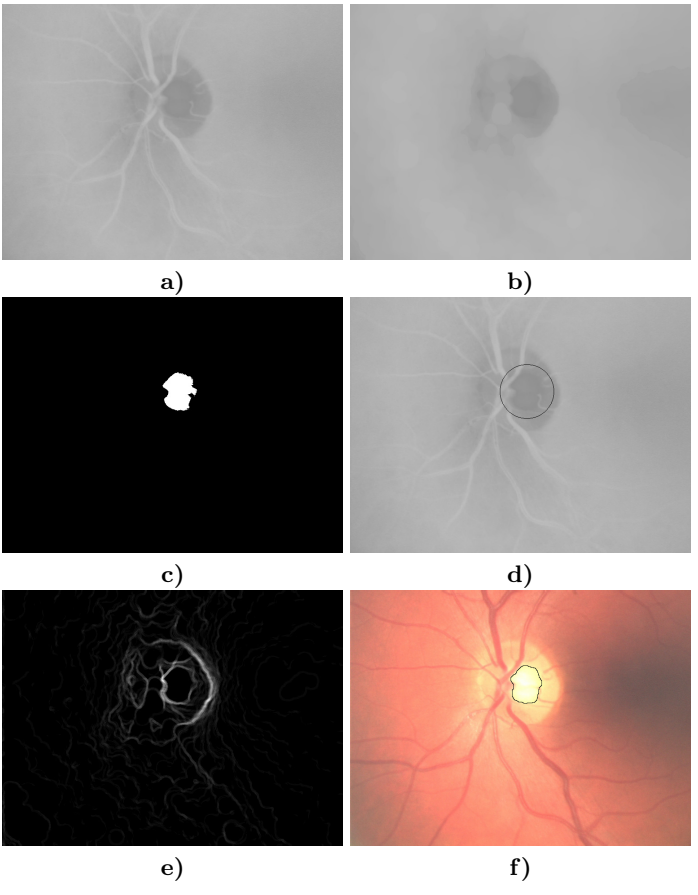


Fig. 1. The detection of the cup region: a) channel a of the input image (image f_a); b) channel a after closing operation (image f_a^2); c) candidate region on the thresholded image (image f_a^4); d) channel a with the external marker imposed; e) morphological gradient of image f_a^2 (imagef_a^5); f) result of the segmentation - watershed line imposed on the input image

5 Results and Conclusions

We tested the algorithm on 50-ty images of size 640x480 that have not been used for the development of the algorithm. The images were taken from the Canon CF-60Uvi fundus-camera. In 48 images we could found exact cup and disc contours. However in two images the contrast was to low, the algorithm failed and the result was not acceptable. It is important to note that the contours of the cup and optic disc obtained as a result of the presented segmentation methods coincide with the contours marked by an ophthalmologist. The results of using the presented method are very encouraging.

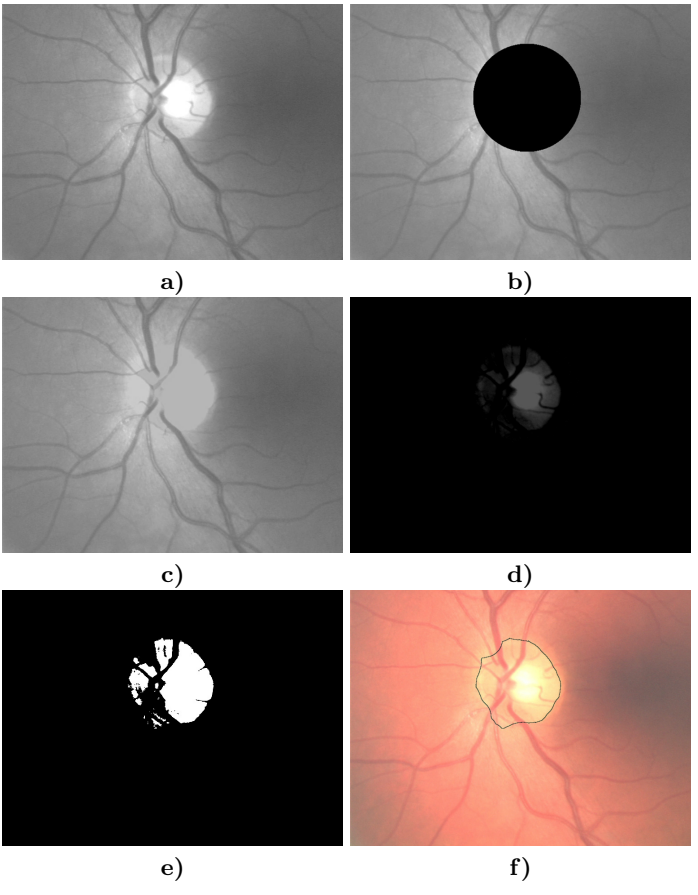


Fig. 2. The detection of the optic disc: a) channel G of the input image (image f_G); b) the marker image $g(x)$; c) the reconstructed image f_G^1; d) the difference between original and reconstructed images; e) the result after thresholding the difference image; f) result of the segmentation - contours of the detected optic disc imposed on the input image

References

1. Beucher S., Meyer F.: The morphological approach to image segmentation: the watershed transformation. E.R. Dougherty (Eds.): Mathematical morphology in image processing (1993) 433-481
2. Goh K.G, et al: ADRIS: an Automatic Diabetic Retinal Image Screening system. K. J. Cios (Eds.): Medical Data Mining and Knowledge Discovery. Springer-Verlag New York (2000) (181-210)
3. Gonzalez R.C., Woods R.E.: Digital image processing. Prentice-Hall (2002)
4. Jonas J. et al.: Ophthalmoscopic evalutation of the optic nerve head. Survey of Ophthalmology. Vol. 43, No. 4 (January - February 1999)

5. Kanski J. et al. Glaucoma: a color manual of diagnosis and treatment. Butterworth-Heinemann (1996)
6. Morris D.T., Donnison C.: Identifying the Neuroretinal Rim Boundary Using Dynamic Contours. Image and Vision Computing, Vol. 17. (1999) 169-174
7. Osareh A., et al.: Classification and localisation of diabetic related eye disease. A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353 (2002) 502-516
8. Pinz A., et al.: Mapping the human retina. IEEE Trans. Medical Imaging, Vol.1, (1998) (210-215)
9. soi Soille P.: Morphological Image analysis: principles and applications. Springer-Verlag, Berlin (1999)
10. Tamura S., Okamoto Y.: Zero-crossing interval correction in tracing eye-fundus blood vessels. Pattern Recognition, Vol.21, No. 3, (1988) (227-233)
11. Vincent L.: Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. IEEE Trans. On Image processing, Vol.2, No. 2, (1993), (176-201)
12. Walter T., Klein J.: Segmentation of color fundus images of the human retina: detection of the optic disc and the vascular tree using morphological techniques. Proc. 2nd Int. Symp. Medical Data Analysis, (1999) (282-287)

Automatic Detection of Glaucomatous Changes Using Adaptive Thresholding and Neural Networks

Katarzyna Stąpor[1], Lesław Pawlaczyk[1], Radim Chrastek[2], and Georg Michelson[3]

[1] Institute of Computer Science, Silesian University of Technology, Akademicka 16, PL-44-100 Gliwice, Poland,
[2] Chair for Pattern Recognition, Friedrich-Alexander-University Erlangen-Nuremberg, Martenstrasse 3, D-91058 Erlangen, Germany,
[3] Department of Ophthalmology, Friedrich-Alexander-University Erlangen-Nuremberg Schwabachanlage 6, D-91054 Erlangen, Germany
delta@ivp.iinf.polsl.gliwice.pl

Abstract. In this paper the new method for automatic classification of fundus eye images into normal and glaucomatous ones is proposed. The new, morphological features for quantitative cup evaluation are proposed based on genetic algorithms. For computation of these features the original method for automatic segmentation of the cup contour is proposed. The computed features are then used in classification procedure which is based on multilayer perceptron. The mean sensitivity is 90%, while the mean specificity: 86%. The obtained results are encouraging.

1 Introduction

Glaucoma is a group of diseases characterized by the proceeding optic nerve neuropathy which leads to the rising diminution in vision field, ending with blindness. The correct optic disk (i.e. the exit of the optic nerve from the eye known as "blind spot") structure contains: neuroretinal rim of pink color and centrally placed yellowish cup [6] (Fig. 2a). The cup is the area within the optic disc where no nerve fibers and blood vessels are present and in 3D image appears as an excavation. The neuroretinal rim is the area between optic disc border and cup border - see Fig. 2a. Glaucomatous changes in retina appearance embrace various changes in neuroretinal rim and cup, as the result of nerve fibers damages.

Optic disc structures evaluation is one of the most important examinations in glaucoma progress monitoring and diagnosis. Searching for glaucoma damages during routine examination is not an easy task and gives uncertain results even with the experienced ophthalmologist [6]. The existing methods of qualitative analysis are very subjective, while quantitative methods of optic disc morphology evaluation (cup to disc ratio, neuroretinal rim area) do not result in full diagnosis. The new methods of morphologic analysis based on scanning-laser-tomography are expensive and accessible only in specialized ophthalmic centers.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 49–55, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In the existing approaches for supporting glaucoma diagnosing [4,7,8] the automatic extraction of the cup region from fei was not the area of interest. Also, automatic classification of single fei acquired from fundus cameras into normal and glaucomatous has received no attention.

That is why we have developed a more objective and cheaper method that enables automatic classification of digital fundus eye images (fei) into normal and glaucomatous ones. The fei images were obtained by classical fundus-camera. We plan to build the proposed methodology into classical fundus-camera software to be used in routine examinations by an ophthalmologist.

2 Methods

The proposed method for automatic detection of glaucomatous changes in fundus eye images is composed of the 3 main stages (shown in Fig. 1):

1. detection of the cup contour,
2. selection of the cup features using genetic algorithms,
3. classification of fundus eye images using neural network classifier

2.1 Automatic Detection of the Cup Contour

Digital fei are acquired from classical fundus camera in RGB additive color model [5]. The color normalization step using histogram specification [5] is performed to decrease the variation in the color of fei from different patients. A copy of the acquired fei is converted into HSV color model [5]. On RGB image blood vessels are detected automatically using a set of contour filters according to a method described in [3].

Based on the detected vessels, the averaging of H,S,V components in HSV image is performed to decrease the contrast. All pixels comprising the detected vessels lying inside the user rectangle belong to the subregion named here $R_{eyecup_vessels}$. First, the input image is converted from RGB to HSV color model [5]. By overlying the image with detected vessels on the input, converted image all border pixels of the detected vessels are found (subregion $R_{eyecup_vessels}$). For each border pixel in $R_{eyecup_vessels}$ its new color components $[H_{avg},S_{avg},V_{avg}]$, being the average of the appropriate components of pixels lying in the 8-connected neighborhood outside of $R_{eyecup_vessels}$ region are found. After recalculation of all border pixels, they are deleted, new border pixels are found and the process is repeated until size of $R_{eyecup_vessels}$ is higher than 0.

This preprocessed HSV image is converted into L*a*b* color model [5]. For further examinations only channel a* is used. Next, the a* component of L*a*b* image is binarized by the proposed adaptive thresholding method which results in white pixels of the cup (i.e. the object) and black pixels of the rest of the image (i.e. the background). In the adaptive thresholding method a local threshold is found by statistically examining the intensity values of a local neighborhood of

each pixel. A window cantered at each pixel is constructed as its local neighborhood. The statistic used is a function:

$$T = M_{mean} - C, \tag{1}$$

where M_{mean} is a mean of gray level values in the window, C is a constans, experimentally set.

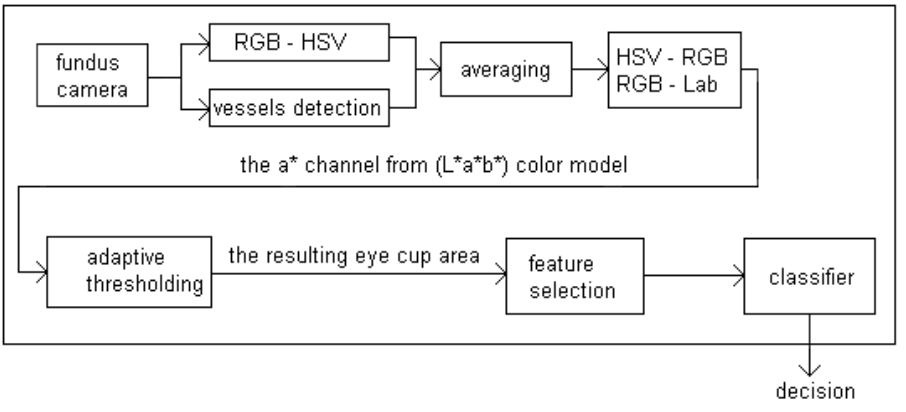
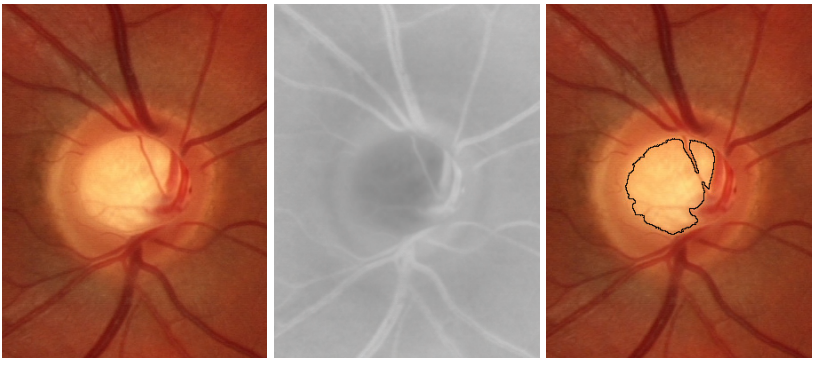


Fig. 1. Stages of the eye cup segmentation method

Due to nerve fibres damages during glaucoma progress, different changes in a shape of the neuroretinal rim (and of the cup) are observed. Proper shape feature selection can reduce not only the cost of recognition by reducing the number of features that need to be collected, but in some cases it can also provide a better classification accuracy due to finite sample size effect In our approach, 29 geometric features are computed on the extracted cup region. These are: seven Hu moment invariants [9], fifteen compound invariant moments [9], two circular coefficients [9], area to perimeter coefficient, Danielsson, Haralick, Blair-Bliss and Feret coefficients [9]. Genetic algorithms [1] are then used to select the most significant features characterizing the shape of cup region.

A given feature subset is represented as a binary string (a chromosome) of length n, with a zero or one in position i denoting the absence or presence of feature i in the set (n is the total number of available features). The initial population is generated in the following way: the number of 1's for each chromosome is generated randomly, then, the 1's are randomly scattered in the chromosome. A population of chromosomes is maintained. Each chromosome is evaluated to determine its "fitness", which determines how likely the chromosome is to survive and breed into next generation. We proposed the following fitness function:

$$Fitness = 10^4 accuracy + 0.4 zeros, \tag{2}$$



a) b) c)

Fig. 2. a) The initial image with the optic disk and the cup in the central part; b) channel a* of the input image; c) the contour of the extracted cup region overlaid on the input image

where accuracy is the accuracy rate that the given subset of features achieves (i.e. the performance of a classifier on a given subset of features), zeros is the number of zeros in the chromosome. Reproduction is based on a random choice according to a fraction with repetitions method [1]. New chromosomes are created from old chromosomes by the process of crossover and mutation [1].

The following 3 dimensional feature vector has been selected from a set of 29 features by genetic algorithm: (FI2, I3, RF), where

$$FI2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2, \tag{3}$$

is Hu invariant moment, where: η_{20}, η_{02}, η_{11} are normalized central moments. Normalized central moment of order (p+q) is defined as [5]:

$$\mu_{pq} = \frac{m_{pq}}{(m_{00})^{\alpha}}, \alpha = \frac{p+q}{2} + 1, \tag{4}$$

where: m_{pq} is a spatial central moment of order $p+q$ of an image f defined as:

$$m_{pq} = \sum_{i=1}^{m}\sum_{j=1}^{n}(i-I)^p(j-J)^q f(i,j), \tag{5}$$

$$I = \frac{m_{10}}{m_{00}}, \tag{6}$$

$$J = \frac{m_{01}}{m_{00}}, \tag{7}$$

$$I3 = \mu_{20}(\mu_{21}\mu_{03} - \mu_{12}^2) - \mu_{11}(\mu_{30}\mu_{03} - \mu_{21}\mu_{12}) + \mu_{02}(\mu_{30}\mu_{12} - \mu_{12}^2), \tag{8}$$

is compound, invariant moment.

$$R_F = \frac{L_h}{L_V} \tag{9}$$

is Feret coefficient, where:
L_h - maximal diameter in horizontal direction
L_V - maximal diameter in vertical direction.

2.2 Classification of Fundus Eye Images Using Neural Network Classifier

The method makes use of the 3-2-2 multilayer perceptron (MLP) [2]. The operation of MLP is specified by:

$$V_j^1 = f(\sum_k w_{jk} V_k^0), \tag{10}$$

$$V_j^2 = f(\sum_k w_{jk} V_k^1), \tag{11}$$

which specifies how input pattern vector V_k^0 is mapped into output pattern vector V_k^2 via the hidden pattern vector V_k^1 in a manner parameterized by the two layers of weights: w_{ij}^1,w_{ij}^2. The univariate function f is set to:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{12}$$

The weights in the network are modified during training to optimize the match between outputs and targets d_i using standard backpropagation rule [2]:

$$w_{ij}^{m-new} = w_{ij}^{m-old} + \eta \delta_i^m V_{ij}^{m-1}, \tag{13}$$

where:

$$\delta_i^M = f'(\sum_j w i_{ij}^M V_j^{M-1})[d_i - V_i^M] \tag{14}$$

delta-error for i_{th} neuron in output layer M,

$$\delta_i^{m-1} = f'(\sum_j w i_{ij}^{m-1} V_j^{m-2}) \sum_j w_{ji}^m \delta j^m \tag{15}$$

$m = M, M-1, \ldots, 2$ delta-error for i_{th} neuron in hidden layer m. The trained network (classifier) can be used to determine which class of pattern in the training data each neuron in the network responds most strongly to. Unseen data can then be classified according to the class label of the neuron with the strongest activation for each pattern.

3 Results

The developed method has been applied into 100 fei of patients with glaucoma and 100 fei of normal patients which where previously examined by conventional methods by ophthalmologist. On the acquired from Canon CF-60Uvi funduscamera images, the cup contour is automatically detected. Next, for the detected

cup the whole set of 29 geometric features is computed. The obtained set of labeled feature vectors is divided into 4 parts: two training and two testing sets. One pair composed of one training and one testing set is used by genetic algorithms for suboptimal feature vector calculation, while the second pair of sets for calculation of a performance of neural network classifier.

The parameters of genetic algorithm used in all experiments are as follows: the length of each chromosome is 29 (equal to the number of features), population size is 120. Genetic algorithm converged to the final solution after 150 generations.

The parameters of a neural network classifier are as follows: the structure is set as 3-2-2 as described above, sigmoidal function is used as activation functions in hidden and output layers. The learning rate η is equal to 1. Weights w_{ij} are initialized to the small random values from $(-1.5, 1.5)$ interval. Classifier performance is tested by k-fold cross validation method. During performance evaluation, the constructed classifier ran 5000 iterations to train and updated the weights each time training data were presented. The following mean results has been obtained: sensitivity 90% and specificity 86%.

4 Conclusions

As far as we know no automatic method for the segmentation and classification of fei acquired from fundus-cameras into normal and glaucomatous has been reported yet. Our method proves that shape of the cup and its numerical characteristics correlate with progress of glaucoma. It also shows that by reducing irrelevant information and using only selected features the classifier performance can be improved significantly which is very important for application supporting glaucoma diagnosing.

The obtained results are encouraging. It is expected that the new method, after clinical tests would support glaucoma diagnosis based on digital fei obtained from fundus-camera.

References

1. Arabas J.: Lectures on genetic algorithms. WNT, Warsaw (2001)
2. Bishop C.M.: Neural networks for pattern recognition. Clarendon Press, Oxford (1995)
3. Chaudhuri S., et al.: Detection of Blood Vessels in Retinal Images Using Two-Dimensional Matched Filter. IEEE Transactions on Medical Imaging, Vol 8, No. 3. (September 1989)
4. Goh K.G, et al: ADRIS An Automatic Diabetic Retinal Image Screening system. K.J. Cios (Ed.): Medical Data Mining and Knowledge Discovery. Springer-Verlag, New York (November 2000) 181-201
5. Gonzalez R.C., Woods R.E.: Digital image processing. Prentice-Hall (2002)
6. Kanski J. et al. Glaucoma: a color manual of diagnosis and treatment. Butterworth-Heinemann (1996)

7. Morris D.T., Donnison C.: Identifying the Neuroretinal Rim Boundary Using Dynamic Contours. Image and Vision Computing, Vol. 17. (1999) 169-174
8. Osareh A., et al.: Classification and localisation of diabetic related eye disease. A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353 (2002) 502-516
9. Trier O., Jain A., Taxt T.: Feature extraction methods for character recognition - a survey. Pattern Recognition, (1996) 641-662

Analytical Design of 2-D Narrow Bandstop FIR Filters

Pavel Zahradnik[1] and Miroslav Vlček[2]

[1] Department of Telecommunications Engineering
Czech Technical University Prague
Technická 2, CZ-166 27 Praha, Czech Republic
Phone: +420-2-24352089, Fax: +420-2-33339810
zahradni@fel.cvut.cz
[2] Department of Applied Mathematics
Czech Technical University Prague
Konviktská 20, CZ-110 00 Praha, Czech Republic
Phone: +420-2-24890720, Fax:+420-2-24890702
vlcek@fel.cvut.cz

Abstract. Novel approach in the design of 2-D extremely narrow bandstop FIR filters is presented. The completely analytical design method is based on the 1-D optimal bandstop FIR filters. The 1-D FIR optimal bandstop filters are based on Zolotarev polynomials. Closed form formulas for the design of the filters are presented. One example demonstrates the design procedure. One application of the 2-D FIR filter with extremely narrow stop bands is presented.

1 Introduction

Two-dimensional narrow bandstop FIR filters play important role in the image and video enhancement/restoration tasks. They are frequently used in order to remove a single frequency component from the spectrum of the signal. The image filtering can be accomplished by both the nonlinear [1], [4], [7], [8], [9], [10] and linear [5], [12], [13], [14], [15] filters. In our paper we are concerned with completely analytical design of 2-D bandstop FIR filters with extremely narrow circularly symmetrical stop bands. The design of the 2-D narrow bandstop FIR filters is based on the 1-D optimal narrow bandstop FIR filters [14]. We introduce the degree formula which relates the degree of the generating polynomial, the length of the filter, the notch frequency, the width of the stopbands and the attenuation in the passbands. Based on the expansion of the generating polynomials into the Chebyshev polynomials, the recursive formula for the direct computation of the impulse response coefficients is presented. The design procedure is recursive one and it does not require any FFT algorithm or any iterative technique.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 56–63, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Polynomial Equiripple Approximation

Let us denote $H(z)$ the transfer function of a 1-D FIR filter with the impulse response $h(m)$ of the length N as

$$H(z) = \sum_{m=0}^{N-1} h(m) z^{-m} . \tag{1}$$

Assuming an odd length $N = 2n+1$ and even symmetry of the impulse response $h(m)$ we can write the transfer function of the bandstop FIR filter

$$H(z) = z^{-n} \left[h(0) + \sum_{m=1}^{n} 2h(m)\, T_m(w) \right] = z^{-n} \left[h(0) + \sum_{m=1}^{n} 2h(m)\, T_m(\cos \omega T) \right] \tag{2}$$

where $T_m(w)$ is Chebyshev polynomial of the first kind and $w = (z+z^{-1})/2$. The

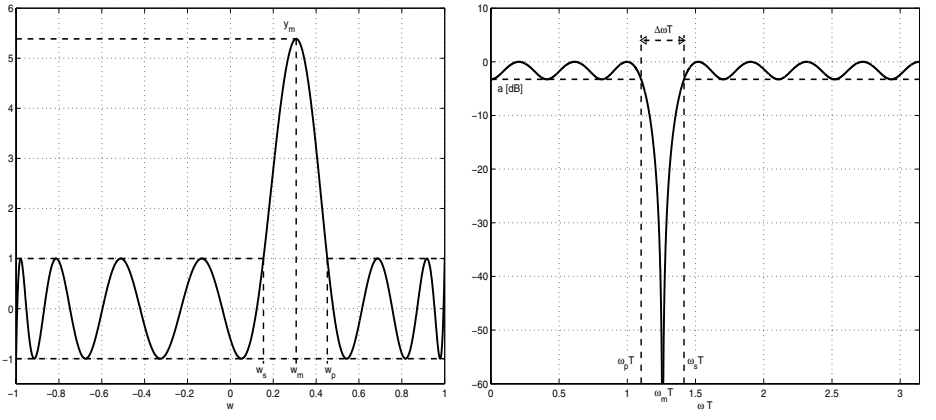


Fig. 1. Zolotarev polynomial $Z_{6,9}(w)$ with $\kappa = 0.6966$, $w_s = 0.1543$, $w_m = 0.3071$, $w_p = 0.4523$, $y_m = 5.3864$ and corresponding amplitude frequency response $20 \log |H(e^{j\omega T})|$ [dB] with parameters $\omega_p T = 0.3506\,\pi$, $\omega_m T = 0.4006\,\pi$, $\omega_s T = 0.4507\,\pi$, $\Delta\omega T = 0.1001\,\pi$ and $a = -3.2634$ dB

1-D equiripple narrow bandstop FIR filter is based on the Zolotarev polynomial $Z_{p,q}(w)$ which approximates constant value in equiripple Chebyshev sense in the two disjoint intervals as shown in Fig. 1. The notation $Z_{p,q}(w)$ emphasizes that p counts the number of zeros right from the maximum w_m and q corresponds to the number of zeros left from the maximum w_m. Zolotarev derived the general solution of this approximation problem in terms of Jacobi's elliptic functions

$$Z_{p,q}(w) = \frac{(-1)^p}{2} \left[\left(\frac{H(u - \frac{p}{n}\, \mathbf{K}(\kappa))}{H(u + \frac{p}{n}\, \mathbf{K}(\kappa))} \right)^n + \left(\frac{H(u + \frac{p}{n}\, \mathbf{K}(\kappa))}{H(u - \frac{p}{n}\, \mathbf{K}(\kappa))} \right)^n \right] \tag{3}$$

where u is expressed by the incomplete elliptical integral of the first kind

$$u = F\left(\mathrm{sn}\left(\frac{p}{n}\ \mathbf{K}(\kappa)|\kappa\right)\sqrt{\frac{1+w}{w+2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)-1}}|\kappa \right) . \tag{4}$$

The function $H\left(u \pm \frac{p}{n}\ \mathbf{K}(\kappa)\right)$ is the Jacobi's Eta function, $\mathrm{sn}(u|\kappa)$, $\mathrm{cn}(u|\kappa)$, $\mathrm{dn}(u|\kappa)$ are Jacobi's elliptic functions, $\mathbf{K}(\kappa)$ is the quarter-period given by the complete elliptic integral of the first kind, $F(\phi|\kappa)$ is the incomplete elliptic integral of the first kind and κ is the Jacobi's elliptic modulus. The degree of the Zolotarev polynomial is $n = p + q$. A comprehensive treatise of the Zolotarev polynomials was published in [14]. It includes the analytical solution of the coefficients of Zolotarev polynomials, the algebraic evaluation of the Jacobi's Zeta function $\mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ and of the elliptic integral of the third kind $\Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ of the discrete argument. The position of the maximum value $y_m = Z_{p,q}(w_m)$ is

$$w_m = 1 - 2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right) + 2\,\frac{\mathrm{sn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)\mathrm{cn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}{\mathrm{dn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}\mathrm{Z}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right) \ . \tag{5}$$

The maximum value y_m useful for the normalization of the Zolotarev polynomial is given as

$$y_m = \cosh 2n\left(\sigma_m \mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\kappa) - \Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\kappa)\right) \ . \tag{6}$$

The degree of the Zolotarev polynomial $Z_{p,q}(w)$ expresses the *degree equation*

$$n \geq \frac{\ln(y_m + \sqrt{y_m^2 - 1})}{2\sigma_m \mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\kappa) - 2\Pi(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\kappa)} \ . \tag{7}$$

The auxiliary parameter σ_m is given by the formula

$$\sigma_m = F\left(\arcsin\left(\frac{1}{\kappa\ \mathrm{sn}\left(\frac{p}{n}\ \mathbf{K}(\kappa)|\kappa\right)}\sqrt{\frac{w_m - w_s}{w_m + 1}}\right)|\kappa\right) \tag{8}$$

where $F(\Phi|\kappa)$ is the incomplete elliptical integral of the first kind. The recursive algorithm for the evaluation of the coefficients $a(m)$ of the Zolotarev polynomial based on the expansion into Chebyshev polynomials of the first kind

$$Z_{p,q}(w) = \sum_{m=0}^{n} a(m) T_m(w) \tag{9}$$

was derived and presented in [14]. It is summarized in Table 1. The impulse response coefficients $h(m)$ of the 1-D equiripple bandstop FIR filter are obtained by the normalization of the coefficients $a(m)$ as follows

$$h(n) = \frac{y_m - a(0)}{y_m + 1} \ , \ h(n \pm m) = -\frac{a(m)}{2(y_m + 1)} \ , \ m = 1 \ldots n \ . \tag{10}$$

3 Analytical Design of 2-D Narrow Bandstop FIR Filter

The goal of the design of the 2-D narrow bandstop FIR filter is to obtain the 2-D impulse response $h(m,n)$ of the filter satisfying the specified notch frequency $\omega_{m1}T$, width of the bandstop $\Delta\omega_1 T$, the attenuation in the passbands a_1 [dB] in the direction ω_1 and the specified values $\omega_{m2}T$, $\Delta\omega_2 T$, a_2 [dB] in the direction ω_2. The design procedure is as follows :

1. For the specified values $\omega_{m1}T$, $\Delta\omega_1 T$ and a_1 [dB] (Fig. 1) in the direction ω_1 design the 1-D FIR narrow bandpass filter. The design procedure consists of the following steps :
   a) Evaluate the Jacobi's elliptic modulus κ

$$\kappa = \sqrt{1 - \frac{1}{\tan^2(\varphi_s)\tan^2(\varphi_p)}} \tag{11}$$

   for the auxiliary parameters φ_s and φ_p

$$\varphi_s = \frac{\omega_{m1} + \Delta\omega_1/2}{2}\,T \;\;,\;\; \varphi_p = \frac{\pi - (\omega_{m1} - \Delta\omega_1/2)}{2}\,T \;. \tag{12}$$

   b) Calculate the rational values $\dfrac{p}{n} = \dfrac{F(\varphi_s|\kappa)}{\mathbf{K}(\kappa)}$ and $\dfrac{q}{n} = \dfrac{F(\varphi_p|\kappa)}{\mathbf{K}(\kappa)}$.

   c) Determine the required maximum value y_m

$$y_m = \frac{2}{1 - 10^{0.05\,a_1[dB]}} - 1\,. \tag{13}$$

   d) Using the degree equation (7) calculate and round up the minimum degree n required to satisfy the filter specification. For the algebraic evaluation of the Jacobi's Zeta function $\mathrm{Z}(\frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ and the elliptic integral of the third kind $\mathit{\Pi}(\sigma_m, \frac{p}{n}\mathbf{K}(\kappa)|\,\kappa)$ in the degree equation (7) use the algebraical procedure [14].

   e) Calculate the integer values $p = \left[n\dfrac{F(\varphi_s|\kappa)}{\mathbf{K}(\kappa)}\right]$ and $q = \left[n\dfrac{F(\varphi_p|\kappa)}{\mathbf{K}(\kappa)}\right]$. The brackets [ ] stand for the rounding operation.

   f) For the integer values p, q and the elliptic modulus κ evaluate the coefficients $a(m)$ (9) of the Zolotarev polynomial $Z_{p,q}(w)$ using recursive algorithm summarized in Tab. 1.

   g) From the coefficients $a(m)$ calculate the M coefficients of the impulse response $h_1(m)$ of the 1-D equiripple bandpass FIR filter using (10).
2. Repeat the first step for the design of the 1-D FIR equiripple narrow bandpass filter in the direction ω_2 specified by $\omega_{m2}T$, $\Delta\omega_2 T$ and a_2 [dB] resulting in the impulse response $h_2(n)$ of the length N coefficients.
3. From the 1-D impulse responses

$$h_1(m)\,,\;\; m = 0,\;\ldots\,, M-1 \quad , \quad h_2(n)\,,\;\; n = 0,\;\ldots\,, N-1 \tag{14}$$

compose the 2-D impulse responses $h_1(m,n)$ and $h_2(m,n)$ by the zero padding. The non-zero coefficients are

$$\begin{aligned} h_1(\frac{M-1}{2},n) &= h_1(m), \quad m=0,\ldots,M-1 \\ h_2(m,\frac{N-1}{2}) &= h_2(n), \quad n=0,\ldots,N-1\ . \end{aligned} \tag{15}$$

4. The 2-D impulse response $h_{BP}(m,n)$ of the dimension $M \times N$ of the narrow bandpass FIR filter is given by the 2-D linear discrete convolution

$$h_{BP}(m,n) = h_1(m,n) * * h_2(m,n)\ . \tag{16}$$

5. The impulse response $h(m,n)$ of the final 2-D bandstop FIR filter is

$$\begin{aligned} h(m,n) = -h_{BP}(m,n) \text{ for } m &\neq \frac{M-1}{2} \\ n &\neq \frac{N-1}{2} \\ h(\frac{M-1}{2},\frac{N-1}{2}) = 1 - h_{BP}(\frac{M-1}{2},&\frac{N-1}{2})\ . \end{aligned} \tag{17}$$

4 Example

Design the 2-D bandstop FIR filter specified in the direction ω_1 by the notch frequency $\omega_{m1}T = 0.4\,\pi$, width of the passbands $\Delta\omega_1 T = 0.1\,\pi$ for the attenuation in the passbands $a_1 = -1$ dB and in the direction ω_2 by the values $\omega_{m2}T = 0.6\,\pi$, $\Delta\omega_2 T = 0.1\,\pi$ for $a_2 = -1$ dB.
Using our recursive design procedure we obtain the two 1-D equiripple narrow band FIR filters with the impulse responses $h_1(m)$, $h_2(n)$ (step 1 and 2 in Sec. 3). The impulse responses $h_1(m)$, $h_2(n)$ of the length $M = N = 41$ coefficients are summarized in Table 2. Their amplitude frequency responses are shown in Fig. 2. The impulse responses $h_1(m)$, $h_2(n)$ are used for the design of the 2-D bandstop FIR filter (step 3, 4 and 5 in Sec. 3). The impulse response $h(m,n)$ of the 2-D narrow bandstop FIR filter consists of 41×41 coefficients. The amplitude frequency response $20\log|H(e^{j\omega_1},e^{j\omega_2})|$ of the 2-D narrow bandstop FIR filter with its contours is shown in Fig. 3.

5 Application of the 2-D Narrow Bandstop FIR Filter

The narrow 2-D bandstop FIR filters were successfully applied for the removal of the unwanted frequency components in the spectrum of the image. Here we present the enhancement of the rastered newspaper picture. The notch frequencies $\omega_1 = 0.32\pi$, $\omega_2 = 0.42\pi$ to be removed were obtained by the evaluation of the spectrum of the input image. The impulse response $h(m,n)$ of the applied filter exhibits 37×37 coefficients. The input and processed image are shown in Fig. 4. The attenuation of the disturbing raster is apparent.

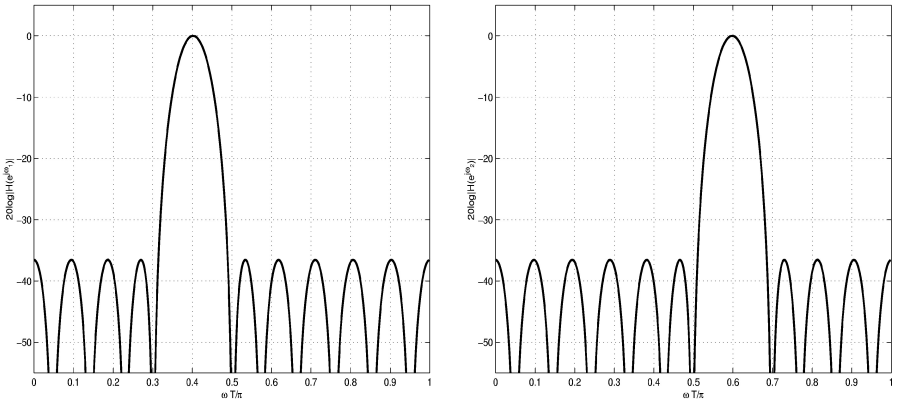


Fig. 2. Amplitude frequency responses $20\log|H(e^{j\omega_1})|$ and $20\log|H(e^{j\omega_2})|$

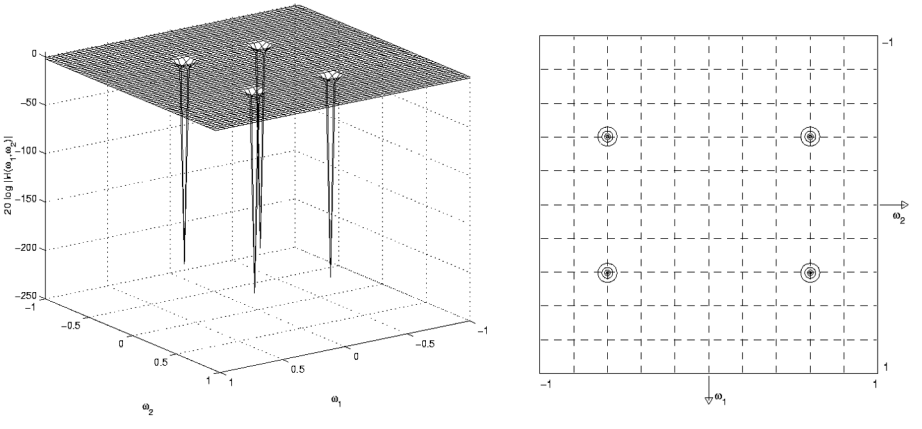


Fig. 3. Amplitude frequency response $|H(e^{j\omega_1}, e^{j\omega_2})|$ with contours

Fig. 4. Input and filtered image

Table 1. Recursive algorithm for the evaluation of the coefficients $a(m)$ of the Zolotarev polynomials $Z_{p,q}(w)$

given

$$p, q, \kappa$$

initialisation

$$n = p + q$$

$$w_p = 2\,\mathrm{sn}^2\left(\frac{q}{n}\mathbf{K}(\kappa)|\kappa\right) - 1$$

$$w_s = 1 - 2\,\mathrm{sn}^2\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)$$

$$w_a = \frac{w_p + w_s}{2}$$

$$w_m = w_s + 2\,\frac{\mathrm{sn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)\mathrm{cn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}{\mathrm{dn}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)}\mathrm{Z}\left(\frac{p}{n}\mathbf{K}(\kappa)|\kappa\right)$$

$$\alpha(n) = 1$$

$$\alpha(n+1) = \alpha(n+2) = \alpha(n+3) = \alpha(n+4) = \alpha(n+5) = 0$$

body

(*for* $m = n + 2$ *to* 3)

$$8c(1) = n^2 - (m+3)^2$$

$$4c(2) = (2m+5)(m+2)(w_m - w_a) + 3w_m[n^2 - (m+2)^2]$$

$$2c(3) = \frac{3}{4}[n^2 - (m+1)^2] + 3w_m[n^2 w_m - (m+1)^2 w_a] - (m+1)(m+2)(w_p w_s - w_m w_a)$$

$$c(4) = \frac{3}{2}(n^2 - m^2) + m^2(w_m - w_a) + w_m(n^2 w_m^2 - m^2 w_p w_s)$$

$$2c(5) = \frac{3}{4}[n^2 - (m-1)^2] + 3w_m[n^2 w_m - (m-1)^2 w_a] - (m-1)(m-2)(w_p w_s - w_m w_a)$$

$$4c(6) = (2m-5)(m-2)(w_m - w_a) + 3w_m[n^2 - (m-2)^2]$$

$$8c(7) = n^2 - (m-3)^2$$

$$\alpha(m-3) = \frac{1}{c(7)}\sum_{\mu=1}^{6} c(\mu)\alpha(m+4-\mu)$$

(*end loop on* m)

normalisation

$$s(n) = \frac{\alpha(0)}{2} + \sum_{m=1}^{n}\alpha(m)$$

$$a(0) = (-1)^p\frac{\alpha(0)}{2s(n)}$$

(*for* $m = 1$ *to* n)

$$a(m) = (-1)^p\frac{\alpha(m)}{s(n)}$$

(*end loop on* m)

Table 2. Coefficients of the Impulse Responses

m,n		$h_1(m)$	$h_2(n)$	m,n		$h_1(m)$	$h_2(n)$
0	40	0.008036	0.008036	11	29	0.020208	-0.020208
1	39	0.003713	-0.003713	12	28	-0.047824	-0.047824
2	38	-0.008856	-0.008856	13	27	-0.055411	0.055411
3	37	-0.013403	0.013403	14	26	0.019492	0.019492
4	36	0.004187	0.004187	15	25	0.075345	-0.075345
5	35	0.023801	-0.023801	16	24	0.026236	0.026236
6	34	0.011174	0.011174	17	23	-0.065318	0.065318
7	33	-0.025726	0.025726	18	22	-0.068525	-0.068525
8	32	-0.033363	-0.033363	19	21	0.025845	-0.025845
9	31	0.010947	-0.010947	20		0.093129	0.093129
10	30	0.050326	0.050326				

References

1. Astola J., Kuosmanen P.: Fundamentals of Nonlinear Digital Filtering. CRC Press, 1997.
2. Abramowitz M., Stegun I.: Handbook of Mathematical Function. Dover Publication, New York Inc., 1972.
3. Achieser N. I.: Über einige Funktionen, die in gegebenen Intervallen am wenigsten von Null abweichen. Bull. de la Soc. Phys. Math. de Kazan, Vol. 3, pp. 1 - 69, 1928.
4. Fischer V., Drutarovsky M., Lukac R.: Implementation of 3-D Adaptive LUM Smoother in Reconfigurable Hardware. Springer Verlag LNCS 2438, pp.720-729.
5. Gonzales R.C., Woods R.E.: Digital Image processing. Wiley Interscience, 2001.
6. Lawden D. F.: Elliptic Functions and Applications. Springer-Verlag, New York Inc., 1989.
7. Lukac R.: Binary LUM Smoothing. IEEE Signal Processing Letters, Vol. 9, No. 12, December 2002, pp. 400-403.
8. Lukac R.: Adaptive Vector Median Filtering. Pattern Recognition Letters, Vol. 24, No. 12, August 2003, pp. 1889-1899.
9. Lukac R.: Simplified Boolean LUM Smoothers. Proceedings of the 4th EURASIP-IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications VIPromCom-2002, Zadar, Croatia, June 16-19, 2002, pp. 159-162.
10. Lukac R.: The Way How to Design and Implement an Adaptive Method Based on Center-Weighted Medians. Proceedings of the IEEE Scientific Workshop Signal Processing 2002, Poznan, Poland, October 11, 2002, pp.9-14.
11. Pitas I., Venetsanopoulos A.N.: Nonlinear Digital Filters : Priciples and Applications. Kluwer Academic Publishers, 1990.
12. Pratt W. K., Venetsanopoulos A.N.: Digital Image processing. Kluwer Academic Publishers, 1990.
13. Vlček M., Jireš L.: Fast Design Algorithms for FIR Notch Filters. Proc. of IEEE International Symposium on Circuits and Systems ISCAS'94, London, Vol. 2, pp. 297 - 300, 1994.
14. Vlček M., Unbehauen R.: Zolotarev Polynomials and Optimal FIR Filters. IEEE Transactions on Signal Processing, Vol. 47, No. 3, pp. 717-730, March 1999.
15. Vlček M., Zahradnik P., Unbehauen R.: Analytic Design of FIR Filters. IEEE Transactions on Signal Processing, Vol. 48, pp. 2705-2709, September 2000.

Analytical Design of Arbitrary Oriented Asteroidal 2-D FIR Filters

Pavel Zahradnik[1] and Miroslav Vlček[2]

[1] Department of Telecommunications Engineering
Czech Technical University Prague
Technická 2, CZ-166 27 Praha, Czech Republic
Phone: +420-2-24352089, Fax: +420-2-33339810
zahradni@fel.cvut.cz
[2] Department of Applied Mathematics
Czech Technical University Prague
Konviktská 20, CZ-110 00 Praha, Czech Republic
Phone: +420-2-24890720, Fax:+420-2-24890702
vlcek@fel.cvut.cz

Abstract. Novel approach to the analytical design of asteroidally shaped two-dimensional FIR filters with arbitrary orientation of the frequency response in the frequency plane is presented. The design consists of two steps. In the first step, the asteroidal 2-D FIR filter in the basic position along the frequency coordinates is designed. The design is based on the analytical contour approximation using the parametric representation of the contour. Closed form formulas for the approximation of the contour with asteroidal shape were derived. In the second step, the asteroidally shaped FIR filter is rotated by the desired angle in the frequency plane. Closed form formulas for the calculation of the impulse response of the filter are presented. One example demonstrates both steps of the design procedure.

1 Introduction

Two-dimensional FIR filters with asteroidal contour of the frequency response represent advantageous choice in the processing of rectangular objects in the image area. This is based on the fact, that spectra of rectangular objects exhibit asteroidal shape. Because of the arbitrary orientation of the objects in the image area, filters with general orientation of the frequency response are desired. The image filtering can be accomplished by both the nonlinear [5], [6], [7], [8], [9] and linear [1], [2], [10] filters. In the design of 2-D FIR filters the McClellan transformation technique [1] became popular. Methods for the design of 2-D FIR filters with circular, elliptical, fan and diamond contour based on the McClellan transformation are available, see e.g. [2]-[4]. Here we present novel analytical method for the design of the coefficients of the McClellan transformation for 2-D FIR filters with asteroidal contour. Using the proposed method other shapes of the contour of the 2-D FIR filter are achievable, too. These are for example

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 64–71, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ellipses, circles, epicycloids, hypocycloids and roses. The consecutive rotation allows to orient the frequency response of the filter in the frequency plane as specified.

2 Transformation Method

The McClellan transformation design technique [1] transforms the 1-D frequency response

$$H(e^{j\omega}) = \sum_{n=0}^{N} a(n)cos(n\omega) = \sum_{n=0}^{N} a(n)T_n\left[cos(\omega)\right] \tag{1}$$

using the transformation function $cos(\omega) = F(e^{j\omega_1}, e^{j\omega_2})$ into the 2-D frequency response

$$H(e^{j\omega_1}, e^{j\omega_2}) = \sum_{n=0}^{N} a(n)T_n\left[F(e^{j\omega_1}, e^{j\omega_2})\right] = \sum_{p=0}^{PN}\sum_{q=0}^{QN} b(p,q)cos(p\omega_1)cos(q\omega_2) \tag{2}$$

where $T_n(x)$ are Chebyshev polynomials of the first kind. In the transformation function

$$cos(\omega) = F(e^{j\omega_1}, e^{j\omega_2}) = \sum_{p=0}^{P}\sum_{q=0}^{Q} f(p,q)cos(p\omega_1)cos(q\omega_2) \tag{3}$$

the coefficients $f(p,q)$ are called coefficients of the McClellan transformation and the integers P, Q represent the order of the McClellan transformation. For constant transformation function

$$cos(\omega_0) = F(e^{j\omega_1}, e^{j\omega_2}) = const \tag{4}$$

the relation between the frequencies ω_1 and ω_2

$$\omega_2 = g(\omega_1) \tag{5}$$

defines the contour of the transformation function. The transformation function $F(e^{j\omega_1}, e^{j\omega_2})$ affects the shape of the 2-D frequency response $H(e^{j\omega_1}, e^{j\omega_2})$, whereas the 1-D frequency response $H(e^{j\omega})$ determines the selectivity of the 2-D filter. The central part in the design of 2-D FIR filter using McClellan transformation represents the determination of the coefficients $f(p,q)$.

3 Contour Approximation

The contour approximation consists in the design of the coefficients $f(p,q)$ of the McClellan transformation. The contour approximation is solved by the minimization of the error

$$E(\omega_1, \omega_2) = cos(\omega_0) - F(e^{j\omega_1}, e^{j\omega_2}) \tag{6}$$

with respect to the desired shape of the contour (5). The minimization of (6) is accomplished using least square method

$$\int_{\mathcal{D}} \frac{\partial E^2(\omega_1, g(\omega_1))}{\partial f(p,q)} d\omega_1 \tag{7}$$

where $\mathcal{D}$ is the region of definition of the desired contour (5). The minimization (7) is usually solved numerically. Instead of the explicit relation (5), we propose the parametric representation of the contour $\omega_1 = g_1(\varphi)$, $\omega_2 = g_2(\varphi)$ or more specifically

$$\omega_1 = g_1(cos\varphi, sin\varphi)\,,\ \omega_2 = g_2(cos\varphi, sin\varphi)\ . \tag{8}$$

Due to the parametric representation of the contour (8), the numerical solution of (7) can be replaced by the analytical solution of the error

$$\int_0^{k\pi} \frac{\partial E^2(g_1(\varphi), g_2(\varphi))}{\partial f(p,q)} d\varphi\ . \tag{9}$$

For the second-order McClellan transformation ($P = Q = 2$), we have to determine nine coefficients $f(p,q)$. Defining the error function in the form

$$\begin{aligned} E(\omega_1,\omega_2) = {} & A_0 + A_1 f(1,0) + A_2 f(0,1) \\ & + A_3 f(1,1) + A_4 f(2,0) + A_5 f(0,2) \\ & + A_6 f(2,1) + A_7 f(1,2) + A_8 f(2,2) \end{aligned} \tag{10}$$

where

$$\begin{aligned} A_0 &= cos(\omega_0) - s & A_1 &= 1 - cos(\omega_1) \\ A_2 &= 1 - cos(\omega_2) & A_3 &= 1 - cos(\omega_1)cos(\omega_2) \\ A_4 &= 1 - cos(2\omega_1) & A_5 &= 1 - cos(2\omega_2) \\ A_6 &= 1 - cos(2\omega_1)cos(\omega_2) & A_7 &= 1 - cos(\omega_1)cos(2\omega_2) \\ A_8 &= 1 - cos(2\omega_1)cos(2\omega_2) \end{aligned}$$

$$s = \sum_{p=0}^{2}\sum_{q=0}^{2} f(p,q)\ , \tag{11}$$

the coefficients of the McClellan transformation $f(p,q)$ are given by the minimization of (9) as a solution of the eight equations

$$\begin{bmatrix} I_1 & I_2 & I_3 & I_4 & I_5 & I_6 & I_7 & I_8 \\ I_2 & I_9 & I_{10} & I_{11} & I_{12} & I_{13} & I_{14} & I_{15} \\ I_3 & I_{10} & I_{16} & I_{17} & I_{18} & I_{19} & I_{20} & I_{21} \\ I_4 & I_{11} & I_{17} & I_{22} & I_{23} & I_{24} & I_{25} & I_{26} \\ I_5 & I_{12} & I_{18} & I_{23} & I_{27} & I_{28} & I_{29} & I_{30} \\ I_6 & I_{13} & I_{19} & I_{24} & I_{28} & I_{31} & I_{32} & I_{33} \\ I_7 & I_{14} & I_{20} & I_{25} & I_{29} & I_{32} & I_{34} & I_{35} \\ I_8 & I_{15} & I_{21} & I_{26} & I_{30} & I_{33} & I_{35} & I_{36} \end{bmatrix} \times \begin{bmatrix} f(1,0) \\ f(0,1) \\ f(1,1) \\ f(2,0) \\ f(0,2) \\ f(2,1) \\ f(1,2) \\ f(2,2) \end{bmatrix} = \begin{bmatrix} I_{37} \\ I_{38} \\ I_{39} \\ I_{40} \\ I_{41} \\ I_{42} \\ I_{43} \\ I_{44} \end{bmatrix}\ . \tag{12}$$

The constants I_i can be for particular shape of the contour (8) expressed analytically by closed form formulas. This analytical solution is based on the expansion of the functions $cos(\alpha cos\varphi)$, $cos(\alpha sin\varphi)$, $sin(\alpha cos\varphi)$, $sin(\alpha sin\varphi)$ into the sum of Bessel functions which enables analytical integration of the terms in the quadratic difference (9). For asteroidal contour defined by

$$\omega_1 = \frac{3}{4}\omega_0 cos\frac{\varphi}{4} + \frac{1}{4}\omega_0 cos\frac{3\varphi}{4} \quad , \quad \omega_2 = \frac{3}{4}\omega_0 sin\frac{\varphi}{4} - \frac{1}{4}\omega_0 sin\frac{3\varphi}{4} \tag{13}$$

the derived constants I_i are summarized in Tab. 1.

4 Rotation of the Frequency Response

The 2-dimensional zero-phase FIR filter is represented by the impulse response $h(m,n)$. Provided the impulse response $h(m,n)$ is of odd length in both directions with central term $h(0,0)$ and with symmetry $h(m,n) = h(-m,-n)$, then the relations between the impulse and frequency response are as follows

$$H(e^{j\omega_1}, e^{j\omega_2}) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} h(m,n) e^{-jm\omega_1} e^{-jn\omega_2} \tag{14}$$

$$h(m,n) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} H(e^{j\omega_1}, e^{j\omega_2}) e^{jm\omega_1} e^{jn\omega_2} \, d\omega_1 \, d\omega_2 \ . \tag{15}$$

I order to rotate the frequency response $H(e^{j\omega_1}, e^{j\omega_2})$ in the frequency plane we propose the transformation of the frequency coordinates

$$\Omega_1 = \omega_1 \cos\phi - \omega_2 \sin\phi \quad , \quad \Omega_2 = \omega_2 \cos\phi + \omega_1 \sin\phi \ . \tag{16}$$

The impulse response $h_r(p,q)$ of the filter with the rotated frequency response $H_r(e^{j\omega_1}, e^{j\omega_2})$ is given by substitution of (16) into (14) and evaluating (15) in the form

$$h_r(p,q) = \frac{1}{(2\pi)^2} \sum_{m=-M}^{M} \sum_{n=-N}^{N} h(m,n) \times \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{j(p\omega_1 - m\Omega_1)} e^{j(q\omega_2 - n\Omega_2)} \, d\omega_1 \, d\omega_2 \tag{17}$$

yielding

$$h_R(p,q) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} h(m,n) \frac{\sin\pi k_1}{\pi k_1} \frac{\sin\pi k_2}{\pi k_2} \quad , \quad |p| \le M \, , \quad |q| \le N \tag{18}$$

where

$$k_1 = p - (m\cos\phi - n\sin\phi) \quad , \quad k_2 = q - (n\cos\phi + m\sin\phi) \, . \tag{19}$$

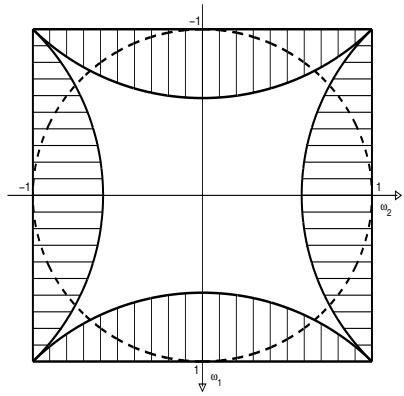


Fig. 1. Limitations of the rotation

The frequency response $H_r(e^{j\omega_1}, e^{j\omega_2})$ given by the impulse response $h_r(p,q)$ is in the frequency domain rotated counterclockwise by the angle ϕ with respect to the frequency response $H(e^{j\omega_1}, e^{j\omega_2})$ given by the impulse response $h(m,n)$. However, we have to take into account two limitations of the rotation denoted in Fig. 1. First, the aliasing from higher spectral periods may occur during the rotation. For aliasing prevention the frequency response $H(e^{j\omega_1}, e^{j\omega_2})$ should be constant in the four dashed border regions demarcated by the arcs with the radius equal $\sqrt{2}$. Second, the regions of the frequency response $H(e^{j\omega_1}, e^{j\omega_2})$ outside the unity circle as indicated in Fig. 1 may disappear during the rotation. The regions of the frequency response inside the unity circle never disappear during the rotation. Due to these limitations it is not possible to rotate the transformation function $F(e^{j\omega_1}, e^{j\omega_2})$ prior the McClellan transformation.

5 Example of the Design

In our example we demonstrate the design of the 2-D FIR low pass filter with asteroidal contour for $\omega_0 = 0.5$ oriented counterclockwise by the angle $\phi = 3\pi/25$ with respect to the frequency coordinates. We assume normalized frequencies in the interval $(-1, 1)$. In the first step we calculate the coefficients of the McClellan transformation

$$f(p,q) = \begin{bmatrix} 0.263308 & 0.092866 & 0.056574 \\ 0.092866 & 0.019525 & 0.075101 \\ 0.056574 & 0.075101 & -0.148859 \end{bmatrix} \tag{20}$$

using formulas summarized in Tab. 1. The corresponding transformation function $F(e^{j\omega_1}, e^{j\omega_2})$ with contours is presented in Fig. 2. The 1-D maximally flat low pass FIR filter with 3dB-decay cut-off frequency $\omega_{stop} = 0.22$ of the length 17 coefficients was designed using the analytical procedure presented in [6]. The impulse response $h(n)$ of the filter is summarized in Tab. 2. Using the transformation coefficients $f(p,q)$ and the 1-D impulse response $h(n)$ we calculate the

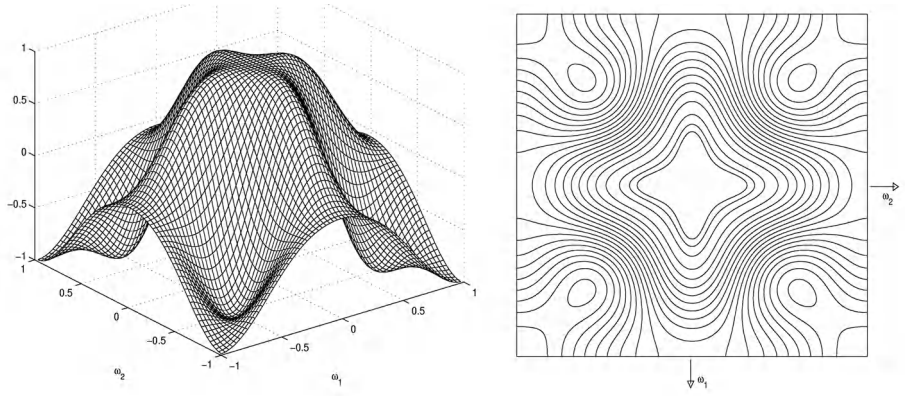


Fig. 2. Transformation function $F(e^{j\omega_1}, e^{j\omega_2})$ with contours

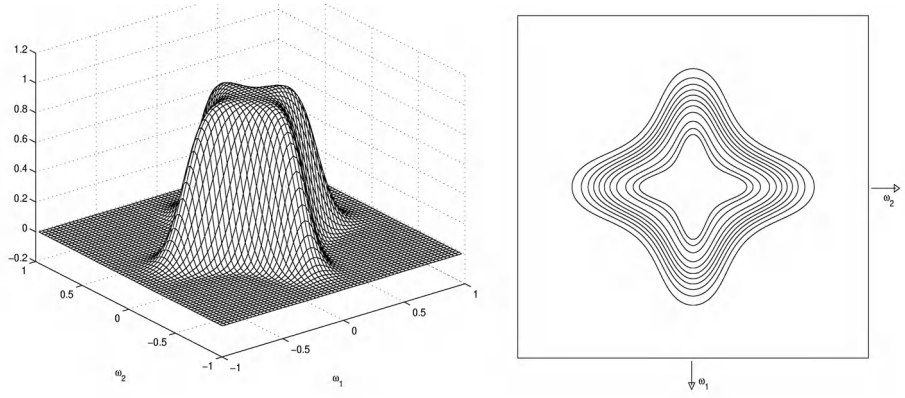


Fig. 3. Amplitude frequency response $|H(e^{j\omega_1}, e^{j\omega_2})|$ with contours

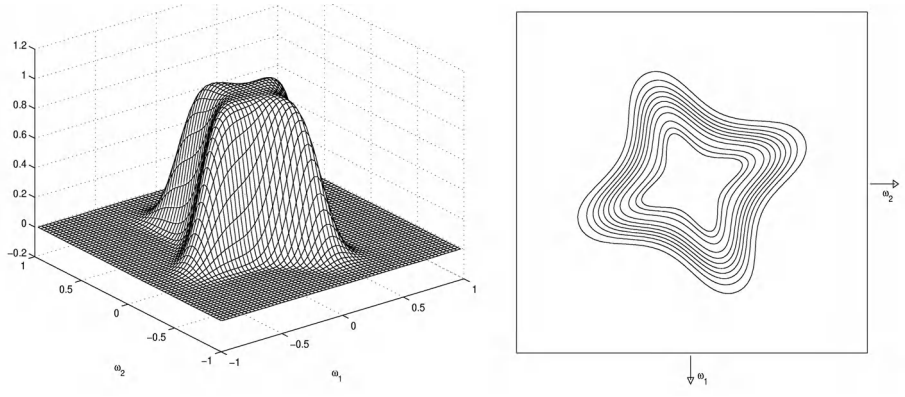


Fig. 4. Rotated amplitude frequency response $|H_r(e^{j\omega_1}, e^{j\omega_2})|$ with contours

Table 1. Constants I_i

$$
\begin{aligned}
I_1 &= 3\pi/2 - 2F(\omega_1) + F(2\omega_1)/2\\
I_2 &= \pi - F(\omega_1) - F(\omega_2) + E(\omega_1,\omega_2)\\
I_3 &= \pi - F(\omega_1) - E(\omega_1,\omega_2) + F(\omega_2)/2 + E(2\omega_1,\omega_2)/2\\
I_4 &= \pi - F(\omega_1)/2 - F(2\omega_1) + F(3\omega_1)/2\\
I_5 &= \pi - F(\omega_1) - F(2\omega_2) + E(\omega_1,2\omega_2)\\
I_6 &= \pi - F(\omega_1) - E(2\omega_1,\omega_2) + E(3\omega_1,\omega_2)/2 + E(\omega_1,\omega_2)/2\\
I_7 &= \pi - F(\omega_1) - E(\omega_1,2\omega_2) + F(2\omega_2)/2 + E(2\omega_1,2\omega_2)/2\\
I_8 &= \pi - F(\omega_1) - E(2\omega_1,2\omega_2) + E(3\omega_1,2\omega_2)/2 + E(\omega_1,2\omega_2)/2\\
I_9 &= 3\pi/2 - 2F(\omega_2) + F(2\omega_2)/2\\
I_{10} &= \pi - F(\omega_2) - E(\omega_1,\omega_2) + F(\omega_1)/2 + E(\omega_1,2\omega_2)/2\\
I_{11} &= \pi - F(\omega_2) - F(2\omega_1) + E(2\omega_1,\omega_2)\\
I_{12} &= \pi - F(\omega_2)/2 - F(2\omega_2) + F(3\omega_2)\\
I_{13} &= \pi - F(\omega_2) - E(2\omega_1,\omega_2) + F(2\omega_1)/2 + E(2\omega_1,2\omega_2)/2\\
I_{14} &= \pi - F(\omega_2) - E(\omega_1,2\omega_2) + E(\omega_1,3\omega_2)/2 + E(\omega_1,\omega_2)/2\\
I_{15} &= \pi - F(\omega_2) - E(2\omega_1,2\omega_2) + E(2\omega_1,3\omega_2)/2 + E(2\omega_1,\omega_2)/2\\
I_{16} &= 5\pi/4 - 2E(\omega_1,\omega_2) + [F(2\omega_1) + F(2\omega_2) + E(2\omega_1,2\omega_2)]/4\\
I_{17} &= \pi - E(\omega_1,\omega_2)/2 - F(2\omega_1) + E(3\omega_1,\omega_2)/2\\
I_{18} &= \pi - E(\omega_1,\omega_2)/2 - F(2\omega_2) + E(\omega_1,3\omega_2)/2\\
I_{19} &= \pi - E(\omega_1,\omega_2) - E(2\omega_1,\omega_2) + [F(3\omega_1) + E(3\omega_1,2\omega_2) + F(\omega_1) + E(\omega_1,2\omega_2)]/4\\
I_{20} &= \pi - E(\omega_1,\omega_2) - E(\omega_1,2\omega_2) + [F(3\omega_2) + F(\omega_2) + E(2\omega_1,3\omega_2) + E(2\omega_1,\omega_2)]/4\\
I_{21} &= \pi - E(\omega_1,\omega_2) - E(2\omega_1,2\omega_2) + [E(3\omega_1,3\omega_2) + E(\omega_1,3\omega_2) + E(\omega_1,\omega_2)]/4\\
I_{22} &= 3\pi/2 - 2F(2\omega_1) + F(4\omega_1)/2\\
I_{23} &= \pi - F(2\omega_1) - F(2\omega_2) + E(2\omega_1,2\omega_2)\\
I_{24} &= \pi - F(2\omega_1) - E(2\omega_1,\omega_2) + F(\omega_2)/2 + E(4\omega_1,\omega_2)/2\\
I_{25} &= \pi - F(2\omega_1) - E(\omega_1,2\omega_2) + F(\omega_1)/2 + E(\omega_1,4\omega_2)/2\\
I_{26} &= \pi - F(2\omega_1) - E(\omega_1,2\omega_2)/2 + E(3\omega_1,2\omega_2)/2\\
I_{27} &= 3\pi/2 - 2F(2\omega_2) + F(4\omega_2)/2\\
I_{28} &= \pi - F(2\omega_2) - E(2\omega_1,\omega_2) + E(2\omega_1,3\omega_2)/2 + E(2\omega_1,\omega_2)/2\\
I_{29} &= \pi - F(2\omega_2) - E(\omega_1,2\omega_2) + E(\omega_1,4\omega_2)/2 + F(\omega_1)/2\\
I_{30} &= \pi - F(2\omega_2) - E(2\omega_1,2\omega_2) + E(2\omega_1,4\omega_2)/2 + F(2\omega_2)/2\\
I_{31} &= 5\pi/4 - 2E(2\omega_1,\omega_2) + [E(4\omega_1,2\omega_2) + F(4\omega_1) + F(2\omega_2)]/4\\
I_{32} &= \pi - E(2\omega_1,\omega_2) - E(\omega_1,2\omega_2) + [E(3\omega_1,3\omega_2) + E(3\omega_1,\omega_2) + E(\omega_1,3\omega_2) + E(\omega_1,\omega_2)]/4\\
I_{33} &= \pi - E(2\omega_1,\omega_2) - E(2\omega_1,2\omega_2) + [E(4\omega_1,3\omega_2) + E(4\omega_1,\omega_2) + F(3\omega_2) + F(\omega_2)]/4\\
I_{34} &= 5\pi/4 - 2E(\omega_1,2\omega_2) + [E(2\omega_1,4\omega_2) + F(2\omega_1) + F(4\omega_2)]/4\\
I_{35} &= \pi - E(\omega_1,2\omega_2) - E(2\omega_1,2\omega_2) + [E(3\omega_1,4\omega_2) + E(\omega_1,4\omega_2) + F(3\omega_1) + F(\omega_1)]/4\\
I_{36} &= 5\pi/4 - 2E(2\omega_1,2\omega_2) + [E(4\omega_1,4\omega_2) + F(4\omega_1) + F(4\omega_2)]/4\\
I_{37} &= (cos\omega_0 - s)(\pi - F(\omega_1))\\
I_{38} &= (cos\omega_0 - s)(\pi - F(\omega_2))\\
I_{39} &= (cos\omega_0 - s)(\pi - E(\omega_1,\omega_2))\\
I_{40} &= (cos\omega_0 - s)(\pi - F(2\omega_1))\\
I_{41} &= (cos\omega_0 - s)(\pi - F(2\omega_2))\\
I_{42} &= (cos\omega_0 - s)(\pi - E(2\omega_1,\omega_2))\\
I_{43} &= (cos\omega_0 - s)(\pi - E(\omega_1,2\omega_2))\\
I_{44} &= (cos\omega_0 - s)(\pi - E(2\omega_1,2\omega_2))
\end{aligned}
$$

$$
F(x) = \omega_0 \left[J_0(x) - 2\sum_{m=1}^{\infty} J_{2m}(x)\frac{1}{4m^2-1} \right]
$$

$$
E(x,y) = \omega_0 \left[J_0\left(\sqrt{x^2+y^2}\right) - 2\sum_{m=1}^{\infty} J_{2m}\left(\sqrt{x^2+y^2}\right) \frac{cos\left(2m \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right)\right)}{4m^2-1} \right]
$$

Table 2. Impulse Response $h(n)$

n		$h(n)$	n		$h(n)$
0	16	-0.000107	5	11	0.022217
1	15	-0.001221	6	10	0.122192
2	14	-0.005981	7	9	0.244385
3	13	-0.015381		8	0.301117
4	12	-0.016663			

2-D impulse response $h(m, n)$ of the asteroidally shaped 2-D FIR filter of the length 37×37 coefficients. The frequency response of the filter is oriented in the basic position along the frequency coordinates. The amplitude frequency response $|H(e^{j\omega_1}, e^{j\omega_2})|$ of the filter with its contours is shown in Fig. 3. In the second step the basically oriented 2-D low-pass filter with asteroidal shape of the passband is rotated counterclockwise by the angle $\phi = \frac{3}{25}\pi$ using (18). The rotated frequency response $|H_r(e^{j\omega_1}, e^{j\omega_2})|$ of the filter with its contours is shown in Fig. 4.

References

1. McClellan J. H.: The Design of Two-dimensional Digital Filters by Transformation. Proc. 7th Ann. Princeton Conf. Inform. Sci. and Syst., March 1973, pp. 247-251.
2. Nguyen D. T., Swamy M. N. S.: Approximation Design of 2-D Digital Filters with Elliptical Magnitude Response of Arbitrary Orientation. IEEE Trans. on Circuits and Systems, Vol. CAS-33, No. 6, June 1986, pp. 597-603.
3. Pei S.-C., Shyu J.-J.: Design of 2-D FIR Digital Filters by McClellan Transformation and Least Squares Eigencontour Mapping. IEEE Trans. on Circuits and Systems-II, Vol. 40, No. 9, September 1993, pp. 546-555.
4. Chen C.-K., Lee J.-H.: McClellan Transform based Design Techniques for Two-dimensional Linear-phase FIR Filters. IEEE Trans. on Circuits and Systems-I, Vol. 41, No. 8, August 1994, pp. 505-517.
5. Fischer V., Drutarovsky M., Lukac R.: Implementation of 3-D Adaptive LUM Smoother in Reconfigurable Hardware. Springer Verlag LNCS 2438, pp.720-729.
6. Lukac R.: Binary LUM Smoothing. IEEE Signal Processing Letters, Vol. 9, No. 12, December 2002, pp. 400-403.
7. Lukac R.: Adaptive Vector Median Filtering. Pattern Recognition Letters, Vol. 24, No. 12, August 2003, pp. 1889-1899.
8. Lukac R.: Simplified Boolean LUM Smoothers. Proceedings of the 4th EURASIP-IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications VIPromCom-2002, Zadar, Croatia, June 16-19, 2002, pp. 159-162.
9. Lukac R.: The Way How to Design and Implement an Adaptive Method Based on Center-Weighted Medians. Proceedings of the IEEE Scientific Workshop Signal Processing 2002, Poznan, Poland, October 11, 2002, pp.9-14.
10. Vlček M., Zahradník P., Unbehauen R.: Analytic Design of FIR Filters. IEEE Transactions on Signal Processing, Vol. 48, Sept. 2000, pp. 2705-2709.

A $\{k, n\}$-Secret Sharing Scheme for Color Images

Rastislav Lukac, Konstantinos N. Plataniotis, and
Anastasios N. Venetsanopoulos

The Edward S. Rogers Sr. Dept. of Electrical and Computer Engineering,
University of Toronto, 10 King's College Road, Toronto, M5S 3G4, Canada
{lukacr,kostas,anv}@dsp.utoronto.ca

Abstract. This paper introduces a new $\{k, n\}$-secret sharing scheme for color images. The proposed method encrypts the color image into n color shares. The secret information is recovered only if the k (or more) allowed shares are available for decryption. The proposed method utilizes the conventional $\{k, n\}$-secret sharing strategy by operating at the bit-levels of the decomposed color image. Modifying the spatial arrangements of the binary components, the method produces color shares with varied both the spectral characteristics among the RGB components and the spatial correlation between the neighboring color vectors. Since encryption is done in the decomposed binary domain, there is no obvious relationship in the RGB color domain between any two color shares or between the original color image and any of the n shares. This increases protection of the secret information. Inverse cryptographic processing of the shares must be realized in the decomposed binary domain and the procedure reveals the original color image with perfect reconstruction.

1 Introduction

Visual cryptography [2],[3],[4],[8],[11] is a popular cryptographic tool used for protection of scanned documents and natural digital images which are distributed via public networks. These techniques are based on the principle of sharing secret information among a group of participants. The shared secret can be recovered only when a coalition of willing participants are polling their encrypted images, the so-called shares, together. Secret sharing schemes are usually termed visual since the secret (original) information can be directly revealed from stacked shares (e.g realized as transparencies) through simple visual inspection, without any computer-guided processing [9],[11].

A $\{k, n\}$-threshold visual cryptography scheme [5],[6],[11] often called $\{k, n\}$-visual secret sharing (VSS), is used to encrypt an input image by splitting the original content into n, seemingly random, shares. To recover the secret information, k (or more) allowed shares must be stacked together.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 72–79, 2004.
© Springer-Verlag Berlin Heidelberg 2004

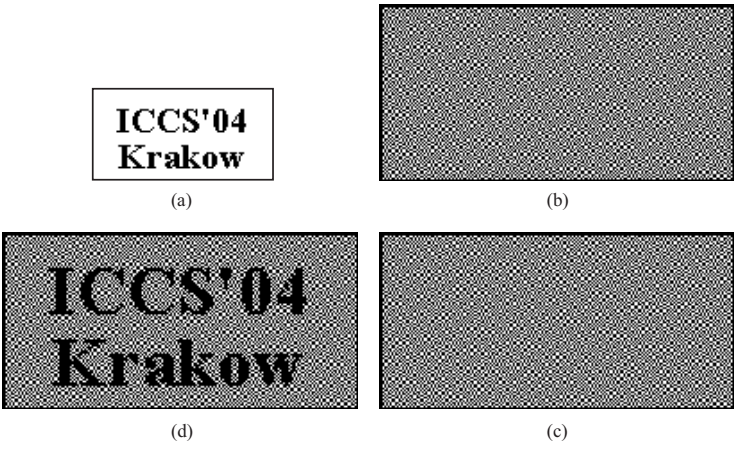


Fig. 1. Visual cryptography: (a) original binary image, (b,c) share images, (d) decrypted, output image

2 $\{k, n\}$-Secret Sharing Scheme

Due to its algorithmic nature, conventional visual cryptography schemes operate on a binary input (Fig. 1) [10]. Assuming a $K_1 \times K_2$ binary image (black and white image with 0 values denoting the black and 1 values denoting the white), each binary pixel $r_{(i,j)}$ determined by spatial coordinates $i = 1, 2, ..., K_1$ and $j = 1, 2, ..., K_2$ is replaced via an encryption function $f_e(\cdot)$ with a $m_1 \times m_2$ block of black and white pixels in each of the n shares. Repeating the process for each input pixel, a $K_1 \times K_2$ input binary image is encrypted into n binary shares each one with a spatial resolution of $m_1 K_1 \times m_2 K_2$ pixels. Since the spatial arrangement of the pixels varies from block to block, the original information cannot be revealed without accessing a predefined number of shares.

Let as assume a basic $\{2, 2\}$-threshold structure which is the basic case designed within the $\{k, n\}$-VSS framework [1],[7]. Assuming for simplicity a basic structure with 2×2 blocks $\mathbf{s}_1 = [s'_{(2i-1,2j-1)}, s'_{(2i-1,2j)}, s'_{(2i,2j-1)}, s'_{(2i,2j)}] \in S_1$ and $\mathbf{s}_2 = [s''_{(2i-1,2j-1)}, s''_{(2i-1,2j)}, s''_{(2i,2j-1)}, s''_{(2i,2j)}] \in S_2$, the encryption process is defined as follows:

$$f_e(r_{(i,j)}) = \begin{cases} [\mathbf{s}_1, \mathbf{s}_2]^T \in C_0 \text{ for } r_{(i,j)} = 0 \\ [\mathbf{s}_1, \mathbf{s}_2]^T \in C_1 \text{ for } r_{(i,j)} = 1 \end{cases} \quad (1)$$

where C_0 and C_1 are the sets obtained by permuting the columns of the $n \times m_1 m_2$ basis matrices A_0 and A_1, respectively [10]. Since $m_1 m_2$ represents the factor by which each share is larger than the original image, it is desirable to make $m_1 m_2$ as small as possible. In the case of the $\{2, 2\}$-VSS the optimal choice m_1 and m_2 leads to $m_1 = 2$ and $m_2 = 2$ resulting in 2×2 blocks $\mathbf{s}_1$ and $\mathbf{s}_2$.

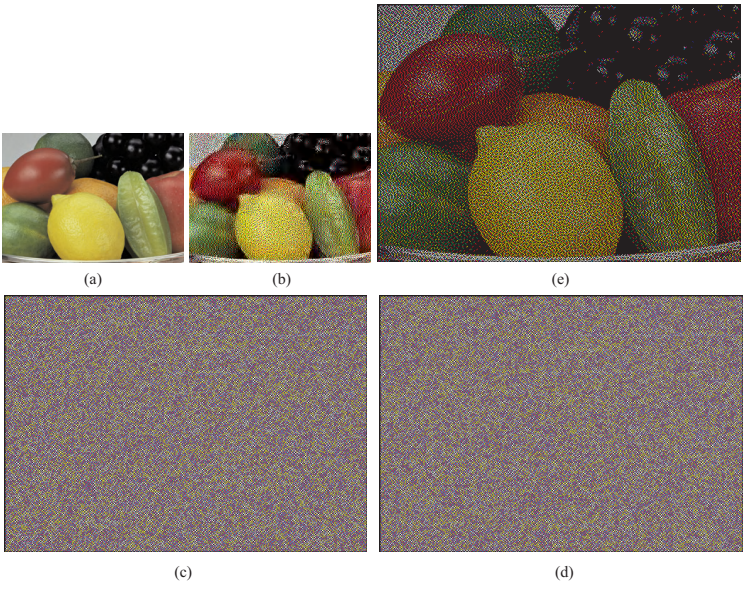


Fig. 2. Halftoning-based secret sharing of color images: (a) original color image, (b) halftone image obtained using Floyd-Steinberg filter [12], (c,d) share images, (e) decrypted, output image

Assuming the $\{2,2\}$-VSS the sets

$$C_0 = \left\{ \begin{bmatrix} 0,1,0,1 \\ 1,0,1,0 \end{bmatrix}, \begin{bmatrix} 1,0,1,0 \\ 0,1,0,1 \end{bmatrix}, \begin{bmatrix} 0,0,1,1 \\ 1,1,0,0 \end{bmatrix}, \begin{bmatrix} 1,1,0,0 \\ 0,0,1,1 \end{bmatrix}, \begin{bmatrix} 1,0,0,1 \\ 0,1,1,0 \end{bmatrix}, \begin{bmatrix} 0,1,1,0 \\ 1,0,0,1 \end{bmatrix} \right\} \tag{2}$$

$$C_1 = \left\{ \begin{bmatrix} 0,1,0,1 \\ 0,1,0,1 \end{bmatrix}, \begin{bmatrix} 1,0,1,0 \\ 1,0,1,0 \end{bmatrix}, \begin{bmatrix} 0,0,1,1 \\ 0,0,1,1 \end{bmatrix}, \begin{bmatrix} 1,1,0,0 \\ 1,1,0,0 \end{bmatrix}, \begin{bmatrix} 1,0,0,1 \\ 1,0,0,1 \end{bmatrix}, \begin{bmatrix} 0,1,1,0 \\ 0,1,1,0 \end{bmatrix} \right\} \tag{3}$$

include all matrices obtained by permuting the columns of the 2×4 basis matrices A_0 and A_1, respectively [10],[11]. The basic matrices considered here are defined as follows:

$$A_0 = \begin{bmatrix} 0\ 1\ 0\ 1 \\ 1\ 0\ 1\ 0 \end{bmatrix}, \ A_1 = \begin{bmatrix} 0\ 1\ 0\ 1 \\ 0\ 1\ 0\ 1 \end{bmatrix} \tag{4}$$

If a secret pixel is white, i.e. $r_{(i,j)} = 1$, then each pixel in $\mathbf{s}_1$ is equivalent to each pixel in $\mathbf{s}_2$, and thus, $[\mathbf{s}_1, \mathbf{s}_2]^T$ can be any member of set C_1. If a secret pixel is black, i.e. $r_{(i,j)} = 0$, then each pixel in $\mathbf{s}_1$ should complement each pixel in $\mathbf{s}_2$ and thus, $[\mathbf{s}_1, \mathbf{s}_2]^T$ should be selected from set C_0. The choice of $[\mathbf{s}_1, \mathbf{s}_2]^T$ is guided by a random number generator, which determines the random character of the shares.

The decrypted block is produced through a decryption function $f_d(\cdot)$. In the case of the $\{2,2\}$-scheme based on the basis matrices of (4), $f_d(\cdot)$ can be defined as follows:

$$y_{2\times2} = f_d(\mathbf{s}_1, \mathbf{s}_2) = \begin{cases} \mathbf{s}_1 & \text{for } \mathbf{s}_1 = \mathbf{s}_2 \\ [0,0,0,0] & \text{for } \mathbf{s}_1 \neq \mathbf{s}_2 \end{cases} \tag{5}$$

Fig. 3. Halftoning-based secret sharing of color images: (a) original color image, (b) halftone image obtained using Floyd-Steinberg filter [12], (c,d) share images, (e) decrypted, output image

where $\mathbf{s}_1 = [s'_{(u,v)}, s'_{(u,v+1)}, s'_{(u+1,v)}, s'_{(u+1,v+1)}] \in S_1$ and $\mathbf{s}_2 = [s''_{(u,v)}, s''_{(u,v+1)}, s''_{(u+1,v)}, s''_{(u+1,v+1)}] \in S_2$, for $u = 1, 3, ..., 2K_1 - 1$ and $v = 1, 3, ..., 2K_2 - 1$, are 2×2 share blocks which are used to recover the output block $y_{2\times 2} = y_{(u,v)}, y_{(u,v+1)}, y_{(u+1,v)}, y_{(u+1,v+1)}$ as $s'_{(u,v)}$ or black pixels described as $[0, 0, 0, 0]$.

The application of a conventional $\{k, n\}$-VSS scheme to a $K_1 \times K_2$ natural image requires halftoning [7],[10]. The image is first transformed into a $K_1 \times K_2$ halftone image by using the density of the net dots to simulate the intensity levels [12]. Applying the procedure for each color channel of the original image (Fig. 2a) independently, each color channel of the halftone image (Fig. 2b) is a binary image and thus appropriate for the VSS. Assuming $\{2, 2\}$-VSS, the two color shares obtained by the procedure are depicted in Fig. 2c,d. Figure 2e shows the $2K_1 \times 2K_2$ decrypted image (result) obtained by stacking the two shares together.

Visual inspection of both the original image (Fig. 1a and Fig. 2a) and the recovered image (Fig. 1d and Fig. 2e) indicates that: i) the decrypted image is darker, and ii) the input image is of quarter size compared to the decrypted output. Moreover, the decrypted color image depicted in Fig. 2e contains a number of color artifacts due to nature of the algorithm. To end this, the conventional $\{k, n\}$-threshold visual cryptography i) cannot provide perfect reconstruction, either in terms of pixel intensity or spatial resolution, and ii) is not appropriate for real-time applications.

Figure 3 shows the images obtained using the conventional $\{2, 2\}$-secret sharing scheme applied to the image with the different color scenario compared to Fig. 2. It can be again observed that the decrypted image depicted in Fig. 3e contains shifted colors which often prohibit correct perception of fine image details.

Is has to be mentioned that the halftoning-based $\{k, n\}$-visual secret sharing schemes are the most popular choice for natural image secret sharing. Another secret sharing approach for color images is based on mean color-mixing [8]. However, this method is not appropriate for practical applications due to significant reduction of color gamut and the extreme increase in the spatial resolution of the shares. Other works, e.g. [9],[13] deals with analytical construction of the secret sharing schemes for color images.

3 $\{k, n\}$-Color Secret Sharing

Let $\mathbf{x} : Z^2 \rightarrow Z^3$ be a $K_1 \times K_2$ Red-Green-Blue (RGB) color image representing a two-dimensional matrix of the three-component color vectors (pixels) $\mathbf{x}_{(i,j)} = [x_{(i,j)1}, x_{(i,j)2}, x_{(i,j)3}]$ located at the spatial position (i, j), for $i = 1, 2, ..., K_1$ and $j = 1, 2, ..., K_2$. Assuming that c describes the color channel (i.e. $c = 1$ for Red, $c = 2$ for Green, and $c = 3$ for Blue) and the color component $x_{(i,j)c}$ is coded with B bits allowing $x_{(i,j)c}$ to take an integer value between 0 and $2^B - 1$, the color vector $\mathbf{x}_{(p,q)}$ can be equivalently expressed in a binary form as follows:

$$\mathbf{x}_{(i,j)} = \sum\nolimits_{b=1}^{B} \mathbf{x}^b_{(i,j)} 2^{B-b} \tag{6}$$

where $\mathbf{x}^b_{(i,j)} = [x^b_{(i,j)1}, x^b_{(i,j)2}, x^b_{(i,j)3}] \in \{0, 1\}^3$ denotes the binary vector at the b-bit level, with $b = 1$ denoting the most significant bits (MSB).

3.1 Encryption

If the c-th component of the binary vector $\mathbf{x}^b_{(i,j)}$ is white ($x^b_{(i,j)c} = 1$), encryption is performed through $[\mathbf{s}_1, \mathbf{s}_2]^T \in C_1$ replacing $x^b_{(i,j)c}$ by binary blocks $\mathbf{s}_1$ and $\mathbf{s}_2$ in each of the two shares. Otherwise, the reference binary component is black ($x^b_{(p,q)c} = 0$), and encryption is defined via $[\mathbf{s}_1, \mathbf{s}_2]^T \in C_0$. This forms an encryption function defined as follows:

$$f_e(x^b_{(i,j)c}) = \begin{cases} [\mathbf{s}_1, \mathbf{s}_2]^T \in C_0 \text{ for } x^b_{(i,j)c} = 0 \\ [\mathbf{s}_1, \mathbf{s}_2]^T \in C_1 \text{ for } x^b_{(i,j)c} = 1 \end{cases} \tag{7}$$

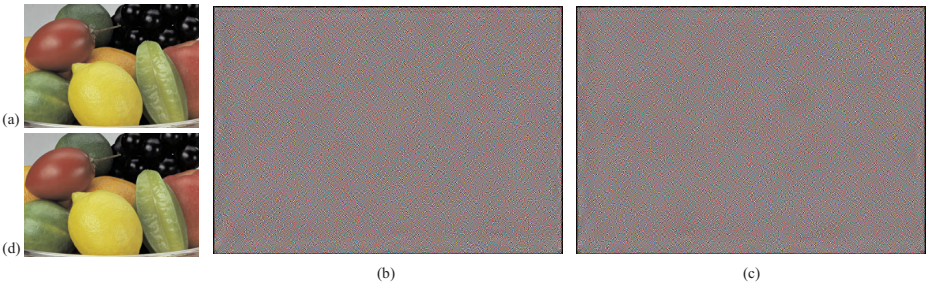


Fig. 4. Proposed $\{2, 2\}$-secret sharing scheme for color images: (a) original color image, (b,c) share images, (d) decrypted, output image

By replacing the binary components $x^b_{(i,j)c}$ with binary blocks $\mathbf{s}_1$ and $\mathbf{s}_2$ for one particular b, the process generates two $2K_1 \times 2K_2$ vector-valued binary shares S^b_1 and S^b_2, respectively. A random number generator guides the choice of $[\mathbf{s}^b_1, \mathbf{s}^b_2]^T$ and determines the random character of S^b_1 and S^b_2. Thus, the process modifies both the spatial correlation between spatially neighboring binary vectors $\mathbf{s}'^b_{(u,v)} = [s'^b_{(u,v)1}, s'^b_{(u,v)2}, s'^b_{(u,v)3}] \in S^b_1$ or $\mathbf{s}''^b_{(u,v)} = [s''^b_{(u,v)1}, s''^b_{(u,v)2}, s''^b_{(u,v)3}] \in S^b_1$, for $u = 1, 2, ..., 2K_1$ and $v = 1, 2, ..., 2K_2$, and the spectral correlation among components $s'^b_{(u,v)c}$ or $s''^b_{(u,v)c}$, for $c = 1, 2, 3$, of the individual binary vectors $\mathbf{s}'^b_{(u,v)}$ or $\mathbf{s}''^b_{(u,v)}$, respectively. Bit-level stacking of the encrypted bit-levels produces the color vectors $\mathbf{s}'_{(u,v)} \in \mathbf{S}_1$ and $\mathbf{s}''_{(u,v)} \in \mathbf{S}_2$ as

$$\mathbf{s}'_{(u,v)} = \sum_{b=1}^{B} \mathbf{s}'^b_{(u,v)} 2^{B-b} \quad \text{and} \quad \mathbf{s}''_{(u,v)} = \sum_{b=1}^{B} \mathbf{s}''^b_{(u,v)} 2^{B-b} \tag{8}$$

Due to random processing taking place at the bit-levels, $\mathbf{S}_1$ and $\mathbf{S}_2$ contain only random, color noise like information (Fig. 4b,c). Since encryption is realized in the decomposed binary vector space, no detectable relationship between the original color vectors $\mathbf{x}_{(p,q)}$ and the color noise of $\mathbf{S}_1$ or $\mathbf{S}_2$ can be found in the RGB color domain. This considerably increases security and prevents unauthorized decryption through brute-force enumeration.

3.2 Decryption

The decryption procedure is designed to satisfy the perfect reconstruction property. The original color data must be recovered from the color shares $\mathbf{S}_1$ and $\mathbf{S}_2$ using inverse algorithmic steps. Therefore, the decryption procedure is applied to the decomposed binary vector arrays of the color shares. Assuming that (i, j), for $i = 1, 2, ..., K_1$ and $j = 1, 2, ..., K_2$, denotes the spatial position in the original image and c denotes the color channel, the corresponding 2×2 binary share blocks are $\mathbf{s}'^b_c = \{s'^b_{(2i-1,2j-1)c}, s'^b_{(2i-1,2j)c}, s'^b_{(2i,2j-1)c}, s'^b_{(2i,2j)c}\}$ and $\mathbf{s}''^b_c = \{s''^b_{(2i-1,2j-1)c}, s''^b_{(2i-1,2j)c}, s''^b_{(2i,2j-1)c}, s''^b_{(2i,2j)c}\}$. Based on the arrangements of the basis matrices A_0 and A_1 in (4) used in the proposed $\{2, 2\}$-secret sharing

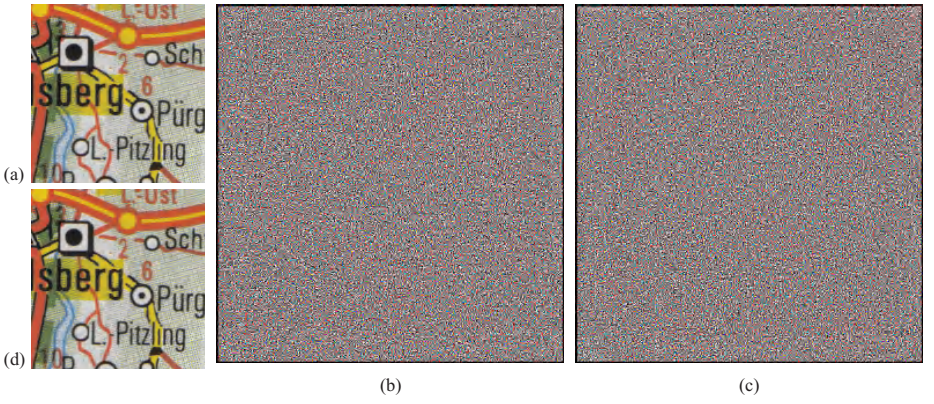


Fig. 5. Proposed $\{2, 2\}$-secret sharing scheme for color images: (a) original color image, (b,c) share images, (d) decrypted, output image

scheme, if both blocks are consistent, i.e. $\mathbf{s}_c'^b = \mathbf{s}_c''^b$, the decrypted original bit $x^b_{(i,j)c}$ is assign white, i.e. $x^b_{(i,j)c} = 1$. Otherwise, the blocks are inconsistent, i.e. $\mathbf{s}_c'^b \neq \mathbf{s}_c''^b$ and the original bit is recovered as black, i.e. $x^b_{(i,j)c} = 0$. This logical comparison forms the following decryption function

$$x^b_{(i,j)c} = f_d(\mathbf{s}_c'^b, \mathbf{s}_c'^b) = \begin{cases} 1 \text{ for } \mathbf{s}_c'^b = \mathbf{s}_c''^b \\ 0 \text{ for } \mathbf{s}_c'^b \neq \mathbf{s}_c''^b \end{cases} \tag{9}$$

which is used to restore the binary vectors $\mathbf{x}^b_{(i,j)}$. The procedure completes with the bit-level stacking (6) resulting in the original color vector $\mathbf{x}_{(i,j)}$.

Figure 4d shows the decrypted color output. Since the proposed method satisfies the perfect reconstruction property, the output image is identical to the original depicted in Fig 4a. Note that perfect reconstruction is demonstrated also in Fig 5, which depicts two full color shares (Fig 5b,c) and the decrypted output (Fig 5d) obtained by the proposed $\{2, 2\}$-secret sharing scheme applied to the test image "Atlas" (Fig 5a).

It has to be mentioned that (9), which is defined for a $\{2, 2\}$-scheme, can be more generally described as follows:

$$o^b_{(i,j)} = f_d(\mathbf{s}^b_1, \mathbf{s}^b_2) = \begin{cases} 1 & \text{for } [\mathbf{s}_c'^b, \mathbf{s}_c''^b]^T \in C_1 \\ 0 & \text{for } [\mathbf{s}_c'^b, \mathbf{s}_c''^b]^T \in C_0 \end{cases} \tag{10}$$

This concept can be further generalized for the share blocks $\{\mathbf{s}_c'^b, \mathbf{s}_c''^b, \mathbf{s}_c'''^b, ...\}$ defined in the specific $\{k, n\}$-threshold schemes.

4 Conclusion

A $\{k, n\}$ secret sharing scheme with perfect reconstruction of the color inputs was introduced. The method cryptographically processes the color images replacing

the bit components with a block of bits for each of the n shares. By separate encrypting each bit plane of the decomposed color image, the method produces color shares with varied spectral and spatial characteristics. Since encryption is realized in the decomposed binary domain, the procedure increases protection against attacks performed in the RGB color domain. This makes the method attractive for secure transmission over the untrusted public channel. Moreover, the perfect reconstruction property allows to reveal the original color image without visual impairments or color shifts.

References

1. Adhikari, A., Sikdar, S.: A new (2,n)-visual threshold scheme for color images. Proc. INDOCRYPT 2003, LNCS **2904** (2003) 148–161
2. Ateniese, G., Blundo, C, de Santis, A., Stinson, D.G.: Visual cryptography for general access structures. Information and Computation **129** (1996) 86–106
3. Ateniese, G., Blundo, C, de Santis, A., Stinson, D.G.: Extended capabilities for visual cryptography. Theoretical Computer Science **250** (2001) 143–161
4. Droste, S.: New results on visual cryptography. Proc. Advances in Cryptology - CRYPTO'96, LNCS **1109** (1996) 401–415
5. Eisen, P.A., Stinson, D.R.: Threshold visual cryptography schemes with specified levels of reconstructed pixels. Design, Codes and Cryptography **25** (2002) 15–61
6. Hofmeister, T., Krause, M., Simon, H.U.: Contrast optimal k out of n secret sharing schemes in visual cryptography. Theoretical Computer Science **240** (2000) 471–485
7. Hou, J.C.: Visual cryptography for color images. Pattern Recognition **36** (2003) 1619–1629
8. Ishihara, T., Koga, H.: A visual secret sharing scheme for color images based on meanvalue-color mixing. IEICE Trans. Fundamentals **E86-A** (2003) 194–197
9. Koga, H., Iwamoto, M., Yakamoto, H.: An analytic construction of the visual secret sharing scheme for color images. IEICE Trans. Fundamentals **E84-A** (2001) 262–272
10. Lin, C.C., Tsai, W.H.: Visual cryptography for gray-level images by dithering techniques. Pattern Recognition Letters **24** (2003) 349–358
11. Naor, M., Shamir, A.: Visual Cryptography. Proc. EUROCRYPT'94, LNCS **950** (1994) 1–12
12. Ulichney, R.A.: Dithering with blue noise. Proceedings of the IEEE **76** (1988) 56–79
13. Yang, C.N.: A note on color visual encryption. Journal of Information Science and Engineering **18** (2002) 367–372

Declarative Modelling in Computer Graphics: Current Results and Future Issues

Pierre-François Bonnefoi, Dimitri Plemenos, and William Ruchaud

In alphabetic order
University of Limoges, MSI laboratory,
83, rue d'Isle,
87000 Limoges, France
{bonnefoi, plemenos, ruchaud}@unilim.fr
http://msi.unilim.fr/

Abstract. A review of declarative scene modelling techniques is presented in this paper. After a definition of the purpose of declarative modelling, some existing declarative modellers are classified according to the manner to manage imprecision in scene description. The aim of this paper is to show the importance of declarative scene modelling for a really computer aided design and some open research problems in order to improve drawbacks of this modelling technique. Some suggestions for possible future extensions of declarative modelling are also given.

1 Introduction

Scene modelling is a very difficult task in computer graphics as traditional geometric modellers are not well adapted to computer aided design. With most of the current modelling tools the user must have quite precise idea of the scene to design before using a modeller to achieve the modelling task. In this manner, the design is not a computer aided one because the main creative ideas have been elaborated without any help of the modeller.

The problem with most of the current scene modellers is that they need, very soon during the modelling process, low-level details which are not important in the creative phase of design. This is due to the lack of levels of abstraction allowing the user to validate general ideas before resolve low-level problems. If the initial very general idea of the designer is, for example, to design a scene comporting a house, a swimming tool in front of the house and a tree on one side, this idea may be realised in many different manners. As the modeller does not offer the user an easy manner to try and test different manners to realise the initial mental idea, he (she) generally tries a small number of possible solutions and chooses the best one. In this manner, the user may lack very interesting possible solutions.

Declarative modelling tries to give intuitive solutions to this kind of problem by using Artificial Intelligence techniques which allow the user to describe high level properties of a scene and the modeller to give all the solutions corresponding to imprecise properties.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 80–89, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 What Is Declarating Modelling

Declarative modelling [1, 2, 3, 4, 17] in computer graphics is a very powerful technique allowing to describe the scene to be designed in an intuitive manner, by only giving some expected properties of the scene and letting the modeller find solutions, if any, verifying these properties.

As the user may describe a scene in an intuitive manner, using common expressions, the described properties are often imprecise. For example, the user can tell the modeller that "the scene A must be put on the left of scene B". There exist several possibilities to put a scene on the left of another one. Another kind of imprecision is due to the fact that the designer does not know the exact property his (her) scene has to satisfy and expects some proposals from the modeller. So, the user can indicate that "the house A must be near the house B" without giving any other precision. Due to this lack of precision, declarative modelling is generally a time consuming scene modelling technique.

There exist two kinds of geometric modellers, general purpose modellers, allowing to design almost everything, and specialised (or dedicated) modellers, offering high level modelling for limited specific modelling areas. In the same manner, there exist two families of declarative modellers: *general purpose* modellers, covering a large set of possible applications, and *dedicated* modellers, covering a specific area (architectural design, mechanical design, …).

The principle of dedicated modelling is to define a declarative modeller each time it is necessary for a well delimited modelling area. Thus, PolyFormes [5] is a declarative modeller designed to generate regular or semi-regular polyhedra. The main advantage of the dedicated declarative modellers is efficiency because their solution generation engine can be well adapted to the properties of the specific modelling area covered by the modeller. On the other hand, it is difficult for such a modeller to evolve in order to be able to process another specific modelling area.

The aim of the general purpose modellers is generality. These modellers include a solution generation engine which can process several kinds of properties, together with a reduced set of pre-defined properties, as general as possible. General purpose declarative modellers could normally be specialised in a specific modelling area by adding to them new properties, corresponding to the specific modelling area we want to cover. In this sense, general purpose modellers can be seen as platforms to generate dedicated declarative modellers. The main advantage of general purpose declarative modellers is generality which allows to specialise a modeller in a specific modelling area without having to modify its solution generation engine. On the other hand, general purpose modellers suffer from their lack of efficiency, because of the generality of the solution generation mechanism. The declarative modeller MultiFormes [2, 3, 9, 10] is a general purpose declarative modeller.

It is generally admitted that the declarative modelling process is made of three phases: the *description* phase, where the designer describes the scene, the *scene generation* phase, where the modeller generates one or more scenes verifying the description, and the *scene understanding* phase, where the designer, or the modeller, tries to understand a generated scene in order to decide whether the proposed solution is a satisfactory one, or not.

3 Imprecision Management in Declarative Modellers

Declarative modelling tries to help the scene designer by allowing intuitive descriptions using a "language" close to the user's one. This kind of description is very often imprecise and can produce many solutions. The modeller has to manage this imprecision in the scene generation phase.Two modes are used by declarative modellers to manage imprecision during the generation phase: *exploration* mode and *solution search* mode.

In exploration mode, the declarative modeller, starting from a user description, performs a full exploration of the solution space and gives the user all found solutions. This mode can be used when the designer has insufficient knowledge of a domain and wants to discover it by an exhaustive exploration or when the designer is looking for new ideas and hopes that the modeller could help him (her) by exploring a vague description. The use of imprecise properties increases the richness of the solution space and allows the user to obtain concrete answers for a vague mental image. So, the use of imprecise properties is very important for the designer. As the exploration mode is based on the use of imprecise properties, it is very important to have techniques to reduce exploration cost by reducing the number of useless tries during the solution search process [9, 10]. A problem with the exploration mode is that the use of general imprecise properties can produce a very important number of solutions and make very difficult the management of these solutions. Furthermore, some families of solutions can be of no interest for the designer and he (she) would like to avoid generation of such solutions in subsequent generations. As the modeller does not know the designer's preferences, interaction is necessary to learn it what kind of scenes are not interesting.

In solution search mode, the modeller generally generates only one solution. To do this, either it interprets the designer's description in a restrictive manner or it asks the designer to precise his (her) choice. So, the designer must have a relatively precise idea of the kind of scenes he (she) would like to obtain.

Declarative modellers working in exploration mode are, generally, able to work in solution search mode if the designer would like to obtain a solution immediately or very quickly from a description using less imprecise properties. As the semantic of a property is often ambiguous and several solutions not satisfactory for the user can be faced by the modeller, the designer must have the possibility to interactively intervene in order to guide the modeller in its search. So, if parts of a solution proposed by the modeller are close to the idea the designer has of the wished scene, the designer should be able to tell the modeller not to modify these parts in proposed solutions. This interaction decreases the solution space because, for a great number of scenes verifying the properties of the initial description, these parts will not satisfy the intuitive idea of the user and these scenes will be avoided.

4 A Classification of Declarative Modellers

In this section will be presented some experimental declarative modellers developed since 1987, date of the beginning of declarative modelling. These modellers will be classified according to the mode - exploration mode or solution search mode - of imprecision management they can support during the scene generation phase.

4.1 Modellers Using Exploration Mode in Scene Generation

PolyFormes [5]. The goal of the PolyFormes declarative modeller is to generate all regular and semi-regular polyhedra, or a part of the whole, according to the user's request.
Requests may be more or less precise and are expressed using dialog boxes. This initial description is then translated in an internal model which will be used during the generation process. This internal model is a knowledge base, made of a rule base and a fact base.
The scene generation process, uses an inference engine which applies rules of the rule base to the facts of the fact base and creates new facts. A solution is obtained by saturation of the fact base. The whole modeller is an expert system on polyhedra. When the initial description is imprecise, all the possible solutions are generated.

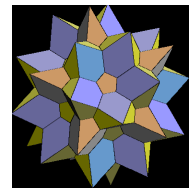


Fig. 1. Scene generated by PolyFormes

In Fig. 1 one can see an example of polyhedron generated by the PolyFormes declarative modeller. PolyFormes is a dedicated declarative modeller as it is specialised in generation of polyhedra.
MultiFormes. The first version of MultiFormes [2] was available in 1991. Several other versions of the modeller have been developed later . The purpose of this modeller was to be a general purpose declarative modeller, able to be specialised in any particular area.
MultiFormes is based on a new conception and modelling technique, declarative modelling by hierarchical decomposition (DMHD) . The DMHD technique can be resumed as follows:

- If the current scene can be described using a small number of predefined high level properties, describe it.
- Otherwise, describe what is possible and then decompose the scene in a number of sub-scenes. Apply the DMHD technique to each sub-scene.

Descriptions in MultiFormes are expressed by means of dialog boxes allowing to represent a tree-like structure, to select nodes and to assign them properties. The initial description is then translated to an internal model to be used during the scene

generation process. In the first version of MultiFormes the internal model was a knowledge base made of a set of rules and a set of facts. In all the other versions of the modeller, the internal model is a set of arithmetic constraints on finite domains (CSP). The scene generation process uses a constraint satisfaction engine which applies CSP techniques [14] to generate all the solutions corresponding to a given description. A special form of primitive arithmetic constraints, CLP (FD) [15], is used to improve the scene generation process.
The tree of the hierarchical description of a scene, used in the scene generation phase, allows scene generation in various levels of detail and reduction of the generation's cost. To do this, the modeller uses a bounding box for each node of the tree. This bounding box is the bounding box of the sub-scene represented by the sub-tree whose the current node is the root. All bounding boxes of the children nodes of a node are physically included in the bounding box of the parent node. This property permits to detect very soon branches of the generation tree which cannot be solutions. In Fig. 2, the spatial relation between the bounding boxes of a scene and its sub-scenes is shown (left), as well as a scene generated by MultiFormes (middle).

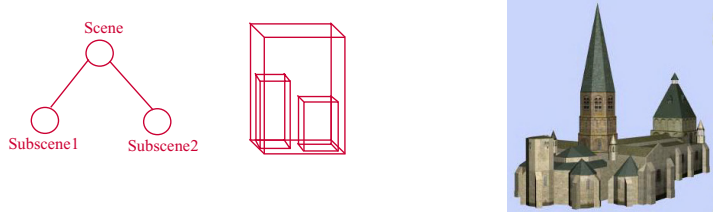


Fig. 2. From left to right: the bounding boxes of the sub-scenes of a scene are inside the bounding box of the parent scene; inside a 3-floor building; Cathedral of Le Dorat (France) designed by W. Ruchaud

MultiFormes is a general purpose declarative modeller which can be specialised by adding new predefined properties. The scene generation phase works in exploration mode, whereas it is possible to use solution search mode by means of user's interaction. MultiFormes can also use a set of geometric constraints [9] instead of arithmetic ones. This set contains constraints like "point P is in the box B" or "Points P1, P2, P3 are aligned". Satisfaction of this kind of constraints is computed using CSP-like techniques and allows more flexibility in creating scenes. An example of scene generated by geometric constraint satisfaction techniques can be seen in Fig. 2. Machine learning techniques based on neural networks [6, 7] have been implemented in MultiFormes [8]. These techniques allow the modeller to select scenes close to the designer's desires in solution search mode, during the scene generation phase.

4.2 Modellers Using Solution Search Mode in Scene Generation

DE2MONS. The DE2MONS declarative modeller [19] is a general purpose modeller whose main properties are:

- A multi modal interface,
- A generation engine limited to the placement of objects,
- A constraint solver able to process dynamic and hierarchical constraints.

The modeller uses a multi modal interface allowing descriptions by means of the voice, the keyboard (natural language), a data glove or 3D captors informing the system of the user's position. The description is translated in an internal model made of linear constraints.The generation engine of DE2MONS uses a linear constraint solver, ORANOS, able to process dynamic constraints (new constraints can be added during generation) and hierarchical constraints. Hierarchical constraints are constraints with priorities assigned by the user. Whenever there is no solution for a given description, constraints with low priority are released in order to always get a solution. The solver computes one solution for a given description.

CCAD. The Cooperative Computer Aided Design (CCAD) paradigm was introduced by S. Kochhar [11,12] to facilitate the integration of *generative* and traditional modelling systems by allowing the designer to guide de generative system through successive rounds of automated geometric modelling.The notion of generative modelling is very close to the notion of declarative modelling, as in both cases imprecise descriptions can generate many solutions. An experimental cooperative scene modeller was implemented for a generative system based on the formal language of schema grammars. The CCAD framework is based on three main premises:

- A generative geometric modelling (GGM) system exists and can be used to generate a set of designs based on some designer-defined constraints or properties.
- The GGM system is supposed not to produce perfect designs, but rather it will be guided to search for better designs by the human designer.
- As the GGM system produces a large set of designs, a specialised browsing system allows the designer to search the set of generated designs in a directed manner.

A typical modelling session using the CCAD system proceeds as follows:

- The designer uses the TGM system to generate a nascent design to be used in the first iteration of automated modelling.
- The designer then uses a dialog with the GGM system to define the constraints to be used during the generation process.
- The GGM system then instanciates all valid geometric designs. These designs are presented as icon-like buttons in a large screen area and the designer can get a large image of a design by clicking on the corresponding button.
- The designer then selects a set of promising designs using the browsing system.
- The selected designs are then returned to GGM system and the 4 four last steps are repeated until the desired design has been constructed.

The CCAD paradigm has also been applied by D. Sellinger [13] in a PhD thesis whose purpose was to integrate declarative and traditional geometric modelling.

VoluFormes [16]. VoluFormes is a *dedicated* declarative modeller allowing the user to quickly define boxes in the space whose purpose is to check the growth of forms. It is made of two modules:

- *Voluboites*, which allows to define boxes where the spatial control is performed.
- *Voluscenes*, which allows to use growth mechanisms applied to elementary germs and to create forms, taking into account the spatial control boxes.

Only Voluboites works in declarative manner. The positions of spatial control boxes are described during the description phase using a natural-like language. Description and generation work in incremental manner. Each box is placed in the 3D space and, if the user does not like the proposed box and placement, another solution

can be given. Once the current box is placed in the space, the same process is applied to define the next one. The generation engine of Voluboites uses a CSP-like constraint satisfaction technique. On the left of Fig. 10, one can see a boxes arrangement obtained by Voluboites.

Voluscenes is used by the designer to put germs in the boxes and to select the growth mechanism among a reduced number of predefined imperative mechanisms. On the right of Fig. 3 one can see an example of scene obtained by form growth.

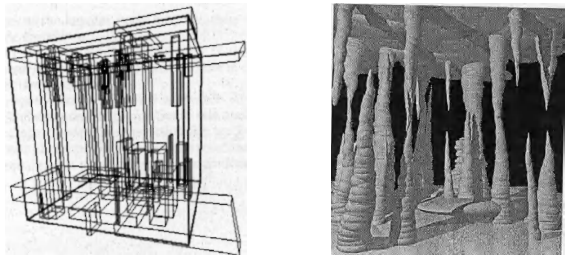


Fig. 3. Boxes arrangement and form growth with VoluFormes

5 Scene Understanding in Declarative Modelling

As declarative modelling generates several solutions and most of them can be unexpected, it is often necessary that the modeller offers scene understanding techniques in order to allow the designer to verify the properties of an obtained solution. Scene understanding can be visual or textual. Most of existing declarative modellers use simple scene display from an arbitrary chosen point of view. Very few declarative modellers use sophisticated scene understanding mechanisms.

PolyFormes uses a "matches-like" display mode allowing the user to better understand the shape of a generated polyhedron. In this kind of display, only the edges of the polyhedron are displayed but they are thickened (see Fig. 4).

MultiFormes uses more sophisticated techniques for scene understanding [20]. These techniques use a good view criterion based on the scene's geometry and automatically compute a good point of view by *heuristic search*. As a single point of view is not always sufficient to understand complex scenes, MultiFormes also proposes an intelligent automatic scene exploration by a virtual camera, moving on the surface of a sphere surrounding the scene (Fig. 4).

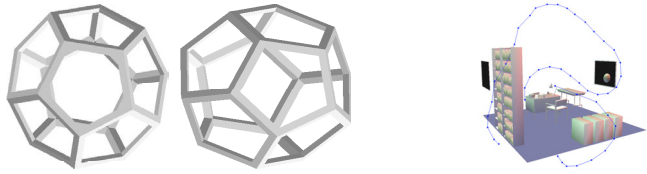


Fig. 4. From left to right: "Matches-like" display of polyhedra; Scene automated exploration by a virtual camera

6 Advantages and Drawbacks of Declarative Modelling

The purpose of declarative modelling is to improve the hard task of scene modelling by allowing the designer to use a high level of abstraction. Taking into account the existing declarative modellers, it is possible to extract the main advantages and drawbacks of declarative modelling.

Declarative modelling is closer to the user than traditional geometric modelling because the user has not to take into account geometric representation and construction of the scene. The declarative modeller produces a scene with the high level properties described by the designer and then translates this solution to a geometric model. In declarative modelling it is possible to describe a scene up to a chosen level of detail. In this manner, even if the designer has not yet determined some parts of the scene, he (she) can get a first draft of the scene and refine it later. Modifications of a scene are easier with declarative modelling because it is easier to replace a property by another one than to modify the scene's geometry. Indeed, it is easier to modify an object by telling the modeller that it is not enough high than by changing the list of vertices in the geometric model of the object. Declarative modelling allows really computer aided design because, from an initial description of a vague idea of the designer it is able to propose several solutions and help the designer in the creative phase of scene design. A declarative modeller can evolve because, if properties are defined in a manner not depending on the scene generation engine, new properties can easily be added to the modeller and improve its designing power.

The main drawback of declarative modelling comes from the possibility to use imprecise properties. When, for a given description, the *search space* is much wider than the *solution space*, the generation process is very time consuming. A research field in the area of declarative modelling is to find efficient methods of reducing the search space. Another drawback is due to the possibility to get many solutions from a scene description. The management of an important number of solutions is always difficult because the designer cannot remember all the solutions during the designing process. A possible solution would be to define a metric and classes of close to each other solutions by the modeller and to show only a representative scene from each class [18]. The problem is that it is not always possible to define such a metric, especially with general purpose modellers.

7 Future Issues

The problem declarative modelling has to face is a hard one. Despite of this complexity, the performances of existing declarative modeller prototypes are quite satisfactory. However, the problem of efficient reduction of the search space is an open research problem.

Current declarative modellers are essentially concerned with geometrical or topological aspects of a scene. However, it is possible to describe in declarative manner non geometric properties of the scene such as ambience (lighting, fog, ...). Some research works have started in this area. In a general manner, if a property may be

translated into constraints on numerical values, it can be described and processed in declarative manner.

Another challenge is to include time in declarative modelling, especially for scenes corresponding to existing things in the real world. For instance, it would be interesting for an architect to show a client not only the current state of a building to be designed but also its appearance 10 or 15 years later, if a small number of parameters such as, construction materials, climate, etc., are known. We call this kind of modelling *predictive declarative modelling*. Some works have started in this area too.

Finally, coupling declarative and traditional scene modelling should give interesting results because it would allow to reduce the cost of declarative modelling, by permitting to first define a draft of the scene by declarative modelling and then refine the draft using an integrated geometric modeller. Such an integrated declarative-traditional geometric modeller has already been implemented [13] but a lot of improvements are still possible in this area.

8 Conclusion

In this paper we have tried to present the challenge of declarative modelling as a tool to make easier the scene designer's work and offering a really computer aided design able to stimulate the designer's creativity. Even if many things have to be improved in declarative modelling, most dedicated declarative modellers are very efficient and, above all, able to produce scenes impossible to obtain by traditional geometric modelling.Declarative modelling has opened a lot of research fields in computer graphics as well as in artificial intelligence. Open research fields currently concern reduction of search space during the generation phase (efficient constraint resolution), management of a big number of solutions (classification, machine-learning, interactivity) and scene understanding (scene visual complexity, path optimisation).

The authors would like to thank Dr. Andres Iglesias for his kind invitation to the CGGM conference.

References

1. Lucas M., Martin D., Martin P., Plemenos D.: The ExploFormes project: some steps towards declarative modelling of forms. Published in BIGRE, no 67, pp 35 – 49, 1990.
2. Plemenos D.: A contribution to study and development of scene modeling, generation and display techniques - The MultiFormes project. Professorial Dissertation, Nantes (France), November 1991 (in French).
3. Plemenos D.: Declarative modeling by hierarchical decomposition. The actual state of the MultiFormes project. International Conference GraphiCon'95, St Petersbourg, Russia, 3-7 of July 1995.
4. Plemenos D., Tamine K.: Increasing the efficiency of declarative modelling. Constraint evaluation for the hierarchical decomposition approach. International Conference WSCG'97, Plzen (Czech Republic), February 1997.
5. Martin D., Martin P.: PolyFormes: software for the declarative modelling of polyhedra. The Visual Computer (1999) 55-76.

6. Mc Culloch W.S., Pitts W.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5, 115 - 133, 1943.
7. F. Rosenblatt: The perceptron: a perceiving and recognizing automaton. Project Para, Cornell Aeronautical Lab. Report 85-460-1, 1957.
8. Plemenos D., Miaoulis G., Vassilas N.: Machine learning for a general purpose declarative scene modeller. International Conference GraphiCon'2002, Nizhny Novgorod (Russia), September 15-21, 2002.
9. Ruchaud W., Plemenos D.: MultiFormes: a declarative modeller as a 3D scene sketching tool. Int. Conference ICCVG'2002, Zakopane (Poland), September 25-29, 2002.
10. Bonnefoi P.-F.: Constraint satisfaction techniques for declarative modelling. Application to concurrent generation of scenes. PhD thesis, Limoges (France), June 1999.
11. Kochhar S.: Cooperative Computer-Aided Design : a paradigm for automating the design and modeling of graphical objects. PhD thesis, Harvard University, Aiken Computation Laboratory, 33 Oxford Street, Cambridge, Mass. 02138, 1990. Available as TR-18-90.
12. Kochhar S.: CCAD : A paradigm for human-computer cooperation in design. IEEE Computer Graphics and Applications, May 1994.
13. Sellinger D., Plemenos D.: Interactive Generative Geometric Modeling by Geometric to Declarative Representation Conversion. WSCG'97 conference, Plzen (Czech Republic), 10-14 February 1997.
14. van Hentenryck P.: Constraint satisfaction in logic programming. Logic Programming Series, MIT Press, 1989.
15. Diaz D.: A study of compiling techniques for logic languages for programming by constraints on finite domains: the clp(FD) system. PhD thesis, 1995, Orleans, France.
16. Chauvat D.: The VoluFormes Project: An example of declarative modelling with spatial control. PhD Thesis, Nantes, December 1994.
17. Poulet F., Lucas M.: Modelling megalithic sites, Eurographics'96, pp 279-288, Poitiers (France), 1996.
18. Champciaux L.: Introduction of learning techniques in declarative modelling, PhD thesis, Nantes (France), June 1998.
19. Kwaiter G.: Declarative scene modelling: study and implementation of constraint solvers. PhD thesis, Toulouse (France), December 1998.
20. Barral P., Dorme G., Plemenos D.: Visual understanding of a scene by automatic movement of a camera. Short paper. Eurographics 2000.

Geometric Snapping for 3D Meshes

Kwan-Hee Yoo[1] and Jong Sung Ha[2]

[1] Dept. of Computer Education and Dept. of Information Industrial Engineering, Chungbuk National University, 48 San Gaesin-Dong Heungduk-Gu Cheongju Chungbuk 361-763 Republic of Korea
khyoo@cbucc.chungbuk.ac.kr

[2] Dept. of Computer Engineering, Woosuk University, 490 Hujongri, Samrae-Up Wanju-Kun Chonbuk 565-701 Republic of Korea
jsha@woosuk.ac.kr

Abstract. Image snapping is the technique to move a cursor position to a nearby feature such as edges in a 2D image when the cursor is located by a user. This paper proposes a new snapping technique called the *geometric snapping* that naturally moves the cursor position to a geometric feature in 3D meshes. The cursor movement is based on the approximate curvatures defined for measuring the geometric characteristics of the 3D meshes. The proposed geometric snapping can be applied to extracting geometric features of 3D mesh models in many CAD and graphics systems.

1 Introduction

The cursor snapping presented in Sketchpad systems is a well-known technique for interactively providing an exact cursor position in graphics user interfaces [12]. Many CAD and graphics systems have adopted the cursor snapping for picking 2D objects such as line segments, circles, polygons, and so on. This technique was extended into picking objects in 3D space by locating the cursor in a 2D screen [1,2,11]. Image snapping is another evolution of the cursor snapping proposed by Gleicher [5], which moves the cursor position to a nearby feature such as edges in an image when the cursor is located by a user. The image snapping can be applied to extracting edge features from an image as well as editing the image efficiently.

In order to visualize the world more realistically in CAD and graphics systems, there have been diverse 3D models describing objects in the real world. Most of the 3D models tend to be represented with 3D meshes for being effectively processed. One of the most important processing in the meshes is to detect the *geometric features* that represent the main boundaries of the 3D meshes, since they are crucial for deciding which parts of the meshes have to be processed or to be preserved in many applications such as mesh simplification, mesh compression, mesh editing, mesh morphing, and mesh deformation [4,13-16]. In mesh simplification and compression, the geometric features have to be maximally preserved. Mesh editing usually processes the parts representing geometric features in a mesh. Mesh morphing is also usually performed by using the

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 90–97, 2004.
© Springer-Verlag Berlin Heidelberg 2004

corresponding geometric features between two meshes. Meshes can be deformed by manipulating their parts representing the geometric features.

In this paper, we propose the *geometric snapping* that can be used as a basic technique for effectively processing meshes. The geometric snapping extends the notion of image snapping to mesh models in the three-dimensional space. In other words, when a user selects an arbitrary vertex or point of a given 3D mesh model with the cursor, the cursor naturally moves to a nearby geometric feature of the mesh. There are two major issues in the extension; defining geometric characteristics on the mesh and moving the cursor onto the surface of the mesh. In this paper, the geometric characteristics are represented with the change of normals of adjacent faces or the radius of the circle passing the centers of three adjacent edges in the mesh. Moreover, we define the movement cost that is required when the cursor moves from a vertex into another vertex. Using the proposed measurements, we develop efficient techniques for the geometric snapping, and then apply them to extracting geometric features from an example mesh model.

2 Geometric Characteristics of 3D Meshes

The effectiveness of geometric snapping depends on the methods of measuring the geometric characteristics of meshes and the strategies of moving the selected vertex to a point on geometric features by the measurement. In general, the definition of geometric characteristics of a mesh may vary according to each application. In this paper, we use the approximate curvatures defined on a mesh to measure the geometric characteristics. This section introduces the methods for computing the approximate curvatures and blurring them.

2.1 Computing Approximate Curvatures

The curvature for a given point in a 3D mesh is defined as the curvature of the curve lying in a plane containing the vector tangent to the surface at the point. The curvature at a specific point on a curve in the plane is defined as the ratio of change in slope on the point. Because it is difficult to calculate exact curvatures on a surface, other curvatures are defined: *principal*, *Gaussian*, and *mean* curvatures. The principal curvatures at a point of a surface are the minimum and maximum of the curvatures at the point, the Gaussian curvature is the product of two principal curvatures, and the mean curvature is the half sum of two principle curvatures [17]. Since it is very difficult to exactly calculate these curvatures on faces of 3D mesh models, there are many attempts to obtain approximate curvatures on a point of the mesh. An important factor in approximating the curvatures is how to explain main geometric features of a mesh model. Many powerful methods [4,9,10,13-16] for obtaining the approximate curvatures have been proposed. This paper proposes new methods for reflecting the geometric characteristics of 3D meshes more exactly. In the first method, we define the

approximate curvature $AC(v)$ on a vertex v in a given mesh by exploiting the normal vectors of faces containing v as:

$$AC(v) = 1.0 - \min_{i=0}^{k-1}(f_i^v \cdot f_{(i+1) \bmod k}^v) \tag{1}$$

In Equation (1), k is the number of faces sharing v, and f_i^v is the normal vector of the i-th face when the adjacency faces are ordered in counter-clockwise, and the operation $\cdot$ is the dot product of two vectors. The approximate curvature $AC(v)$ is defined as the subtraction of the minimum value among the inner products of normal vectors for all pairs of adjacent faces from 1.

As the second method, the curvature for a vertex v is the average of approximate curvatures on the edges incident to v. Let the ordered vertices adjacent to v be nv_i for all $i = 0, \ldots, k–1$, where k is the number of vertices adjacent to v. We denote the edge connecting v and nv_i with ne_i. Then, the curvature $C(ne_i)$ on the edge ne_i is defined as.

$$C(ne_i) = \frac{r_i^v + r_i^{nv}}{2} \tag{2}$$

In Equation (2), r_i^v is the radius of the circle passing the center of ne_i, and each center of two edges that are adjacent to ne_i while sharing v. Similarly, the radius r_i^{nv} of another circle is defined by ne_i and the two edges that are adjacent to ne_i while sharing nv_i. In addition to the proposed methods, we can compute approximate curvatures by applying other methods such as the quadric error metric [4], the inner product of two adjacent vertices [9], and the mean or Gaussian curvatures of edges or vertices [10].

2.2 Blurring Approximate Curvatures

If we regard approximate curvatures as height maps, the cursor movement of geometric snapping can be explained with a ball rolling down to valleys. The rolling ball may fall into local minimums before reaching at the deepest valley, which is caused by the limitation of computing method or the geometric characteristics themselves. In order to avoid this undesirable phenomenon, we soften up local minimums and emphasize global minimums by weighting the approximate curvature of each vertex on its nearby vertices. This technique is called the *blurring*. In this paper, we blur the approximate curvatures by using a well-known weighting factor called the *Gaussian smoothing filter*. That is, the approximate curvature on a vertex v is redefined as.

$$BAV(v) = \sum_{i=0}^{k-1} AC(nv_i) \times \frac{1}{\sqrt{2\pi}\sigma} \times e^{\frac{-(dx^2+dy^2+dz^2)}{2\sigma^2}} \tag{3}$$

In Equation (3), the vector (dx_i, dy_i, dz_i) is $(v^x – nv_i^x, v^y – nv_i^y, v^z – nv_i^z)$ for a vertex $v = (v^x, v^y, v^z)$ and its adjacent vertex $nv_i = (nv_i^x, nv_i^y, nv_i^z)$. Since the smoothing degree of the Gaussian filter is determined by the size of σ, we assign appropriate values to σ according to the size of k; $\sigma = 0.85$ if $k \leq 7$, $\sigma = 1.7$ if $7 < k \leq 16$, $\sigma = 2.5$ if $16 < k \leq 36$, and $\sigma = 3.5$ if $k > 36$.

3 Geometric Snapping in 3D Meshes

Assume that a user selects a vertex on the 3D mesh in which each vertex has the approximate curvature computed by one of the methods described in Section 2. The cursor pointing to the selected vertex should be moved into other vertex appearing geometric features. In order to process the movement, we express a given mesh as a connected graph whose vertices and edges are just the ones of the mesh. Each vertex of the graph corresponds to that of the mesh, and has 3D coordinates and the approximate curvature. In this section, after establishing the cost function that is used as a criterion for moving the cursor vertex to other vertex, we develop several strategies how to move the cursor by using the cost function.

3.1 Move Cost Function

We estimate the cost required for moving the cursor from a vertex to another vertex over a mesh. Let u and v, respectively, be the current vertex and the next vertex to be chosen. The cost function for moving from u to v is denoted by $cost(u, v)$, which is defined similarly to that of an image pixel [8] as:

$$movecost(u, v) = \omega_z f_z(v) + \omega_d f_d(u, v) + \omega_g f_g(v). \tag{4}$$

In Equation (4), the three functions of f_z, f_d, and f_g are Laplacian zero-crossing, curvature direction, and curvature magnitude respectively. The Laplacian zero-crossing $f_z(v)$ is used for representing whether or not a vertex v is on geometric features such as edges. From experimental results, we use the critical value of approximate curvatures for determining whether a vertex u represents the geometric features as; if $AC(v) > 2$ then $f_z(v) = 1$ else $f_z(v) = 0$. Since the vertex with a larger curvature represents the geometric feature better than other vertices with smaller curvatures, the curvature direction $f_d(u, v)$ is defined as $f_d(u, v) = AC(u) - AC(v)$. If $f_d(u, v) > 0$, the cursor tends to move from u to v. Otherwise, the movement occurs conversely. The last function of the curvature magnitude $f_g(v)$ is the approximate curvature $AC(v)$ itself. Each ω is also the weight of the corresponding function. We set the weights as $\omega_z = 0.43$, $\omega_d = 0.43$, and $\omega_g = 0.14$ respectively from the experimental results, that is, the Laplacian zero-crossing and the curvature direction play important roles while the curvature magnitude has a little effects relatively. For nonadjacent two vertices u and v, we consider the cost function $movecost(u, v)$ for moving from u to v by using the shortest path $sp(u, v)$ from u to v. If $sp(u, v)$ consists of a sequence of k vertices, $u = v_1, \cdots, v_k = v$, then the cost function $movecost(u, v)$ can be defined as:

$$movecost(u, v) = \sum_{i=1}^{k-1} movecost(v_i, v_{i+1}) \tag{5}$$

3.2 Strategies for Moving the Cursor

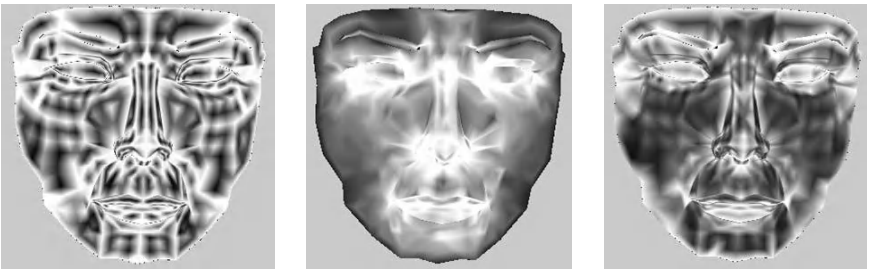
We consider three strategies for moving the cursor to a nearby geometric feature by computing the cost function. The first is to check the vertices adjacent to current vertex v. If the largest cost of the adjacent vertices is greater than 0, the cursor moves to the vertex with the largest. This movement is iterated until the costs of the vertices adjacent to current vertex are all zeros. This is a simple and convenient method, but it would take too much time in a dense mesh, i.e., lots of vertices are connected near to each other.

To enhance the performance of moving the cursor in a dense mesh, it is possible to check farther vertices with a certain range from v instead of its adjacent vertices. The second strategy is to use the range of Euclidean distance, while the third one is to use the range of path length. The Euclidean distance d may be determined by the adjacent vertex that is the farthest from v. An appropriate integer value n may be selected for the path length that is the minimum number of edges connecting two vertices. Hence, we check the vertices inside a sphere at origin v with the radius d, or the vertices whose path length to v is less than n. The cursor movement iterates same as the first method.

4 Experimental Results

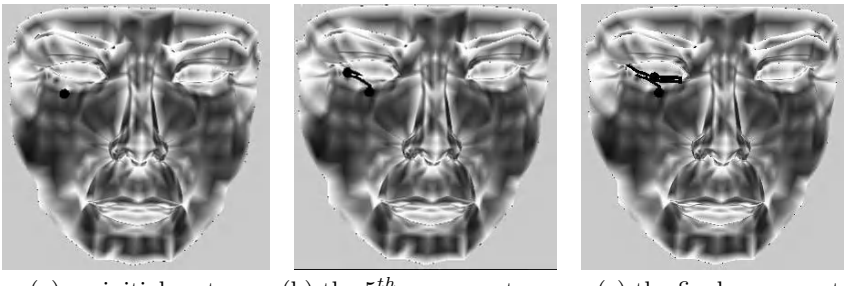
The proposed geometric snapping has been implemented on PC environments with the libraries of Microsoft Foundation Class (MFC) and OpenGL. The half-edge data structures are adopted for representing 3D meshes. We tested the implementation in the mesh of a face model. The approximate curvatures computed with Equation (1) and (2) in all vertices of the face model are visualized as grey colors in Fig. 1 (a) and (b) respectively, where the brightness depends on the magnitude of the curvatures. However, the computed approximate curvatures are discontinuous and noisy in some regions. To remove these phenomena, the approximate curvatures were blurred with Equation (3). Fig. 1 (c) illustrates the result from blurring the approximate curvatures in Fig. 1(a). For geometric snapping, a user first selects any vertex over a mesh that is preprocessed like Fig. 1(c). The located cursor will be moved into a nearby geometric feature within a certain neighboring range by using the *movecost* function in Equation (4) or Equation (5). Fig. 2 demonstrates the steps of the cursor movement in the first strategy using the adjacency for checking neighbors of the selected vertex: the 1^{st} movement (a), the 5^{th} movement (b), and the final movement (c). The final movements obtained by other strategies using a Euclidean length and a path length are shown in Fig. 2 (d) and (e) respectively. The cursor settled down after 3 movements for a determined Euclidean length, while 2 movements are needed for the path length 3. In these figures, the initial vertex selected by a user and the vertices passed by the cursor are black-colored. Solid lines represent the whole paths along which the cursor moved by the geometric snapping.

The problem of edge extraction in a 2D image [3,6,8] is very important for extracting feature boundaries in many applications. Similarly to the edge extraction in an image, we consider the extraction of geometric features such as

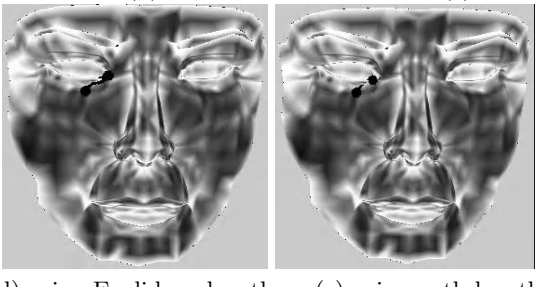


(a) Equation (1) (b) Equation (2) (c) Equation (3)

Fig. 1. Computing and blurring approximate curvatures



(a) an initial vertex (b) the 5^{th} movement (c) the final movement

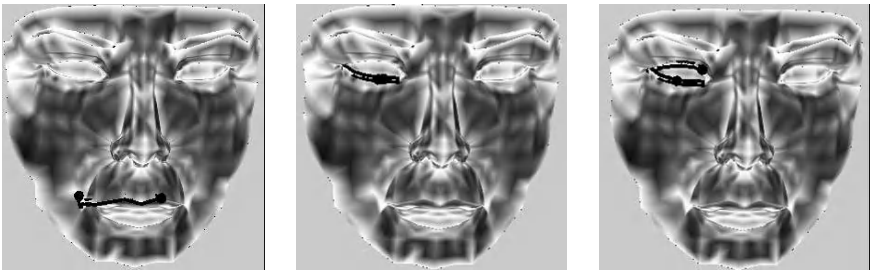


(d) using Euclidean length (e) using path length

Fig. 2. Moving the selected cursor

eyes, eyebrows, noses, and lips in a facial mesh model. Various applications such as facial deformation and facial animation in the facial model have needed the effective process of the geometric features. To extract the geometric features for a 3D mesh model, Lee and Lee [7] proposed the geometric snake as one method that is the extension of an image snake [6]. In this paper, we use the procedure of the geometric snapping for the geometric extraction; a sequence of vertices selected in this procedure can be identified as the geometric features. Fig. 3 (a) shows the result from applying the geometric snapping to extracting the boundary of lips. The black vertex is the one selected by a user, and the black solid lines represent the trace along which the cursor moves from the selected vertex when the geometric snapping is applied iteratively. Fig. 3 (b) and (c) are the

results from iteratively applying the geometric snapping to extracting the lower boundary and the whole boundary of a left eye respectively.



(a) lips (b) a left eye I (c) a left eye II

Fig. 3. Extracting geometric features

5 Conclusion

This paper proposed the technique of geometric snapping that naturally moves the cursor from a selected vertex to other vertex representing a geometric feature in a 3D mesh. We applied it to extracting geometric features from a face model. In the future, it is required to develop another technique of geometric snapping, which considers the relations among several vertices selected by a user.

The geometric features obtained by iteratively applying our geometric snapping are represented with a set of open or closed polygonal lines connecting a selected vertex and other vertices passed during the iteration. This is because a mesh consists of vertices, edges and faces. Hence, the obtained geometric feature may have the shape of staircases. It is also very important to remove these staircases of the extracted geometric features.

References

1. Bier, E., Snap-Dragging in Three Dimensions. Proc. Of Symposium on Interactive 3D Graphics, ACM Press, (1990), 193-204.
2. Bier, E., Stone, M., Snap-Dragging, Proc. Of SIGGRAPH'86, ACM Press, (1986), 223-240.
3. Falcao, A.X., User-Steered Image Segmentation Paradigms: Live Wire and Live Lane, Graphical Models and Image Processing **60**, (1998), 223-260.
4. Garland, M., Hecbert, P.S., Surface Simplification using Quadric Error Metric, ACM Computer Graphics (Proc. Of SIGGRAPH'97), (1997) 209-216.
5. Gleicher, M., Image Snapping, ACM Computer Graphics (Proc. of SIGGRAPH'95), (1995,) 183-190.
6. Kass, M., Witkin, A., Terzopoulos, D., Snakes, Active contour models. Int. Journal of Computer Vision **1**, (1987), 321-331.

7. Lee, Y., Lee, S., Geometric Snakes for Triangular Meshes, EuroGraphics Forum, (2002).
8. Mortensen, E., Barrett, W.A., Intelligent scissors for image composition, ACM Computer Graphics (Proc. of SIGGRAPH '95), (1995), 191-198.
9. Rosenfeld, A., Johnston, E., Angle Detection in Digital Curves, IEEE Transactions on Computers **22**, (1973), 875-878.
10. Smith, A.D.C., The folding of the human brain: from shape to function, PhD Dissertations, University of London, (1999).
11. Stork, A., An Algorithm for Fast Picking and Snapping using a 3D Input Device and 3D Cursor, CAD Tools and Algo-rithms for Product Design, (1998), 113-127.
12. Sutherland, I., Sketchpad: A Man Machine Graphical Communication System, PhD Dissertations, MIT, (1963).
13. Kobbelt, L.P., Bischoff, S., Botsch, M., Kehler, K., Ressl, C., Schneider, R., Vorsatz, J., Geometric modeling based on polygonal meshes, EUROGRAPHICS 2000 Tutorial, (2000).
14. Gu X., Gortler S., Hoppe H., Geometry images, Proceedings of SIGGRAPH' 02, (2002), 355-361.
15. Vorsatz, J., Rossl, C., Kobbelt, L., Seidel, H., Feature Sensitive Remeshing, Proc. of EUROGRAPHICS '01, (2001), 393-401.
16. Alliez, P., Cohen-Steiner, D., Levoy, B., Desbrun, M., Anisotropic Polygonal Remeshes, Proceedings of SIGGRAPH '03, (2003), 485-193.
17. Yamaguchi, F., Curves and surfaces in Computer Aided Geometric Design, Springer-Berlag, (1988).

Multiresolution Approximations of Generalized Voronoi Diagrams

I. Boada, N. Coll, and J.A. Sellarès

Institut Informàtica i Aplicacions, Universitat de Girona, Spain
{imma, coll, sellares}@ima.udg.es

Abstract. A framework to support multiresolution approximations of planar generalized Voronoi diagrams is presented. Our proposal is: (1) A multiresolution model based on a quadtree data structure which encodes approximations of a generalized Voronoi diagram at different levels of detail. (2) A user driven refinement strategy which generates from the quadtree a continuous polygonal approximation of the Voronoi diagram.

1 Introduction

The generalized Voronoi diagram of a set of sites partitions the plane into regions, one per site, such that all points in a region have the same closest site according to some given distance function [3,4,12]. Voronoi diagrams are widely used in many scientific fields and application areas, such as computer graphics, geometric modeling, solid modeling, geographic information systems, ... (see [17]). Although there are different algorithms to compute exact generalized planar Voronoi diagrams, they usually suffer from numerical robustness problems and are time-consuming. To avoid these problems researchers have proposed techniques to compute approximated Voronoi diagram within a predetermined precision. Among the existing techniques, we focus our interest on adaptive Voronoi diagram approximation methods based on hierarchical structures [11,16,15, 14,7,8,9].

One of the main advantages of hierarchical methods relies on their ability to support multiresolution. Multiresolution approaches can effectively control the tradeoff between quality and speed extracting models in which the resolution varies over the domain of the dataset. Multiresolution approximations of Voronoi diagrams are interesting in their own right. They are useful tools to solve problems in robot path planning [10,5], curve and surface reconstruction [2], and region approximation in GIS [1]. Unfortunately, algorithm for obtaining multiresolution approximations of Voronoi diagrams are still scarse. In this paper we propose a method to obtain them. The proposed approach is an extension of the algorithm we presented for the adaptive approximation of generalized planar Voronoi diagrams [7]. In this algorithm we built a quadtree, called the Voronoi Quadtree (VQ), to encode the information of the boundaries of the Voronoi regions in its leaf nodes. Then we group leaf nodes into five different patterns that determine how to generate the polygonal approximation of the part of the

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 98–106, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Voronoi diagram contained in the leaf. This approximation is represented by a DCEL structure.

Taking into account the capabilities of quadtrees to support multiresolution, we propose to use the VQ as a multiresolution model, i.e. the model that maintains the diagram approximations at different levels of detail. Our multiresolution approach starts with the construction of a VQ. Once all the information of the diagram is encoded in the VQ leaf nodes we apply a refinement process. The user defines a region of interest (ROI) and a degree of accuracy. Then, the process selects or generates, when it is required, the nodes from which the polygonal approximation of the Voronoi diagram satisfiying user requirements has to be obtained. The method guarantees that the approximation is continuous, the ROI is represented with the user desired degree of accuracy and out of the ROI the representation is as simple as possible.

2 Definitions and Notation

In this section we present the definitions and notation used int he paper. Let $\mathcal{S} = \{s_1, \cdots, s_n\}$ be the set of input sites. Each site s is represented by $s =< G_s, D_s, P_s >$, where G_s defines the geometry of the site s, D_s is the function that gives the distance from any point p to s and P_s (the base point of s) is a point such that $D_s(P_s) = 0$ and $P_s \in K$, where K is a rectangle. Each site $s_i \in \mathcal{S}$ has associated a Voronoi region $VR(s_i) = \{p \mid D_{s_i}(p) \leq D_{s_j}(p) \text{ for all } j \neq i\}$. The generalized Voronoi diagram of $\mathcal{S}$, denoted $VD(\mathcal{S})$, is defined as the partition of the plane induced by the Voronoi regions. Our goal is to obtain a multiresolution approximation of the part of $VD(\mathcal{S})$ included in K.

3 A Multiresolution Framework

Multiresolution representations permit extracting models in which the resolution varies over the domain of the dataset. In this way the user may choose to approximate with highest detail only some parts of the dataset, for example the ones considered of maximal interest. To define a framework able to support multiresolution Voronoi diagram approximations two issues must be resolved. First of all, it has to be defined a model capable of maintaining approximations of the Voronoi diagram at different levels of detail, taking into account that the level of detail may be different in distinct areas of the diagram. Secondly, it has to be defined a strategy able to generate from the information encoded in the multiresolution model the polygonal approximation of the Voronoi diagram that satisfies user requirements. The strategy has also to detect and solve the cracks (i.e discontinuities typical of domain decompositions that are fine in certain regions and coarse in the others).

4 A Quadtree-Based Multiresolution Model

Our multiresolution model must allow us to extract approximations of the Voronoi diagram suitable for diverse circumstances and it must also allow us to change the level of detail without excessive overhead. To satisfy all these requirements we propose to exploit the capabilities of the Voronoi quadtree data structure presented in [7]. Thus, the first phase of our approach consist on the creation of a Voronoi quadtree (VQ).

4.1 The Voronoi Quadtree

The VQ make use of the divide-and-conquer power of binary subdivision of quadtrees to encode the information required to obtain a polygonal approximation of a Voronoi diagram. Differently of other adaptive methods, that always consider all the set of sites, in the VQ approach at each step of the process we only take into account the sites related to each node, reducing in this manner the complexity of the diagram approximation with respect to other adaptative related methods.

To construct the VQ a set of basic definitions is introduced. Let N be a node and s a site. We say that: (i) s is a *I-site* with respect to N when $P_s \in N \cap VR(s)$; (ii) s is a *V-site* with respect to N when some vertex v of N verifies $v \in VR(s)$ and (iii) s is a *E-site* with respect to N when it is not a V-site and there exist some edge e of N that verifies $e \cap VR(s) \neq \emptyset$.

A node of the VQ is a leaf node when its level is L_M (the maximal subdivision level) or it is completely contained in a Voronoi Region, i.e. the total number of V-sites, I-sites and E-sites contained in the node is one. The VQ construction process is based on a breadth first traversal of the quadtree which uses a priority queue Q. The process starts with the creation of the root node assigning to it the rectangular region K and all the sites of S as I-sites. Then the V-sites of the root are computed and the root is sent to Q. In a loop over the Q nodes, for every node N we actualize its V-sites with the nearest of its sites. To maintain the coherence of the quadtree we evaluate adjacent nodes of N, modifying the information of its sites when it is required, and sending the nodes to Q if it is convenient. The construction process ends when Q is empty [7].

As the VQ encodes the information of the boundaries of the Voronoi regions in the leaf nodes we consider this phase as the initialization of the multiresolution model.

5 DCEL Based Multiresolution Polygonal Approximations of the Voronoi Diagram

To generate the polygonal approximation of the Voronoi diagram we will use the DCEL data structure [6]. This data structure uses three types of records, vertices, halfedges and faces, to maintain the adjacency between vertices, edges and faces of a planar subdivision. In [7] we describe how to obtain a DCEL based polygonal approximations of the Voronoi Diagram from the information encoded

in the VQ leaf nodes. We assign a pattern to each VQ leaf node according to the distribution of its V-sites. This pattern determines the position of the DCEL-vertices and how they have to be connected (see Fig. 1). The accuracy of the diagram approximation obtained from leaf nodes of level L_M is $\frac{\sqrt{a^2+b^2}}{2^{L_M}}$, where a and b are edge lengths of K.

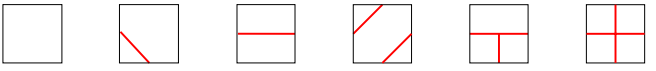


Fig. 1. Leaf node patterns of a Voronoi-Quadtree

To obtain a multiresolution approximation of the Voronoi diagram our idea is to extend the DCEL generation strategy proposed in [7]. The user defines the ROI, by a simple subregion of the dataset domain, and introduces the desired degree of accuracy ϵ. Then, since the VQ obtained in the previous phase has all leaf nodes at level L_M, we apply a refinement process that determines the set of nodes of the VQ from which the polygonal approximation of the Voronoi diagram that satisfies the user requirements has to be obtained.

The refinement process classifies VQ leaf nodes as outer ROI nodes if their four V-sites are out of the ROI and as inner ROI nodes on the contrary. The ϵ parameter determines the level $L_\epsilon = \left\lceil \log_2 \left(\frac{\sqrt{a^2+b^2}}{\epsilon} \right) \right\rceil$ of the quadtree at which nodes of the ROI have to be represented. Obtain the inner ROI representation is straightforward, it is only required the same subdivision process applied for the VQ construction described in the previous section.

The main difficulty of the refinement process is on the representation of outer ROI nodes, its representation has to be simplified as much as possible while preserving the continuity. Hence, the critical point is how to guarantee the continuity of the diagram approximation. To solve this problem we propose the crack solving strategy presented in the next section.

5.1 Crack Solving

To generate the DCEL approximation we apply the policy based on a set of predefined patterns (see Fig. 1) and the location of DCEL-vertices always on the midpoint of the intersected edges of the node or on the node's center. When an intersected edge is common to nodes of different level a crack situation arises. To define the crack solving strategy we have analyzed the possible crack situations. We detect three different cases, each one characterized by the distribution of the V-sites onto the intersected edge that contains the crack.

Case 1. The first case is given when the V-sites of the edge from one side are the same that the V-sites of the other one, the edge has a unique DCEL-vertex and the nodes that contain this DCEL-vertex have different levels. This case has been illustrated in Fig. 2(a.1). In this case to solve the crack we force the coarse

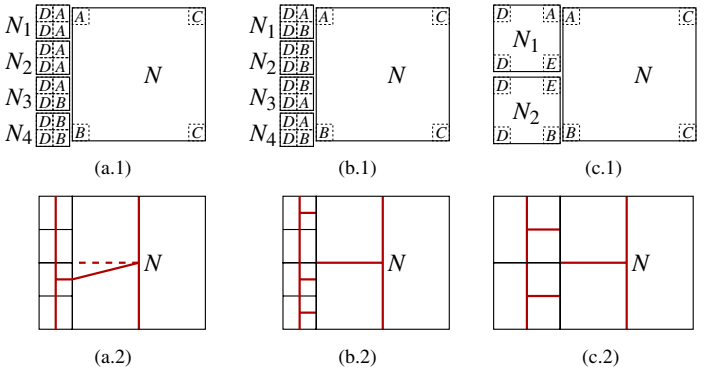


Fig. 2. (a.1) If the DCEL vertex is computed with respect to N has a different position that if it is computed with respect to N_3. (a.2) To solve the crack we always force the coarse leaf node representation to meet with the finer one.(b.1) Observe that all the V-sites located on the intersected edge are A or B. (b.2) Onto the common edge node there are different DCEL-vertices, one if we consider the edge from the N side, another another for the N_1 side, another for N_3 and another for N_4. To solve this crack node N has to be subdivided. (c.1) Node N has V-sites A and B while nodes N_1 and N_2 have V-sites A, E and B. (c.2) Onto the common edge there are three DCEL-vertices. To solve the crack the V-site E has to be introduced as an E-site of N, and N has to be subdivided

leaf node representation to meet with the finer one (see Fig. 2(a.2)).

Case 2. The second situation is given when, although the set of V-sites located on a common edge are the same, we identify more than one intersection point. This case has been illustrated in Figs. 2(b.1), 2(b.2). The correct approximation of the DCEL only can be obtained if the node of upper level is subdivided.

Case 3. The last case is given when the set of V-sites located on the common edge are different from one side to the other. This case has been illustrated in Fig. 2(c.1). To solve the discontinuity generated by this configuration (see Fig. 2(c.2)) the V-site not common to the two nodes has to be introduced as an E-site and the node has to be subdivided.

5.2 VQ Refinement Process

Once the situations of crack have been detected and we know how to solve them we define the refinement strategy. This refinement strategy uses a new leaf node criterion which varies according the position of the node with respect to the ROI. If the node is inside the ROI the criterion is the same used in the VQ construction phase, i.e. the node is a leaf if its number of sites is one. A node outside the ROI is a leaf if it has no E-sites.

Driven by this new leaf node criterion the refinement process starts with an initialization phase that detects all the VQ leaf nodes contained in the ROI and sends them to the queue Q. In a loop over Q nodes, for every node N its V-sites are actualized with the nearest of its sites. Then: (i) if N is a leaf node we apply

to it the propagation procedure. This procedure sends to Q the adjacent nodes of each vertex v of N that have to be reprocessed. Let N' be one of these nodes. N' has to be reprocessed if one of three situations represented in Fig. 3 is given. At the end of the propagation procedure if N and its brothers contain only one site, they are pruned. (ii)If N is not a leaf its four son nodes are created, the I-sites and E-sites of N are properly distributed to them and the V-sites of the sons are computed considering the sites of N. The propagation procedure is applied to each one of the descendant nodes and if a son is not a leaf either it is send to Q. The process ends when Q is empty.

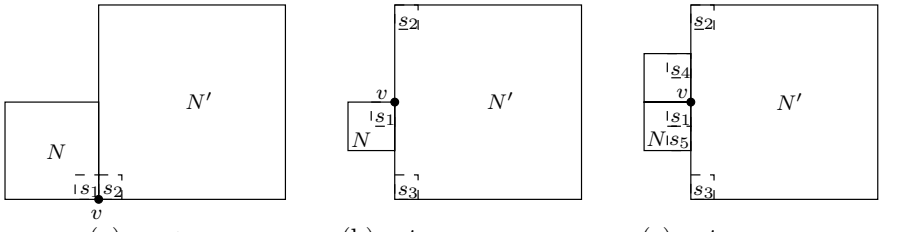


Fig. 3. N' has to be reprocessed if: (a) N' has a different V-site in v; (b) v lies on an edge of N', N' is inside the ROI, and the N V-site of v is different to one of the N' V-sites on the edge; (c) v lies on an edge of N', N' is outside the ROI, and the N V-site of v is different to each one of the nearest V-sites to v on the edge

At the end of the refinement process, leaf nodes of the ROI are all at level L_ϵ while nodes outside the ROI may be distributed at different levels of the quadtree. The different distribution of leaf nodes affects the DCEL obtainment. Now when a DCEL-vertex is common to leaf nodes of different level nodes we always force the coarse leaf node representation to meet with the finer one.

6 Computational Cost

In this section the computational costs of the processes related with the proposed method are summarized.

Let n be the number of sites, n_{ROI} be the number of sites whose Voronoi region intersects the ROI, $\overline{\partial VD(\mathcal{S})}_K$ be the length of the part boundary of $VD(\mathcal{S})$ included in K, and $\overline{\partial VD(\mathcal{S})}_{ROI}$ be the length of the part boundary of $VD(\mathcal{S})$ included in the ROI. There are some important considerations to be observed in order to obtain the computational cost of the VQ construction algorithm and the refinement process: (i) The algorithm applies the subdivision process to nodes that contain a piece of $VD(\mathcal{S})$. (ii)A curve of length $\overline{C}$ generates $O(\overline{C}\,2^l)$ nodes of level l in a quadtree, and $O(\overline{C}\,2^{L_M+1})$ nodes in a quadtree of maximum level L_M [13]. (iii) For each level we distribute the n sites to some

nodes as I-sites. (iv) For each node we need to locate its neighbor nodes. This can be done in L_M worst time, but the expecting time for locating neighbors is approximately $O(4)$ [13].

According to the previous considerations we have the next results. The number of nodes generated by the VQ construction algorithm is: $O(\overline{\partial VD(\mathcal{S})}_K 2^{L_M+1})$. The running time of the VQ construction algorithm is: $o(nL_M + \overline{\partial VD(\mathcal{S})}_K 2^{L_M+1})$. The number of nodes generated by the refinement process in a the ROI with accuracy ϵ is: $O(\overline{\partial VD(\mathcal{S})}_{ROI} 2^{L_\epsilon - L_M})$. The running time of the refinement process in the ROI with accuracy ϵ is: $o(n_{ROI}(L_\epsilon - L_M) + \overline{\partial VD(\mathcal{S})}_{ROI} 2^{L_\epsilon - L_M})$.

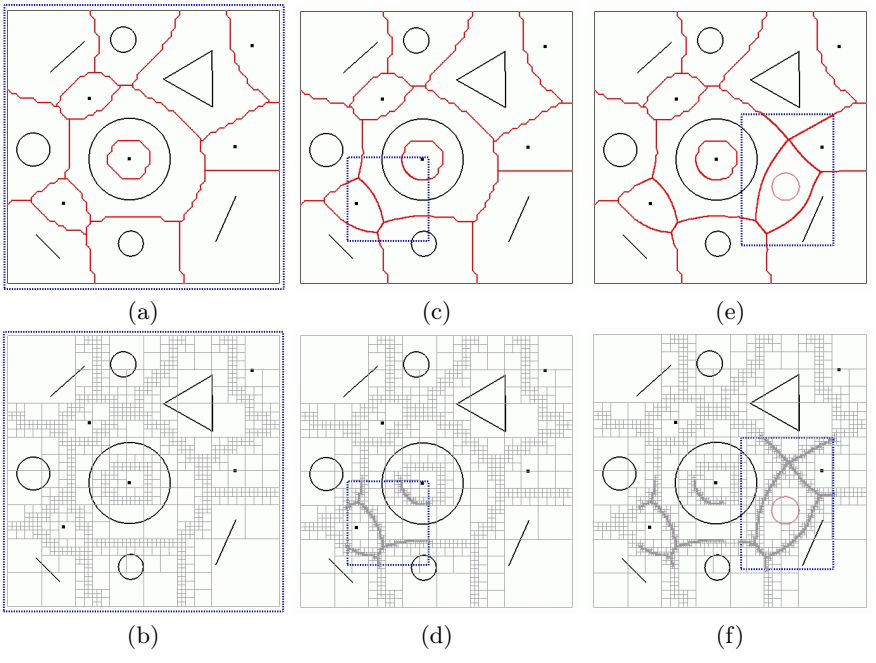


Fig. 4. Voronoi diagrams approximations obtained with the proposed approach are represented in the first row. The last row represents the distribution of nodes used to generate these polygonal approximation

7 Results

In this section we present the results obtained with our proposed approach when experimenting with a set of 13 sites. All these results have been represented in Fig. 4. The first image (see Figs. 4(a)) corresponds to the polygonal approximation of the Voronoi diagram obtained at the end of the VQ initialization phase.

In Fig. 4(b) we have also illustrated the distribution of leaf nodes. Next figures (see Figs. 4(c)(d)) correspond to the approximations of the diagram once a ROI has been defined. Observe that the refinement process only affects nodes of the ROI and some on the boundary of the ROI. It can be seen that there are no cracks on the polygonal approximation. We want to remark that our multiresolution approach supports dynamic maintenance, under insertion and deletion of sites, by using the strategy we describe in [9]. To illustrate this property we have introduced a new site to the diagram and once the method has properly updated the Voronoi diagram approximation we have selected a ROI around the new site. The obtained diagrams are represented in Figs. 4(e)(f)).

Acknowledgments. This work has been funded in part with grant numbers TIC2000-1009, TIC2001-2226-C02-02, TIC2001-2392-C03-01 and 2001-SGR-00296.

References

1. Alani, H., Jones, C.B., Tudhope, D.: Voronoi-based region approximation for geographical information retrieval with gazetteers. Int. J. Geographical Information Science, 15(4). (2001) 287-306
2. Amenta, N., Bern, M. Kamvysselis, M.: A new Voronoi-based surface reconstruction algorithm. Proceedings of Siggraph '98. ACM (1998) 415-421
3. Aurenhammer, F.: Voronoi diagrams: A survey of a fundamental geometric data structure. ACM Computer Surveys, 23(3). (1991) 686-695
4. Aurenhammer, F. Klein, R.: Voronoi diagrams. In: Sack, J.R., Urrutia, J. (eds.): Handbook of Computational Geometry. Elsevier (2000) 201-290
5. Behnke, S.: Local Multiresolution Path Planning. Proceedings of RoboCup 2003 International Symposium. (2003)
6. de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.: Computational Geometry; Algorithms and applications. 2nd edn. Springer-Verlag, Berlin Germany (2000)
7. Boada, I., Coll, N., Sellarès, J.A.: The Voronoi-Quadtree: construction and visualization. Eurographics 2002 Short Presentations. (2002) 349-355
8. Boada, I., Coll, N., Sellarès, J.A.: Hierarchical Planar Voronoi Diagram Approximations. Proceedings of 14th Canadian Conference on Computational Geometry. (2002) 40-45
9. Boada, I., Coll, N., Sellarès, J.A.: Dynamically maintaining a hierarchical planar Voronoi diagram approximation. In: Kumar, V. et al. (eds.): ICCSA 2003, Lecture Notes in Computer Science, 2669. Springer-Verlag (2003) 836-846
10. Kambhampati, S., Davis, L.S.: Multiresolution Path Planning for Mobile Robot's, IEEE Journal of Robotics Automation, RA-2(3). (1986) 135-145
11. Lavender, D., Bowyer, A., Davenport, J., Wallis, A., Woodwark, J.: Voronoi diagrams of set-theoretic solid models. IEEE Computer Graphics and Applications, 12(5). (1992) 69-77
12. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Application of Voronoi Diagrams. John Wiley (2000)
13. Samet, H.: Applications of Spatial Data Structures: computer graphics, image processing, and GIS. Addison-Wesley (1993)

14. Teichmann, T., Teller, S.: Polygonal approximation of Voronoi diagrams of a set of triangles in three dimensions. Technical Report 766. Laboratory of Computer science, MIT (1997).
15. Telea, A.C., van Wijk, J.J.: Visualization of Generalized Voronoi Diagrams. Proceedings of IEEE VisSym '01. Springer (2001) 165-174
16. Vleugels, J., Overmars, M.: Approximating Generalized Voronoi Diagrams in Any Dimension. Int. J. on Computational Geometry and Applications, 8. (1998) 201-221
17. Gold, C.: Voronoi Diagrams page on the Web: Applications. http://www.voronoi.com/section_1.htm

LodStrips: Level of Detail Strips

J.F. Ramos and M. Chover

Departamento de Lenguajes y Sistemas Informáticos
Universitat Jaume I, Campus de Riu Sec, 12071, Castellón, Spain
{jromero,chover}@uji.es

Abstract. Meshes representation at different levels of detail is an important tool in the rendering of complex geometric environments. Most works have been addressed to the multiresolution model representation by means of triangle meshes. Nowadays, models that exploit connectivity have been developed, in this paper a multiresolution model that uses triangle strips as primitive is presented. This primitive is used both in the data structure and in the rendering stage, decreasing the storage cost and accelerating the rendering time. Model efficiency is measured by means of a set of tests and results compared to Progressive Meshes and Multiresolution Triangle Strips multiresolution models, obtaining better rendering times and spatial cost.

1 Introduction

One of the main problems in graphics applications is the bottlenecks that take place in the graphics pipeline. These bottlenecks reduce the performance of the application and can vary even from frame to frame. The identification and elimination of these bottlenecks will be fundamental for the optimization of the application.

In each stage of the rendering process from the CPU to the GPU, there are different locations where these problems can appear. If the problem is analyzed from the point of view of geometry, the basic drawback is how to draw a huge number of triangles per frame. In this case, the problem is the number of vertices that are sent to the GPU. The traditional solution to this problem has been to use discrete level of detail (LOD), in an attempt to avoid overloading the CPU. Nevertheless, the use of discrete LODs has the disadvantage of popping and the need to use blending techniques.

In this paper, we present a continuous multiresolution model, called LodStrips, which has the following characteristics:

- *Continuity*. Transitions between levels of detail are smooth. The changes mean eliminating or adding one vertex.
- *Connectivity exploitation*. The model is based on the use of triangle strips. This leads to reduction in the storage and rendering costs.
- *Fast extraction*. It avoids the intensive use of the CPU that usually takes place with the continuous multiresolution models.
- *Cache use*. The use of strips means having at least one cache of two vertices.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 107–114, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Previous Work

Continuous multiresolution models are widely used because they are able to solve the problems of interactive rendering, progressive transmission, geometric compression and variable resolution. These models have been developed to represent, chiefly, triangle meshes. A characterization of these models can be found in [10]. Nevertheless, at the present time, some of the models presented exploit connectivity information using primitives like triangle strips or triangle fans. Fast rendering and a smaller storage cost can be achieved using these primitives. The rendering time decreases when fewer vertices are sent to the GPU and the connectivity of the mesh is stored implicitly.

One of the first models to use triangle strips is VDPM by Hoppe [6] After calculating the set of triangles to be rendered, this model performs an on-the-fly determination of the strips to be rendered. This is a time-consuming task but the final rendering time is reduced because triangle strips are faster than triangles.

Later, El-Sana et al. introduces the Skip-Strips model [2]. This is the first model to maintain a data structure to store strips, thus avoiding the need to calculate them on-the-fly.

Ribelles et al. introduced the MOM-Fan[9] This is the first model that no longer uses triangles, but instead another primitive that exploits connectivity. This model uses the triangle fan primitive both in the data structure and in the rendering stage. The main drawback of this model is the high number of degenerated triangles used in the representation. Another drawback to the model is that the average number of triangles in each triangle fan is small.

Following this approach, MTS by Belmonte et al. appeared. This is a model that uses the strip primitive in the storage and in the rendering stage [1]. The model is made up of a collection of multiresolution strips. Each multiresolution strip represents a triangle strip at every LOD, and this is coded as a graph. Only the strips that are modified between two consecutive LOD extractions are updated before rendering.

Recently, some works based on the triangle strip primitive have been presented. These focus on the dynamic simplification of the triangle strips for each demanded LOD. The model by Shafae et al. called DStrips [11] manages the triangle strips in such a way that only those triangle strips that are being modified are processed, while the rest of the triangle strips in the model remain unmodified. This updating mechanism reduces the extraction time. However, results published from this work still show a high extraction time.

Another approach to the use of triangle strips in a multiresolution model is the work carried out by A. James Stewart [12], and extended by Porcu [7]. This work uses a tunneling algorithm to connect isolated triangle strips, thus obtaining triangle strips with high numbers of triangles while reducing the number of triangle strips in the model as it is simplified. Again, its main drawback is the time consumed by the stripification algorithm.

Improvements of multiresolution models are applied in many ways. In [15] vertex connectivity exploitation is applied to implement a multiresolution scheme and in [14] a method is applied to maximize vertex reuse.

3 The LodStrips Model

The LodStrips model represents a mesh as a set of multiresolution strips. We denote a triangle strip mesh M as a tuple (V;S), where V is a set of vertices v_i with positions $v_i \in R^3$, and S is a collection of sub-triangulations $s^1,\ldots,s^m$, so each $s^i \in S$ is an ordered vertex sequence (1) also called a strip

$$\left\{S_1^i \cdots S_q^i\right\} \qquad S = \left\{ \begin{array}{c} S_1^1 \cdots S_k^1 \\ \vdots \\ S_1^m \cdots S_r^m \end{array} \right\} \qquad V = \left\{v_1 \ldots v_n\right\} \tag{1}$$

Each row inside the S matrix represents a triangle strip. After some modifications, this matrix will be adapted to become a multiresolution triangle strip data structure in order to be used in our model. In this way, this data structure will change during level of detail transitions, as described in the sections below.

The model has been built in order to optimize data access as well as the vertices sent to the graphics pipeline. In this way, it manages the triangle strips both in the data structure and in the rendering stage.

A set of vertices with their 3D coordinates and a set of multiresolution strips are needed to support the multiresolution model. Moreover, an auxiliary structure is used to improve level of detail transitions.

3.1 Data Structures

The purpose of the data structure is to store all the information necessary to recover every level of detail on demand. Three data structures are used: lVerts, lStrips and lChanges.

We denote a lVerts structure as a set V which contains an ordered vertex sequence and where each $v_i \in V$ consists of four items (2). The first three items are vertex positions: $(x_i, y_i, z_i) \in R^3$ and the last one, v_{k_i}, is the vertex into which v_i collapses, where $k > i$.

$$V = \left\{v_1, \ldots, v_n\right\} \qquad v_i = \left(x_i, y_i, z_i, v_{k_i}\right) \tag{2}$$

Thus, the lVerts data structure stores 3D coordinates and information about vertex simplification of each vertex in the mesh.

In order to collapse a vertex onto another, two kinds of simplification can be applied: external and internal edge collapses. External edge collapses consist of a vertex simplification where the destination vertex is on the external edge of the strip. Internal edge collapses are applied when the destination vertex is on the opposite edge. Simplification is achieved by means of two external vertex collapses.

Transitions between levels of detail involve vertex collapses. This information is stored in the lVerts data structure and when a vertex v_i has to be collapsed, it is replaced by v_{k_i} in every strip where it appears.

V is ordered according to its simplification order, that is, v_0 will be the first vertex to collapse, v_1 will be the second, and so on. Assuming that a vertex simplification is a level of detail change, when a transition from LOD i to LOD i+1 is demanded by the application, vertex v_i is replaced by v_{k_i} in all occurrences of that vertex and in every multiresolution triangle strip, in other words, in the data structure lStrips. Thus, an initial mesh M^1 can be simplified into a coarser M^n by applying a sequence of successive vertex collapse operations.

The sequence of vertex collapses is selected from simplification algorithms, since it determines the quality of the approximating meshes.

The multiresolution strip set is stored by the lStrips data structure. It consists of a collection L, where each $L^i \in L$ is an ordered vertex sequence, which denotes a multiresolution strip.

$$L = \begin{Bmatrix} v_1^1 & \dots & v_r^1 \\ \vdots & \vdots & \vdots \\ v_1^m & \dots & v_t^m \end{Bmatrix} \tag{3}$$

Each row $L^i \in L$, or each strip in the L collection, changes dynamically with vertex collapses and with strip resizing.

Vertex collapses are performed by replacing vertices in the data structure L by others that simplify them. These collapses can give rise to situations where repeated sequences appear in the data structure and these repetitions have to be removed, which involves resizing some strips.

Thus, as vertex collapses are applied, the lStrips data structure will become smaller, which allows us to have a very simple data structure for coarse levels of detail.

The lodStrips model incorporates a new data structure, lChanges, which allows us to quickly recover the positions of the vertices that are changed in each level of detail transition; this also allows the quick removal of consecutive vertex repetitions.

We denote a lChanges data structure as a level of detail ordered set, C, where each tuple c_j^i has the structure (l_j^i, k_j^i), where $l_j^i \in L$, which represents a position in L and k_j^i is another scalar that determines whether l_j^i values are for collapsing a vertex or for removing a set of consecutive vertices.

$$C = \begin{Bmatrix} c_1^1 & \dots & c_s^1 \\ \vdots & \vdots & \vdots \\ c_1^m & \dots & c_t^m \end{Bmatrix} \tag{4}$$

This data structure increases model performance because it allows us to quickly apply level of detail changes between transitions. Without this data structure it would be very expensive to apply these changes.

3.2 Algorithms

Multiresolution models need algorithms to be able to support multiresolution capabilities. The LodStrips model and most multiresolution models have two main algorithms to do these tasks, i.e. a level of detail recovery algorithm and a drawing algorithm. We assume the rendering stage to be a stage that contains these two algorithms, which are applied in a sequential order, first extraction and then drawing.

The level of detail recovery algorithm goes into action when a level of detail change is induced by the application. Then, data structure C is traversed from $C^{currentLOD}$ to C^{newLOD}, applying changes stored in each tuple Cij ∈ C, where i is in the interval [currentLOD , newLOD]. It is important to notice that, depending on whether the level of detail is bigger or smaller than the current one, splits or collapses will be applied to the lStrips data structure, although the information stored in c_j^i referring to collapses is also used to perform splits. The pseudo-code is shown below.

Level of detail recovery algorithm.

```
for lod=currentLOD to newLOD
   if newLOD>currentLOD   //To a more coarse mesh
    for change=lChanges.Begin(lod) to lChanges.End(lod)
     if (change.isCollapse()) then
      lStrips.Collapse(lod,change);
     else
      lStrips.Resize(lod,change);
   else  //To a more detailed mesh
    for change=lChanges.Begin(lod) to lChanges.End(lod)
     if (change.isSplit()) then
      lStrips.Split(lod,change);
     else
      lStrips.Resize(lod,change);
```

After the level of detail recovery algorithm has processed multiresolution strips the drawing algorithm takes over, traversing each strip to obtain their vertices in order to send them to the graphics system.

3.3 Model Performance

LodStrips model consists of data structures, with a pre-process that fills them, and algorithms for online extraction of the level of detail demanded:

- *Stripification*. Using the STRIPE algorithm [3] the lVerts data structure is built and lStrips filled with the highest level of detail.
- *Simplification*. We get the vertex collapsing order by means of the QSLIM algorithm [4].
- *Arrangement*. Once we have the vertex collapsing order, data structures must be adapted to the simplification order obtained from QSLIM.

- *Collapse*. For each v_i E V, v_{k_i} is calculated, as described in section 3. This item represents the vertex v_i to be collapsed. It is calculated taking into account the results of the simplification process.

4 Results

The LodStrips model has been submitted to several tests, all of which are aimed at evaluating the rendering time in a real time application.

All these tests were carried out to PM [5], MTS [1] and LodStrips, and results were compared. The first model has been and still is a reference model in the multiresolution world. The second is the most recent multiresolution model that makes complete use of multiresolution strips.

To carry out the tests, three well-known meshes from the Stanford 3D Scanning Repository were taken as a reference, so as to make it easy to compare this model with other well-developed models.

Tests were carried out with a PC with an Intel Xeon 2.4 Ghz processor and 512 Mb of main memory, using an ATI Fire GL E1 64 Mb graphics card.

Table 1 shows triangle mesh costs of the objects used in the tests and the sizes of the three models compared. It can be seen how the model presented here has a spatial cost that is lower than the rest of the models compared. This allows more objects to be placed in memory if necessary.

Table 1. Spatial cost comparison in Mb.

Mesh	Tr. Mesh	PM	MTS	LS	Ratio PM	Ratio MTS	Ratio LS
Cow	0.100	0.272	0.252	0.186	2.7	2.5	1.9
Bunny	1.193	3.282	2.963	2.111	2.8	2.5	1.8
Phone	2.850	7.863	6.765	4.844	2.8	2.4	1.7

Tests designed to compare multiresolution models follow the ones introduced by [8]. The tests carried out are the linear test and exponential test.

The linear test consists of extracting the LODs in a linear and proportionately increasing or decreasing way. The Exponential test consists of extracting LODs in an exponential way, that is, in the beginning it extracts very distant levels of detail and, later, it extracts closer levels.

The following tables show the results of applying the linear and exponential tests to models PM [5], MTS [1] and the one presented here, LodStrips.

As can be seen in Table 2, corresponding to the linear and exponential tests, the total rendering time is shown first. The lower part of the table, shows the percentage of this time used in extracting the level of detail and in drawing the resultant mesh.

Table 2. Linear and exponential tests

	LINEAR TEST						EXPONENTIAL TEST					
	PM		MTS		LodStrips		PM		MTS		LodStrips	
	Render (ms)		Render (ms)		Render (ms)		Render (ms)		Render (ms)		Render (ms)	
	% rec	% drw	% rec	% drw	% rec	% drw	% rec	% drw	% rec	% drw	% rec	% drw
Cow	0.917916		0.934682		0.231398		1.234464		0.53519		0.298161	
	6.43	93.57	24.36	57.38	24.36	75.64	5.88	94.12	37.6	62.4	20.73	79.27
Bunny	10.792452		6.304261		3.077063		16.164691		6.998482		4.129842	
	0.59	99.41	21.85	78.15	2.57	97.43	0.49	99.51	17.54	82.46	1.89	98.11
Phone	32.983562		14.812924		8.301228		48.922801		16.735283		11.756625	
	0.24	99.76	16.29	83.71	1.65	98.35	0.17	99.83	12.81	87.19	1.15	98.85

As we can see in both tests, the LodStrips model offers better rendering times than MTS and PM. LodStrips spends a small percentage of time on extracting the level of detail, which leads to good rendering times. In the opposite case, MTS spends quite a lot of time on extraction, and this slows down the total rendering time for this model.

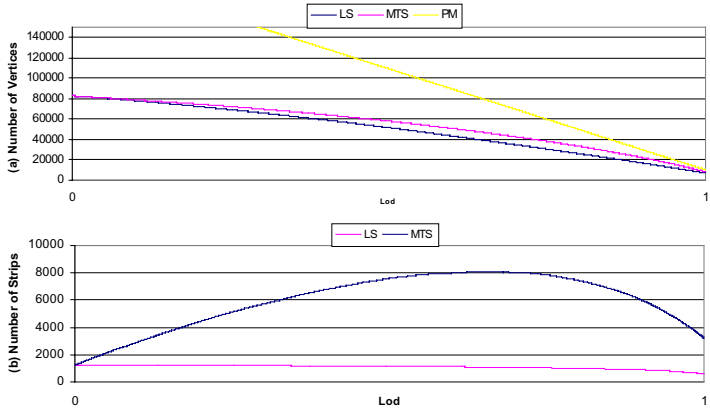


Fig. 1. a) Vertices sent to the graphics system by the bunny object in a linear test, b) Strips sent by MTS and LodStrips model in a linear test for the bunny object.

Vertices sent to the graphics system are directly related to the rendering time. The number of vertices sent by the model can be seen in Figure 1 a).

Strips are a way for organizing vertices that allows us to have a virtual two-vertex cache. As we can see in Figure 2 b), the LodStrips model has a low variation of strips sent, whereas the MTS model has a high variation of them in the progression of levels of detail. In any case, it seems that there is no relation between the vertex sent and the number of strips sent.

5 Conclusions

The LodStrips model offers many advantages and it should be underlined that it is a model with only three simple data structures and it is easy to implement. Moreover, it offers a fast LOD extraction which allows us to obtain smooth transitions between LODs, as well as very good rendering times because extraction is usually an important

part of the total rendering time. This model is wholly based on the triangle strips, which leads to an important reduction in storage and rendering costs.

This work was supported by the Spanish Ministry of Science and Technology grants TIC2001-2416-C03-02 and TIC2002-04166-C03-02, and FEDER funds.

References

1. O. Belmonte, I. Remolar, J. Ribelles, M. Chover, M. Fernández. Efficient Use Connectivity Information between Triangles in a Mesh for Real-Time Rendering, Future Generation Computer Systems, 2003.
2. El-Sana J, Azanli E, Varshney A. Skip strips: maintaining triangle strips for view-dependent rendering. In: Proceedings of Visualization 99, 1999. p.131-7.
3. F. Evans, S. Skiena and A. Varshney, Optimising Triangle Strips for Fast Rendering, IEEE Visualization '96, 319-326, 1996. http://www.cs.sunysb.edu/~stripe
4. M. Garland, P. Heckbert, Surface Simplification Using Quadratic Error Metrics, SIGGRAPH'97, 209-216, 1997.
5. Hoppe H. Progressive Meshes. Computer Graphics (SIGGRAPH), 30:99-108, 1996.
6. Hoppe H. View-dependent refinement of progressive meshes. SIGGRAPH, 1997.
7. Massimiliano B. Porcu, Riccardo Scateni. An Iterative Stripification Algorithm Based on Dual Graph Operations. EUROGRAPHICS 03.
8. J. Ribelles , M. Chover, A. Lopez and J. Huerta. A First Step to Evaluate and Compare Multirresolution Models, Short Papers and Demos EUROGRAPHICS'99, 230-232, 1999.
9. J. Ribelles, A. López, I. Remolar, O. Belmonte, M. Chover. Multiresolution Modelling of Polygonal Surface Meshes Using Triangle Fans. Proc.of 9th DGCI 2000, 431-442, 2000.
10. J. Ribelles, A. López, Ó. Belmonte, I. Remolar, M. Chover, Multiresolution modeling of arbitrary polygonal surfaces: a characterization, Computers & Graphics, vol. 26, n.3 2002.
11. Michael Shafae, Renato Pajarola. DStrips: Dynamic Triangle Strips for Real-Time Mesh Simplification and Rendering. Proceedings Pacific Graphics Conference, 2003.
12. A. James Stewart: Tunneling for Triangle Strips in Continuous Level of Detail Meshes. Graphics Interface 2001: 91-100.
13. L. Velho, L.H. de Figueiredo, and J. Gomes.: Hierarchical Generalized Triangle Strips. The Visual Computer, 15(1):21-35, 1999.
14. A. Bogomjakov, C. Gostman.: Universal Rendering Sequences for Transparent Vertex Caching of Progressive Meshes. Proceedings of Graphics Interface 2001.
15. Leif P. Kobbelt, Thilo Bareuther. Hans-Peter Seidel.: Multiresolution Shape Deformations for Meshes with Dynamic Vertex Connectivity. Computer Graphics Forum. vol. 19, 2000.

Declarative Specification of Ambiance in VRML Landscapes

Vincent Jolivet, Dimitri Plemenos, and Patrick Poulingeas

Laboratoire MSI.
83, rue d'Isle. 87000 Limoges. France.
{jolivet,plemenos}@unilim.fr
poulinge@msi.unilim.fr

Abstract. Nowadays, VRML has grown to a Web3D standard. But there is still a lack of high-level tools to help a VRML designer in his/her conception of a virtual world (in the same way as declarative modellers make easier the conception than classical modellers). The purpose of our work is to allow a VRML designer to specify an ambiance in a declarative manner. The method described in this paper takes into account imprecision and uncertainty in ambiance descriptions with the use of fuzzy subsets theory. With this framework, the designer can introduce some modifiers (like: "very little" or "extremely") in his/her objectives in order to refine the wished ambiance. Two characteristics have been studied: the foggy aspect and natural lighting. Natural lighting is introduced with the aim of a natural description: the lived time.

1 Introduction

VRML [12] has become a popular standard for building virtual worlds. Its success is likely due to its open text-based and powerful format. Unlike other 3D scene files, VRML files can be treated with simple text-processing tools (such as Unix shell commands) or any script language (such as Perl or Python). Moreover, VRML allows the scene designer to model his virtual world with high-level components (called "nodes" in VRML) ordered in a scene graph. VRML users don't have to bother with a low-level description model such as classical B-rep [7]. Since VRML97 standard, some nodes can even add a dynamic aspect to artificial worlds (reactions to user's actions, temporal changes in the scene, scripts written in ECMAScript or Java, etc.)

There are two kinds of tools useful for a VRML designer in his conception work:

1. Geometric modellers (for the static aspect of a world).
2. Behavioral modellers (for the dynamic aspect of a world).

Our aim is to propose an improvement for geometric modellers in the case where the virtual world is a landscape. Classical VRML geometric modellers do not allow to specify in an intuitive and declarative manner the ambiance wished for a scene. The user has to choose numerical values in order to obtain a wished property in an empirical manner. The purpose of this method is the one of declarative modelling [11,1]: give to the designer some means to create several scenes with a set of high-level properties. These properties are the foggy aspect and the natural lighting of the VRML landscape.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 115–122, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In section 2, declarative modelling will be presented and we will locate our work in its field. We will then introduce a process for managing declarative specifications of an ambiance in a landscape, and a new method for introducing uncertainty in descriptions will be presented. In section 3, we will explain how to process the density of fog in a scene with the paradigm developed in the previous section. Then, in section 4, the natural lighting of a scene will be studied. A discussion on the drawbacks and the future developments of our works will be presented in section 5, and we will finally conclude in section 6.

2 Declarative Modelling and Declarative Specifications with Fuzzy Subsets Theory

Declarative modelling [11][1] is a useful technique allowing the designer to describe in an intuitive manner his/her idea of a scene. This kind of high-level modellers generates then the scenes corresponding to the eventually imprecise description given by the user. A lot of work usually done by the designer is then done by the declarative modeller; so the designer can spend more time on the conception phase.

There are two kinds of declarative modellers: general purpose declarative modellers which can build almost every scene, and dedicated declarative modellers which have a specific modelling domain. As we are studying ambiance in a VRML landscape, our researches belong to the area of dedicated declarative modellers.

[3][4][5][6] have proposed to represent a property not just as a standard interval, but as a fuzzy interval. This is to manage the imprecision of words like “big” or “weak” in declarative specifications given by the user to a declarative modeller. [3] introduced a taxonomy for properties and some types of linguistic modifiers that can be applied to basic properties.

A new type of linguistic modifiers will be introduced here. They allow the introduction of uncertainty in descriptions. The designer will be able to use such descriptions: “It's rather certain that the weather is foggy”.

2.1 Simple Properties

[3] has started with a simple property associated to a concept (e.g. the property “foggy” associated to the concept “weather” in the following description: “The weather is foggy”). A simple property is represented by a fuzzy subset of the domain D associated to its concept, i.e. a membership function f. To define this membership function, [3] uses a quadruple (α,a,b,β) and two functions L and R called form functions.

To check a property, the α-support set A_v associated to f must be computed (v is a fixed threshold value and $v \in]0,1]$). The α-support set A_v is given by the formula:

$$A_v = [a-\alpha L^{-1}(v) ; b+\beta R^{-1}(v)]. \quad (1)$$

To easily calculate the inverse functions of L and R, in [8], L and R are trapezoidal functions. To generate numeric values for a property, a discretization of the α-support set is finally done.

2.2 Parametric Properties

[3] introduced parametric properties, properties depending on a numerical value (e.g. "The level of snow is 2 meters"). There is still a little vagueness with these properties, so the previous fuzzy representation based on L-R membership functions can be applied to them.

2.3 Modifiers

Modifiers are operators that act on simple properties. They produce a new simple property, and they sometimes can be applied once again. The following set of modifiers has been selected: {"extremely little", "very little", "rather little", ∅, "rather", "very", "extremely"}. ∅ means that no modifier is used. With the simple property: "The weather is foggy", it's then possible to build a new property with a modifier (e.g. "The weather is *very* foggy").

Modifiers can only be used with simple properties and not with parametric properties.

[3] and [4] have developed some techniques to determine the membership function f' associated to a property like "x is *m* P" where m is a modifier and P a simple property. The function f' depends on the membership function f associated to P and the modifier m, but also on other semantic parameters of the property P.

In practice, the application of a modifier makes a translation and a contraction (or a dilatation) on the membership function. For our work, the most important thing is the facility (or difficulty) to compute quickly an α-support set with the new membership function f'. The membership function f' associated to the property mP and the α-support set are given in [8].

2.4 Fuzzy Operators

Fuzzy operators act both on simple and parametric properties. They produce a new property, and they can't be applied once again. The following set of fuzzy operators have been chosen: {"exactly", "really", ∅, "neighboring", "more or less", "vaguely"}. ∅ means that no modifier is used. With the parametric property: "The visibility limit is 70 meters", it is then possible to build a new property with a fuzzy operator (e.g. "The visibility limit is *really* 70 meters").

Let's consider a property like "X is o P" where o is a fuzzy operator. A contraction (or a dilatation) will be applied to the membership function f of P. However, the kernel of the membership function will not change.

A fuzzy operator is characterized by a fuzzy coefficient $j \in \mathbf{R}^{+*}$ which does not depend on the semantic of the property P.

The membership function f' representing the property oP and its associated α-support set is given in [8].

2.5 Uncertainty in Descriptions

The previous works have focused on the management of imprecision in declarative descriptions. We will now introduce a new kind of descriptions: descriptions with uncertainty. These descriptions take the form of: "It's *rather* certain that the size of X is 3 meters". The word "rather" acts as an uncertainty operator which increases or decreases the uncertainty of the proposition.

We suggest the following set of uncertainty operators: {∅, "rather", "enough", "little", "very little"}. ∅ means that the description is certain (but there is still a degree of imprecision). Uncertainty operators can be used with every type of properties.

To model a proposition with an uncertainty, fuzzy intervals can still be used. The action of an uncertainty operator on the membership function will be a reduction of the size of the kernel and a contraction of the function.

An uncertainty operator is characterized by a couple of reals $(j,k) \in]0,1] \times \mathbf{R}^{+}$. j acts on the form of the new membership function and k on the size of the kernel. The case j=1 corresponds to the operator ∅.

The parameters of the new membership function f' are given by the formulas:

$$\alpha' = \begin{cases} \alpha\left(1 + \frac{c}{j}\right) & if \quad j \in]0,1[\\ \alpha & if \quad j = 1 \end{cases}$$

$$\beta' = \begin{cases} \beta\left(1 + \frac{c}{j}\right) & if \quad j \in]0,1[\\ \beta & if \quad j = 1 \end{cases}$$

$$a' = a + \frac{kc(b-a)}{2}$$

$$b' = b - \frac{kc(b-a)}{2}$$

$$t' = t$$

$$L' = L^{j}$$

$$R' = R^{j}$$

Remark: In fuzzy sets theory, the usual value for the coefficient c is 10%.

As the form functions L and R have changed, the α-support set A_v associated to the new function f' will be:

$$A_v = [a' - \alpha' L'^{-1}(v) \; ; \; b' + \beta' R'^{-1}(v)] \tag{2}$$

with:

$$L'^{-1}(v) = R'^{-1}(v) = 1 - v^{1/j} \tag{3}$$

3 Foggy Scenes

VRML allows us to change the foggy aspect of a scene with the Fog node [12]. This node has a visibility field which determines the visibility limit. Two types of properties are proposed to the designer: a simple property and a parametric property.

The simple property is expressed as: “The weather is foggy”. The parametric property introduces the notion of visibility limit. These parametric properties can be expressed as: “The visibility limit is 50 meters”.

For every property, a solution interval is computed using the method exposed in section 3, with $\alpha=\beta=10$. Wishing to center the segment [a,b] on the value d specified in a parametric property (like “The visibility limit is d meters”), a heuristic has been introduced.

The values a = d-e and b = d+e have been chosen, with:

$$e=\frac{M-m}{10} \tag{4}$$

where: [m,M] is the domain of the membership function.

In figures 1 and 2 are shown some results found by our method.

First, we will show how an operator can modify a property. The processed parametric property is: “The visibility limit is 70 meters”, and we will then apply the fuzzy operator “exactly”. There are several solutions for each description. The minimum of the interval found has been chosen to generate the pictures in figure 1.

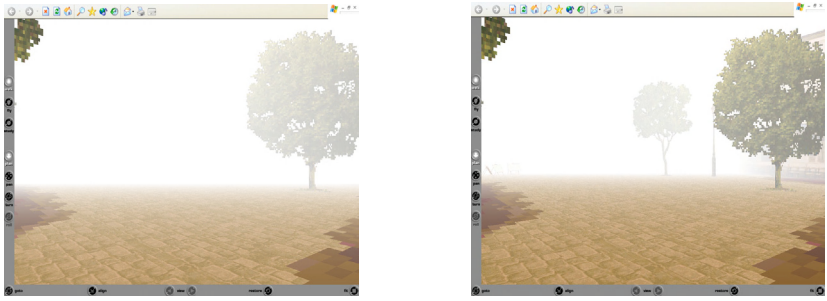


Fig. 1. First image: “The visibility limit is 70 meters”. Second image: “The visibility limit is exactly 70 meters”

For a property, our method generates a lot of solutions (depending on the discretization). The scenes in figure 2 correspond to the extreme values computed for the same description as in figure 1. The solution interval computed is: [50,90].

Remark:

The scenes generated by the discretization of the interval corresponding to the user's description are often very similar. It seems not obvious to find the fine parameters for the membership function in order to produce very dissimilar scenes. A solution to this problem would be a tool for classifying the different scenes. This tool could select only a few scenes with a significant difference between them.

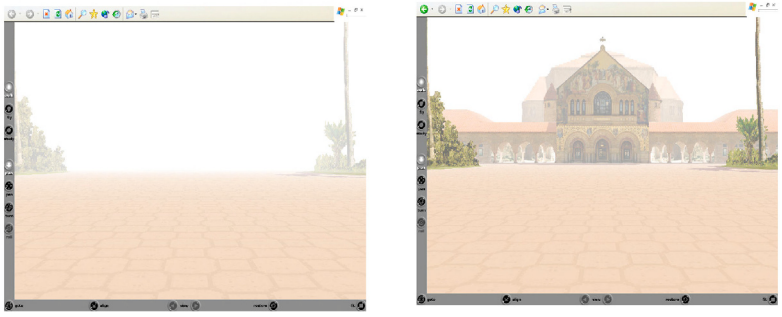


Fig. 2. The first image comes from the first scene found for the property: “The visibility limit is 70 meters”. The second image comes from the last scene found for the same property

4 Natural Lighting in a Scene

To model natural lighting, VRML proposes a node called DirectionalLight. This node allows the designer to introduce a directional light in his/her scene, light corresponding to the sun. Our method can find the sun direction satisfying a declarative specification of sunlighting. For coherence, our method changes the skyAngle and skyColor fields of the Background node to create an appropriate sky.

The sunlighting property is expressed with the concept of lived time introduced by [13]. This concept allows the use of most of semantic descriptions of time intervals. The designer can specify sunlighting with terms like: “the end of morning in spring”.

[14] proposes some formulas (coming from astronomical studies) to translate these descriptions into numerical intervals giving the direction of the sun (The direction is specified with the azimuthal angle A and the height H of the sun, as shown in figure 3). These intervals correspond to a part of the surrounding hemisphere containing the entire VRML world. Figure 3 shows the part of the hemisphere for a lived time T (the point p is the centre of the VRML world in our case).

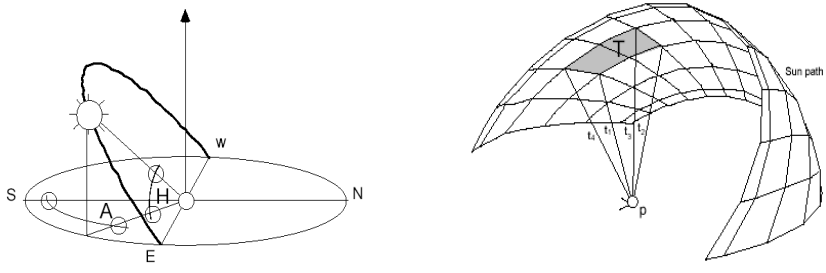


Fig. 3. Left image: Determination of the position of the sun [14]. Right image: Intervals describing the lived time [13]

The possibilities of an imprecise description of the time come from the composition of an element of {“beginning”, “the middle of”, “the end of”} with an element of {“the day”, “the morning”, “the afternoon”}, from the special word

"midday" and from the direct specification of the time. The possibilities of an imprecise description of a day come from the composition of an element of {"beginning", "the middle of", "the end of"} with a month, from the seasons: {"Summer", etc.} and from the special words: {"equinoxes", "solstices"}.

In figure 4 are shown scenes with different kinds of natural lighting of the same world. The minima of the intervals corresponding to the description of the lived time have been taken for generating the following pictures:

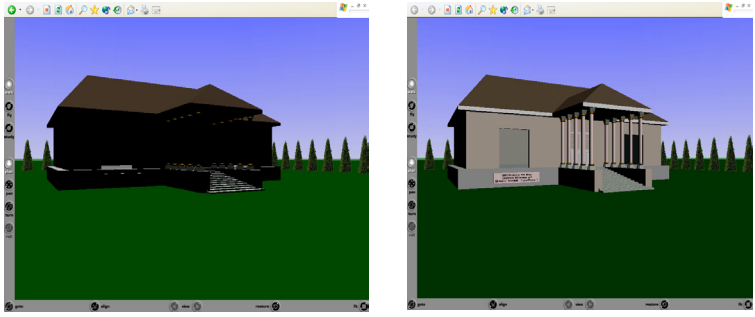


Fig. 4. Examples of natural lighting: "midday" (left) and "the middle of the afternoon" (right)

5 Future Researches

Our method presents several drawbacks most due to VRML limitations. The difficulties resulting from VRML are:

- The absence of shadows, which damages the realism of solar lighting.
- The lighting included in some textures may be inconsistent when the solar direction changes.
- The impossibility to take into account some complex physical parameters influencing the atmosphere in a landscape. This impossibility comes from the VRML format limitations and the lack of control in the rendering process.

To overcome these drawbacks, there are several possibilities:

- The rendering of VRML viewers is often based on OpenGL. The new version of OpenGL (OpenGL 1.5 [10]) will introduce shadow functions. So, in the future, VRML may be able to manage shadows in a scene.
- Lighting in textures can be recovered and changed with the method exposed in [9]. This method may be used in our framework as a pre-process stage and with a VRML parser (to locate the texture fields in the VRML file).
- The emergence of X3D will normally offer a better rendering than VRML. The extensions of VRML proposed by Contact [2] seem to be very useful too.

In the future, we will try to develop the scene understanding phase (which is still reduced to the simple display of the VRML world). A classification of the different scenes obtained from a description would be a powerful tool. It could avoid generating a lot of similar scenes with the presentation of some reference scenes.

6 Conclusion

In this paper, a tool for creating ambiance effects in VRML landscapes has been proposed. The spirit of this tool follows the ideas of declarative modelling: the designer can specify in a declarative manner his/her wishes with high-level properties.

To model and manipulate these properties, a fuzzy subsets-based representation has been introduced. This representation can take into account the necessary imprecision of the specifications and can manage uncertainty in the user's propositions. One of the key concepts is to introduce several modifiers and operators that can build different properties starting from a basic property.

Two aspects of VRML worlds have been studied: foggy appearance and sunlighting. The first one is a direct application of the concepts and ideas associated to the fuzzy subset representation of a property. The second aspect uses a natural description of sunlighting: the lived time.

The lack of realism in VRML scenes limits dramatically the extension of our method to other ambiance effects. Some extensions of VRML like [2] could allow us to treat other ambiance properties in a similar context.

References

1. P.-F. Bonnefoi. Techniques de Satisfaction de Contraintes pour la Modélisation Déclarative. Application à la Génération Concurrente de Scènes. Ph.D. Thesis (in french). Université de Limoges, June 1999.
2. Contact, a VRML plug-in of Blaxxun Technologies. http://developer.blaxxun.com
3. E. Desmontils. Formalisation des Propriétés en Modélisation déclarative à l'Aide des Ensembles flous. Rapport de Recherche IRIN 106. Nantes, December 1995.
4. E. Desmontils, D. Pacholczyk. A qualitative Approach to fuzzy Properties in Scene Description. CISST'97. Las Vegas, USA, 1997, p.139-148.
5. E. Desmontils, J.-Y. Martin. Properties Taxonomy in Declarative Modeling. CISST'97. Las Vegas, USA, 1997, p.130-138.
6. E. Desmontils. Expressing Constraint Satisfaction Problems in Declarative Modeling using natural Language and fuzzy Sets. Computer & Graphics 24 (2000). p.555-568.
7. J. Foley, A. van Dam, S. Feiner, J. Hughes. Introduction to Computer Graphics. Addison Wesley Longman, Inc, 1994 .
8. V. Jolivet, D. Plemenos, P. Poulingeas. Declarative Approach of Inverse Direct Lighting Problems. 3IA'2002. Limoges, 14-15 of May, 2002.
9. S. Marschner, D. Greenberg. Inverse Lighting for Photography. IS&T/SID Fifth Color Imaging Conference. November 1997.
10. OpenGL 1.5 Specifications: http://www.opengl.org
11. D. Plemenos. Declarative Modelling by Hierarchical Decomposition.The Actual State of the MultiFormes Project. GraphiCon'95, St Petersbourg, 1-5 of July 1995.
12. VRML 97. ISO/IEC 14772:1997. http://www.web3d.org
13. D. Siret. Sunlighting Design: an Inverse Approach of Simulation for CAD Tools. Advances in Computer-Aided-Design, CADEX'96. Hagenberg, September 1996.
14. D. Siret. Propositions pour une Approche Déclarative des Ambiances dans le Projet Architectural. Application à l'Ensoleillement. Thèse de Doctorat. Université de Nantes, June 1997.

Using Constraints in Delaunay and Greedy Triangulation for Contour Lines Improvement

Ivana Kolingerová[1], Václav Strych[2], and Václav Čada[2]

[1]Department of Computer Science and Engineering
[2]Department of Mathematics
University of West Bohemia, Plzeň, Czech Republic
kolinger@kiv.zcu.cz, http://iason.zcu.cz/~kolinger

Abstract. Automatic computation of contour lines on a triangulation is a difficult problem because due to input data digitization and the type of triangulation used, some triangles can be a source of a strange behaviour of the contour lines. In this paper, we show what problems can appear in contour lines when Delaunay or greedy triangulations are used and how the contour lines can be improved using constraints in the triangulation. We improved contour lines by manually imposing constraints in a triangulation editing program. Automation of this process is a next step of our work.

1 Introduction

A computation of contour lines on a triangulation is a necessary part of GIS programs. However, automatically obtained contours are rarely satisfactory. For an experienced expert, it is easy to recognize the problems and to correct them manually, but at present it is nearly impossible to make this recognition and correction automatic.

This paper makes one step to a future fully automatic solution: it shows results of using constraints – constrained edges prescribed into the triangulation on the places where contour lines computed on the original triangulation are not satisfactory. Detection of these places is manual, local recomputation of triangulation and of contour lines is automatic.

Section 2 describes background. Section 3 shows main problems in contour lines on Delaunay and greedy triangulations. Section 4 shows how to improve the contour lines using constraints, Section 5 presents results, Section 6 concludes the paper.

2 Triangulations and Contour Lines Computation

Def. 1. A triangulation. A triangulation T(P) of a set P of N points in the Euclidean plane is a set of edges E such that

- no two edges in E intersect at a point not in P
- the edges in E divide the convex hull of P into triangles.

There exist many types of triangulations. The most popular one - due to good properties of its triangles and simple and well-studied computation, usually in $O(N)$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 123–130, 2004.
© Springer-Verlag Berlin Heidelberg 2004

expected time and $O(N \log N)$ worst case – is the Delaunay triangulation. Another well-known triangulation is the greedy triangulation.

Def. 2. Delaunay triangulation (DT). The triangulation DT(P) of a set of points P in the plane is a Delaunay triangulation of P if and only if the circumcircle of any triangle of DT(P) does not contain a point of P in its interior.

Delaunay triangulation ensures maximizing minimum angle of each triangle as well as of the whole triangulation. Therefore, it produces the most equiangular triangulation of all. Other positive features exist, as described in [2, 3, 9]. Algorithms for DT computation can be found in [6, 12, 13].

Def. 3. Greedy triangulation (GT). The triangulation GT(P) of a set of points P in the plane is a greedy triangulation if it consists of the shortest possible compatible edges where a compatible edge is defined to be an edge that crosses none of those triangulation edges which are shorter than this edge.

The main problem connected to GT is its time complexity. It can be computed in $O(N^2 \log N)$ time using demanding algorithms. $O(N)$ expected time solution is possible only for uniform data. Effective algorithms can be found in [4, 8, 10].

DT and GT use for computation only planar information given by x and y coordinates of the input points, however, resulting meshes are of good quality, with one exception: a terrain with a steep slope. Such a kind of data needs a special triangulation, taking into consideration also heights of points (z-coordinates), angles between triangle normals, etc. These triangulations are called **data dependent triangulations (DDT)** and were established in [5].

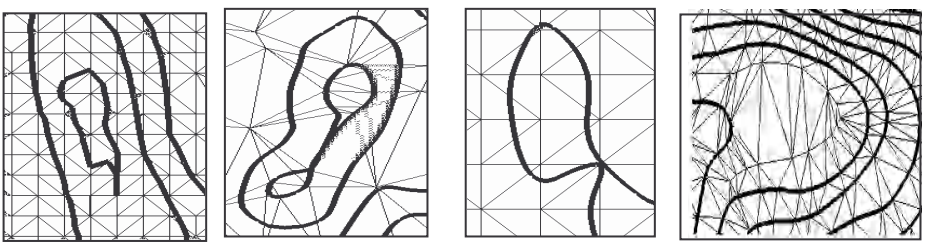
One more concept used in triangulations is a **constrained edge** (a **constraint**). It is an edge prescribed to be used in a triangulation. This approach is useful to adjust a triangulation to some domain limitations, such as a break, a polygonal boundary, etc. DT as defined in Def. 2 is not able to consider constraints - its definition has to be modified: in the **constrained Delaunay triangulation (CDT)**, only non-constrained edges are checked on the empty circumcircle property. For the CDT algorithms see [11, 1]. It is very easy to use constrained edges in GT: such edges are accepted into the triangulation first, then a usual greedy construction continues.

There exist more possible algorithms how to compute **contour lines** on a triangulation, e.g. [7]. As the approach described in this paper is independent of the contour line computation method, we will not explain details. Generally, we are looking for intersections of the triangulation with the height range $<z_{min};z_{max}>$ on which the contour lines are to be computed. Contour lines computed on triangulation are piecewise linear - polylines, later they are usually smoothed out by an interpolation. For one particular $z=z_h$, the contour lines may be formed by open segments – polylines starting and ending at the triangulation boundaries, and by closed segments - polylines surrounding some extreme sites in the terrain model.

3 Main Problems in Contour Lines on DT and GT

One of the main criteria when producing triangulations for contour lines is avoiding too long and skinny triangles. As GT does not consider shape and angles of the triangles, we expected problems with the contour lines computed on this triangulation. Surprisingly, we came across the situations where the contour lines on GT were better than on DT. Now we will present the problems which we identified in the countour lines computed on DT. We should note that for more pleasant figures, the contour lines have been smoothed by splines, however, the described effects are present also on non-smoothed contour lines, so they are not caused by the interpolation.

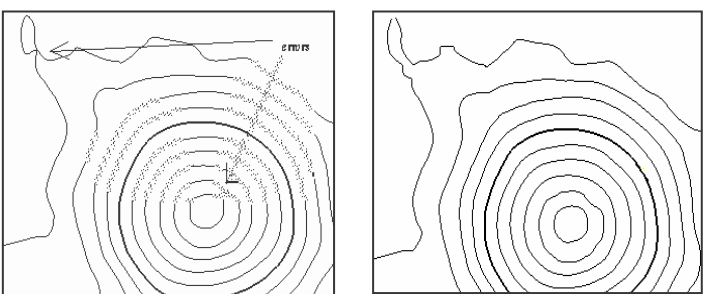
While testing DT, we found the following types of errors in contour lines. Fig. 1a) shows the situation where the contour lines lie on one edge of the triangle and then return back on the same edge. Fig. 1b) shows bad contour lines where two tops of hills were connected into one contour line. Fig. 1c) shows a problem of „too close lines". An interesting type of detected errors is in Fig. 1d). This error is usually caused by the triangles which have all their vertices in the same height.



a) Identical parts b) Wrong connection c) Improper closeness d) An error obtained due to horizontal triangles

Fig. 1. Errors detected on contour lines when using DT

Fig. 2 shows contour lines generated from a regular grid. It can be seen that in case of DT there appear two problematic places which are correct when using GT. Fig. 3a) documents again a problem with nearly horizontal triangles. Contour lines on GT in Fig. 3b) are a bit better but not completely correct, either.



a) DT used b) GT used

Fig. 2. Comparison of contour lines on DT and GT

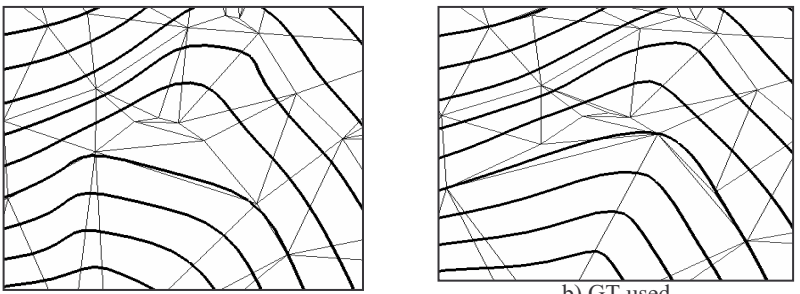


Fig. 3. Another problematic situation with nearly horizontal triangles

According to these two cases, it might seem that GT enables to create better contour lines. Not each time this is true. In most cases, contour lines are similar in both triangulations. There appears a problem for GT if a narrow triangle is created. In such a case the contour lines twist unpleasantly, see Fig. 4b), or break, see Fig. 4d), Figs. 4a), 4c) show DT for comparison.

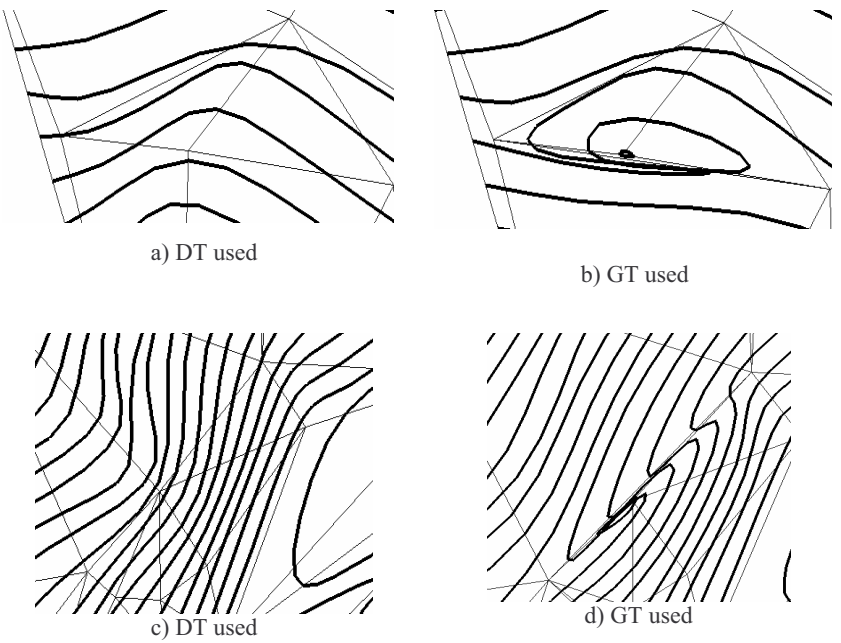


Fig. 4. Also GT may cause contour line problems: narrow triangles and breaks

If we want to evaluate the triangulations according to their handling of problem places, it is necessary to say that GT creates less of such errors. The reason is that DT triangles are usually oriented by their longer edges against the slope direction. It causes a creation of horizontal or nearly horizontal triangles. In GT, some triangles are oriented by their longer edges into all possible directions and vertices with different

height can be connected. It prevents from horizontal triangles creation. This is not the only problem that may appear but it is the most visible one. This visibility is on one hand a disadvantage because several such places cause bad appearance of the whole map; on the other hand, it allows to detect such places easily and to edit them. Skinny triangles, more often appearing on GT than DT, have the opposite property; they are not that much visible at first sight, and, therefore, also more difficult to find, and, what is more important, they are more often than horizontal triangles.

For comparison, Fig. 5-7 show contour lines computed by the Erdas Imaging software, by DT and GT in our implementations. It can be seen that even a professional software sometimes produces problems, see a 'tooth' on Fig. 5. Boundaries of Figs. 6 and 7 are wrong due to convex hulls in the triangulation. For good contour lines, convex hulls should be avoided and replaced by domain boundary preservation over constraints. If we do not take into account boundaries, the contour lines obtained on DT are comparative to the professional software and they do not contain the 'tooth effect'. GT results are worse – there are two places where thin triangles spoil the shape of contour lines.

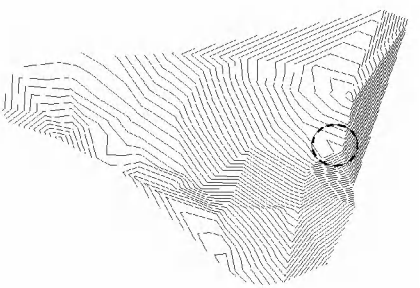


Fig. 5. Erdas

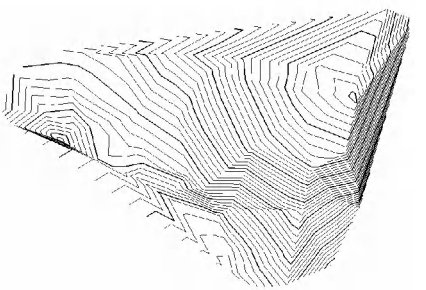


Fig. 6. DT

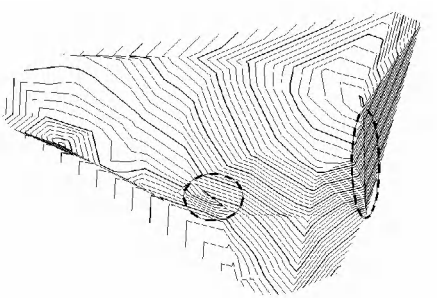


Fig. 7. GT

As according to our experiments, neither GT nor DT can provide perfect contour lines, DT should be recommended as a better choice as its computation is cheaper and

some of problems on contour lines are less often and easier to be corrected. However, it is visible that none of the considered triangulations is perfect; therefore, other ways to improve the contour lines have to be searched. Next section will present how the contour lines quality can be improved using constrained edges.

4 Improving Contour Cines by Constraints

Let us recall that constrained edges (or constraints) are edges which are prescribed in advance to be used within a triangulation. They are necessary to achieve an outer shape of the triangulated area different from the convex hull of the triangulated points or to include some breaks or ridges into the triangulation. We will show now that such 'enforced' edges are also important to increase a quality of contour lines.

The constraints are to be used to change the triangulation on the places where contour lines are incorrect or improper. Such places must be identified manually by inspecting the (automatically computed) contour lines. It would be also possible to detect wrong contour lines automatically by a set of general criteria, however, improper places are for experienced cartographic experts easy to be seen but difficult to be quantified, and formulation of criteria for automatic processing needs quantification. This is still an open problem for our future contour lines research. Therefore, our solution uses an editing program which displays the triangulation together with contour lines computed on it. The user – a cartographic expert – inspects the contour lines and if he detects some problems as described in Section 3, he may prescribe a constrained edge between a pair of triangulation vertices.

Insertion of a constrained edge means that the triangulation edges intersecting the newly inserted edge have to be deleted and the hole in the triangulation retriangulated. Then, in case of DT, the newly obtained triangles have to be checked whether they hold empty circumcircle property; eventual changes may spread into the whole triangulation, although usually they are only local. After the triangulation changes, also contour lines on changed triangles have to be recomputed.

A fundamental step for constraints insertion is an edge flip. Let us suppose we have two triangles ($v_1v_2v_3$), (v_2,v_4,v_3) sharing an edge (v_2v_3) and we need to replace this edge by the constraint v_1v_4. If ($v_1v_2v_4v_3$) is a convex quadrangle, the flip is possible, and vice versa. This condition ensures that the new edge will not intersect other edges, not participating in the flip operation.

Retriangulation of the area intersected by a constrained edge according to [11] has steps as follows (see also an example in Fig.8):

1. Find a triangle where the constrained edge starts (the triangle which has one of the endpoints of the constrained edge as its vertex and at the same time is intersected by the constrained edge).
2. By walking from the starting triangle over its neighbouring triangles, find all edges intersecting the constrained edge and store them in a queue.
3. Loop until the queue is empty: pull one edge out of the queue, if the edge is a diagonal of a convex quadrangle, flip the edge. If the quadrangle is not convex, return the non-flipped edge into the queue to be solved later.

In case of CDT:

4. Insert all newly made edges except the constrained one into another queue.
5. Loop until the queue is empty: pull out an edge and check it on empty circumcircle criterion. If the criterion does not hold, flip the edge and insert the newly created edge into the queue.

After retriangulation, the contour lines for $<z_1;z_2>$ has to be recomputed where z_1 and z_2 are the minimum and the maximum heights on the triangles changed in retriangulation ($z_{min} \leq z_1 \leq z_2 \leq z_{max}$).

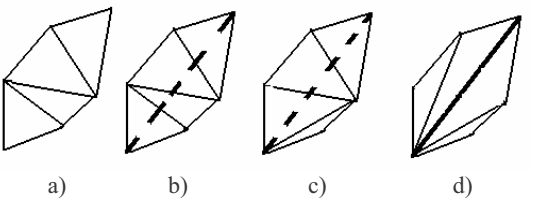


Fig. 8. Retriangulation, a) The original triangulation, b) The inserted constrained edge is dashed, c) After one edge flip, d) The resulting triangulation

5 Experiments and Results

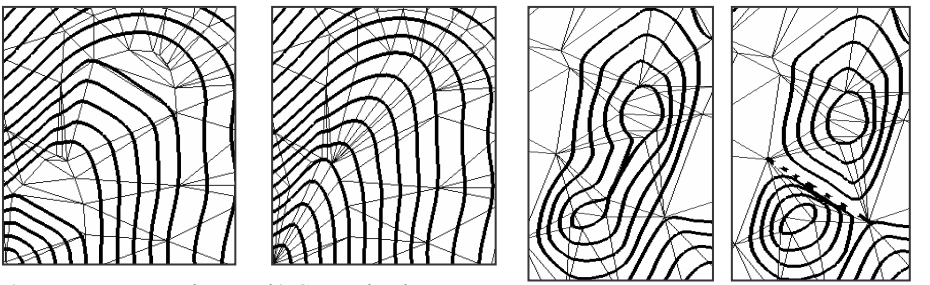
The described methods – DT, GT, CDT, contour lines computation and triangulation editor - were implemented as a set of programs in Delphi under Windows NT. The triangulation editor, besides others, allows inserting constraints on user request, as already described. Our DT and CDT implementation is of $O(N)$ expected time and memory complexity and allows to compute triangulations up to about 1mil. of points. As our GT implementation uses only brute force algorithm with $O(N^3)$ time and $O(N^2)$ memory complexity, we are able to test data sets up to 2000 points in GT. However, such size of data sets was enough to detect potential differences against DT results.

The use of constrained edges in DT as well as in GT brings a substantial improvement of contour lines shapes in critical places. See examples: Fig. 9a) shows two problems in the contour lines. Fig. 9b) shows correction of these situations thanks to the constrained edges. Fig. 9c) shows two wrongly connected hills, Fig. 9d) indicates correction of the situation by a constrained edge.

In the future, we would like to detect the problems in a triangulation automatically, so that the contour lines were computed on a mesh already optimized by constraints, without the need of manual corrections and of recomputation.

6 Conclusion

While inspecting the DT, GT and contour lines computed on them, we came to conclusion that GT brings more problems than advantages (time and memory complexity, skinny triangles) and, therefore, we recommend to use DT as a triangulation for contour lines computation. However, constrained edges are necessary to obtain satisfactory results. So far, we have to find problematic places manually, then insert constrained edges, recompute the attached part of the triangulation and of the contour lines. Automation of the contour lines improvements is our next research goal.



a) Two erroneous places in contour lines

b) Correction by constrained edges

c) Incorrect connection of two hills

d) Correction by a constrained edge insertion

Fig. 9. Two errors in contour lines and their correction with the use of constrained edges

References

1. Anglada, M.V. : An Improved Incremental Algorithm for Constructing Restricted Delaunay Triangulations, Computers & Graphics 21 (1997) 215-223
2. Aurenhammer, F.: Voronoi Diagrams - a Survey of a Fundamental Geometric Data Structure, ACM Computing Survey, Vol.23, No.3 (1991) 345-405
3. de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.: Computational Geometry. Algorithms and Applications, Springer-Verlag (1997)
4. Dickerson, M.T., Drysdale, R.L.S., McElfresh, S.A, Welzl, E.: Fast Greedy Triangulation Algorithms, Proc.10th Annual Symp. on Comp.Geom., ACM (1994) 211-220
5. Dyn, N., Levin, D., Rippa, S.: Data Dependent Triangulations for Piecewisse Linear Interpolation, IMA Journal of Numerical Analysis 10 (1990) 137-154
6. Kolingerová, I., Žalik, B.: Improvements to Randomized Incremental Delaunay Insertion, Computers & Graphics 26 (2002) 477-490
7. Krcho, J.: Modelling of Georelief and Its Geometrical Structure Using DTM: Positional and Numerical Accuracy (In Slovak), Georeliéf a geologické procesy, Q111, Bratislava (2001) 269-574
8. Levcopoulos, C., Lingas, A.: Fast Algorithms for Greedy Triangulation, BIT 32 (1992) 280-296
9. Okabe A., Boots B., Sugihara K.: Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, John Wiley & Sons, Chichester New York Brisbane Toronto Singapore (1992)
10. Preparata, F.P., Shamos, M.I.: Computational Geometry: An Introduction, Springer Verlag New York Berlin Heidelberg Tokyo (1985)
11. Sloan, S.W.: A Fast Algorithm for Generating Constrained Delaunay Triangulations, Computers & Structures 47 (1993) 441-450
12. Su P., Drysdale R.L.S.: A Comparison of Sequential Delaunay Triangulation Algorithms, In: Proc. 11-th Annual Symp. on Comp.Geom., ACM (1995), 61-70
13. Žalik, B., Kolingerová, I.: An Incremental Construction Algorithm for Delaunay Triangulation Using the Nearest-point Paradigm, Int. J. Geographical Information Science 17 (2003) 119-138

An Effective Modeling of Single Cores Prostheses Using Geometric Techniques

Kwan-Hee Yoo[1] and Jong Sung Ha[2]

[1] Dept. of Computer Education and Dept. of Information Industrial Engineering, Chungbuk National University, 48 San Gaesin-Dong Heungduk-Gu Cheongju Chungbuk 361-763 Republic of Korea
khyoo@cbucc.chungbuk.ac.kr

[2] Dept. of Computer Engineering, Woosuk University, 490 Hujongri, Samrae-Up Wanju-Kun Chonbuk 565-701 Republic of Korea
jsha@woosuk.ac.kr

Abstract. There has been a great demand for the artificial teeth prostheses that are made of materials sintered at 1500^oC, such as Zirconia ceramic. It is, however, very difficult for dental technicians to manually fabricate the prostheses with such materials since their degree of strength is very high. For that reason, the prostheses of strong materials have been fabricated by using CNC (computer numerical control) machines. The most important subject in the CNC fabrication is how accurately the prostheses can be modeled in three-dimensional space according to dentists' requirements. In this paper, we propose effective methods for modeling single cores, such as single caps and Conus cores, which are the principal ones of the artificial teeth prostheses. Our approach employs the 2D Minkowski sum, a developed ZMap algorithm, and other geometric techniques such as the tiling of contours. We also present and analyze the 3D visual examples of the single cores modeled by the proposed methods.

1 Introduction

Computers have been diversely applied to the area of dental surgery. Some examples are the digitalization of dental data, the 3D visualization of dental information, the automatic fabrication of artificial teeth using the CAD/CAM, and the measurement of accuracy during impression or die castings [2,3,5,9,10,14,15]. The reconstruction of artificial teeth prostheses is one of the most important processes in dental treatment. Most of the prostheses are still fabricated manually by dental technicians, but the manual labor causes various problems. The accuracy of the prostheses depends on the skills of dental technicians. Sometimes the prostheses cannot be fabricated within a limited time since the manual fabrication takes long time. Furthermore, there has lately been an increasing demand for prostheses that are made of materials sintered at 1500^oC, such as Zirconia ceramic, which cannot be handled by human hands. In order to resolve these problems, CAD/CAM techniques began to be applied to the fabrication systems of the prostheses [5,15].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 131–138, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In general, the fabrication of artificial teeth prostheses is composed of four successive processes. First, a plaster cast is fabricated to make a pattern for a patient's teeth. This process is the same as the one that has been performed in the previous dental surgery, while the following other processes can be performed by using CAD/CAM techniques. Second, the 3D surface of the teeth is modeled by scanning the fabricated plaster cast with 3D scanners. Third, exploiting the 3D information of the teeth model, they design the 3D models that exactly correspond to the prostheses satisfying dentists' requirements. Finally, the designed models of prostheses can be fabricated by cutting ceramic materials with milling machines. When artificial teeth prostheses are fabricated, the most important subject is how accurately the prostheses can be modeled according to dentists' requirements. For the accurate modeling of the prostheses, we may be able to use the existing 3D modeling tools such as 3DMax, Maya, CATIA, and SolidWorks. But it may be impossible or it takes too long to model the prostheses with the existing tools so that the required conditions can be satisfied. Hence, we need to develop a dedicated 3D modeling tool for the efficient and accurate modeling of the prostheses.

Artificial teeth prostheses are composed of *core* prostheses and *crown* prostheses [13]. The crown prostheses are visible when the artificial teeth prostheses are put in. On the other hand, the core prostheses directly touch the original teeth ground by dentists, and the important function of them is to increase the adhesive strength between the original teeth and the crown prostheses. A *single core* means the prosthesis that is composed of one piece, and the typical examples are single caps and single Conus cores. In this paper, we first analyze the requirements that are necessary for modeling the prostheses of single cores, and then propose the methods for effectively modeling the single cores using geometric techniques [1,8] such as 2D Minkowski Sum, a developed ZMap algorithm, and the tiling of contours. We also provide and analyze the practical examples of the single cores modeled by using our techniques.

2 Requirements Analysis

If there are any problems with a tooth due to decay, non-vitality, etc., a dentist grinds the original tooth, designs artificial tooth prostheses, and then fabricates the prosthesis put in above the ground tooth called the *abutment*. When a dentist grinds the original tooth, she/he makes the shape of an abutment so that the artificial teeth prosthesis can normally be mounted on it. The prosthesis is modeled after the coping design so that it is secure and it adheres to the abutment. A result of the coping design is just the single cap. The surfaces in both sides of the single cap adhere to the abutment and the crown, respectively. The surface adhering to the abutment is called the *inner surface*, while the surface adhering to the crown is called the *outer surface*.

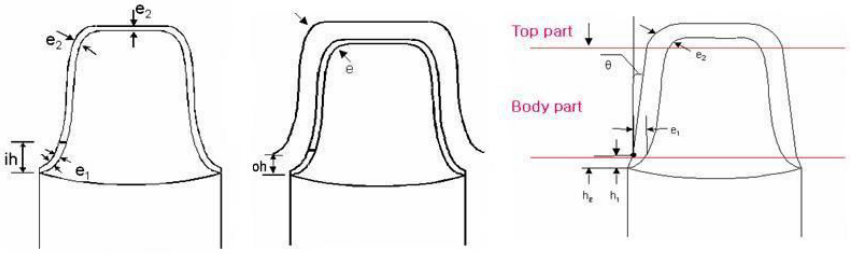
With the sectional diagrams of a single cap, we explain the requirements for single caps in the area of prosthodontics. Fig. 1(a) illustrates the values that are required for guaranteeing the adhesive strength of the inner surface of the single

cap. The inner surface is determined by expanding the surface of the abutment; the one part within a fixed height ih from the margin is expanded as much as a particular value e_1, while the other part is expanded as much as another particular value e_2. Fig. 1(b) illustrates the values that are required for the outer surface of the single cap. The outer surface is determined by expanding a part of the inner surface; the part above a fixed height oh from the margin is also expanded as much as a particular value e. The space between abutments may be wide when a large number of teeth are lost and only a few original teeth survive. The wide space causes the surviving original tooth to sustain a great load and the abutments are easily damaged. An alternative plan for resolving the problem was proposed by Korber [6], which is a method of dental treatment using the new prosthesis called the *Conus*. The Conus is designed to avoid the damage to the abutments and to sustain the load additionally with the mucosa of the mouth. The Conus core has an inner surface and an outer surface, as a single cap has the two surfaces. The inner surface should well adhere to the abutments. Hence, the inner surfaces of the Conus cores can be modeled by expanding the surface of the abutment the same as those of single caps were done: the one part within a fixed height h from the margin and the other parts expanded as much as particular values respectively. The outer surface of a single Conus core should be able to control the power sustaining the crown, and it has to be completely fixed at the time of the Conus's function. It also should be able to make the surviving teeth almost never get the stress. The values for satisfying these requirements are illustrated in Fig. 1(c), which is the sectional diagram of a Conus core. For determining the outer surface, the one part of the abutment within a fixed height h_1 from the margin is increasingly expanded to a particular value e_1, and then another part between the heights h_1 and h_2 from the margin is modeled as a conic type. The conic angle θ of the conic type is very important due to separation and adhesion between Conus cores and crown, which is recommended as 4^o, 6^o, and 9^o in the area of prosthodontics [13]. These two parts are defined as the *body part*, while the remaining part above the body part is defined as the *top part*. The outer surface corresponding to the top part is also determined by expanding the inner surface as much as a particular value e_2 similarly to the one of a single cap.

3 Modeling of Single Cores and Its Examples

3.1 The Inner Surface of Single Cores

For designing the inner surface of the single core for a given abutment, we have to first scan the given abutment impression, and then extract the tooth model lying above the margin from the scanned 3D data. The 3D scanners of the touch type were used for the scanning, and we choose a small value $0.05mm$ as input intervals on X- and Y-axes for guaranteeing the accuracy of the 3D data. In general, the tooth model has the same characteristics as a *terrain model* has since the abutment is scanned by the 3D scanner of touch type. Hence, the modeling of the inner surfaces of single cores can be transformed into the



(a) (b) (c)

Fig. 1. Designing requirements (a) the inner surfaces of a single core (b) the outer surfaces of a single cap (c) the outer surfaces of a single Conus cap

problem of expanding terrain models as much as given fixed values. In other words, the modeling can be formulated as the problem of Minkowski sum [1] in the plane.

MinkowskiSum$(\mathcal{T}, e) = \{\bar{v} + \bar{w} \parallel \bar{v} \in \mathcal{T},\ \bar{w} \in \mathcal{S}\}$, where $\mathcal{S}$ is an arbitrary sphere with radius e, whose center exists on the surface of $\mathcal{T}$. In our case of modeling the inner surfaces of a single core, e_1 is assigned to e if the height value for a point v is less than a given value h. Otherwise, we assign e_2 to e.

The Minkowski sum of a general model in 3D space can be solved by applying the results of Lozano-Perez [7] or Ghosh [4]. Lozano-Perez decomposes an original object into convex objects, and computes convex Minkowski sums for each pair of the convex objects. As a result, the original problem is represented with the union of all the convex Minkowski sums, but it is difficult to accurately implement the decomposition and union operations for this method. Assuming that the intersection of line/line, plane/line, and plane/plane can be exactly computed, Ghosh proposed an algorithm for computing the Minkowski sum of a general model by using the unions of face/face, face/edge, face/vertex and edge/edge Minkowski sums. This method is numerically error-prone when floating point arithmetic is used as well as it takes heavy load in the union operations.

In order to make the problems simpler, we review the characteristics of the tooth model and its inner surface in detail. In short, let $\mathcal{M}$ and $\mathcal{TM}$ represent a detected margin and an extracted tooth model lying above the margin respectively. For increasing the adhesive strength of a single core to be designed, its inner surface is determined by expanding the one part of $\mathcal{TM}$ within the height $0.5mm$ from $\mathcal{M}$ as much as $0.0120mm$, and expanding the other part as much as $0.0125mm$. From the observation that the expanding values are much smaller than the input interval $0.05mm$ on X- and Y-axes, we are going to solve the problem of modeling the inner surface by transforming it into the problem of 2D Minkowski sum [1]. Without the loss of generality, a given $\mathcal{TM}$ can be regarded as a terrain located at XY plane of the 3-axis coordinate system. After determining the minimum and maximum values of Y coordinates of $\mathcal{TM}$, we construct the 2D polygonal lines that have the X and Z coordinate values sequentially

in the input interval $0.05mm$ on Y-axis. Geometrically speaking, the polygonal lines are the intersection of the tooth surface and the planes that are parallel with XZ plane. Clearly, the end points of the 2D polygonal lines are on $\mathcal{M}$. Before applying the Minkowski sum to the constructed polygonal lines, a particular contour $\mathcal{C}$ of $\mathcal{TM}$ has to be computed to lie at a fixed height h from $\mathcal{M}$. We can obtain $\mathcal{C}$ by simultaneously traversing the edges and vertices of $\mathcal{M}$ and $\mathcal{TM}$. Then $\mathcal{TM}$ can be divided into two parts with $\mathcal{C}$, which are expanded by e_1 and e_2 respectively, as discussed in Sect. 2. For efficiency, we determine where the vertices of $\mathcal{TM}$ lie between the two parts of $\mathcal{TM}$ during computing $\mathcal{C}$. After completing the vertex marking, we process the Minkowski sum operation with respect to the constructed polygonal lines.

3.2 The Outer Surfaces of a Single Cap

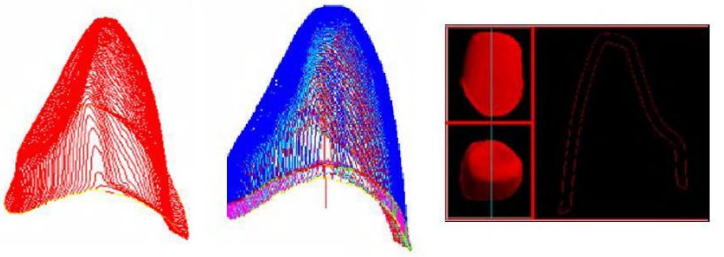
The outer surface of a single cap is modeled by expanding a part of $\mathcal{TM}$ above a fixed height h as much as a particular value e. If we apply the techniques in Sect. 3.1 to this model, the expansion does not occur in the points that have the minimum and maximum values of Y-coordinate. This is because the outer surface to be modeled includes the part of $\mathcal{TM}$ above the height h. Furthermore, since the expansion value $0.4mm$ is relatively large with respect to the input interval $0.05mm$ on X- and Y-axes, the polygonal lines obtained by applying the 2D Minkowski sum may be composed of the vertices that have the intervals greater than $0.05mm$.

For avoiding the limitation caused by a large expansion value, we develop a ZMap algorithm that can be applied to the part of $\mathcal{TM}$ above the height h. Our ZMap algorithm performs a geometric transformation for the set Vtx of all original vertices of $\mathcal{TM}$ above the height h. For each vertex $\bar{v} = (x, y, z) \in Vtx$, it builds the sphere with the origin $\bar{v}$ and the radius e, and then it finds the set $Vtx^n(\bar{v})$ of vertices that are contained in the sphere. For each vertex $\bar{v}^n = (x^n, y^n, z^n) \in Vtx^n(\bar{v})$, we calculate the Z coordinate value zz such that the point (x^n, y^n, zz) is on the sphere. Comparing the value z^n with the calculated value zz, it is replaced with the value zz if only $z^n < zz$.

Fig. 2(a) shows a set of polygonal lines of the tooth model lying above the margin that was detected by Yoo's algorithm [12]. Fig. 2(b) is the inner surface and the outer surface obtained from the given model, where the fixed height is $0.5mm$, and the expansion values below and above the height are $0.012mm$ and $0.0125mm$ respectively. Since a tooth model has the characteristics of a terrain, it is clear that the intersection takes the shape of a polygon called a cross-sectional polygon. Fig. 2(c) presents the cross-sectional polygon of the single cap modeled like Fig. 2(b) and the XZ-plane containing the center of the margin.

3.3 The Outer Surfaces of a Single Conus Core

The outer surface of a single Conus core was divided into the top part and body part as defined in Sect. 2. The top part can be modeled by employing the ZMap algorithm in Sect. 3.2, but the body part has to be modeled carefully



(a) (b) (c)

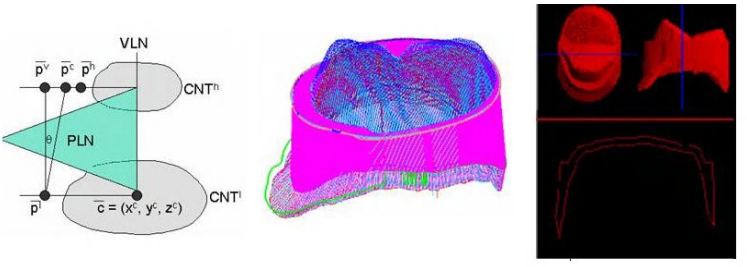
Fig. 2. An example of modeling a single cap (a) an original tooth model above a detected margin (b) a single cap model with the height $0.5mm$ and the two expansion values $0.012/0.0125mm$ below/above the height respectively (c) the visualization of a cross-sectional polygon

for satisfying the condition of a given conic angle. We develop a method for constructing the surface of the body part with 3D contours, which is known as the *tiling* problem. Fig. 3(a) shows the techniques of treating the body part for a given conic angle θ. First we determine the center $\bar{c} = (x^c, y^c, z^c)$ of $\mathcal{M}$, and the vertical line VLN passing the point $\bar{c}$. Next we extract the 3D contours of $\mathcal{TM}$ at the points of the fixed heights h_1 and h_2. The extracted contours are denoted by $\mathcal{C}^l$ and $\mathcal{C}^h$ which are called the *low region* and the *high region* of the body part respectively. The two contours $\mathcal{C}^l$ and $\mathcal{C}^h$ are, respectively, expanded as much as the values that dentists recommend, which were represented as e_1 and e_2 in Fig.3(a). After defining the plane PLN defined by $\bar{c}$, VLN, and a vertex on $\mathcal{M}$, we find the intersection point $\bar{p}^l$ ($\bar{p}^h$) between PLN and the expansion of $\mathcal{C}^l$ ($\mathcal{C}^h$). Let $\bar{p}^v$ be the point that has X/Y coordinates of $\bar{p}^l$ and Z coordinate of $\bar{p}^h$. Then, the line determining the conic angle θ from $\bar{p}^l$ will intersect the line segment between $\bar{p}^v$ and $\bar{p}^h$. This intersection point is represented as $\bar{p}^c$ in Fig. 3(a), and the intersection is iteratively computed around 360^o by rotating PLN with a fixed angle about the axis of VLN. Finally, we can describe an algorithm for modeling the body part by triangulating the line segments that connect all the found points $\bar{p}^l$, $\bar{p}^c$, and $\bar{p}^h$.

If we apply the modeling techniques in the above with the same conic angle to every region of $\mathcal{TM}$, however, the thickness of the modeled Conus will not be uniform since the inclined angle of $\mathcal{TM}$ is varied. In this paper, we divide the surface of an abutment into four regions by considering separation and adhesion between Conus cores and crown. The four conic angles which may be different from each other are assigned to each of the four regions. The assignment of different conic angles may cause another problem in the surface of the modeled Conus. A folding phenomenon may occur in the polygonal lines connecting the set $\{\bar{p}_i^c\}$ of intersection points that are obtained with different conic angles. Hence, this phenomenon will occur also on the surface of the body part that is formed with the point sets $\{\bar{p}_i^l\}$, $\{\bar{p}_i^c\}$, and $\{\bar{p}_i^h\}$. In order to resolve the folding phenomenon, we employ the B-Spine approximating curve algorithm to the

consecutive points p_i^c for all $i = 1, \cdots, n$. The B-Spline approximating curve is iteratively generated until the folding phenomenon disappears [8,11].

Fig. 3(b) illustrates the single Conus core modeled for an abutment, where the fixed height is $0.3mm$, the conic angles in four regions are 10^o, 2^o, 2^o and 4^o, and the height of the high region of the body part is $0.6mm$. Fig. 3(c) shows the cross-sectional polygon generated by intersecting the modeled single Conus core with the XZ-plane containing the center of the margin.



(a) (b) (c)

Fig. 3. An example of modeling a single Conus core (a) the body part of a single Conus core (b) the single Conus model with two height values $0.3/0.6mm$ and four conic angles $10^o/2^o/2^o/4^o$ (c) the visualization of a cross-sectional polygon

4 Conclusion

This paper analyzed the requirements for modeling single caps and single Conus cores, which are the mostly used artificial teeth prostheses. Based on the analyzed requirements, we developed geometric algorithms for effectively modeling the prostheses. We adopted 2D Minkowski sum to modeling the inner surfaces of single cores, while a ZMap algorithm was developed for modeling the outer surfaces of single caps. The outer surfaces of single Conus cores were modeled by using the combination of the Minkowski sum of contours and the tiling of the expanded contours. We also presented the examples of the single cores modeled through simple interactions. In the future, it is required to develop more techniques for the efficient modeling of other prostheses such as bridge cores, single crowns, bridge crowns, etc.

References

1. Berg M.D., Kreveld W.V., Overmars M., Schwarzkopf O., Computational Geometry - Algorithms and Applications, Springer-Verlag, (1997).
2. Besimo C., Jeger C., Guggenheim R., Marginal Adaptation of Titanium Frameworks produced by CAD/CAM Techniques, The International Journal of Prosthodontics **10**, 6, (1997), 541-546.

3. Duret F., Blouin J.L., Duret B., CAD/CAM in dentistry, Jorunal Am. Dent. Assoc.**117**, 11, (1988), 715-720.
4. Ghosh P., A Unified Computational Framework for Minkowski Operations, Computer and Graphics **17**, 4, (1993), 357-378.
5. Gurke S., Generation of Tooth Models for Ceramic Dental Restorations, The 4th International Conference on Computer Integrated Manufacturing, Singapore, (Oct 1997).
6. Korber K.H., Konuskronen, Das rationelle Teleskopsystem, Einfuhrung Inklinik und Technik, Auflage, Huthig, Heidel-berg, (1983).
7. Lozano-Perez T., Spatial Planning: A Configuration Space Approach, IEEE Transaction on Computers **32**, 2, (1983), 108-120.
8. Meyers D., Skinner S., and Sloan K., Surfaces from Contours, ACM Transactions on Graphics **11**, 3, (1992), 228-258.
9. Rekow D. E., The Minnesota CAD/CAM System DentiCAD, Technical report, University of Minnesota, (1989).
10. Rekow D. E., CAD/CAM in Dentistry:Critical Analysis of Systems, In Computers in Clinical Dentistry, Quintessence Co. Inc., (1991), 172-185.
11. Yamagichi F., Curves and Surfaces in Computer Aided Geometric Design, Springer-Verlag, (1988).
12. Yoo K.H., An Efficient Algorithm for Extracting the Margin of a Tooth, Manuscript, (2003).
13. Yoon C.G., Kang D.W., Chung S.M., State-of-arts in Fixed Prosthodontics, Jongii Press, Korea, (1999).
14. Willer J., Rossbach A., and Weber H., Computer Assisted Milling of Dental Restorations using a new CAD/CAM Data Acquisition System, The International Journal of Prosthodontics **80**, 3, (1988) 346-353.
15. Jef M. van der Zel, Ceramic-fused-to-metak Restorations with a new CAD/CAM system, Quintessence International **24**, 11, (1993) 769-778.

GA and CHC. Two Evolutionary Algorithms to Solve the Root Identification Problem in Geometric Constraint Solving

M.V. Luzón[1], E. Barreiro[1], E. Yeguas, and R. Joan-Arinyo[2]

[1] Escuela Superior de Ingeniería Informática. Universidade de Vigo,
Av. As Lagoas s/n, E-32004 Ourense
{luzon,enrique}@uvigo.es
[2] Escola Técnica Superior d'Enginyeria Industrial de Barcelona.
Universitat Politècnica de Catalunya,
Av. Diagonal 647, 8^a, E-08028 Barcelona
robert@lsi.upc.es

Abstract. Geometric problems defined by constraints have an exponential number of solution instances in the number of geometric elements involved. Generally, the user is only interested in one instance such that, besides fulfilling the geometric constraints, exhibits some additional properties.
Selecting a solution instance amounts to selecting a given root every time the geometric constraint solver needs to compute the zeros of a multi valuated function. The problem of selecting a given root is known as the *Root Identification Problem*.
In this paper we present a comparative study of a basic genetic algorithm against the CHC algorithm. Both techniques are based on an automatic search in the space of solutions driven by a set of extra constraints. A number of case studies illustrate the performance of the methods.

Keywords: Evolutionary algorithms, Constructive geometric constraint solving, Root identification problem, Solution selection.

1 Introduction

Modern computer aided design and manufacturing systems are built on top of parametric geometric modeling engines. The field has developed sketching systems that automatically instantiate geometric objects from a rough sketch, annotated with dimensions and constraints input by the user. The sketch only has to be topologically correct and constraints are normally not yet satisfied. The core of those sketching systems is the geometric constraint *solver*.

Geometric problems defined by constraints have an exponential number of solution instances in the number of geometric elements involved. Generally, the user is only interested in one instance such that besides fulfilling the geometric constraints, exhibits some additional properties. This solution instance is called the *intended solution*.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 139–146, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Selecting a solution instance amounts to selecting one among a number of different roots of a nonlinear equation or system of equations. The problem of selecting a given root was named in [1] the *Root Identification Problem*.

Several approaches to solve the Root Identification Problem have been reported in the literature. Examples are: Selectively moving the geometric elements, conducting a dialogue with the constraint solver that identifies interactively the intended solution, and preserving the topology of the sketch input by the user. For a discussion of these approaches see, for example, references [1,7,20] and references therein.

In [16,15] we reported on a new technique to automatically solve the Root Identification Problem for constructive solvers, [1,9]. The technique over-constrains the geometric problem by defining two different categories of constraints. One category includes the set of constraints specifically needed to solve the geometric constraint problem. The other category includes a set of extra constraints or predicates on the geometric elements which identify the intended solution instance. Once the constructive solver has generated the space of solution instances, the extra constraints are used to drive an automatic search of the solution instances space performed by a genetic algorithm, [11,24]. The search outputs a solution instance that maximizes the number of extra constraints fulfilled.

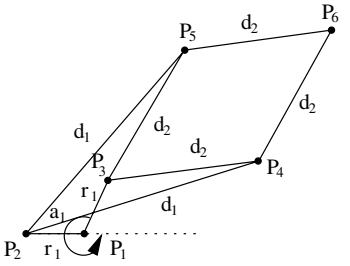
In this paper we study the performance of two evolutive algorithms applied to solve the Root Identification Problem: the basic genetic algorithm and the CHC algorithm. In both cases, the automatic search in the space of solutions is driven by the set of extra constraints.

2 Constructive Geometric Constraint Solving

In two-dimensional constraint-based geometric design, the designer creates a rough sketch of an object made out of simple geometric elements like points, lines, circles and arcs of circle. Then the intended exact shape is specified by annotating the sketch with constraints like distance between two points, distance from a point to a line, angle between two lines, line-circle tangency and so on. A geometric constraint solver then checks whether the set of geometric constraints coherently defines the object and, if so, determines the position of the geometric elements. Figure 1 shows an example sketch of a constraint-based design.

Many techniques have been reported in the literature that provide powerful and efficient methods for solving systems of geometric constraints. For example, see [4] and references therein for an extensive analysis of work on constraint solving. Among all the geometric constraint solving techniques, our interest focuses on the one known as *constructive*.

Constructive solvers have two major components: the *analyzer* and the *constructor*. The analyzer symbolically determines whether a geometric problem defined by constraints is solvable. If the problem is solvable, the output of the analyzer is a sequence of construction steps which places each geometric element in such a way that all constraints are satisfied. This sequence is known



$$
\begin{aligned}
&distance(P_1, P_2) && = r_1\\
&distance(P_1, P_3) && = r_1\\
&distance(P_2, P_4) && = d_1\\
&distance(P_2, P_5) && = d_1\\
&distance(P_3, P_4) && = d_2\\
&distance(P_3, P_5) && = d_2\\
&distance(P_4, P_6) && = d_2\\
&distance(P_5, P_6) && = d_2\\
&angle(line(P_1, P_3), line(P_2, P_1)) && = a_1
\end{aligned}
$$

Fig. 1. Geometric problem defined by constraints

as the *construction plan*. Figure 2 shows a construction plan generated by the ruler-and-compass geometric constraint solver reported in [14] for the problem depicted in 1. After assigning specific values to the parameters, the constructor interprets the construction plan and builds an object instance, provided that no numerical incompatibilities arise.

1. $P_1 = point(0, 0)$
2. $P_2 = point(r_1, 0)$
3. $\alpha_1 = direction(P_2, P_1)$
4. $\alpha_2 = adif(\alpha_1, a_1)$
5. $P_3 = rc(line(P_1, \alpha_2), circle(P_1, r_1))$
6. $P_4 = cc(circle(P_2, d_1), circle(P_3, d_2))$
7. $P_5 = cc(circle(P_2, d_1), circle(P_3, d_2))$
8. $P_6 = cc(circle(P_4, d_2), circle(P_5, d_2))$

Fig. 2. Construction plan for the problem in Fig. 1

Function names in the plan are self explanatory. For example function $adif$ denotes subtracting the second angle from the first one and $asum$ denotes the addition of two angles while rc and cc stand for the intersection of a straight line and a circle, and the intersection of two circles, respectively.

In general, a well constrained geometric constraint problem, [10,13,18], has an exponential number of solutions. For example, consider a geometric constraint problem that properly places n points with respect to each other. Assume that the points can be placed serially, each time determining the next point by two distances from two already placed points. In general, each point can be placed in two different locations corresponding to the intersection points of two circles. For n points, therefore, we could have up to 2^{n-2} solutions. Possible different locations of geometric elements corresponding to different roots of systems of nonlinear algebraic equations can be distinguished by enumerating the roots with an integer index. For a more formal definition see [7,22].

In what follows, we assume that the set of geometric constraints coherently defines the object under design, that is, the object is generically well constrained and that a ruler-and-compass constructive geometric constraint solver like that reported in [14] is available. In this solver, intersection operations where circles are involved, rc and cc, may lead to up to two different intersection points,

depending on whether the second degree equation to be solved has no solution, one or two different solutions in the real domain. With each feasible *rc* and *cc* operation, the constructor in the solver associates an integer parameter $s_k \in \{-1, 1\}$, that characterizes each intersection point by the sign of the square root in the corresponding quadratic equation. For details on how to compute s_k, the reader is referred to [22].

3 The Root Identification as a Constraint Optimization Problem

We will solve the Root Identification Problem by over-constraining the geometric constraint problem: The intended solution instance to a well constrained problem is specified by defining a set of extra constraints or predicates on the geometric elements. As extra constraint, the user can apply the usual geometric constraints or specific topological constraints like $PointOnSide(P, line(P_i, P_j), side)$, which means that point P must be placed on one of the two open half spaces defined by the straight line through points P_i, P_j, oriented from P_i to P_j. Parameter *side* takes values in {*right, left*}.

Recall that we consider ruler-and-compass constructive geometric constraint solving. In this context, geometric operations correspond to quadratic equations, thus each constructive step has at most two different roots. Let s_j denote the integer parameter associated by the solver with the j-th intersection operation, either *rc* or *cc*, occurring in the construction plan. Since we are interested only in solution instances that actually are feasible, that is, solution instances where no numerical incompatibilities arise in the constructor, we only need to consider integer parameters s_j taking value in the set of signs $D_j = \{-1, 1\}$ that characterizes each intersection point.

Assume that n is the total number of *rc* plus *cc* intersection operations in the construction. We define the *index* associated with the construction plan as the ordered set $I = \{s_1, \ldots, s_j, \ldots, s_n\}$ with $s_j \in D_j, 1 \leq j \leq n$. Therefore the Cartesian product of sets $\mathcal{I} = D_1 \times \ldots \times D_n$ defines the space where the solution instances to the geometric constraint problem belong to.

A construction plan which is solution to a geometric constraint problem can be seen as a function of the index I. Moreover, the construction plan can be expressed as a first order logic formula, [15]. Let $\Psi(I)$ denote this formula. Clearly, the set of indexes $\{I \in \mathcal{I} \mid \Psi(I)\}$ is the space of feasible indexes, that is the set of indexes each defining a solution to the geometric constraint problem. This set of indexes is the *allowable search space*, [5].

Let $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$ be the set of extra constraints given to specify the intended solution instance and let $\Phi = \mathcal{C}_1 \wedge \mathcal{C}_2 \wedge \ldots \wedge \mathcal{C}_m$. Let f be a (possibly real-valued) function defined on $\Psi(I) \wedge \Phi$ which has to be optimized. Then, according to Eiben and Ruttkay, [5], the triple $< \mathcal{I}, f, \Psi(I) >$ defines a *constraint optimization problem* where finding a solution means finding an index I in the allowable search space with an optimal f value.

4 Evolutionary Algorithms

Evolutionary algorithms which model natural evolution processes were already proposed for optimization in the 1960s. The goal was to design powerful optimization methods, both in discrete and continuous domains, based on searching methods on a population, members of which are coded problem solutions, [2].

In previous works [16,15] we shown that evolutionary algorithms are a feasible technique to solve the Root Identification Problem. In this work we compare the performance of two kinds of evolutionary algorithms: Genetic Algorithms and CHC algorithms.

4.1 The Genetic Algorithm

Genetic algorithms (GA) are search algorithms that model sexual reproduction. Sexual reproduction is characterized by recombining two parent strings into an offspring. This recombination is called *crossover*. Crossover is the recombination of traits of the selected members in the hope of producing a child with better fitness levels than its parents. Crossover is accomplished by swapping parts of strings representing two members in the population. GAs were invented by Holland, [12]. Recent surveys can be found in [8] and [11].

4.2 The CHC Algorithm

The main drawback of GAs is the premature convergence: After a few generations, the current population reaches an state where the goal function is not optimal and no longer improves.

The heterogeneous recombination and cataclysmic mutation algorithm (CHC) is an evolutive algorithm with binary coding which tries to avoid the premature convergence by getting a suitable balance between the ability to explore the search space for diversity and the ability to explote the local properties of the search for an appropriate selection. Instead the mutation operation of GAs, CHC includes a restart process that offers many of the benefits of a great size population without the cost of a slower search, [6].

5 Experimental Study

To assess and compare the performance of GA and CHC algorithms when applied to solve the Root Identification Problem, we considered eleven different problems. For each problem a number of extra constraints to select the intended solution instance were defined. The goal was to select one index such that the number extra constraints fulfilled by the associated solution instance was maximum.

The number of indexes in the initial population was always 20 and the maximum number of generations allowed was 30. For GAs the crossover and mutation probabilities were 0.3 and 0.1 respectively. In the search reinicialization, the CHC

Table 1. T test results

Problem	*Algorithm*	*Mean*	*StDev*	*SE mean*	*t*	*Sig. level*
1	CHC	32.70	1.251	0.125	-1.879	0.062
	GA	32.32	1.588	0.159		
2	CHC	28.38	1.324	0.132	-2.738	0.007
	GA	27.84	1.461	0.146		
3	CHC	38.48	1.467	0.147	-2.221	0.028
	GA	37.96	1.825	0.183		
4	CHC	34.17	1.484	0.148	-1.079	0.282
	GA	33.94	1.530	0.153		
5	CHC	39.12	3.059	0.306	-2.109	0.036
	GA	37.87	5.076	0.508		
6	CHC	37.14	1.826	0.183	-2.175	0.031
	GA	36.53	2.129	0.213		
7	CHC	46.74	2.473	0.247	-2.537	0.012
	GA	45.71	3.220	0.322		
8	CHC	47.92	1.857	0.186	-2.546	0.012
	GA	47.18	2.236	0.224		
9	CHC	26.91	0.944	0.094	-0.596	0.552
	GA	26.83	0.954	0.095		
10	CHC	31.73	2.178	0.218	-2.429	0.016
	GA	30.60	4.110	0.411		
11	CHC	23.16	1.779	0.178	-1.828	0.069
	GA	22.53	2.952	0.295		

algorithm kept the 3 best fitting indexes. The remaining 17 indexes in the population were generated using these 3 individuals as templates and randomly changing the 35% of the binary signs, [6].

GA and CHC algorithms were applied to each problem one hundred times. The initial population in each run was the same for both algorithms and the value of the goal function at the end of the run was recorded.

Table 1 summarizes the experimental results. The fourth and fifth column give respectively the mean and the standard deviation of the number of extra constraints fulfilled in the set of one hundred runs. In all cases, the mean for the CHC algorithm was higher that for the GA while the standard deviation for CHC was smaller that for GA. Therefore, CHC showed a better performance in finding the intended solution instance.

To assess whether the mean of the goal functions yielded by each evolutive algorithm are statistically different from each other, we applied a t-test, [23]. Columns labeled t and *Sig. level* in Table 1 give the t value for the t-test and the significance level associated respectively. Problems 1, 4, 9 and 11 show a significance level higher than 0.05, the usually accepted level value. Notice, however, that only in problems 4 and 9 the significance level is clearly higher than 0.05. Therefore, we conclude that, in average, the instance solution selected by the CHC algorithm fulfills more extra constraints than that selected by the GA.

6 Conclusions and Future Work

The Root Identification Problem can be solved using GA and CHC evolutive algorithms. The idea is to over-constrain the problem and use the extra constraints to drive the algorithm search.

Experimental results from a significative benchmark show that performance of CHC is better than GA's. On the one hand, CHC algorithms do no show premature convergence. On the other hand, in the average and with a significance level higher than 0.05, the instance solution selected by the CHC algorithm shows a better fitting than that selected by the GA.

Currently we are working in two different directions. One focuses on applying new evolutive algorithms to solve the Root Identification Problem and to study the relative performance. The goal of the other line of research is to define strategies to automatically define values for evolutive parameters (population size, crossover and mutation probabilities, etc) as a function of the geometric constraint problem at hand.

Acknowledgements. This research has been partially supported by FEDER and CICYT under the project TIC2001-2099-C03-01.

References

1. W. Bouma, I. Fudos, C. Hoffmann, J. Cai, and R. Paige. Geometric constraint solver. *Computer-Aided Design*, 27(6):487–501, June 1995.
2. H.J. Bremermann, J. Roghson, and S. Salaff. Global properties of evolution processes. In H.H. Pattee, E.A. Edelsack, L. Fein, and A.B. Callahan, editors, *Natural Automata and Useful Simulations*, pages 3–42. Macmillan, 1966.
3. B.D. Brüderlin. *Rule-Based Geometric Modelling*. PhD thesis, Institut für Informatik der ETH Zürich, 1988.
4. C. Durand. *Symbolic and Numerical Techniques for Constraint Solving*. PhD thesis, Purdue University, Department of Computer Sciences, December 1998.
5. A.E. Eiben and Zs. Ruttkay. Constraint-satisfaction problems. In T. Bäck, D.B. Fogel, and Z. Michalewicz, editors, *Handbook of Evolutionary Computation*, chapter C5.7, pages C5.7:1–C5.7:5. Institute of Physics Publishing Ltd and Oxford University Press, 1997.
6. L.J. Eshelman. The CHC adaptive search algorithm: How to safe search when engaging in nontraditional genetic recombination. *Foundations of Genetic Algorithms*, pages 265–283, 1991.
7. C. Essert-Villard, P. Schreck, and J.-F. Dufourd. Sketch-based pruning of a solution space within a formal geometric constraint solver. *Artificial Intelligence*, 124:139–159, 2000.
8. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo, CA, 1993. Morgan Kaufmann.
9. I. Fudos and C.M. Hoffmann. Correctness proof of a geometric constraint solver. *International Journal of Computational Geometry & Applications*, 6(4):405–420, 1996.

10. I. Fudos and C.M. Hoffmann. A graph-constructive approach to solving systems of geometric constraints. *ACM Transactions on Graphics*, 16(2):179–216, April 1997.
11. D.E. Goldberg. *Genetic Algorithms in Search, Optimization , and Machine Learning*. Addison Wesley, 1989.
12. J. H. Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, 1975.
13. R. Joan-Arinyo and A. Soto-Riera. Combining geometric and equational geometric constraint solving techniques. In *VII Congreso Español de Informática Gráfica*, pages 309–324. Eurographics, June 1997.
14. R. Joan-Arinyo and A. Soto-Riera. Combining constructive and equational geometric constraint solving techniques. *ACM Transactions on Graphics*, 18(1):35–55, January 1999.
15. R. Joan-Arinyo and M.V. Luzón and A. Soto. Constructive geometric constraint solving: a new application of genetic algorithms. In *Parallel Problem Solving from Nature-PPSN VII*, volume 1, pages 759–768, 2002.
16. R. Joan-Arinyo and M.V. Luzón and A. Soto. Genetic algorithms for root multiselection in constructive geometric constraint solving. *Computer & Graphics*, 27:51–60, 2003.
17. S.C. Kleene. *Mathematical Logic*. John Wiley and Sons, New York, 1967.
18. G. Laman. On graphs and rigidity of plane skeletal structures. *Journal of Engineering Mathematics*, 4(4):331–340, October 1970.
19. L. Lovász and Y. Yemini. On generic rigidity in the plane. *SIAM Journal on Algebraic and Discrete Methods*, 3(1):91–98, March 1982.
20. M.V. Luzón. *Resolución de Restricciones geométricas. Selección de la solución deseada*. PhD thesis, Dpto. de Informática. Universidade de Vigo., Septiembre 2001.
21. N. Mata. Solving incidence and tangency constraints in 2D. Technical Report LSI-97-3R, Department LiSI, Universitat Politècnica de Catalunya, 1997.
22. N. Mata. *Constructible Geometric Problems with Interval Parameters*. PhD thesis, Dept. LSI, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain, 2000.
23. W. Mendenhall and T. Sincich. *Statistics for engineering and the sciences*, 4th Edition. Prentice-Hall, 199.
24. Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, 1996.

Manifold Extraction in Surface Reconstruction

Michal Varnuška[1] and Ivana Kolingerová[2]

Centre of Computer Graphics and Data Visualization
Department of Computer Science and Engineering
University of West Bohemia, Pilsen, Czech Republic
miva@kiv.zcu.cz, kolinger@kiv.zcu.cz

Abstract. Given a finite point set in R^3 scanned by special devices from the object surface, a surface model interpolating or approximating the points set has to be obtained. We use for the reconstruction a CRUST algorithm by Nina Amenta, which selects surface triangles from the Delaunay tetrahedronization using information from the dual Voronoi diagram. This set of candidate surface triangles does not form a manifold, so the manifold extraction step is necessary. We present two improvements for this step, the former is limited to the used algorithm and the latter can be used with any other reconstruction algorithm.

1 Introduction

Many applications from various areas of science or industry need to work with the piecewise interpolation of the real objects. One of often-used ways to obtain the model is the points cloud reconstruction. The task of the reconstruction is not simple, we have only points in 3D space without any additional data (such as normal vectors).

Four kinds of approaches exist based on warping, distance function, incremental surface reconstruction and spatial subdivision. Warping works on the basic idea that we deform some starting surface to the surface that forms the object. The idea of warping is relatively old and is used in Müller's approach [17] or by Muraki [18].

The incremental surface reconstruction is the second huge group of algorithms. Boissonat's approach [8] begins on the shortest edge from all possible edges between points and incrementally appends the edges to create a triangle mesh. Mencl and Müller [19] developed a similar algorithm. It creates an extended minimum spanning tree, extends it to the surface description graph and extracts typical features.

Hoppe [16] presented an algorithm, where the surface is represented by the zero set of a signed distance function. The function sign is plus if the point lies inside the closed surface and minus otherwise, the value is the distance to the surface. Curless and Levoy [9] gave an effective algorithm using the signed distance function on a voxel grid, it is able to reconstruct eventual holes in a post-processing.

The fundamental property of the methods based on spatial subdivision is the space division into independent areas. The simplest division is presented by the voxel grid,

[1] The author was supported by the project FRVŠ G1/1349
[2] The author was supported by the project MSM 235200005

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 147–155, 2004.
© Springer-Verlag Berlin Heidelberg 2004

which Algorri and Schmitt [1] use in their effective algorithm. The voxels containing points from the input set are chosen and the surface is extracted. The most often used division is the Delaunay tetrahedronization (DT) because the surface forms a subgraph of the tetrahedronization. Edelsbrunner and Mücke [14, 15] developed an α-shape algorithm for uniform sample sets, Bernardini and Bajaj [6] extended it. They use the binary search on the parameter alpha to find the surface subcomplex. Bernardini [7] presented a very fast and efficient ball pivoting algorithm.

Amenta introduced the concept of CRUST in [2, 3, 4]. Dey extended the ideas of Amenta, giving an effective COCONE algorithm. The extension of COCONE algorithm can handle large data [10], detect boundaries [11], undersampling and oversampling [12]. These methods are based on the observation that the places with a changing of point density can be detected using shape of Voronoi cells in these places. Both authors gave then algorithms for watertight surface creation, Amenta's POWERCRUST [5] and Dey's TightCOCONE [13].

As mentioned in abstract, we use the CRUST algorithm. It works on spatial subdivision achieved by DT. Auxiliary subdivision, Voronoi diagram (VD), is obtained by dualization from DT. There exist two versions of the algorithm based on onepass or twopass tetrahedronization. We have chosen these methods because they have strong theoretical background, are not so much sensitive to the sampling errors (the sampling criterion is based on *local feature size (LFS)*, closer in [3]) and we have a fast and robust implementation of DT. Due the sampling LFS criterion the CRUST algorithm is not sensitive to big changes in sampling density, the data need not to be uniformly sampled, but it has problems with outliers and sharp edges. Then the surface normals estimated using poles (explained bellow) point to bad directions, the reconstruction fails and a lot of bad triangles appear.

The details of these methods can be found in [2, 3, 4], we concentrate only to the information necessary for later understanding. The positive pole $p+$ is the furthest Voronoi vertex (VV) of the Voronoi cell around some point p, the negative pole $p-$ is the farthest VV on the "second side" (the dot product of the vectors ($p-$, p) and ($p+$, p) is negative). For successfully sampled surface all Voronoi cells are thin, long, the poles lay on the medial axis and vectors to the poles approximate the normal vectors.

The first shared step of both versions is the DT creation followed by its dualization to VD and poles computation. Then the versions differ, the twopass algorithm takes the poles as an approximation of the medial axis while the onepass takes the poles as the approximation of the normal vectors. We use the onepass version because it is more than three times faster and less memory consuming. Three conditions must hold for the surface triangles: their dual Voronoi edges must intersect the surface, the radius of the circumcircle around the surface triangle is much smaller than the distance to the medial axis at its vertices and the normals of surface triangles make small angles with the estimated surface normals at the vertices.

We can compute the set of surface triangles as follows. For each point p we have an approximation of its normal vector $n = p+ - p$. Each triangle in DT has an edge e dual in VD. For the triangles on the surface, this edge has to pass through the surface P. Let us denote the vertices of the edge e as w_1, w_2, the angles $\alpha = \angle(w_1 p, n)$ and $\beta = \angle(w_2 p, n)$. When the interval $<\alpha, \beta>$ intersects the interval $<\pi/2 - \theta, \pi/2 + \theta>$ and this condition holds for each vertex p of the triangle, then the triangle is on the surface. The parameter θ is the input parameter of the method.

2 Manifold Extraction

The result of the CRUST algorithm is the set of surface triangles T (we call it the primary surface). These triangles passed conditions of the CRUST algorithm but they do not form the manifold yet. There can be more than two triangles incident on some edges or some triangles may miss on the places of local discontinuity. For example, very flat tetrahedra in the smooth part of the surface (Fig. 1a) or the tetrahedra on the surface edge (Fig. 1b) may have all faces marked as surface triangles. The number of overlapped triangles differs from model to model and depends on the surface smoothness. For smooth surface it is in tens percent and when the surface is rough, the rate decreases.

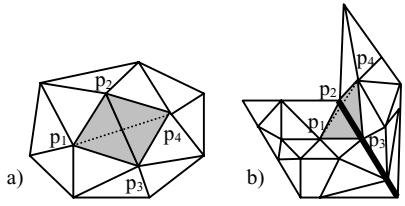


Fig. 1. a) Flat part of the surface, b) the part with the sharp edge (bold line). One pair of triangles is ($p_1 p_2 p_3$) and ($p_2 p_3 p_4$), the second pair is ($p_1 p_2 p_4$) and ($p_2 p_3 p_4$)

That is why the surface extraction step must be followed by a manifold extraction step. The input to the manifold extraction step is just the set of triangles. Manifold extraction step is independent of the reconstruction method, therefore it could be combined with other algorithms than CRUST. We have developed our own algorithm. The reason was that the manifold extraction methods were explained very briefly in the papers, however, this step is important. Our approach uses breadth-first search for appending triangles on free edges and has a linear time complexity. The algorithm is presented in [20], for clarity of the text we will briefly explain it.

The preprocessing step of the extraction is creation of two structures, a list of incident triangles for each point and a multiple neighbors mesh containing for each triangle the list of incident triangles on the edges. More then two triangles sharing one edge can exist as the manifold is not ensured yet (e.g. Fig. 2). First, we have to find the starting triangles using these structures, they will form the root of the searching tree. These triangles form a triangle fan; no triangles from the fan overlap when we project them to the plane defined by the point and the normal at this point.

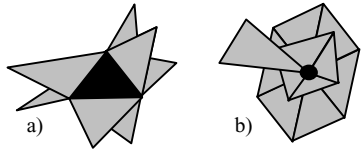


Fig. 2. a) An example of multiple neighbors to some triangle, b) an example of incident triangles to one point

Next, we add to the already extracted triangles their neighbors on the non-connected edges. These neighbors form next level of the tree. Because we can have multiple neighbors, we have to find just one triangle of them. We assume that the triangles must be small to form a correct surface, so we take the one, which has the shortest sum of edge length. We need only 2 levels of the tree at one moment, older levels can be safely deleted. We continue recursively until all edges are processed.

3 Prefiltering

We already mentioned that the CRUST algorithm has very good results for smooth surfaces. However, even with datasets of smooth objects, sometimes small triangle holes appear in the reconstructed surface. It is not a problem to find and fill them in the postprocessing step, but the question is why they appear. Each tetrahedron has four faces – triangles. The CRUST marks them whether they belong to the set of the primary surface *T*. We have found that the triangle holes appear in the smooth places where very flat tetrahedron lies whose three faces are marked as surface triangles. See Fig. 3a) for an example: the dark gray triangles are already extracted and we are looking for the triangle neighbor on the bold edge of the triangle 1. The light gray triangles are marked triangles from one tetrahedron (there are three overlapping triangles), two of them are incident with the bold edge of triangle 1 and we have to choose only one of them. When we select bad triangle then in the next step of extraction the triangle hole occurs (Fig. 3b). Fig. 3c) shows a correct configuration.

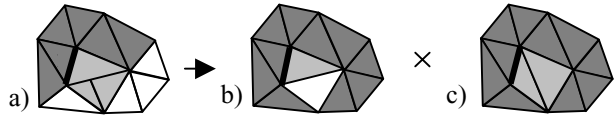


Fig. 3. Two configurations in the manifold extraction of the tetrahedron with three marked surface triangles, a) initial status, b) wrong choice, c) correct choice

In order to avoid such situations it is necessary before the manifold extraction step to detect the tetrahedra, which have three marked faces, and remove one overlapped face. So we take one tetrahedron after another and we mark surface triangles (faces) using the CRUST algorithm. If there are three marked faces on one tetrahedron, we preserve only these two faces whose normals make the smallest angle (the tetrahedron is flat, so the triangles on the other edges make sharp angle together), the third face is deleted. We have to be careful with the orientation of the triangle normals, they have to be oriented in the direction of the tetrahedron centre of gravity (see an example in Fig. 4). The best configuration is in Fig. 4d), the angle between triangles normals incident on the edge is the smallest (the dot product of the normals is close to one, 4b) and 4c) are close to minus one).

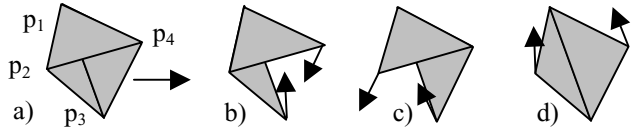


Fig. 4. Tetrahedron with three marked faces ($p_1p_2p_4$, $p_1p_2p_3$, $p_1p_4p_3$) and three possibilities which two triangles to choose. Arrows present the triangle normals

This approach converts tetrahedra with three marked triangles to tetrahedra with two marked triangles. We can use it to filter tetrahedra with four marked triangles too. Besides removal of problematic places, the prefiltering approach reduces the number of triangles in the primary surface. After converting all tetrahedra with four and three good faces to tetrahedra with two good faces, the set of primary surface triangles is ready for extraction.

4 Postfiltering

When we have the data, which are not uniformly sampled, with some noise or some features missing due to undersampling, the manifold extraction may fail because the CRUST selects bad surface triangles and unwanted triangle configurations occur (see Fig. 8a). This detail is taken from a dataset which is not uniformly sampled and contains some noise. The highlighted part presents the erroneous place after the manifold extraction – missing and overlapping triangles.

Missing and overlapping triangles appear there due to bad normal vectors arisen from the incorrect shape of Voronoi cells. We have analyzed triangle fans around the points obtained after the reconstruction. These configurations may be detected using an undirected graph. The nodes of the graph correspond to the fan triangles. A graph edge *e* exists in the graph if the nodes of the edge *e* correspond to neighboring triangles (see Fig. 5a), 5b).

There exist two acceptable configurations of the triangle fan. Fig. 5c) presents a full fan around a point. It can be detected as the graph cycle which contains all nodes. Fig. 5d) is just one single triangle, which can appear, e.g. on the corners of the surface with the boundary. Detection of these configurations is simple.

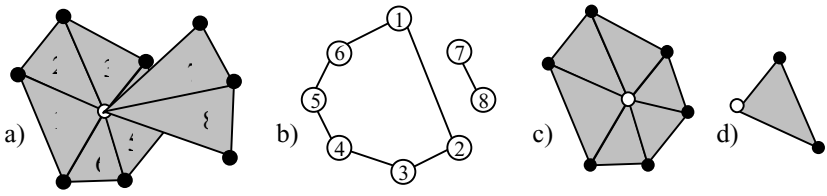


Fig. 5. a) Some fan configuration, b) a graph corresponding to the fan. Pictures c) and d) present acceptable fan configurations, c) a full fan, d) a point with one triangle

Other configurations are incorrect and some triangles have to be deleted. When we are able to find one cycle there, we can delete all triangles whose graph nodes are not included in the cycle. The most common configuration is shown in Fig. 6a), one full triangle fan with one separated triangle. Fig. 6b) is a hypothetic situation with more then one cycle but we did not find any occurrence of this.

The configurations presented in the Fig. 6c), 6d), 6e) are more problematic. When there are only subfans (we denote the fan as subfan if it does not form a cycle), the finding good fan configuration is not so simple and it will be explained in the following text. Here we can not avoid the use of the normal vectors (we are testing these configurations in the projected plane), and it can bring problems. The normal vectors have good estimation only on the smooth parts of the surface, but the places, where these problematic configurations of the fans appear, are on the places where the sampling is not correct.

All the triangles around the fan are projected to the plane given by the point (centre of the fan) and its normal vector (although the normal direction probably has not correct direction). The detection is simpler for the configuration in the Fig. 6c) and 6d) than 6e) because the triangles create only one subfan. When the sum of angles of the projected triangles (angle between two edges incident with the point) has less then 2π (Fig. 6c) the configuration is accepted and no changes in the triangle mesh is done. When it is more (Fig. 6d) we delete triangles from one end of the subfan until the angle is less then 2π. We have implemented just the removing from one end but it is better to remove these triangles in order to choose the sum of angles closer to 2π.

The Fig. 6e) represents the worst case, a set of more subfans. This configuration occurs fortunately very rarely and we remove all tringles except the subfan with the largest sum of angles.

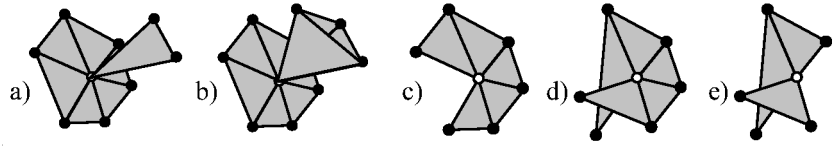


Fig. 6. Illustrations a) and b) present two types of configuration with a full fan, a) full fan with another separated triangle, b) more fans. Illustrations c) d) e) show some fan configurations (in a projection) without fans, c) one subfan, d) one overlapping subfan, e) more subfans

5 Results

The implementation of the CRUST algorithm and all of our improvements was done in Borland Delphi under the Windows XP system running on AMD Athlon XP+ 1500MHz processor with 1GB of memory. We have tested our implemented algorithm (Amenta's CRUST with our manifold extraction) together with Dey's COCONE algorithm [10, 11, 12], which is similar to the CRUST.

When we ran our algorithm without the prefiltering improvements, several triangle holes appeared. The number was not so high but when looking closer to the reconstructed object, it can disturb the visual perception and the object does not form the manifold. The same occurs in the Dey's algorithm (Fig. 7a). After prefiltering, the situation changed and our algorithm was able to reconstruct the surface with much less triangles holes (Fig. 7b), 7c). Some triangle holes still appear but the cause is different, the missing triangles did not pass the surface triangle test (recall Section 1).

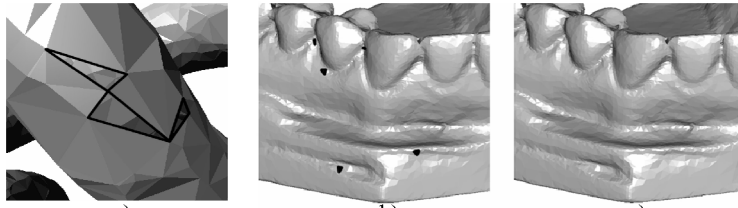


Fig. 7. a) Detail of the reconstructed surface by Dey's COCONE, black are highlighted triangle holes in the surface. The picture b) and c) shows the reconstruction using our approach, b) missing triangles are black, c) the same part of the surface after prefiltering applied

The next consequence of this prefiltering improvement was the reduction of the amount of triangles in the primary surface. We have measured (Table 1) the number of redundant triangles, which it is necessary to remove from the triangulation. The row "without" presents the number of redundant triangles marked as surface triangles without the prefiltering applied. The number of redundant marked surface triangles computed with the help of the prefiltering is in the row "prefilter". The last row presents the rate in percents of the number of marked triangles before applying prefiltering and the number of triangles after prefiltering. It can be seen that 38-99 percent of the redundant triangles are removed by prefiltering.

Table 1. Number of points ("N") in datasets used for testing, Number of triangles marked as surface without prefiltering ("without"), number of triangles with prefiltering ("prefilter") and the percent rate of the removed triangles using the prefiltering ("rem")

	bone	bunny	x2y2	engine	hypshet	knot	mann	nascar	teeth
N	68537	35947	5000	22888	6752	10000	12772	20621	29166
withou	8106	11937	358	9835	1451	2017	926	992	4642
prefilte	111	71	122	33	898	70	54	297	145
% rem	98	99	65	99	38	96	94	70	96

Now we will present the result of postfiltering. In Fig. 8a) we can see the case where some bad fan (or umbrella) configurations appear, in Fig. 8b) the same part of the surface after applying the postfiltering is shown. The overlapped "flying" triangles disappear and the remaining triangle holes are filled with the triangles.

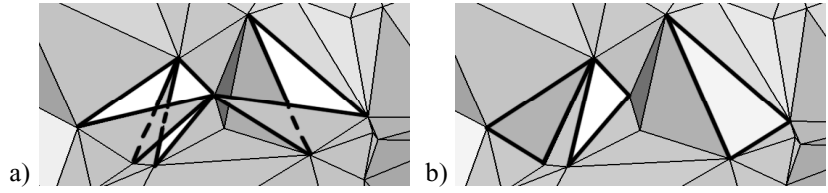


Fig. 8. a) The part of the surface with and b) without bad fans after postfiltering

Our implementation of manifold extraction is able to find all holes in the surface, but the holes filling is now limited to the triangles holes (as presented in Fig. 9). Not all holes are so small, we are planning in the future to develop or apply some simple algorithm for holes triangulation.

The same problem occurs using the Dey's algorithm, we found overlapping triangles on the surface of the reconstructed objects, too (Fig. 9a), 9b). In this case, we were not able to reproduce Fig. 9a), 9b) by our program, because although the algorithms are similar, the code is not the same and the reconstructed meshes differ a little for the same models. Fig. 9c) and 9d) shows the same part of the reconstructed model using our approach and the same part after postfiltering.

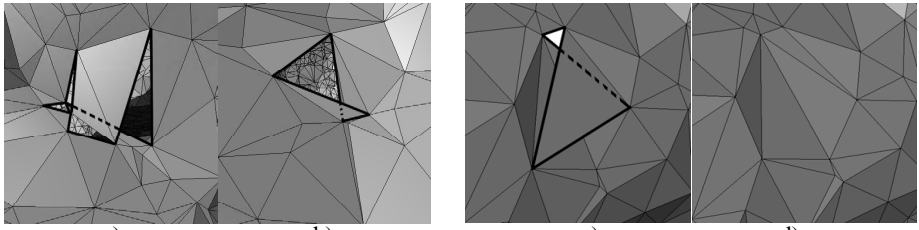


Fig. 9. a), b) The overlapping triangles in the surface reconstructed using COCONE, c) the overlapping triangles from our reconstruction without and d) with postfiltering

6 Conclusion

We have presented two improvements to the manifold extraction step in surface reconstruction problem. When the surface is not well sampled or a noise is present, some surface triangles are missing or other hybrid triangles appear. Our tests show that it is not a problem only of our algorithm. The prefiltering improvement helped us with the missing triangles in the smooth places and it makes the manifold extraction a little faster. The postfiltering improvement prevents from the creation of overlapped triangles, the holes are later triangulated. That would be the next step of our development, to use the existing structures and to better develop this step, or to use some existing algorithm, for a hole retriangulation.

References

1. M. E. Algorri, F. Schmitt. Surface reconstruction from unstructured 3D data. Computer Graphic Forum (1996) 47 - 60
2. N.Amenta, M.Bern, M.Kamvysselis.A new Voronoi-based surface reconstruction algorithm. SIGGRAPH (1998) 415 - 421
3. N.Amenta, M.Bern. Surface reconstruction by Voronoi filtering. Discr. and Comput. Geometry 22 (4), (1999) 481 - 504
4. N. Amenta, S. Choi, T. K. Dey, N. Leekha. A simple algorithm for homeomorphic surface reconstruction. 16th. Sympos. Comput. Geometry (2000)

5. N.Amenta, S.Choi, R.Kolluri. The PowerCrust. Proc. of 6th ACM Sympos. on Solid Modeling (2001)
6. F.Bernardini, C.Bajaj. A triangulation based. Sampling and reconstruction manifolds using a-shapes. 9th Canad. Conf. on Comput. Geometry (1997) 193 - 168
7. F.Bernardini, J.Mittleman, H.Rushmeier, C.Silva, G.Taubin. The ball pivoting algorithm for surface reconstruction. IEEE Trans. on Vis. and Comp. Graphics 5 (4) (1999)
8. J.D.Boissonat. Geometric structures for three-dimensional shape representation. ACM Trans. Graphics 3, (1984) 266 - 286
9. B.Curless, M.Levoy. A volumetric method for building complex models from range images. SIGGRAPH (1996) 302 - 312.
10. T.K.Dey, J.Giesen, J.Hudson. Delaunay Based Shape Reconstruction from Large Data. Proc. IEEE Sympos. in Parallel and Large Data Visualization and Graphics (2001)
11. T.K.Dey, J.Giesen, N.Leekha, R.Wenger. Detecting boundaries for surface reconstruction using co-cones. Intl. J. Computer Graphics & CAD/CAM, vol. 16 (2001) 141 - 159
12. T.K.Dey, J.Giesen. Detecting undersampling in surface reconstruction. Proc. of 17th ACM Sympos. Comput. Geometry (2001) 257 - 263
13. T.K.Dey, S.Goswami. Tight Cocone: A water-tight surface reconstructor. Proc. 8th ACM Sympos. Solid Modeling application (2003) 127 - 134 [27]
14. H.Edelsbrunner, E.P.Mücke. Three-dimensional alpha shapes. ACM Trans. Graphics 13 (1994) 43 - 72
15. H.Edelsbrunner. Weighted alpha shapes. Technical report UIUCDCS-R92-1760, DCS University of Illinois at Urbana-Champaign, Urbana, Illinois (1992)
16. H.Hoppe, T.DeRose, T.Duchamp, J.McDonald, W.Stuetzle. Surface reconstruction from unorganized points. Computer Graphics 26 (2) (1992) 71 - 78
17. J.V.Müller, D.E.Breen, W.E.Lorenzem, R.M.O'Bara, M.J.Wozny. Geometrically deformed models: A Method for extracting closed geometric models from volume data. Proc. SIGGRAPH (1991) 217 - 226
18. S.Muraki. Volumetric shape description of range data using "Blobby model". Comp. Graphics (1991). 217 - 226
19. R.Mencl, H.Müller. Graph based surface reconstruction using structures in scattered point sets. Proc. CGI (1998) 298 - 311
20. M.Varnuška, I.Kolingerová. Improvements to surface reconstruction by CRUST algorithm. SCCG Budmerice, Slovakia (2003) 101-109

Expression of a Set of Points' Structure within a Specific Geometrical Model

Jean-Luc Mari and Jean Sequeira

LSIS Laboratory - LXAO Department
Méditerranée University
ESIL, Campus de Luminy, case 925, 13288 Marseille cedex 9, France,
jlmari@esil.univ-mrs.fr,
: http://esil.univ-mrs.fr/~jlmari/

Abstract. We present a new approach based on a multi-layer model to represent the structure of an object defined by a cloud of points. This technique focuses on the ability to take into account both the global characteristics and the local specificities of a complex object, on topological and morphological levels, as well as on the geometric level. To do that, the proposed model is composed of three layers. We call the boundary mesh the *external layer*, adding a multiresolution feature. We enhance this representation by including an internal structure: the *inner skeleton*, which is topologically equivalent to the input object. In addition to that, a third layer links the structural entity and the geometrical crust, to induce an intermediary level of representation. This approach, which overcomes the limitations of skeleton based models and free-form surfaces, is applied to classical and medical data through a specific algorithm.

1 Introduction

The two main classes of approaches to represent a shape are skeleton based models and free-form surfaces, which are manipulated with control points. The first class allows us to determine the structure of an object, limiting the surface considerations. The second class permits to control the boundary of the shape precisely, but often neglects the general vision we expect. Most of the modeling techniques compensate these lacks by adding features which make the underlying model more complex, but none of them really takes into account both the global and the local specificities of the shape.

In this paper, we focus on designing a modeling method which integrates the local geometric characterization of one class, and the ability to represent the topology and the morphology of the other class.

We use this specific model to reconstruct a set of points. On the opposite of classical methods, our aim is not only to characterize the boundary of the related object. Even if this is sufficient to represent the solid, we want a topological and a morphological descriptor of the object, as well as a coherent link between the various structures. To do that, we introduce three layers whose roles are to take into account these features.

In Sect. 2, we skim over different models' formalizations, to extract the key characteristics of our approach. In Sect. 3, we describe the principle of a new modeling approach. We define the model by developing its three main entities: the *inner skeleton*,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 156–163, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the *external layer* and the *transition layer*. Then we detail the reconstruction process itself in Sect. 4. We finally validate the approach with examples (one of them being from real medical data) in Sect. 5.

2 Overview

Implicit Surfaces and Skeletons. Most of the skeleton based techniques use the formalism of implicit surfaces. These surfaces, whose skeleton is usually a set of geometrical primitives, have been more and more employed in computer graphics over the past 15 years. These approaches have several advantages, such as providing a compact and volume-based representation of the object. Moreover, the skeleton supplies the topological prior assumption and the structural information. However these surfaces are rather dedicated to represent smooth objects, sharp edges being more difficult to obtain. Moreover, the characterization of details implies to take into account a large number of primitives. These arguments point out the fact that it is difficult to get a *local control* with such surfaces because of the growing number of primitives to consider.

Surfaces Defined by Control Points. The most common formalism to represent a free-form object consists in using a parametric surface. Among these kinds of surfaces, defined by control points, there are the classical Bézier, B-splines and NURBS. The control points imply an intuitive and precise appreciation of the shape to model. Particularly, it is for this reason these surfaces are frequently used in CAD/CAM: they permit an intrinsic *local control* on the object's geometry. However, the modeling of objects with complex topologies (like with branching shapes or holes) keeps on being a traditional problem. Moreover, this kind of representation is hard to apprehend on a global level. If we wish to deform an initial model by stretching it, we have to move the right set of points, and to verify the induced transformations on the shape. Although high level operators exist to solve this problem (like *warping*), these manipulations are fussy and hard to calibrate (taking into account a small set of points or the whole object).

Similar Techniques. Among the approaches that emerge from the problematics to take into account both global and local characteristics of a 3D object, there are *hierarchical B-splines* (*H-splines*) [1], *simplex meshes* [2] and enhanced implicit methods (*skins* [3]). Multiresolution techniques are frequently used according to the complexity of the geometrical details and the size of the data. Most of the time, it consists in representing the object within several levels of detail [4,5,6]. In this frame, the formalism of *subdivision surfaces* is more and more used, as it presents a powerful multiresolution feature. It is nowadays used into a large panel of applications in computer graphics [7,8].

An Interesting Compromise. The approaches with skeletons show that instead of considering the skeleton's instability, we should take into account that it is not well adapted to surface phenomena, but rather to shape description. The control points approaches provide a very good control on differential properties, but hardly any on topology and morphology. We need a model that integrates in a coherent way the global and local characteristics of these two approaches.

3 The Multilayer Model

The model must allow us to control these three concepts in a coherent framework: the topology (to be able to model complex shapes with no prior assumption), the morphology (including a shape descriptor) and the geometry (integrating a crust entity for precise handling). Moreover, according to the size of the data to model, we wish to include the multiresolution feature. We can define the aims of our modeling approach in two points: we want both the global structure and the boundary surface; and we want to detach the surface representation from the global shape's.

The model is composed of three layers (see Fig. 1-a). The first layer, the internal one, that we call *inner skeleton* (layer L_I). It defines the global structure of the shape, on topological and morphological levels. The *external layer* (layer L_E) characterizes the local variations of the shape's surface, regardless of the skeleton. The *transition level* L_T represents the articulation between the internal and external layers.

The goal of this split between local and global characterization is that the local perturbations on the surface do not deteriorate the global shape descriptor which is the inner skeleton, and that the transformations on this inner structure are propagated on the external layer.

Inner Skeleton. The inner skeleton L_I is a *homotopic kernel* enhanced by morphological features. We define the inner skeleton L_I on a structural level as a 3-*complex*, i.e. a set of tetrahedrons, triangles, segments and points. The edges define the connectedness relations between vertices. When three neighbors are connected, we obtain a triangle, and when four neighbors are connected, we obtain a tetrahedron (Fig. 1-b). A cycle of edges defines a one-holed surface.

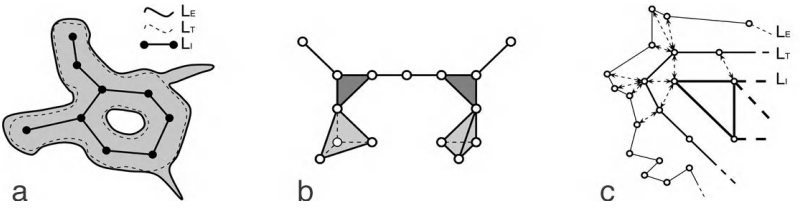


Fig. 1. (a) Scheme of the 3-layer model. (b) Example of 3D structure for the inner skeleton. (c) The three layers and the links between the primitives

External Layer. The external layer L_E is a simple triangulation. The vocation of this layer is to define the geometry of the object, given as a set of unstructured points. The multiresolution feature is supported by this layer. Considering $L_E^{100\%}$ the maximal level of detail for this layer, the various resolution levels are defined by $L_E^{r\%}$, with $r_{min} \leq r \leq 100$, r_{min} being the percentage below which the layer becomes non-manifold.

Transition Layer. The transition layer L_T represents an intermediate geometrical level and a structural entity which makes the link between the global definition and the local characterization of an object. The inner and external representation levels are both as important and we want to characterize the articulation between them.

In the geometrical frame, we define the transition layer as an intermediary triangulation between the two other entities. It induces a structural link allowing us to go from one layer to another (i.e. an element of the external layer can refer to an element of the inner layer and vice versa). Furthermore, we set the equality $L_T = L_E^{r_{min}\%}$ between the transition layer and the minimal resolution level of the external layer. The fact that L_T is the most simplified level of L_E provides a natural evolution from $L_E^{100\%}$ to the transition layer L_T by mesh reduction in the reconstruction process.

Transition Graph. In addition to the previous geometrical definition, the transition layer includes a particular data structure: the underlying graph G_T linking the two skeletons allows us to set coherent relations within the object (see Fig. 1-c). The edges of this graph are defined by a *shortest distance* criterion [9].

4 Reconstruction

In this section, we develop the reconstruction process related to the proposed model in order to express how the three layers are obtained, starting from a 3D set of points.

The main idea is to get an expression of the structure of an unorganized cloud of points, given as input data. We do not simply want to characterize the boundary of the shape with the external layer. Even if it is sufficient to represent the related solid, we attempt to exhibit a topological and morphological descriptor of the object. This point is fundamental, because efficient techniques of reconstruction are numerous, but they principally focus on the surface reconstruction without taking the structure into account.

4.1 Specific Process

The process is composed of two stages, themselves being most of the time well known techniques in computer graphics. However, concerning the first stage, we developed an original method to obtain the inner skeleton by doing an *homotopic peeling* applied to an octree.

The following algorithm consists of two independent steps: the extraction of the inner skeleton and the characterization of the crust (from the external layer to the transition layer). We illustrate the process by considering the *Stanford bunny*[1] data in the initial form of a cloud of points (cf. Fig. 2-a) . It is a simple object, of genus zero, but more complex examples are given in the next section.

```
1. Inner skeleton extraction
   a- Embedding the cloud of points into a digital volume
   b- Coding the digital volume into an octree
   c- Peeling the octree (topological conservation)
      Interactive selection of the ending points (morphological characterization)
   d- Computing the complex related to the reduced octree
         ⟹ Inner skeleton
2. Computing the external layer and the transition layer
   a- Polygonizing the cloud of points
         ⟹ External layer
   b- Mesh simplification
         ⟹ Transition layer
   c- Computing the transition graph
```

[1] available at the URL: `http//www-graphics.stanford.edu/data/`

4.2 Inner Skeleton Extraction

The inner skeleton extraction goes through a conversion of the data into the discrete space $\mathbb{Z}^3$. We wish to characterize the topology and the global shape, by keeping only a small set of relevant voxels (or groups of voxels).

Step 1: Embedding the Cloud of Points into a Digital Volume. Let C be the initial cloud of points. We adopt the same principle found in [10] to embed C into a digital volume V (Fig. 2-b).

Step 2: Octree Conversion. The digital volume is then converted into an octree [11], to keep in mind the notion of *voxels grouping*. A single voxel does not represent an important morphological detail, but on the contrary, a block of voxels defines a large area that has to be included into the structure of the inner skeleton (Fig. 2-c).

Step 3: Interactive Thinning of the Octree. We adopt a classical thinning process, whose asset is to supply a homeomorphic entity related to the object. There is no algorithm dealing with octrees in such a way, so we use a simple algorithm initially designed for digital volumes, and we adapt it to work on octrees (by modifying the neighborhood relationship and the local thinning criterion). The thinning (or the *peeling*) problematics first appeared in $\mathbb{Z}^2$ [12]. The principle is to delete the pixels that do not affect to topology of the object. Such points are said *simple*: when erased, no holes are created and no components are disconnected. [13] extended this concept to $\mathbb{Z}^3$. To extend this criterion to an octree, we defined in [9] the ω-neighborhood as the equivalent to the 26-neighborhood in $\mathbb{Z}^3$, and the $\overline{\omega}$-neighborhood as the equivalent to the 6-neighborhood (for the complementary object, i.e. the background).

We expose the algorithm used to peel an octree, which is derived from the initial thinning process into a digital volume. In our case, the size of the octree elements intervenes.

```
PeelingGridSize := 1
Repeat
   FindTargets()
   Repeat
      finish := TRUE
      For each target o of Ω
         If Simple(o)
         Then
            Delete(o)
            finish := FALSE
   Until finish
   PeelingGridSize := PeelingGridSize * 2
Until PeelingGridSize > 2^max order
```

The routine `FindTargets()` computes the list of the octree elements of the object whose size equals the current peeling grid size (i.e. elements that can be erased). The Figs. 2-d,e illustrate two steps of the bunny peeling with this algorithm. We enhance this algorithm by adding the interactive feature that allows the user to guide the process. It consists in setting representative elements, which contribute to the morphology as *ending points*. When the current element is such a point, it cannot be removed. The Figs. 2-f,g show the result when four ending points are selected by the user: the ears, the nose and the tail. The resulting octree is really homeomorphic to the initial object, and it supplies a good morphological characterization.

Step 4: Computing the Complex Related to the Reduced Octree. The last step to get the inner skeleton L_I consists in computing the reduced octree into a complex. This is done thanks to the ω-neighborhood. Edges, triangles and tetrahedrons are created according to the adjacency of the octree elements (Fig. 2-h).

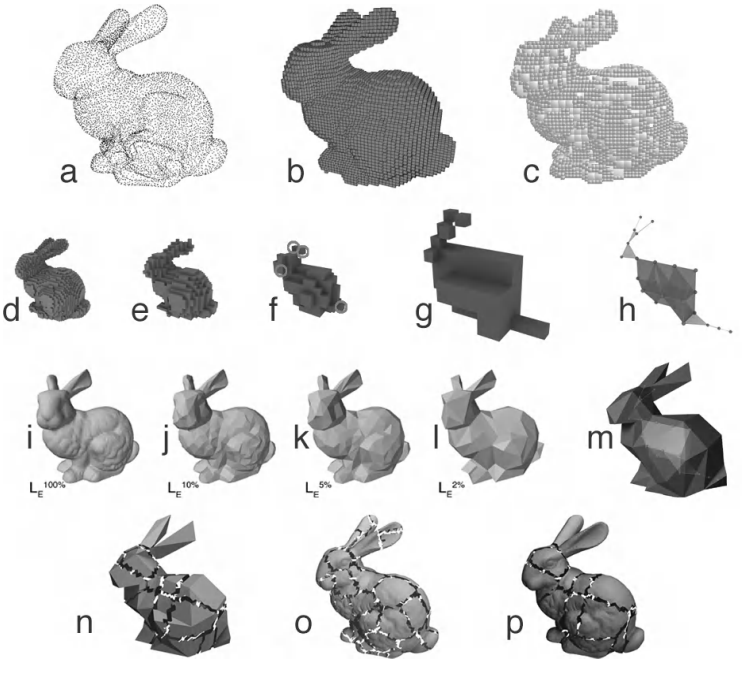


Fig. 2. (a) Initial cloud of points. (b) Digital volume. (c) Octree of the bunny. (d,e) Two peeling steps. (f,g) Setting four ending points. (h) Inner skeleton. (i,j,k,l) Some LoDs of the bunny. (m) The transition layer of the bunny. (n,o,p) Distribution of the points between the three layers

4.3 From the External Layer to the Transition Layer

Step 1: Polygonizing the Cloud of Points. The finest geometrical characterization of the external layer $L_E^{100\%}$ is a classical triangulation of the data points (see Fig. 2-i). We used the *Cocone* module described in [14].

Step 2: Multiresolution and Transition Layer. The Figs. 2-j,k,l show some reduced meshes of the bunny, until $L_E^{1\%}$ which is the last step before the triangulation becomes non-manifold (cf. Fig. 2-m). We go from the external to the transition layer, and we set $L_T = L_E^{1\%}$ to define the geometrical characterization of the transition layer. All the mesh simplifications in this paper have been done using *QSlim* module [5].

Step 3: Computing the Transition Graph. To make the link between structural and geometrical levels of the model, the last step of the whole process is the computation of the transition graph G_T. The Figs. 2-n,o,p show the distribution of the points of L_T

according to the points of L_I (n), the points of L_E according to the points of L_T (o), and the points of L_E according to the points of L_I (p).

5 Validation and Examples

In addition to the bunny's example, we go further into the validation of our approach by taking three other examples. The horse and the dragon (a one holed object) are classical clouds of points from the Stanford database, and the last example comes from medical imaging (a *foetus's heart*). Such an organ presents a complex structure. The table below illustrates the number of vertices according to the layers for each example. The Fig. 3 shows the inner skeleton, the transition layer and the external layer of the three objects.

Number of vertices	L_I	L_T	$L_E^{100\%}$
Bunny	24	76	7030
Horse	97	244	48485
Dragon	63	108	44315
Heart	1310	1596	35260

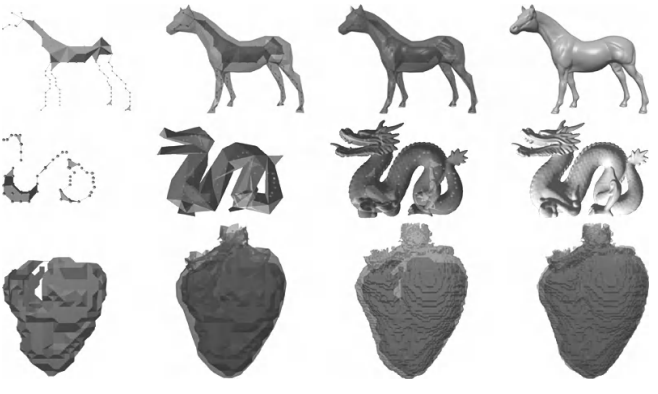


Fig. 3. Expression of the structure of four objects within the reconstruction process

6 Future Work

At the current level of the approach's elaboration, we envisage several points which we feel are necessary to be dealt with as future work.

The good morphological properties of the inner skeleton could be used in a *shapes' recognition* module. The aim being to classify an object on topological and morphological criterions, the process could lie on a catalog of typical objects arranged according to shape indications. This analysis perspective could be applied to computer vision, robotics, etc.

The major work to do in the future is the *animation* of a reconstructed object. This can only be done if the object is well positioned (as an evidence, it cannot work on the bunny, as legs are not defined by the inner skeleton). For example, to animate a character

expressed by the 3-layer model in a standing position (limbs being well defined), specific rules have to be determined to move external and transition layers after handling the skeleton.

7 Conclusion

We have presented a new shape formalism, which aims to give an explicit expression of an unstructured cloud of points, through three entities. The external layer defines the crust of the object in a classical way. In addition to that, the transition layer and, above all, the inner skeleton, allow us to get the structure of the object. This is done not only by characterizing the global shape, but also by a specifying a relevant topological entity. Moreover, we have validated our approach on typical data and on complex medical data. We believe the underlying model can be applied to several applicative domains, taking part of the inner skeleton's assets.

References

1. D. R. Forsey, R. H. Bartels: Hierarchical B-spline refinement. Computer Graphics (Proceedings of SIGGRAPH'88) (1988) **22**:4 205–212
2. H. Delingette: Simplex Meshes: a general representation for 3D shape reconstruction. INRIA Reseach Report 2214 (1994)
3. L. Markosian, J. M. Cohen, T. Crulli, J. Hugues: Skin: a constructive approach to modeling free-form shapes. Computer Graphics Proceedings (SIGGRAPH'99) (1999) 393–400
4. W.J. Schroeder, J.A. Zarge, W.E. Lorensen: Decimation of triangle meshes. ACM Computer Graphics (SIGGRAPH '92 Proceedings) **26** (1992) 65–70
5. M. Garland, P.S. Heckbert: Surface Simplification Using Quadric Error Metrics: Computer Graphics (Annual Conference Series) **31** (1997) 209–216
6. M. Eck, T. DeRose, T. Duchamp, H. Hoppe, M. Lounsbery, W. Stuetzle: Multiresolution analysis of arbitrary meshes. SIGGRAPH'95 Proceedings (1995) 173–181
7. M. Lounsbery, T. DeRose, J. Warren: Multiresolution analysis for surfaces of arbitrary topological type. ACM Transactions on Graphics **16**:1 (1997) 34–73
8. A. Lee, W. Sweldens, P. Schrǫder, L. Cowsar, D. Dobkin: MAPS: Multiresolution Adaptive Parameterization of Surfaces. Computer Graphics Proceedings (SIGGRAPH'98) (1998) 95–104
9. J.-L. Mari: Modélisation de formes complexes intégrant leurs caractéristiques globales et leurs spécificités locales. PhD Thesis, Université de la Méditérranée - Aix-Marseille II (France) (2002). `http://www.esil.univ-mrs.fr/~jlmari/`
10. E. Bittar, N. Tsingos, M.-P. Gascuel: Automatic reconstruction of unstructured 3D data: Combining a medial axis and implicit surfaces. Computer Graphics Forum (Eurographics'95 Proc.) **14** (1995) 457–468
11. D. Meagher: Geometric modeling using octree encoding. IEEE Computer graphics and Image Processing **19**:2 (1982) 129–147
12. A. Rosenfeld: A characterization of parallel thinning algorithms. Information Control **29** (1975) 286–291
13. G. Bertrand: Simple points, topological numbers and geodesic neighborhoods in cubic grids. Patterns Recognition Letters **15** (1994) 1003–1011
14. N. Amenta, S. Choi, T.K. Dey, N. Leekha A simple algorithm for homeomorphic surface reconstruction. 16th ACM Symposium on Computational Geometry (2000) 213–222

Effective Use of Procedural Shaders in Animated Scenes

Polina Kondratieva, Vlastimil Havran, and Hans-Peter Seidel

MPI Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany.
{polina,havran,hpseidel}@mpi-sb.mpg.de

Abstract. Complex procedural shaders are commonly used to enrich the appearance of high-quality computer animations. In traditional rendering architectures the shading computation is performed independently for each animation frame which leads to significant costs. In this paper we propose an approach which eliminates redundant computation between subsequent frames by exploiting temporal coherence in shading. The shading computation is decomposed into view-dependent and view-independent parts and the results of the latter one are shared by a number of subsequent frames. This leads to a significant improvement of the computation performance. Also, the visual quality of resulting animations is much better due to the reduction of temporal aliasing in shading patterns.

1 Introduction

Creation of photo-realistic images with low computational requirements is one of the main goals of computer graphics. Procedural shaders can be used as an effective mean for rendering high-quality realistic images due to some distinct advantages, such as simplicity of procedural shading for arbitrarily complex surfaces and the possibility to change the shaded surface with time, viewing angle or distance [7].

The approach presented in this paper extends the research in Havran et al. [3]. We show that a significant part of shading computation can be reused in subsequent frames. There are two different techniques to prepare the shader data for reusing. While the first approach is based on the 3D-texture notion, the second one is related to analytical splitting of the procedural shader into the view-dependent and view-independent parts. The algorithm of reusing the view-independent data for both techniques is similar.

A key aspect of most procedural shading is the use of a shading language which allows a high-level description of the color and shading of each surface. Shaders written in the RenderMan Shading Language can be used by any compliant renderer, no matter what rendering method it uses [8]. For this reason all examples of shaders in this paper are similar to the *RenderMan* shaders.

The paper is organized as follows. Section 2 discusses the properties of three rendering architectures. The algorithm of reusing the view-independent data is

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 164–172, 2004.
© Springer-Verlag Berlin Heidelberg 2004

presented in Sect. 3. Section 4 describes preprocessing techniques for the preparation of shading data for reusing. Examples of such a preparation are also presented in this section. The achieved results are shown in Sect. 5. Finally, Sect. 6 concludes the paper and proposes some directions for future work.

2 Related Work

Here we discuss the advantage of Efficient Spatio-Temporal Architecture for Rendering Animation (ESTARA) [3] compared to the well-known rendering architectures using procedural shaders, such as REYES [2] and Maya renderer [5].

Rendered images have the property of similarity between the consecutive frames known as *temporal coherence*, which can be used to accelerate the rendering. Both Maya and REYES architectures compute images of animation sequence frame by frame. On the other hand, ESTARA exploits the property of temporal coherence by splitting the shading function into the view-independent and view-dependent parts, whereas the first one is computed only once for a given sample point and the second one is recomputed for each frame. In this way, ESTARA outperforms both REYES and Maya by considerable reducing the computational cost of shading computation as well as the temporal aliasing (flickering). ESTARA can be used for pixel based renderers including bidirectional path tracing, ray tracing, etc. Here we extend the ideas in [3] for ray tracing with procedural shaders.

3 Algorithm of Reusing the View-Independent Data for Procedural Shaders

3.1 Notions of View-Dependent and View-Independent Shader Parts

Before discussing the features of the algorithm of reusing the data, we give a brief definition of *view−dependent* (VD) and *view−independent* (VI) data with respect to procedural shaders. The symbolic notation used throughout the rest of the paper and adopted from RenderMan Shading Language is shown in Table 1.

The computation of the color for each sample can be split into two parts: VD and VI. The VI data do not change when the camera moves (see Fig. 1(a)). A simple example of the VI data is the *diffuse surface color* which can be computed according to Lambertian law as follows:

$$diffuseCol \;=\; Kd \;\cdot Cs \;\cdot cos(\angle \mathbf{N}, \mathbf{L}) \;. \tag{1}$$

In contrast to the VI data, VD data change whenever the camera moves. More precisely, the VD data depend on the reciprocal location of the surface point (hit point) and camera (viewer) in 3D space. A simple example of the VI data is the *specular shading color* of the surface. According to the well-known Phong model, specular color can be computed as following:

$$specularCol \;=\; Ks \;\cdot\; specular \cdot cos^{n}(\angle \mathbf{R_m}, \mathbf{V}) \;. \tag{2}$$

In the following chapters we show examples of shader decompositions into the VD and VI data.

Table 1. Symbolic notation used in the document

Symbol	Description
Ka, Kd, Ks	coefficients for the ambient, diffuse and specular color
$rough$	roughness coefficient
Kt, Kr	refraction and reflection coefficients
$\mathbf{T}$, $\mathbf{R}$, $\mathbf{R_m}$	transmitted, reflected and mirror-reflected ray directions
Cs, Os	surface self color and opacity
Ci, Oi	incident ray color and opacity
$specular$	specular color of the surface
$\mathbf{N}$, $\mathbf{N_f}$	geometric and face-forwarded normals
$\mathbf{L}$, $\mathbf{V}$	incoming light and opposite view directions
$\mathbf{P}$	hit point position on the surface of an object
du, dv	changes in surface parameters
$\mathbf{dPdu}$, $\mathbf{dPdv}$	derivatives of surface position along u, v directions.

3.2 Algorithm of Reusing the Shading Data

The problem of wasting time for unnecessary recomputations of unchanged data is addressed by the algorithm of reusing the shading data described in detail here. The main idea of the algorithm is to save the VI data into the cache and then reuse it for subsequent frames.

The aliasing artifacts caused by poor sampling of procedural textures as well as ordinary 2D-textures can be reduced by increasing the number of samples, thus increasing rendering time. By reusing the VI data for shading in ESTARA we can decrease the computational cost for one sample. In addition, since the algorithm of reusing the VI data spreads the shading data into the time domain (Fig. 1(b)), similarly to [6], the temporal aliasing known as flickering and scintillation is highly reduced.

Let us describe the algorithm of reusing in more detail. For a sequence of N frames (N camera positions) the VI component of the pixel color in a given frame can be reused for subsequent frames. In this way, for the first frame of a sequence the ray with ID number unique for each hit point is shot and the color of the hit point is computed. The VI data of the computed pixel color is saved into the cache with the corresponding ray ID as a search key.

Afterwards the hit point is reprojected onto the image plane of subsequent frames. Note, that due to the fact that the camera position can change within the sequence of frames, the positions of the correspondent reprojected pixels on the image plane for subsequent frames are different. After the reprojection the hit point is checked for occlusion. If the ray is not occluded, the VI data can be used for shading computation.

For a range of frames the VI data are possibly combined with recomputed VD data to get the pixel color. This reusing of the VI data for shading is performed for all the pixels in the sequence of frames. Since the samples are obtained in a low cost, the total time of computations decreases.

An example of hit point reprojection for subsequent camera positions is shown in Fig. 1(b). For more details, see original paper [3]. Since any pixel

of an image in a sequence of frames can be accessed by the reprojection, all the image data for a sequence of frames are saved in the main memory.

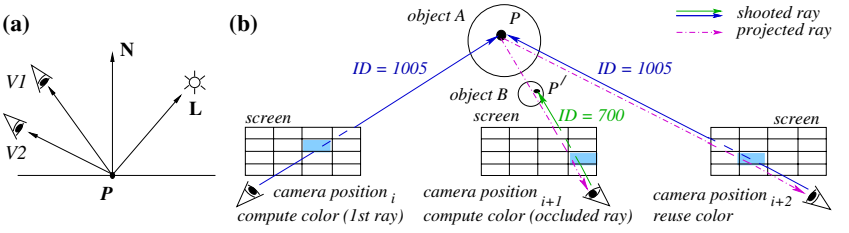


Fig. 1. Illustrations for the algorithm of reusing the data: (a) Example of the incoming/outgoing vectors for a given hit point **P**, (b) Hit point reprojection for subsequent camera positions followed by reusing or recomputation the shading data

3.3 Speedup Analysis

Let us now compare the timings for shading computation required for the traditional frame-by-frame approach and for the proposed algorithm of reusing the VI data. Suppose the pixel color computation for each camera position in this case takes T_s^1 time. Since, without reusing, the color of each sample for each camera position should be recomputed, the time T^1 required to compute the pixel color for $\boldsymbol{n}$ camera positions is:

$$T^1 = T_s^1 \cdot \boldsymbol{n} \ . \tag{3}$$

If the algorithm of reusing the data is involved, the situation changes. For the first camera position times T_{vd} and T_{vi} are required to compute the VD and VI data correspondingly, $T_{combine}$ to combine these parts, and T_{save} to save the VI data into the cache. For the remaining $\boldsymbol{n}-1$ camera positions times T_{vd}, $T_{combine}$ are needed as above, and $T_{extract}$ is required to extract (reuse) the VI data from the cache. Thus, the time needed to compute the pixel color for the first (T_s^2) and for the remaining $\boldsymbol{n}-1$ (T_s^3) camera positions can be calculated as follows:

$$T_s^2 = T_{vd} + T_{vi} + T_{combine} + T_{save}, \quad T_s^3 = T_{vd} + T_{extract} + T_{combine} \ .$$

The total time T^2 with reusing the VI data is then:

$$T^2 = T_s^2 + T_s^3 \cdot (\boldsymbol{n} - 1) \ . \tag{4}$$

The speedup of shading computation achieved by ESTARA with the algorithm of reusing the VI data can be evaluated from (3) and (4). It is clear that $T_s^1 < T_s^2$ and $T_s^1 > T_s^3$. Hence, $T_s^1 \cdot (\boldsymbol{n} - 1) > T_s^3 \cdot (\boldsymbol{n} - 1)$. If $\boldsymbol{n} > \lceil (T_s^2 - T_s^3)/(T_s^1 - T_s^3) \rceil$, then $T^1 > T^2$. Therefore, the maximum theoretical speedup achieved by applying the algorithm of reusing the data for $\boldsymbol{n}$ camera positions can be evaluated as follows:

$$\lim_{n \to \infty} speedup = \boldsymbol{n} \ \cdot \ T_s^1/(T_s^2 \ + \ (\boldsymbol{n} - 1) \cdot T_s^3) = T_s^1/T_s^3 \ . \tag{5}$$

Formula (5) shows that the computational cost of rendering can be reduced by the algorithm of reusing the shading data. Note that the main point is to use a fast data structure for saving the VI data, otherwise no speedup can be achieved. For this purpose some kind of the fixed-size cache with LRU replacement policy [4] is used in ESTARA.

4 Preprocessing Techniques for Shading Data Reusing

We distinguish two different procedural shader classes: representable as 3D-textures and non-representable as 3D-textures.

4.1 Shaders Representable as 3D-Textures

The main feature of these shaders is that all the properties of the shader are defined by the complex color (Cs in formula (1)), which represents some pattern on the surface and it is independent of both $\mathbf{V}$ and $\mathbf{N}$.

Analysis of the VD and VI data described in Sect. 3.1 allows to define whether a given shader is representable as a 3D-texture. So, the shader can be represented as a 3D-texture if it has the following properties:

- its VD data contain only the computation of the glossy specular color,
- the computed complex diffuse color does not depend on $\mathbf{V}$ and can be used together with the function $diffuse()$ in the same way as Cs color,
- it does not create any surface displacements ($\mathbf{N}$ perturbations).

Good examples of such shaders are *CWoodTexture* [11], *CStoneTexture* [12], and *CCobbleStoneTexture* [10]. The pseudo-code of the function which computes the complex diffuse color for the *CCobblestoneTexture* is shown in Fig. 2.

```
CobbleStone_DiffuseColor(float Kd, jitter, sscale, ttscale, txtscale; color selfCol, varCol)
  {
     Scale P with txtscale, u with sscale, and v with ttscale
     Compute Voronoi noise voronoi(ss, tt,jitter, f1, spos1, tpos1, f2, spos2, tpos2)
     paintColor = selfCol,    Cgrout = selfCol · 0.5
     paintColor += (cellnoise([ spos1-67,tpos1+55 ])+1.5)· varCol · 0.5
     //Create cellular pattern with f2-f1.
     blendval = smoothstep( 0.03, 0.07, f2-f1)
     diffuseColor = Kd · (paintColor · blendval + Cgrout · (1— blendval))
  }
```

Fig. 2. Example of procedural 3D-texture

If the shader is representable as a 3D-texture the complex diffuse color can be saved into the cache as a simple Cs color and then reused for the next frames.

4.2 Shaders Non-representable as 3D-Textures

The shaders of this class have implicit VD and VI data closely interacting with each other; the shading computation for them is decomposed into *layers*. There is a great variety of non-representable as 3D-textures shaders: some of them have

```
RCSkin( float Kd, eta, thickness, angle, Xrough, Yrough, maxfreq, blemishfreq, blemishthresh,
      blemishopac, oily, brightness, poresfreq, poresthresh, poresdepth; color sheen, Cs,Os)
 {
      //--- layer 0 - pores ------------------
      Spread the pores over the surface, compute displaced normal (N_N)

      //--- layer 1 - skin main color --------
      color_skin = Cs, Oi = Os

      //--- layer 2 - blemishes subsurface ---
      PP = transform(object, P)· blemishfreq;   turb = 0;
      for (f = 1; f < maxfreq; f *= 2)   turb += abs(snoise(PP· f)) / f;
      blemishColor = spline(color_1,...,color_n,turb^blemishtresh)
      Compute Kr, Kt, R, T for the view ray
      illuminance cycle(P, N_f, π/2)
      {
        if(cos(∠(L + V), N_f) > 0)
          glossy = Kr · sheen · Cl · cos(∠L, N_f) · cos^4(∠(L+V),L)
        glossy += 2 · Kr · sheen · Cl · abs(cos(∠L, N_f))
        Compute Kr_2, Kt_2, R_2, T_2 for L, and single scattering approximations s_1, s_2, s_3
        glossy += blemishColor · Cl · cos(∠L, N_f) · Kt · Kt_2 · (s_1+s_2+s_3)
      }
      Mix color glossy with color_skin

      //- layer 3 - anisotropic Ward model ---
      Compute anisotropic directions anisDir(dPdu, N_f,angle), XaDir, YaDir
      illuminance cycle(P, N_f, π/2)
      {
        Compute Ward coefficient rho(XaDir, N_f, YaDir, L, V)
        if(Light source is specular) Canis = (Cl · cos(∠L, N_f) · rho)/(4 · Xrough · Yrough)
      }
      Diff = Kd · diffuse(N_f))
      color_skin += (Canis · 0.1 · oily + Diff) · brightness
      Save Diff, blemishColor, Os, N_N, P, XaDir, YaDir into the cache
 }
```

Fig. 3. An example of non-representable as 3D-texture 4-layer shader

only one layer, such as *velvet* [11], some others consist of many complicated layers involving Fresnel function, Ward reflection model, and/or some other functions for anisotropic reflection, such as *RCSkin* [11]. Despite of their complexity, even these shaders can be usually split into VD and VI data.

At this point, let us consider an example of complicated shader - *RCSkin*. It consists of four layers and the computation of the color for each layer is highly time consuming. The pseudo-code of the function which calculates the pixel color for this shader is presented in Fig. 3.

The *RCSkin* shader presented in Fig. 3 computes a number of specific VD functions. For example, the Ward reflection model and the Fresnel function are quite computationally demanding. Fortunately, after the careful analysis the following components of the shader can be considered as the VI data:

- At ***layer 0*** the displaced normal $\mathbf{N_N}$ for pores,
- At ***layer 1*** the skin color *color_skin* and *Oi*,
- At ***layer 2*** the *blemishColor* computed by the *spline* function for 3D-vectors,
- At ***layer 3*** the anisotropic directions (**XaDir**, **YaDir**).

In the same way all the other non-representable as a 3D-texture shaders can be split into the VD and VI data. The main point is that the time required to compute the VI data should be greater than the time required to insert/extract the data from the cache.

5 Results

We have verified the efficiency of the described algorithm of reusing the shading data embedded by ESTARA on three scenes applying different shaders. A computer with processor Intel(R) Xeon(TM) CPU 1.706MHz and 1024MB of memory was used for rendering. All the shaders were taken from the [1], [9], or RenderMan sites [10],[11], [12] and adapted for our renderer, as described above.

At the first step, the speedup of the shading color computation for each shader was evaluated for a simple scene avoiding the visibility test. The timing results in seconds for all the shaders are shown in Table 2. In Table 2 column NOREUSE presents the shading time results for the traditional frame-by-frame approach. Column REUSE shows the timing results for the algorithm of reusing the shading data. Column SPEEDUP depicts the speedup (SPEEDUP = NOREUSE/REUSE).

At the next step, the speedup evaluation was accomplished for two more complex scenes: scene *Face*, containing *RCSkin* and *greenmarble* shaders, and scene *Interior*, containing all the shaders from Table 2 except *RCSkin*, performing the visibility test. Note, that speedup was achieved by the combined reusing of the VI data and visibility information. The timing results in seconds for both scenes are presented in Table 3.

Table 2. Timing and speedup results for shaders applied to the simple scene

	50 camera positions			100 camera positions		
SHADER	NOREUSE	REUSE	SPEEDUP	NOREUSE	REUSE	SPEEDUP
blocks	25.47	8.02	3.18	52.00	15.96	3.26
carpet	27.35	9.84	2.78	55.14	18.92	2.91
cmarble	77.89	18.24	4.27	154.76	34.93	4.43
colormarble	73.36	18.29	4.01	148.43	34.93	4.25
cobblestone	26.99	11.83	2.28	62.25	27.20	2.29
greenmarble	50.97	17.44	2.92	97.91	33.06	2.96
spatter	21.57	12.33	1.75	47.55	27.17	1.75
stone	15.68	13.13	1.19	32.91	25.99	1.27
velvet	12.27	12.44	0.99	24.45	23.84	1.03
wood	25.44	15.27	1.67	51.34	29.85	1.72
RCSkin	97.12	46.98	2.07	185.75	78.05	2.38

Table 3. Timimg results for scenes *Interior* and *Face*

	Time		
Scene	REUSE	NOREUSE	SPEEDUP
Face	2.24e+03	2.73e+03	1.22
Interior	1.25e+05	3.28e+05	2.62

The resulting images for the scene *Face* with different values of parameters for *RCSkin* shader are presented in Fig. 4(a) and 4(b). The resulting images for the scene *Interior* with procedural shaders are depicted in Fig. 4(c).

6 Conclusion and Future Work

In this paper we have described techniques, which significantly reduce the computational cost of procedural shading in animation rendering, while improving

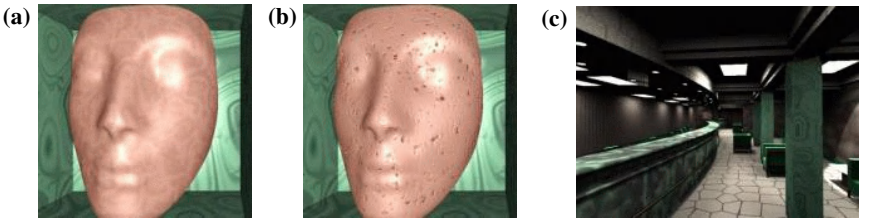


Fig. 4. Images rendered by ESTARA with reusing: (a), (b) scene *Face* with different parameter settings for *RCSkin* shader; (c) scene *Interior*

the quality of resulting images in the context of ESTARA rendering architecture [3]. The speedup is achieved by splitting the shader into two parts: the view-dependent (VD) and the view-independent (VI).

Applying the algorithm of reusing the shading data for ray tracing a moderately complex scene with procedural shaders, we received significant speedup up to a factor of 2.62. Since the VI data of the color are the same for the pixels corresponding to the reprojection of the shaded point in the object space to the image plane of subsequent frames, the temporal aliasing (flickering) is reduced.

The main disadvantage of the proposed algorithm is the fact that all shaders should be split into the VD and VI data manually. Intuitively, this time consuming and laborious process could be done by computer. We envision the automation of the splitting process as the next step in the development of the algorithm of reusing the shading data.

Acknowledgments. The authors would like to thank Karol Myszkowski for helpful discussions and suggestions during the preparation of the paper.

References

1. Apodaca, A.A., and Gritz, L. Advanced RenderMan. Morgan Kaufmann, 1999
2. Cook, R.L., Carpenter, L., Catmull, E.: The Reyes Image Rendering Architecture. ACM Computer Graphics SIGGRAPH'97 Proc. (1987) 95-102
3. Havran, V., Damez, C., Myszkowski, K., and Seidel, H.-P.: An Efficient Spatio-temporal Architecture for Animation Rendering. Eurographics Symposium on Rendering (2003)
4. Knuth, D.E.: The Art of Computer Programming, Vol.3 (Sorting and Searching). Addison-Wesley Series (1973).
5. Sung, K., Craighead, J., Wang, C., Bakshi, S., Pearce, A., and Woo, A.: Design and implementation of the Maya Renderer. Pacific Graphics'98 Proc. (1998) 150-159
6. Martin, W., Reinhard, E., Shirley, P., Parker, S. and Thompson, W.: Temporally coherent interactive ray tracing. Journal of Graphics Tools **2** (2002) 41-48
7. Olano, M.: A Programmable Pipeline for Graphics Hardware. PhD dissertation, University of North Carolina, Chapel Hill (1998)

8. Olano, M., Lastra, A.: A Shading Language on Graphics Hardware: The PixelFlow Shading System. ACM Computer Graphics SIGGRAPH'98 Proc (1998) 159-168
9. Upstill, S.: The RenderMan Companion. A programmer's Guide to realistic Computer Graphics. Addison-Wesley publishing company (1990)
10. http://www.cs.unc.edu/~{}stewart/comp238/shade.html
11. http://www.renderman.org/RMR/Shaders/
12. http://www-2.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15462/arch/sgi_65/prman/lib/shaders/stone.sl

Real-Time Tree Rendering

I. Remolar, C. Rebollo, M. Chover, and J. Ribelles

Departamento de Lenguajes y Sistemas Informáticos,
Universitat Jaume I, Castellón
{remolar,rebollo,chover,ribelles}@uji.es

Abstract. Interactive rendering of outdoor scenes is currently one of the most important challenges in computer graphics. This article presents a new method of real-time visualization of trees and plants that combines both multiresolution modeling techniques and the dynamic generation of impostors. In our method, trees and plants are represented by continuous view-dependent levels of detail. This enables us to represent many complex tree models with variable resolution. The number of primitives rendered per tree is reduced according to their importance in the scene without loss of leafiness. Furthermore, trees are visualized using dynamic impostors that take advantage of the frame-to-frame coherence inherent in tree-dimensional scenes. The impostors avoid the need to redraw all the geometry of the scene continuously. This method permits visualization of outdoor scenes with a high number of trees in interactive applications such as computer games or virtual reality, adapting the level of detail to the capability of graphic systems.

1 Introduction

Rendering outdoor scenes with a great number of trees or plants cannot be accomplished in real-time with present graphic hardware. Many current interactive applications such as flight simulators, virtual reality environments or computer games happen in outdoor scenes and one of the essential components in these scenes is the vegetation. Tree modeling has been widely investigated [1], [2], and very realistic representations have been demonstrated. However, tree models are formed by such a vast number of polygons that real-time visualization of scenes with trees are practically impossible.

Various solutions have been researched for interactive visualization of complex models. Multiresolution modeling [3] is one of them. It makes possible the increase or reduction of the number of polygons in a geometric model according to their importance in the scene. Another approach is image-based rendering approximations [4],[5],[6] which substitutes some of the geometry of the scene by images textured on polygons. In this way, the models are represented in two dimensions. When the observer moves toward the object, the lack of details causes the realism of the scene to be lost. Point-based rendering is used in order to achieve interactive visualization of complex objects and is based on substituting the rendering primitive triangle mainly

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 173–180, 2004.
© Springer-Verlag Berlin Heidelberg 2004

by points [7],[8],[9]. But if the trees situated very close to the observer are rendered with points or lines, the details of the leaves can not be appreciated.

This paper presents a general scheme for rendering outdoor scenes with vegetation in real-time. It combines both multiresolution modeling and image-based rendering techniques. Our approach allows us to represent both near and distant trees with a high level of detail. The geometry of the trees is represented by means of dynamic view-dependent levels of detail. In this way, the number of the rendered polygons vary depending on the importance of the tree in the scene. The used multiresolution schemes accept a polygonal description of the tree, not exclusively restricted to the generation with L-systems.

Furthermore, in order to make possible real-time rendering, dynamically generated impostors [10] are combined with the geometric representation. These impostors have been successfully used in real-time cloud rendering [11]. They used to replace distant geometric models in the past, but in our method they are also used to represent some of the closer parts of nearby trees. This approach avoids redrawing the entire scene geometry in every frame because they can be re-used over multiple ones.

2 Previous Work

Research aimed at vegetation can be divided in two major fields: the generation of plants and trees, and their visualisation. Vegetation modelling has been explored extensively. The most important works in this field are Lindermayer-systems [1], used for generating realistic models of trees. Other solutions combine grammar based modelling with a graph description [2]. Apart from the great number of works that have appeared in the literature, some commercial applications have been developed for modelling trees. Some of the most important are OnyxTree (*www.onyxtree.com*), AMAP (*www.bionatics.com*), Xfrog (*www.greenworks.de*) and SpeedTreeRT (`www.idvinc.com`).

The analysis of previous work related with our approach can be divided as geometric representation and image-based rendering methods.

Geometric representation: Level of detail rendering [5] is one of the most popular methods to reduce the complexity of polygonal data sets in a smart manner.

The continuous multiresolution models presented thus far deal with general meshes and cannot be applied effectively to such scenes. Hoppe [13], Xia [14] and El-Sana [15] use methods based on the union of pairs of vertices in the construction process simplification. Luebke [16] uses a method based on vertex clustering: a set of vertices is collapsed into one vertex. These methods can not process the foliage without degradation of appearance [9].

Another technique in interactive visualization of complex plant models uses point-based rendering based on the idea of substituting the basic primitive triangle by points or lines. Reeves and Blau [17] rendered trees using small disks representing the foliage, and Weber and Penn [18] used sets of points for the leaves and lines for the tree skeleton. Stamminger and Dettrakis [8] visualize plants with a random sample set of

points. One of the most recent works in this field has been presented by Deussen et al. [9]. Their approach combines geometry with points and lines.

Image-Based Rendering methods: Billboarding is one of the most frequently used techniques due to its simplicity. The trees are reduced to images textured on polygons, which always maintain their orientation towards the observer. However this technique has great deficiencies, because the models are represented in two dimensions. When the observer moves toward the object, the lack of details produces a loss of realism in the scene.

Layered depth images [19], LDI, store in each pixel of the image a 2D array of depth pixels. In each depth pixel are stored, in proximity order to the point of view, the surfaces that appear in that image. But the LDI files created for trees are excessively large. Another similar method however using Z-buffers is presented by Max [4].

Jakulin [5] presents a method based on images with alpha-blended textured polygons. Lluch et al. [6] present a method based on a hierarchy of images obtained from pre-processing the botanical tree structure (a L-system) and storing the information in a texture data tree.

3 System Overview

The trees used in our study were modelled by the Xfrog application [2]. They are very realistic, but generally consist of more than 50.000 polygons each. This is a disadvantage when it comes to generating images in an interactive way.

Trees can be separated in two different parts:

- the solid component of the tree, the trunk and the branches, and
- the sparse component, the foliage.

In this work, two different multiresolution schemes have been used to represent each of the parts. The trunk and branches are represented by a set of connected surfaces, usually triangular meshes. There are many multiresolution models in the literature that deal with this kind of mesh. In this work the Multiresolution Ordered Meshes "*MOM"* [20] has been used to model them. An efficient implementation of the View Dependent Multiresolution Model for the Foliage, "*VDF"* [21], is used for the crown representation.

In order to construct a multiresolution model, a simplification method must be used. Qslim [22] is used to obtain the sequence of vertex simplification required by *MOM*. *VDF* is created from a sequence of leaf collapses determined off-line by the simplification algorithm called Foliage Simplification Algorithm, "*FSA"* [12]. The key to this simplification algorithm is *leaf collapse*. Two leaves are transformed into a single one, so that the area of the new leaf is similar to the area formed by the two leaves initially. This reduces loss of appearance at the end of the process. An error function is used to determine which pair of leaves will be simplified to create a new one.

The data obtained are stored once and many representations of this object can be visualized by instancing. A distribution editor is used to distribute every one of this

instances in a scene, in order to obtain a tree population. Fig. 1 displays the system outline.

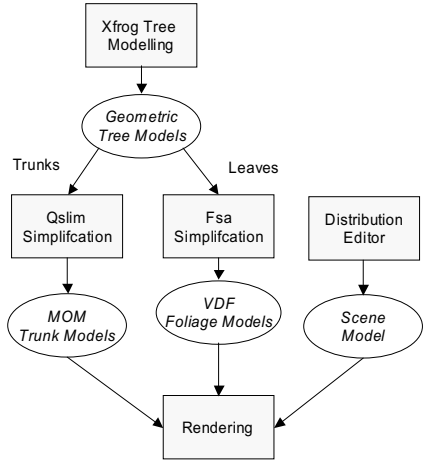


Fig. 1. System Overview

The tree instances are visualized with variable resolution depending on some criteria, such as distance from the observer or importance of the object in the scene. Furthermore, in order to increase the frame rate of the scene, dynamically generated impostors [10] are automatically combined with geometry. The zone nearest to the viewer will be represented by geometry, and the zone farther away, by an impostor. This impostor moves backwards when the viewer zooms in, disappearing when the observer is very close to the tree. In this case, the tree is represented only by geometry. On the contrary, when the viewer is moving away from the object, the impostor moves towards until the tree is represented only by the impostor.

Multiresolution modelling allows us to reduce the generation time of the impostors because the trees are visualized with less number of polygons in accordance to the importance in the scene.

4 Rendering Algorithm

Dynamically generated impostors have been combined in our rendering algorithm with multiresolution modelling in order to obtain real-time rendering of scenes. Impostors have been mainly used for objects situated far from the viewer. In our scheme, they are also used for close objects. Another novelty is that we use movable impostors: each one can move from the front part of the tree to the back (Fig 2). When it is situated in the front part of the foliage, tree is represented only by a textured image, and when the impostor is getting the back limit of the foliage, only by geometry. The movement of the impostor is conditioned on the distance to the observer. As the viewer moves towards the tree, the impostor moves backwards. The part of the tree

nearest the viewer is represented by geometry, and the rest of it, textured on the impostor.

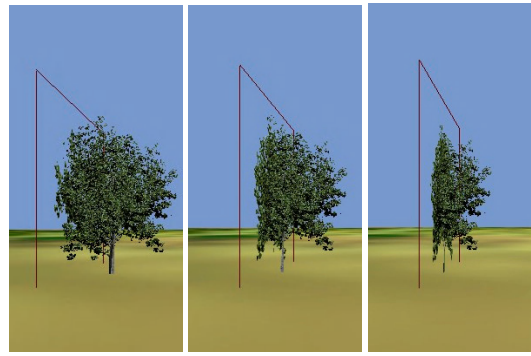


Fig. 2. Example of movable impostor

As far as the use of the dynamic impostors and the resolution variable in the tree, it is necessary to determine when they are no longer valid. We have used the same error measures used by Harris and Lastra in their work on real time clouds visualisation [11]. An impostor stops being valid when the tolerance of one of the following errors is surpassed:

Translation Error. This measures the validity of the representation in transferring movements of the observer. This error is calculated taking the angle that forms the present position of the observer with the position when the impostor was generated.

Resolution Error. This measures the validity of the resolution of the texture of the impostor. It is calculated by the following equation:

$$\text{resTexture} = \text{resImage} \times \frac{\text{objSize}}{\text{objDistance}} \tag{1}$$

While the impostors are valid, they are rendered as billboards in order to avoid unnecessary popping effects.

5 Results

The presented work has been implemented with OpenGL on a PC with Windows 2000 operating system. The computer is a dual Pentium Xeon at 1.8GHz. with a graphical processor NVIDIA Quadro2 with 64MB.

The trees used in our experiments is formed by 88.443 polygons. Their foliages were initially formed by 20.376 leaves, that is, 40.752 triangles and their trunks by 47.691 polygons.

The tests measure the frames per second in a scene where the number of trees is increased. The camera follows a random path in this scene. The trees are rendered using three methods. Fig 3 shows the results for these test.

- *Geometry:* every tree in the scene is represented with the maximum level of detail.
- *Multiresolution Modelling*: the level of detail of the tree is adapted to its importance in the scene. Trees close to the viewer are represented with high detail and it degrades as the observer is moving backwards.
- *Impostors + Multiresolution*: Trees are rendered using both multiresolution modelling and dynamically generated impostors.

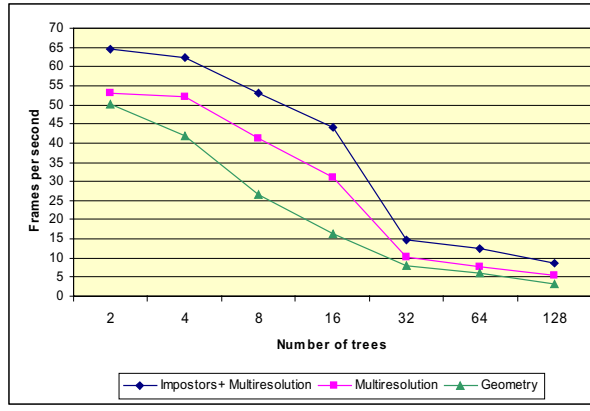


Fig. 3. Results of the experiments

As we can see in Fig 3, multiresolution modelling increases the frame rate. This is because the number of polygons that are drawn diminishes without reducing the realism of the scene. This chart shows that frame rate improves remarkably via the use of impostors. This allows us to render a scene with a high number of trees in outdoor scenes. In Fig 4, a scene of our test is shown, where the trees are represented by impostors. Fig 5 shows an example of the detail that can be obtained with our rendering method.

6 Conclusions and Future Work

In this paper, we have presented a system for the realistic visualisation of trees and plants in real time. The developed technique combines two suitable methods of acceleration for current graphics hardware: multiresolution modelling techniques and dynamically generated impostors.

The main contributions of our work can be summarized as:

- A general framework for the representation of outdoor scenes with trees. The method accepts a polygonal description of the tree, not exclusively restricted to the generation with L-systems.

- Efficient implementation of *VDF*, a multiresolution model specifically designed for foliage. It supports view-dependent visualization and multiples instances of a tree model can coexist.
- Use of movable, dynamically generated impostors both for distant and near trees. These are combined with geometry without producing artifacts. The management of the texture memory is optimized using this type of impostors. In our approach no images of different views of the tree are stored in memory as a part of preprocessing, as is the case in other methods [6].
- This general framework can easily be extended to other modeling representations such as particle systems.

The next step to improve the realistic representation of the trees, is to take account of illumination. We are developing solutions based on the use of light maps for trunks and terrain and per-vertex lighting for leaves.

Furthermore, the visualisation of scenes with multiple trees requires occlusion culling methods and multi-layered impostors.

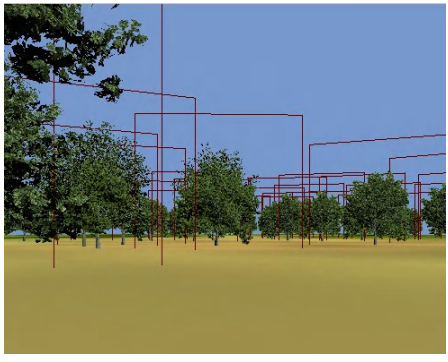


Fig. 4. Scene where impostors are visualized

Fig. 5. Detail of the trees

Acknowledgements. This work has been supported by the Spanish Ministry of Science and Technology grants TIC2001-2416-C03-02 and TIC2002-04166-C03-02, the Fundació Caixa-Castelló Bancaixa grant P1-1B2002-12 and the Integrated Action HU2000-0011.

References

1. P. Prusinkiewicz, A. Lindenmayer, “The algorithmic beauty of plants”, Ed. Springer-Verlag, New York, 1990.

2. B. Lintermann, O. Deussen. “Interactive modeling of plants”, IEEE Computer Graphics and Applications, vol. 19(1), 1999.
3. J. Ribelles, A. López, Ó. Belmonte, I. Remolar, M. Chover. "Multiresolution Modeling of Arbitrary Polygonal Surfaces: A Characterization", Computers & Graphics, 26(3), pp. 449-462, 2002.
4. N. Max, K. Ohsaki. “Rendering trees from precomputed Z-buffer views”. Eurographics Workshop on Rendering 1996, pp. 165-174, 1996.
5. A. Jakulin. “Interactive Vegetation Rendering with Slicing and Blending”. Eurographics’2000, Short presentations 2000.
6. J. Lluch, E. Camahort, R. Vivó. “An Image-Based Multiresolution Model for Interactive Foliage Rendering”, Journal of WSCG04, vol. 12(3), pp. 507-514, 2004.
7. J. Weber, J. Penn. “Creation and rendering of realistic trees”, Proc of SIGGRAPH’95, 1995, pp. 119-128.
8. M. Stamminger, G. Drettakis. “Interactive sampling and rendering for complex and procedural geometry”, Rendering Techniques 2001, Eurographics, Springer-Verlag, pp. 151-162, 2001
9. O. Deussen, C. Colditz, M. Stamminger, G. Dettrakis. “Interactive Visualization of Complex Plant Ecosystems”, Proc. of the IEEE Visualization Conference, 2002.
10. G. Schaufler, “Dynamically Generated Impostors”, GI Workshop, Modeling - Virtual Worlds - Distribute Graphics 1995, pp. 129-136, 1995.
11. M. J. Harris, A. Lastra, “Real-Time Cloud Rendering”, Eurographics’2001,vol. 20(3), 2001.
12. I. Remolar, M. Chover, O. Belmonte, J. Ribelles, C. Rebollo, “Geometric Simplification of Foliage”, Eurographics'02 Short Presentations, pp. 397-404, 2002.
13. H. Hoppe, "View-dependent refinement of progressive meshes", Proc. of SIGGRAPH'97, pp. 189-198, 1997.
14. J. Xia., A. Varshney, “Dynamic view-dependent simplification for polygonal models”, Proc. of IEEE Visualization’96, pp. 327-334, 1996.
15. J. El-Sana, A. Varshney, “Generalized View-Dependent Simplification”, Eurgographics’99, pp. 131-137, 1999.
16. D. Luebke and C. Erikson, “View-Dependent Simplification of Arbitrary Polygonal Environments”, Proc of SIGGRAPH’97, pp. 202-210, 1997.
17. W. T. Reeves, R. Blau. “Approximate and probabilistic algorithms for shading and rendering structures particle systems”, Computer Graphics, Proc. of SIGGRAPH’85, vol. 19, pp. 313-322, 1985.
18. J. Weber, J. Penn. “Creation and rendering of realistic trees”, Proc. of SIGGRAPH’95, pp. 119-128, 1995.
19. J. W. Shade, S. J. Gortler, L. He, R. Szeliski. “Layered depth images”, Proc. of SIGGRAPH’98, pp. 231-242, 1998.
20. J. Ribelles, A. López, O. Belmonte, I. Remolar, M. Chover. "Variable Resolution Level-of-detail of Multiresolution Ordered Meshes" , Proc. of 9-th International Conference in Central Europe on Computer Graphics, vol. 2, pp. 299-306, 2001.
21. I. Remolar, M. Chover, J. Ribelles, O. Belmonte. “View-Dependent Multiresolution Model for Foliage”, Journal of WSCG03, vol. 11(2), pp. 370-378, 2003.
22. M. Garland, P.S. Heckbert, “Surface simplification using quadric error metrics”; Proc of SIGGRAPH’98, pp. 209-216, 1998.

A Brush Stroke Generation Using Magnetic Field Model for Painterly Rendering

Lee Soo Yeon, Yong Han Soon, and Yoon Kyung Hyun

221, HukSeok-Dong, DongJak-Gu, Seoul, Korea, 156-756
Computer Graphics Lab, CS&E, Chung-Ang University
{henim, hansoon, khyoon}@cglab.cse.cau.ac.kr
http://cglab.cse.cau.ac.kr

Abstract. In this paper, we propose a new method for generating stroke which is the core of creating an image with a hand-painted appearance from a photograph. Magnetic Field Model is used in generating strokes and it is possible to create unique and natural brush strokes. In order to determine the direction of a stroke, a Magnetic Force Direction Map is made in the form of various circular and emissive shapes based on this model. We present two methods for generating stroke according to this direction map. We are able to create idiomatic and characteristic brush styles of a real painter using these methods.

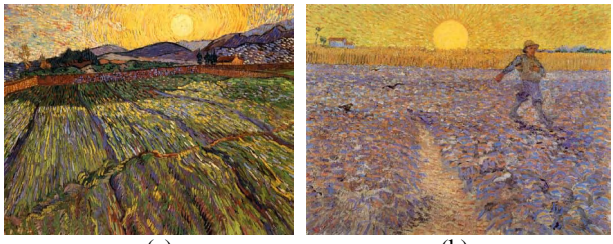
1 Introduction

There are a lot of existing techniques to create a painted version of a digital photograph using computer. Figure 1 is a painting of Van Gogh which consists of brush strokes. Brush stroke gives a painting characteristic into the image. Therefore, many algorithms are developed to express strokes. Existing methods generated strokes using both straight lines and curved lines. So far, the algorithms[1,2,3] proposed for creating brush strokes using image gradients. But these methods have disadvantage that the direction of the strokes may appear artificial and cannot overcome the goal of creating strokes that express the character of a painter. Cha[4] created strokes with the consistent directionality by the region. This algorithm also has the same consistent directionality by each segmented area and does not express the various forms of the strokes. Therefore, it becomes difficult to express the brush strokes of real paintings using the image gradients only.

We suggest a magnetic field model for creating idiomatic strokes that can be found on paintings. The magnetic field model is created by the physical force between the magnetic objects, and the Magnetic Force Direction Map(MFDM) used to determine the direction of the stroke is acquired from the model. There are two different approaches to create strokes with specific styles using this map.

A vector field is created using the magnetic field model to create strokes freely with various styles. And then the MFDM is constructed to be applied on the direction of the stroke with it. We introduce two different approaches to create strokes using it.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 181–188, 2004.
© Springer-Verlag Berlin Heidelberg 2004



(a) (b)

Fig. 1. Vicent van Gogh Oil on canvas, (a) wheat Rising with Sum, (b) Sower with Setting Sun, strokes around the sun are assumed the form of circular and emissive

The first approach is an automatic rendering method by the size of the brush. And the process is in the following. First, we acquire the gradient from the input image, and we reestablish it using MFDM which follows the magnetic field model. The reestablished direction draws a cubic B-Spline curve to create the stroke. In the second approach, the user adds strokes created by the MFDM on to the intermediate result image. The intermediate result image used for this process can be any image that was rendered by the existing algorithm. With the two approaches mentioned above, we can express the various types of strokes created by a human artist.

2 Magnetic Field Model

2.1 Shape of Magnetic Field

Magnetic force is a force that occurs between two magnetic objects and the space that is affected by it is called a magnetic field. A magnetic line of force is a virtual line used to indicate the magnetic force and the direction of the magnetic field is determined by the direction of the tangent line following the law of Ampere [5]. The magnetic force (B) at the location that is perpendicularly away by a distance r which is proportional to the force at the center of the magnetic field (I) and is disproportional to the distance from the conducting wire (r). It can be expressed by the following equation (1).

$$B = k\frac{I}{r} \tag{1}$$

A circular negative electric charge or positive electric charge grazing magnetic field is constructed around conducting wire where the current flows. The shape of the vector field also varies according to the number of conducting wire. Therefore, various shapes of vector fields can be created based on the magnetic field theory and the MFDM can be structured by this vector field to decide the direction of the stroke.

2.2 Vector Field Generation

The location of the conducting wire is set up as the seed point and the currency is used as the input for creating the vector filed of the magnetic field model.

Circular and Emissive

To create a circular vector field, the current at the seed point (I) and the conducting wire are used as the input. The distance affected by the current is calculated first and the directions of the tangent line for the magnetic field model are acquired to form the vector field. The direction of the magnetic field is acquired by using the fact that the tangent line of the circle is perpendicular to the half-circle crossing the point of contact and the characteristics of the vector. The direction and force of the user input currency become the parameter for creating the various number of vector field according to the number of seed points.

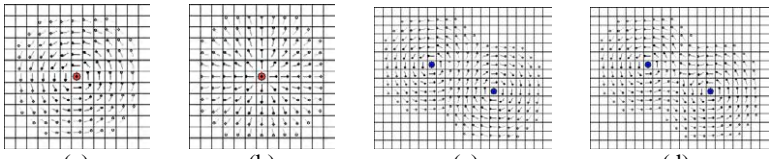


Fig. 2. Vector field. (a) Circular, (b) Emissive, (c),(d) User defined shape vector Field

Emissive vectors have either a positive electric charge or negative electric charge magnetic field as its model. The vector from a random point within the magnetic field to the input seed point can easily be acquired. (b) from Figure 2 shows the emissive vector fields and the magnetic field becomes larger as the currency becomes higher.

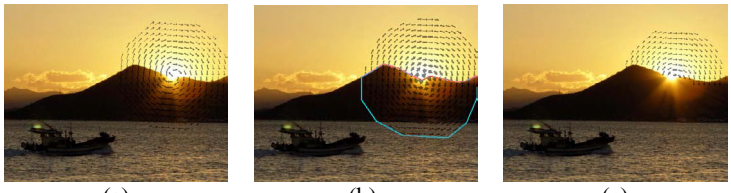
User defined shape

The conducting wire where the currency flows is used as the model, and the currency value is used as the seed point. All the vectors created by each currency from a random location P are obtained for calculating the vector to a new direction. The number of seed points, direction and force of the currency are acquired through user input for creating various vector fields. (c) and (d) from Figure 2 each show vector fields created by setting two seed points with different direction.

The acquired vector has size and direction. The directions follow the magnetic field model, and the size at each location is express by the force of the magnetic field. The force of the magnetic field is reduced as the distance from the seed point increases until it is not affected at all at one point. The force of the magnetic field at a random point within the magnetic field is assigned as a gray-level value and is calculated as the ratio of the magnetic force at its current location against the maximum magnetic force.

2.3 Magnetic Force Direction Map

The vector field created above becomes a main part of the magnetic force direction map. The MFDM is the same size as the input image, and it is consists of vector fields created by the magnetic force. Also, the user designates the area where the circular or emissive shape is applied. The created MFDM should go through a filtering stage for removing unnecessary images. Figure 3-(a) shows the result after a circular magnetic force direction map has been applied to the input image. Since there is no need for circular shaped strokes in drawing a mountain, the vectors are removed during the filtering stage. The removed area is replaced with gradients acquired from the input image and may be modified by the user.



(a) (b) (c)

Fig. 3. (a) Result image which is apply MFDM to the source image, (b) select the region that will be removed unnecessary vector, (c) result image of removing the vector for the selected region

3 Stroke Rendering

We introduce two algorithms for creating strokes. They are common in that they use the magnetic force direction map through referencing the magnetic field model. The first method applies an automatic method for rendering the whole canvas through converting the gradient of the selected region to the direction of the magnetic force direction map. The second method applies rendering by adding an extra stroke that follows the magnetic force direction map onto the original image. The two methods differ in the rendering method used.

3.1 Curved Brush Stroke

The user selects the area for either a circular or emissive stroke from the input image. A magnetic force direction map is created for the selected area and is applied to the input image to re-establish the direction of the gradient through the magnetic force direction map. The brush stroke is created using the cubic B-Spline curve. The direction of the stroke follows the magnetic force direction map and the other areas refer the direction of the gradient. The parameters of the stroke such as the starting point, color, length, and depth are applied by changing the curved line brush stroke creation algorithm[3].

3.2 Magnetic Field Based Stroke

The strokes around the sun as shown in Figure 9 (a), a work of Van Gogh, have different directions and colors from that used to draw the sky. The two new algorithms in this study, produce various shapes of strokes by adjusting the force of the magnetic field through changing the parameters. The result image using this method can reflect the original image, compared to other methods using image gradients.

Figure 4 (a) shows the process for creating a circular brush stroke. The starting point is set at a location with a distance of r from the seed point and creates strokes by locating a control point every l to the direction of the vector. The distance between the strokes θ is set randomly, and the condition for finishing the strokes from the circle is set as a pre-defined angle, which in this case was set to $30°$.

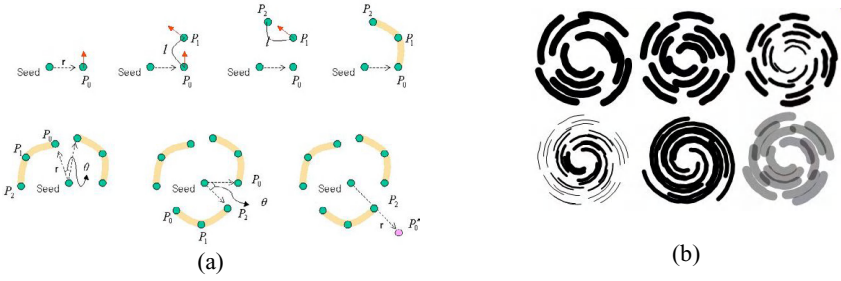


Fig. 4. (a) The process of creating circular stroke, (b) circular stroke

Figure 5 (a) shows the process for creating the emissive brush stroke, and the resulting brush stroke is shown in (b). The starting point is randomly selected from the created magnetic field, and the length and the thickness is determined by the force of the magnetic field. The strokes are created bi-directionally from the starting point and ends when it contacts the seed point.

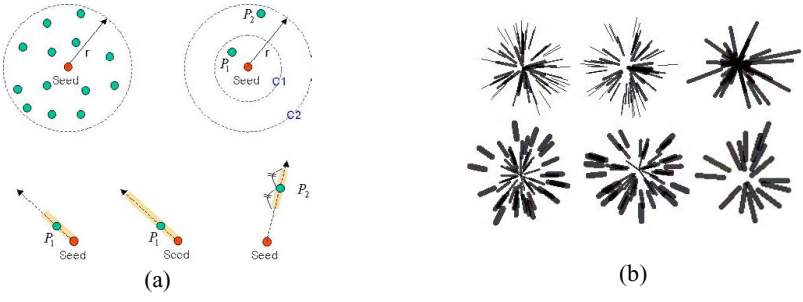
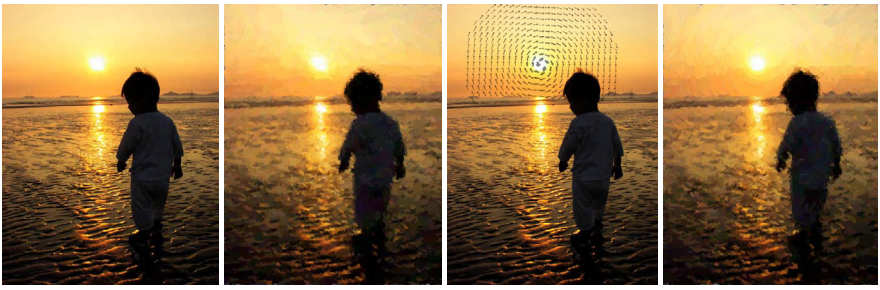
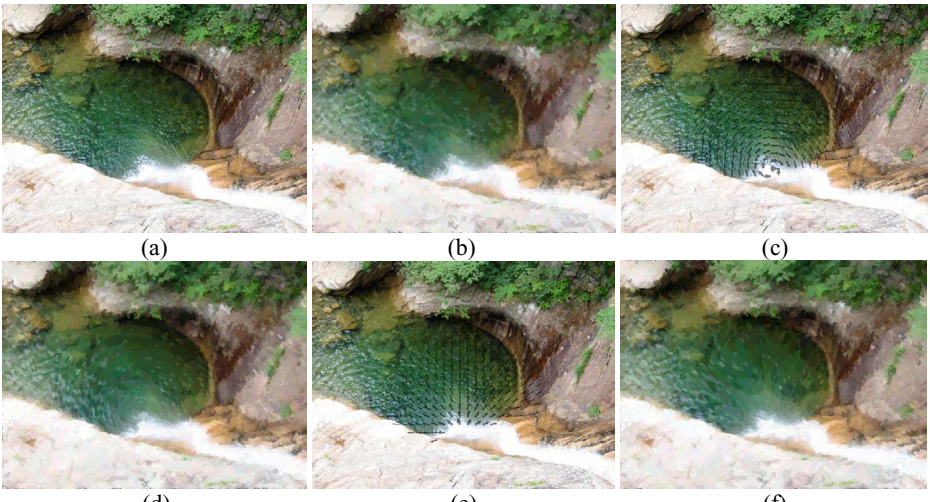


Fig. 5. (a) The process of creating emissive stroke, (b) emissive stroke



(a) (b) (c) (d)

Fig. 6. The process of applying circular MFDM to the source image. (b) the result of using gradient, (d) is the result image when the direction map of (c) was applied



(a) (b) (c)

(d) (e) (f)

Fig. 7. Result images of applying circular and emissive MFDM to the source image. (a) input image, (b) result image using gradient, (c),(e) are the results of applying circular and emissive MFDM to the source image, and the result rendered images are (d),(f)

(a) (b)

Fig. 8. Result which is rendered by using user defined MFDM. The result of (a) is each (b)

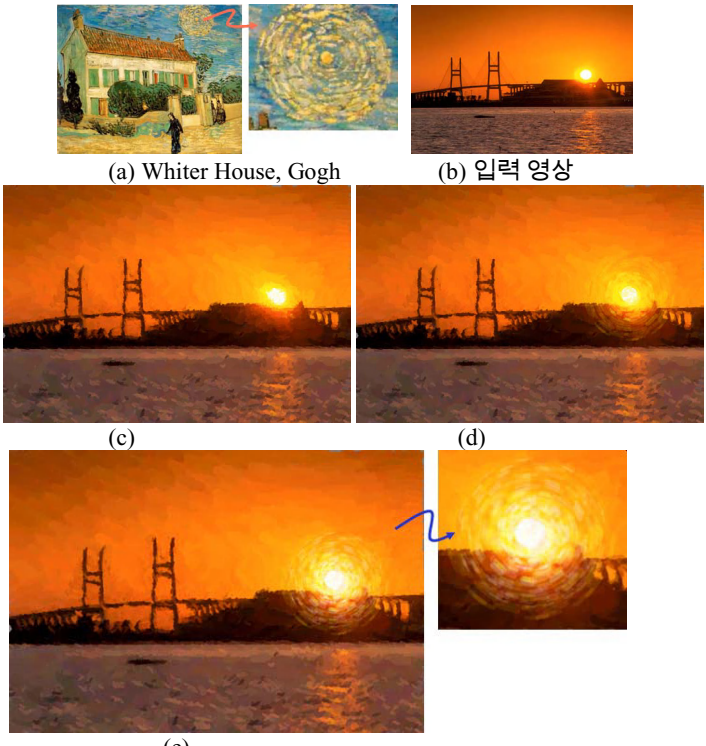


Fig. 9. (a) The painting of Gogh, (c) intermediate result image, (e) is created of circular stroke through (d)

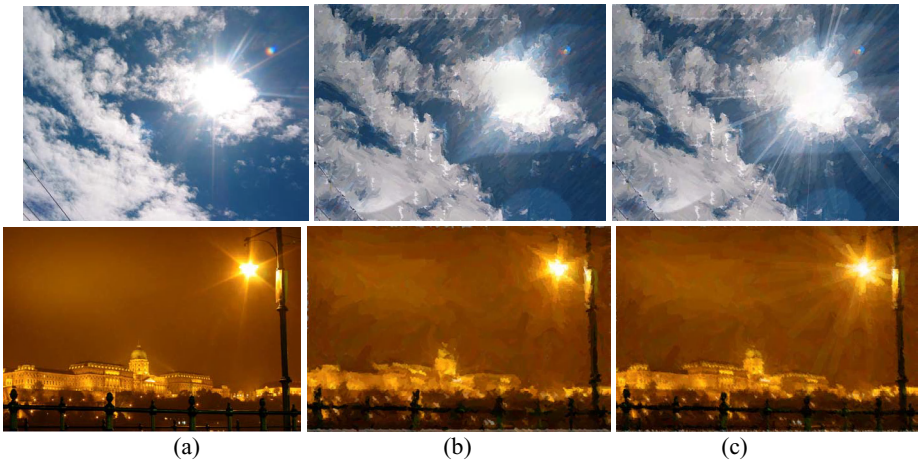


Fig. 10. (a) Rows are input image, (c) rows are the result image of applying emissive stroke to the result image which is rendered by using gradient as (b) rows

4 Result and Future Work

In this paper, a Magnetic Field Model is proposed to create strokes to make a painterly rendering image. As the result, we can represent the circular and emissive stroke and furthermore, we can illustrate idiomatic brush style of artists with various MFDM.

There is an advantage that the effect of a source image is reflected in the result image as in the figure 9 and 10. However, several troublesome tests have to be gone through in order to achieve these good result images.

It is complicated but possible to produce 2-dimension animations: the scene for the intensity of the light is gradually stronger or weaker and for the movement of the clouds, by strokes only from one source image. The images which are including streetlights or stars and sunset or sunrise are proper with the source image. And the blending phenomenon which is observed in actual paintings and illustration of quality and texture need to be simulated for the further study.

This work was partially supported by the National Research Laboratory program of the Korean Ministry of Science and Technology (No. 2000-N-NL-01-C-285) and by the ITRI of Chung-Ang University in 2004

References

1. Paul Haeberli, Paint by Numbers : Abstract image representations, SIGGRAPH90 Proceeding (1990) 207–214
2. Peter Litwinowicz, Processing Images and Video for an Impressionist Effect, In SIGGRAPH 97 (1997) 407–414
3. Aaron Hertzmann, Painterly rendering with curved brush strokes of multiple sizes, SIGGRAPH98 Proceeding (1998)
4. Jeong Seob Cha, A Directional Stroke Generation Algorithm for Painterly Rendering, LNCS2669 (2003) 326–335
5. Zoya Popovic, Branko D. Popovic, Introductory Electromagnetics, Prentice Hall (2000)

Reuse of Paths in Final Gathering Step with Moving Light Sources

Mateu Sbert and Francesc Castro

Institut d'Informàtica i Aplicacions, Universitat de Girona.
Edifici P4. Campus Montilivi. 17071 Girona, Spain.
{mateu,castro}@ima.udg.es

Abstract. In this paper we extend our previous result on reusing shooting paths when dealing with moving light sources in radiosity to the final gathering step algorithm, which combines a shooting solution and a gathering step. Paths are reused for each frame in both shooting and gathering steps, obtaining a noticeable reduction of cost in the computation of the solutions for every frame. Moreover, the correlation between consecutive frames happens to be higher than in the case of independent solutions, resulting in a clear reduction of the temporal aliasing effects in the animation.

Keywords: Radiosity, Random Walk, Animation

1 Introduction

The radiosity solution [3] can be obtained by combining shooting and gathering random walks [1]. One of these hybrid algorithms is the *final gathering step* [5]. From a radiosity solution obtained by shooting, a simple gathering step allows to obtain a final solution for a finer level of subdivision.

We present in this paper an acceleration strategy applied to this final gathering step algorithm and valid for animations involving moving light sources. Our strategy is based on the reuse of paths [2], which is done for both shooting and gathering ones. The reuse of shooting paths has been already presented in [6], and it will be described in next section. Here we incorporate the reuse of gathering paths, based on the use of each path for all the frames (that is, for all positions of the light source). The reuse of both shooting and gathering paths permits to calculate the solution for all frames with a noticeable reduction of cost regarding to the independent computation of each solution (classic approach).

Moreover, the reuse of the paths (in both shooting and gathering) produces a correlation between the solutions of consecutive frames. This fact clearly reduces the flickering effect observed in the video sequences. This is not applicable when computing each frame independently, as done in the classic approach.

This paper is organized as follows. In next section we will refer to previous work, including the reuse of shooting paths. The description of our contribution appears in Sect. 3. Section 4 presents the results, with both error graphs and images. Finally, in last section we present the conclusions and future work.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 189–196, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Previous Work

2.1 Final Gathering Step

Given a radiosity solution at a coarse level, say patches, the final gathering step is meant to solve at a finer level, say elements or pixels, for the interesting (or important) part of the scene. Rays can be cast from the eye to get a view dependent solution (usually to obtain a glossy solution enhancing a first radiosity solution) for the pixels in the screen plane as in the photon-map technique [7] , or from the patches of interest to get a new radiosity solution for the elements. On the other hand the first radiosity solution can be obtained by a deterministic method (for instance classic hierarchical radiosity) or a Monte Carlo method (random walk, hierarchical Monte Carlo) (see [5] for a short survey). When gathering from a complete radiosity solution, the gathering path can be limited to length one. The foundation in this case is the following one. Consider the radiosity equation:

$$B_i = R_i \Sigma_{j=1}^{n_p} F_{ij} B_j + E_j \tag{1}$$

where B_i, E_i, R_i are respectively the radiosity, emissivity and reflectance value for patch (element) i, and F_{ij} the form factor from patch (element) i to patch (element) j. We can solve with Monte Carlo equation (1), by choosing N patches j with a given probability vector (or pdf in the continuous case). Taking as probabilities the form factors F_{ij}, and considering a first step radiosity approximation B_j^*, the final gathering estimator for B_i results in:

$$\hat{B}_i = \frac{1}{N} R_i \Sigma_{j=1}^{N} B_j^* + E_j \tag{2}$$

2.2 Reuse of Shooting Paths in Light Source Animation

Shooting paths can be reused in the context of radiosity when dealing with moving light sources [6]. The main idea is that a path exiting from a point of the light source can be used to distribute power from every light source position, or in other words, to calculate the illumination for every frame. This means that each path will be used to build n paths, n being the number of light source positions (see Fig. 1). This involves a noticeable reduction of cost.

Let Φ be the power of a source taking n positions. M points are considered on the source. Let x be a point on the source repeated in each different position $x_1, x_2, ..., x_n$. Let j be the position taken into account and $F(x_j, y)$ the form factor from x_j to a point y (y being any point to receive direct illumination from x_j). The unbiased estimator for the power arriving to point y from x_j is (see [6])

$$\frac{F(x_j, y)}{F(x_1, y) + F(x_2, y) + ... + F(x_n, y)} \times \frac{\Phi}{M} \tag{3}$$

The analysis of the cost shows that the theoretical acceleration factor due to the reuse of the shooting paths is bounded by l, l being the average length of

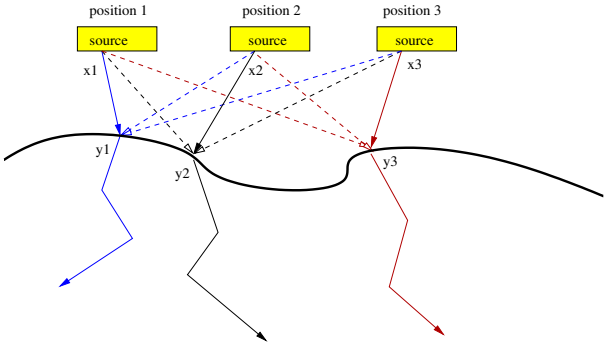


Fig. 1. 2D representation of a shooting random walk with reuse of the paths. Path from light source in position 1 (blue) is reused to distribute power from positions 2 and 3 of the light source, and the same with paths from position 2 (black) and 3 (red)

a path, when n grows. On the other hand, the expected error using the same number of paths per position will be the same in the best case. This corresponds to the limiting case in which all the light source positions were the same. In general, the more distant the positions, the bigger the error, due to the fact that when positions get distant, the sampling pdf for y goes away from the importance sampling function.

3 Reuse of Gathering Paths in Light Source Animation

The new contribution presented in this paper consists in the reuse of paths in the final gathering step algorithm (reviewed in Sect. 2.1) for the case of moving light sources. We will reuse both shooting and gathering paths. The reuse of shooting paths is done in the same way as explained in Sect. 2.2. The reuse of gathering paths is based on the fact that each gathering path can be used to gather the radiosity for each of the n frames.

3.1 Dealing with Direct Illumination

The final gathering step algorithm presents a problem when dealing with small light sources. This problem concerns the distribution of the direct illumination. Such illumination is dealt with when a gathering ray hits a light source. But the usual small size of the light source makes the number of gathering rays that hit the source from a given patch to be small, and thus the error for the received direct illumination is big. This results in an unacceptable aliasing in the final image, only avoided if the number of gathering rays is very large.

This problem can be solved by calculating direct illumination in the shooting step instead of in the gathering one. Thus, the gathering step will only deal with the indirect illumination. Final radiosity values will be obtained as the sum of

direct and indirect values. In this approach, the shooting step will work with two different levels of subdivision. The finest level will be used to store direct illumination, while the coarse one will be used for the complete shooting solution, which will be gathered in the final step. From here on, we will refer to the finest level subdivision units as *elements*, and to the roughest level ones as *patches*.

Since direct illumination solution will be calculated at the element level, and a complete shooting solution will be calculated at the patch level, the number of paths needed to get a determinated accuracy is not the same in each case. The number of shooting paths used to calculate direct illumination must be higher than the one for the complete shooting solution. A compromise consists of eliminating part of the paths after the first reflection, using a parameter λ $(0, 1]$ as the probability for a path to continue after the first hit. This means that each shooting path will have a double function. On the one hand, it will be used to distribute primary power from the sources to the elements, computing in this way the direct illumination. On the other hand, it will be used to distribute power (not only primary) from the sources to the patches in the scene. After the first hit, the continuation of paths will be decided according to λ. If a path survives, its carried power has to be multiplied by $1/\lambda$ to compensate for the reduction of the number of paths.

3.2 Reusing Gathering Paths

In the gathering step one-length paths are cast from each element i in the scene, and the nearest intersected patch j is obtained for each path. Then, radiosity of patch j is gathered by i. The reusing strategy proposed here consists of using every path to gather at once the radiosity for each of the n solutions corresponding to the n frames (n positions of the light source). Note that each element and each patch in the scene store an array of n radiosity accumulators (1 per frame). All n gathering solutions are computed at the cost of a single one. As the expected error for our solutions is exactly the same as from n results obtained in n independent executions, the expected speed-up factor of our strategy for the gathering step will be n.

Also, the reuse of the paths will provide us with an important additional gain: the correlation between consecutive frames, due to the reuse of the paths, makes the transition between frames smoother, avoiding the annoying flickering effect observed with independent solutions. This is clearly visible in the videos accompanying the paper.

3.3 Comparing the Expected Cost

Next we compare the cost of a radiosity solution S for n frames obtained by reusing both shooting and gathering paths against another radiosity solution S' obtained without reusing any path (that is, from independent executions for each frame in both shooting and gathering). Both solutions S and S' have been obtained using the same number of shooting and gathering paths per frame.

Let l be the average length of a path. Let N be the total number of shooting paths per frame, and let N_g be the total number of gathering paths per frame, being $N_g = kN$ for a positive value k (note that finding an optimal value of k is not an objective of this paper). We have to remark that, regarding to the number of shooting paths per frame, their reuse permits to reduce their number by a factor of n; that is, to obtain N effective shooting paths per frame, only $M = N/n$ actual paths have to be cast from each position of the source.

For the sake of simplicity in the following formulas, the cost has been expressed as the number of nearest intersection computations (actually, most of the real cost is due to such computations).

The results in [6] can be now generalized by considering the parameter λ (see Sect. 3.1). A simplification has been done by considering the cost of a visibility computation (used in the point-to-point form factor computation) to be the same as the cost of computing the nearest intersection (actually, and depending on the implementation, the cost of a visibility computation can be lower or, in the worst case, equal to the nearest intersection one). The cost of the shooting step with reusing will be $N(n + \lambda(l-1))$, while the cost without reusing will be $Nn(1 + \lambda(l-1))$.

The total cost of the gathering step when computing independently the n frames (classic approach) is given by nkN, since we have to cast kN paths for each frame. Conversely, if we reuse the gathering paths (new approach) we just have to cast kN one-length paths in total, so this is the cost in this case.

The total expected acceleration factor, that considers the addition of both shooting and gathering cost, is given by

$$\frac{nN(1+\lambda(l-1)) + nkN}{N(n+\lambda(l-1)) + kN} = \frac{n(1+\lambda(l-1)) + kn}{n+\lambda(l-1)+k} \tag{4}$$

When the number of light source positions (frames) grows to infinity, we have

$$\lim_{n\to\infty} \frac{n(1+\lambda(l-1)) + kn}{n+\lambda(l-1)+k} = k + 1 + \lambda(l-1) \tag{5}$$

Observe incidentally that the acceleration factor for indirect illumination can be obtained by putting $\lambda = 1$ in (5).

3.4 Expected Error

In the best case, using the same number of paths per position in both classical and new approach will produce the same error. This best case corresponds to the limiting case in which all the light source positions were the same. In general, the more distant the positions, the bigger the error. This is due to the fact that when positions get distant, $p(y)$ goes away from the importance sampling function corresponding to each source position j.

This means that the new algorithm presents a theoretical speed-up factor bounded by the expression (5).

4 Implementation and Results

We have used in our implementation a collision random walk estimator (see [1]). Parameter λ (see Sect. 3.1) has been set to 0.25 in all our tests. The reason for this value is that we have subdivided each patch in 4 elements, so that the area of an element is approximately 1/4 of the area of the patch, and the variance of the radiosity random walks is known to be proportional to the area of the elements.

Regarding to the gathering step, the value of k (see Sect. 3.3) that indicates the relation between the number of shooting and gathering paths, has been set to 1 in all our tests.

The new algorithm implies some additional storage: for each patch, the accumulated values for each frame must be stored, and also for each element in which patches are subdivided. This corresponds to store an array of n triplets (r,g,b) per patch and element, that is, a total of $n(P+4P) = 5nP$ triplets, where P is the number of patches. Note that using the classical approach we just have to store $5P$ triplets (one per patch and element) for each of the n executions. However, no storage for the shooting or gathering paths is required, due to the immediate update of values for each position when dealing with a path.

The tested scene (see Fig. 3) has been discretized in 2686 patches and 10744 elements. We have considered 30 very close light source positions that are tracing a diagonal route near the ceiling.

We have obtained an acceleration factor of about 6, much higher than the theoretical prevision (about 2.8 according to (4)) due to the lower cost of the visibility computations. On the other hand, we have noted that the Mean Square Error (MSE) using the new method is lower for frames placed in the middle of the frame sequence than for frames placed in the extremes. This behavior can be theoretically justified by the fact that central positions have an importance sampling function closer to the pdf that we have used than the extreme positions.

We present in Fig. 2 the time vs. MSE graph corresponding to frame 17, including also the results of reusing only gathering paths (from independent shooting solutions).

Figure 3 shows two frames in the animation. The one on the left has been obtained using the classic algorithm, in which no reuse of paths is done. The one on the right has been obtained using the new algorithm, involving reuse of both shooting and gathering paths. The number of shooting and gathering paths is the same in both classic and new approach, but in the new approach the number of shooting paths includes the reuse. No remarkable difference is observed between the quality of left and right images, but the cost is much lower using the new algorithm.

Animation sequences accompanying the paper [1] show the reduction of the flickering effect when reusing the paths (see Sect. 3.2). Animation 1 involves no reuse of paths (classic approach). Each image has been obtained in about 200 seconds. That means a total of $200 \times 30 = 6000$ sec.$= 100$ min. Animation 2

[1] ima.udg.es/∼castro/videoCGGM/anim1.gif|anim2.gif|anim3.gif

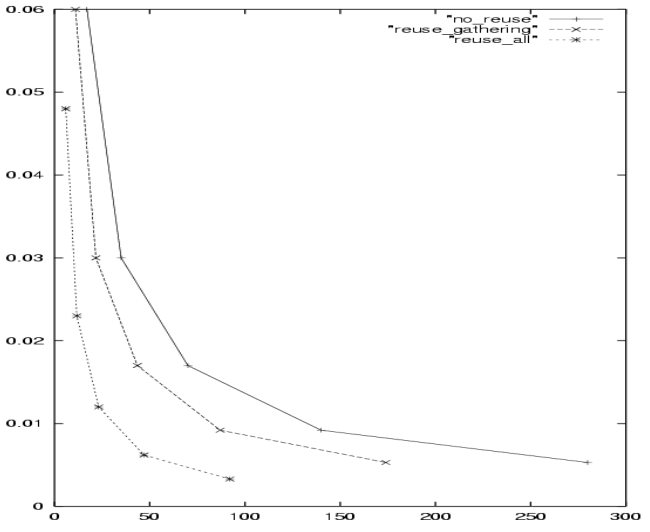


Fig. 2. Graph of time in seconds (horizontal axis) vs. MSE (vertical axis). This graph corresponds to frame 17. Note the clear advantage of new method in front of the classical one, and also in front of reusing only gathering paths

Fig. 3. (left) Frame 17 without reusing paths. 600.000 paths. Time= 70 sec. (right) Frame 17 reusing shooting and gathering paths. 600.000 paths. Time= 11.7 sec. MSE is a bit higher in the second image. Speed-up factor about 4.4

involves reuse of the gathering paths, but not of the shooting ones. Each shooting has spent about 114 sec., and the combined gathering has spent about 86 sec. This is a total of $114 \times 30 + 86 = 3506$ sec.$= 58.5$ min. Animation 3 has been obtained by reusing both shooting and gathering paths. The combined shooting

has spent 1017 sec., and the combined gathering about 86 sec. This means a total of $1017 + 86 = 1103$ sec. $= 18.4$ min.

5 Conclusions and Future Work

We have extended in this paper our previous results on reusing shooting paths to the final gathering step algorithm, which combines a shooting solution and a gathering step. It has been applied in the context of radiosity to the case of moving light sources. Both shooting and gathering paths have been reused for each frame, so that different frames in an animation (corresponding to different positions of the light source) have been obtained with a considerable gain respect to the classical algorithm.

The gain will come not only from the reusing strategy, but also from the inner-frame coherence obtained by eliminating the annoying flickering due to temporal aliasing.

The scope of the idea presented in this paper is not limited to radiosity. It could be applied to different techniques in global illumination, like virtual light sources [4], and photon maps [7].

Acknowledgments. This project has been funded in part with a Catalan-Flemish Joint Action n. ACI2002-52 from the Catalan Government, and with grant number TIC2001-2416-C03-01 from the Spanish Government.

References

1. P.Bekaert *Hierarchical and Stochastic Algorithms for Radiosity.* Ph.D. thesis. Katholic Univ. of Leuven. 1999.
2. P.Bekaert, M.Sbert, and J.Halton *Accelerating Path Tracing by Re-Using Paths.* Proceedings of Workshop on Rendering 125-134. 2002.
3. M.Cohen, and J.Wallace *Radiosity and Realistic Image Synthesis.* Academic Press Professional, Cambridge. 1993.
4. A.Keller *Instant Radiosity.* Computer Graphics Proceedings, Siggraph'97. 49-56. 1997.
5. F.Perez *Global Illumination Techniques for the Computation of High Quality Images in General Environments.* Ph.D. thesis. Universitat Politècnica de Catalunya. 2003.
6. M.Sbert, F.Castro, and J.Halton *Reuse of Paths in Light Source Animation.* Computer Graphics International Proceedings (to appear as short paper). 2004.
7. H.Wann Jensen *Global Illumination using Photon Maps.* Rendering Techniques. p. 21-30. 2004.

Real Time Tree Sketching

Celso Campos[1], Ricardo Quirós[2], Joaquin Huerta[2], Emilio Camahort[3], Roberto Vivó[3], and Javier Lluch[3]

[1] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Vigo, Spain
ccampos@ei.uvigo.es
[2] Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Spain
{quiros, huerta }@lsi.uji.es
[3] Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Spain
{camahort, rvivo, jlluch}@dsic.upv.es

Abstract. Modeling and rendering of synthetic plants and trees has always received a lot of attention from computer graphics practitioners. Recent advances in plant and tree modeling have made it possible to generate and render very complex scenes. Models developed so far allow low quality and photorealistic rendering as well as a fine control on the amount of geometry sent to the graphics pipeline. Recently, non-photorealistic rendering techniques have been proposed as an alternative to traditional rendering. In this paper we present a method for interactive rendering of vegetation silhouettes. Our goal is to expressively render plant and tree models. Previous methods are either too slow for real-time rendering or they do not maintain the general appearance of a given vegetable species. We solve these two problems in our work.

1 Introduction

Modeling and rendering of synthetic plants and trees has always received a lot of attention from computer graphics practitioners. Plant and tree models are fundamental to the representation of forests, gardens and interior scenes. Still, these models are geometrically complex and their description and rendering require a lot of resources in traditional computer graphics. A tree model may contain hundreds of thousands of polygons, and a forest scene may contain thousands of trees. Therefore, interactive rendering of such scenes requires specific modeling and acceleration techniques for applications such as outdoor walkthroughs and fly-by´s.

Recent advances in plant and tree modeling have made it possible to generate and render very complex scenes. The models developed so far allow low quality and photorealistic quality rendering as well as a fine control on the amount of geometry sent to the graphics pipeline. The most important plant and tree models are based on L-systems [1] [2].

Recently, non-photorealistic rendering techniques have been proposed as an alternative to traditional rendering [3]. Their goal is to increase the expressiveness of rendering, using techniques similar to those used in the arts [4], in animated movies (toon

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 197–204, 2004.
© Springer-Verlag Berlin Heidelberg 2004

shading), and in sketches representing buildings and interiors [5]. All these applications usually include plant and tree models. Hence, there is a need for specific techniques to efficiently render expressive images from these models.

In this paper we present a method for interactive rendering of vegetation silhouettes. Our goal is to expressively render plant and tree models. Previous methods are either too slow for real-time rendering or they do not maintain the general appearance of a given vegetable species. We solve these two problems in our work. First, we survey previous work in non-photorealistic rendering of plants and trees. Then, we present a method that generates simplified tree models and we describe our silhouette rendering algorithm. We conclude our paper with some results and directions for future work.

2 Previous Work

2.1 Non-photorealistic Rendering of Plants and Trees

The structural complexity of plants and trees requires using specific techniques to render them non-photorealistically. The first methods for automatic illustration of vegetable species were introduced by Yessios [6] and Sasada [7]. They both produce tree sketches for architectural applications.

Kowalski et al. [8] also create abstract tree sketches using geometric primitives that approximate the tree's foliage. Those primitives are used in a two-step rendering algorithm. The first step creates gray-scale reference images of the trees. The second step improves the references by adding so-called *graftals*, small objects that represent leaves, grass or hair.

Markosian et al. [9] improve on Kowalki's work by using a static scheme for graftal definition during the modeling stage. At rendering time a graftal may or may not be rendered depending on the viewing parameters. Some graftals, known as tufts, are stored using a multiresolution representation that allows any graftal to become a set of tufts when the viewer is close enough. Other improvements take advantage of frame-to-frame coherence and add control to the appearance and behavior of the graftals. Still, the rendering algorithm is slow for very complex scenes.

Deussen [10] presents a method that creates pen-and-ink illustrations of plants and trees. The method starts with a detailed tree model that includes a branch and leaf skeleton. This skeleton is later used to compute the silhouettes necessary to draw the tree's contour. Rendering is accomplished by combining a large set of primitives instead of using graftals. This allows the representation of specific plants and trees and not just generic trees like those in [8] and [9].

More recently, Di Fiore [11] proposes a method that renders cartoon shaded trees. It uses tree models generated from L-systems. The models contain no leaves, but only the hierarchical structure of the trunk and branches. Given that information, the artist develops a picture library to represent branches and leaf groups. The final image is obtained by rendering the pictures corresponding to the branches and adding at the branch joints the pictures that represent the leaf groups.

2.2 Stroke-Based Rendering of Plants and Trees

[12] and [13] present a stroke-based method for non-photorealistic rendering of plants and trees. This method improves on Meier's algorithm [4] by supporting interactive frame rates. The method can be applied to videogames, virtual walkthroughs and other real-time computer graphics applications.

The method models trees using random parametric L-systems (RL-systems) [14]. These are an improvement on L-systems that associates to each system a set of random variables. This approach has several advantages over the surface patch algorithm of Meier. For example, it supports the simultaneous generation of both the tree's geometry and its stroke particles for future rendering.

The stroke particles are distributed using the same RL-system that is used for modeling the tree's geometry. To achieve this goal we use a *shape instantiation* process. This process represents every instantiable object, a branch or leaf, using both 3D geometry and a cloud of strokes. The latter are used for expressive rendering. Fig. 1 shows some results obtained using our software to render different plants and trees.

Fig. 1. Stroke-based rendering results

3 Tree Generalization

We present our approach to generalizing a geometric tree model to an abstract model that maintains the visual appearance of the tree and supports its silhouette-based rendering. In the following Section we describe how we render these models.

Modeling using RL-systems requires two steps: *system derivation* and *graphical interpretation*. Given a system, derivation generates a *parametric chain*. This chain is interpreted to obtain the geometric model to be rendered. In this paper, we use an RL-system that generates a ternary tree model [12].

RL-systems allow us to model plants and trees keeping their structural complexity. This is good for photorealistic rendering, where a lot of detail may be required. But this may be too much detail for expressive rendering. We propose a *generalized model* for plants and trees. This model maintains an approximated representation that keeps the tree's visual appearance at a higher level of abstraction.

Using a generalized model has several advantages. We can correctly represent the branching structure of the tree by using contour lines for each branch. We can define an abstract model for the leaves that supports different types of leaves and different ways of rendering them. When rendering the leaves we can apply different illumina-

tion models. We use our generalized model to obtain the information needed to generate the contours and render them using a suitable illumination model. We describe our generalized model for both branches and leaves.

3.1 Modeling the Branches

Branches are typically modeled using separate geometric primitives like truncated cones, generalized cylinders and polygonal meshes. In order to avoid discontinuity and visibility problems at the branch joints, we use a single polygonal model for the entire branching structure of the tree [15]. This is illustrated in Fig. 2. Using a single model we can easily apply geometric simplification methods to build a multiresolution representation of the branches with different LODs.

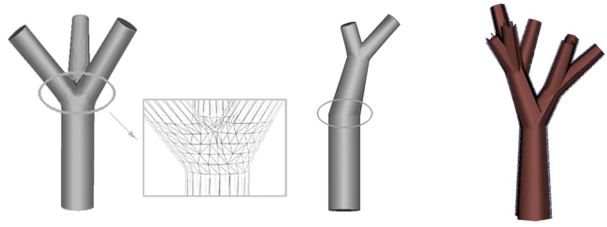


Fig. 2. Using a single polygonal mesh to represent all the branches of a tree

3.2 Modeling the Leaves

We propose a generalized model for the leaves that preserves the visual appearance of their geometric representation. To build a model we start with an RL-system and follow these steps: (i) we compute the convex hulls of each of the branches and its leaves, (ii) we compute oriented bounding boxes for the convex hulls, and (iii) we replace the bounding boxes with substitution shapes for final rendering.

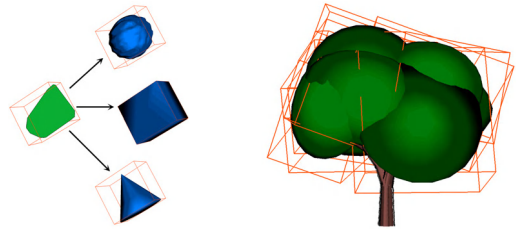


Fig. 3. Left, a convex hull and its bounding box can be replaced by one of several substitution shapes. Right, our sample tree rendered using spheres as substitution shapes

For step (i) we assume that the leaves are made of strokes, whose convex hull can be easily computed. For each branch and its sub-branches and leaves we compute the

convex hull using an algorithm due to O'Rourke [16]. In step (ii) we compute an oriented bounding box for each convex hull. An oriented bounding box is the bounding box with minimal volume that encloses the convex hull. In step (iii) of our algorithm we replace each bounding box by a generic *substitution shape*. A substitution shape can be any object that we want to render in place of a branch and its sub-branches and leaves. Fig. 3 left shows three examples. Once the substitution shapes have been generated we can render the tree (see Fig. 3 right).

4 Rendering

We compute the silhouettes using an enhanced version of the algorithm by Raskar and Cohen [18]. We apply an illumination model like the one proposed by Hamlaoui [19]. The original algorithm by Raskar and Cohen computes the wire-frame silhouette of a polygonal mesh. We modified this algorithm to draw the silhouette using texture mapped polygons. We start by computing all the polygons that belong to the silhouette. Then, we replace each of the edges of those polygons by a new polygon whose scale depends on the normal to the edge. Finally, we texture map and render those silhouette polygons.

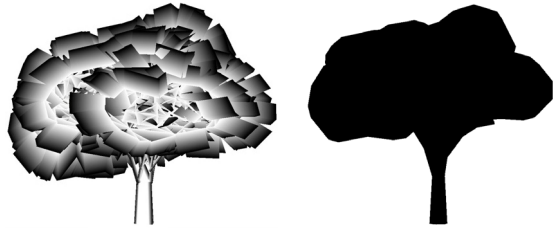


Fig. 4. Left, polygonal model generated from the silhouette polygons of our sample tree. Right, mask to remove interior edge polygons

The main drawback of this algorithm is that we need to remove those polygons that have been generated for interior hidden edges of the original polygon. A simple solution to this problem draws the mesh during a second pass, once the hidden polygons have been removed.

This method yields an analytical description of the silhouette, which can be used to create and texture map new polygons. The method is fast enough to allow interactive frame rates. Fig. 4 left shows the polygons generated from the silhouette of our sample tree. In order to remove the polygons associated to the interior edges of the silhouette we generate a mask for the tree (see Fig. 4 right). Then, we choose a texture for the silhouette polygons and we render the final image by composing it with a suitable background (see Fig. 5).

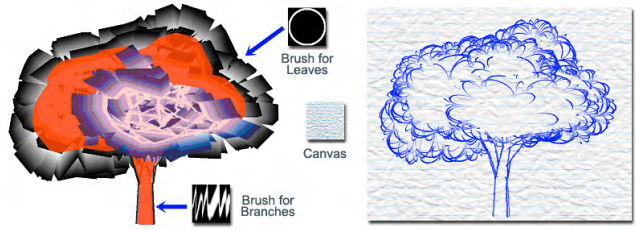


Fig. 5. Composing the final image for silhouette/based rendering

We use a modified version of Hamlaoui's illumination model [19]. Our version supports both flat and gradient shading. The idea is to apply a 1D texture to the polygons of the silhouette. The 1D texture contains a gradation of the same color. Given a vertex normal and its view and light vectors we apply an illumination model and determine which texture coordinate to use in order to obtain the desired shading. The texture element stored for that coordinate determines the final color of the vertex.

The final image is the combination of the results of a two-pass algorithm. In the first pass the algorithm computes the silhouette of the model. In the second pass, the silhouette is shaded using the appropriate illumination, color and tone. The result is a toon-shaded tree like the one shown in Fig. 6.

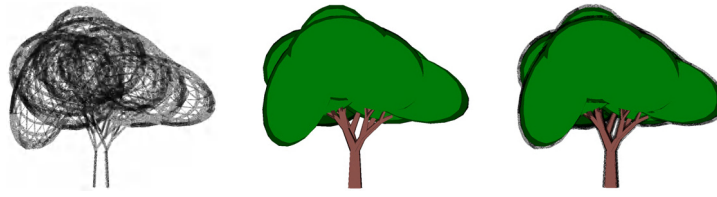


Fig. 6. Left, 1st pass - silhouette. Middle, 2nd pass - shading. Right, combined result

We demonstrate our rendering algorithm by running it on three different graphics cards, a Creative GeForce 2 GTS Pro 64 Mb, an ATI Radeon 9200 128 Mb, and an nVidia GeForce FX 5200 128 Mb. We render our sample tree at five different LODs. We use two rendering algorithms: silhouette rendering (see Fig. 7 up) and toon shading (see Fig.7 down). Our silhouette rendering algorithm runs at interactive rates (see Table 1) making it suitable for interactive walkthroughs. Our toon shading algorithm runs as fast, as shown in Table 1.

5 Conclusions

In this paper, we introduce the generalized model for representing trees in an abstract way suitable for expressive rendering. Our model stores a single polygonal mesh and preserves the visual appearance of any given tree. We can abstract the leaves' representation to obtain different leaf rendering styles. Our leaf representation supports multiple illumination models.

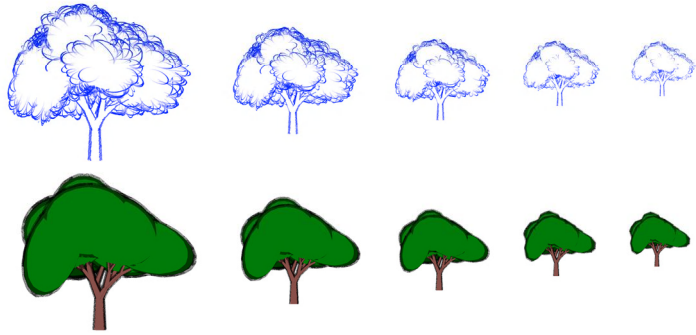


Fig. 7. Our sample tree rendered at different LODs using silhouette rendering (up) and toon shading (down)

Table 1. Frame rates achieved by our rendering algorithms

	Silhouette Rendering			Toon Shading		
	Creative	**ATI**	**nVidia**	**Creative**	**ATI**	**nVidia**
Original	35	48	60	58	60	98
LOD 1	45	58	75	60	62	99
LOD 2	65	78	105	100	102	127
LOD 3	105	95	140	125	126	142
LOD 4	125	128	155	133	135	155

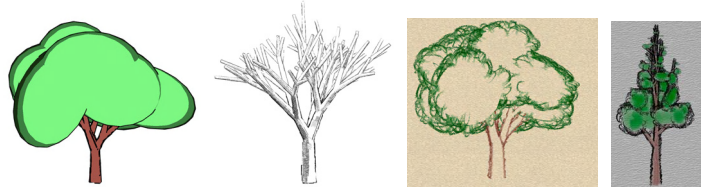


Fig. 8. Other results obtained with different rendering styles

References

1. P. Prusinkiewicz and A. Lindenmayer, The algorithmic beauty of plants, Ed. Springer-Verlag, 1990.
2. P. Prusinkiewicz, M. James, and M. Mech, "Synthetic Topiary," Computer Graphics, pp 351-358, 1994.
3. T. Strothotte and S. Schlechtweg, Non-photorealistic computer graphics : modeling, rendering, and animation. San Francisco, CA, Morgan Kaufmann, 2002.
4. B. J. Meier, "Painterly rendering for animation", Proceedings of SIGGRAPH 96, pp 477-484, Agosto 1996. New Orleans, Louisiana.
5. M. Webb, E. Praun, A. Finkelstein, and H. Hoppe, "Fine Control in Hardware Hatching", Proceedings of SIGGRAPH 02, 2002.

6. C. I. Yessios, "Computer drafting of stones, wood, plant and ground materials." Proceedings of SIGGRAPH'79 (Computer Graphics), pp 190-198, 1979.
7. T. T. Sasada, "Drawing Natural Scenery by Computer Graphics.," Computer-Aided Design, vol. 19, pp 212-218, 1987.
8. M. A. Kowalski, L. Markosian, J. D. Northrup, L. D. Bourdev, R. Barzel, L. S. Holden, and J. F. Hughes, "Art-Based Rendering of Fur, Grass ant Trees", Proceedings of SIGGRAPH 99, pp 433-438, Agosto 1999. Los Angeles, California.
9. L. Markosian, B. J. Meier, M. A. Kowalski, L. S. Holden, J. D. Northrup, and J. F. Hughes, "Art-based Rendering with Continuous Levels of Detail", in NPAR 2000, Annecy, France, 2000.
10. O. Deussen and T. Strothotte, "Computer-generated pen-and-ink illustration of trees", Proceedings of SIGGRAPH 2000, pp 13-18, Julio 2000.
11. F. Di Fiore, W. Van Haevre, and F. Van Reeth, "Rendering Artistic and Believable Trees for Cartoon Animation", in CGI2003, 2003.
12. C. Campos, R. Quirós, J. Huerta, M. Chover, J. Lluch, and R. Vivó, "Non Photorealistic Rendering of Plants and Trees", in International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging., Grecia, 2001.
13. C. Campos, E. Camahort, R. Quirós, J. Huerta, and I. Remolar, "Acceleration Techniques for Non-Photorealistic Rendering of Trees", Iberoamerican Symposium on Computer Graphics, Guimaraes, Portugal., 2002.
14. J. Lluch, M. J. Vicent, R. Vivó, and R. Quirós, "GREEN: A new tool for modelling natural elements", in WSCG'2000 International Conference on Computer Graphics and Visualization, Plzen, Checz Republic, 2000.
15. J. Lluch, M. J. Vicent, C. Monserrat, and S. Fernández, "The modeling of branched structures using a single polygonal mesh," IAESTED Visualization, Imaging, and Image Processing, 2001.
16. J. O´Rourke, Computational Geometry in C, Cambridge University Press, 1998.
17. G. Barequet and S. Har-Peled, "Efficiently Approximating the Minimum-Volume Bounding Box of a Point Set in 3D," Proccedings 10th ACM-SIAM Symposium on Discrete Algorithms, 1999.
18. R. Raskar and M. Cohen, "Image Precision Silhouette Edge", In Proc. 1999 ACM Symp. on Interactive 3D Graphics, 1999.
19. S. Hamlaoui, “Cel-Shading”, GameDev.net., 2001
http://www.gamedev.net/reference/ /programming/features/celshading

Facial Expression Recognition Based on Dimension Model Using Sparse Coding

Young-suk Shin

Department of Information and telecommunication Engineering, Chosun University, #375 Seosuk-dong, Dong-gu, Gwangu, 501-759, Korea
ysshin@mail.chosun.ac.kr

Abstract. We present an expression recognition system based on dimension model of internal states that is capable of identifying the various emotions using automated feature extraction. Feature vectors for facial expressions are extracted from a hybrid approach using fuzzy c-mean clustering algorithm and dynamic linking based on Gabor wavelet representation. The result of facial expression recognition is compared with dimensional values of internal states derived from semantic ratings of words related to emotion by experimental subjects. The dimensional model recognizes not only six facial expressions related to six basic emotions (happiness, sadness, surprise, angry, fear, disgust), but also expressions of various internal states. In this paper, with dimension model we have improved the limitation of expression recognition based on basic emotions, and have extracted features automatically with a new approach using FCM algorithm and the dynamic linking model.

1 Introduction

Face is an important social stimulus in human interactions. Specially, facial expression plays a major role in human communication. If a computer can understand emotions from human's facial expressions, it is possible to help humans in various situations dynamically. Currently, most facial expression recognition systems use the six principle emotions of Ekman [1]. Ekman considers six basic emotions: happiness, surprise, fear, anger, disgust, sadness; and categorizes facial expressions with these six basic emotions. Most research on facial expression recognition includes studies using the basic emotions of Ekman[2, 3, 4, 5], therefore these studies have limitations for recognition of natural facial expressions which consist of several other emotions and many combinations of emotions. Here we describe research extended on the dimension model of internal states for recognizing not only facial expressions of basic emotions but also expressions of various emotions.

Previous work on facial expression processing includes studies using representation based on optical flow from image sequences [6, 7], principle components analysis of single image [7,8], physically based models [9], and wavelets transformation[10]. These methods are similar in that they first extract some features from the images, then these features are used as inputs into a classification system.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 205–212, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper, we use a hybrid approach for automatic feature extraction. The first stage detects the edges of major face components, using the average value of the image's 2-D Gabor wavelet coefficient histogram on all the images. The second stage, FCM clustering algorithm is used to select sparse pixels from edges of major facial components extracted previously from a neutral face of each expressor. The third stage is an application of the Dynamic Link Architecture [11]. This capability is used here to detect sparse local features on expression images from preselected points in the neutral face. Finally, we show the recognition of facial expressions based on the dimension model of internal states using a multi-layer perceptron.

2 Facial Expression Database Based on Internal States

The images used in this study were obtained from the Korean facial expression database for mapping of facial expressions into internal states [12]. This database consists of 500 facial expression images of males and females under well controlled lighting condition. Expressions were divided into two dimensions (pleasure-displeasure and arousal-sleep dimension) according to the study of internal states through the semantic analysis of words related with emotion by Kim et al. [13] using expressive 83 words. For experiment we used 11 expressions in a set of 44 internal state expressions from each of 6 person. The 11 expressions are happiness, surprise, sadness, disgust, fear, satisfaction, comfort, distress, tiredness, worry (including neutral face). A few of these are shown in Fig. 1. The result of the dimension analysis of 44 emotion words related to internal emotion states is shown in Fig. 2. The dimension model expresses a degree of expression in each of the two dimensions on a nine point scale. Our paper shows the recognition of facial expressions on dimension model.

3 Sparse Coding of Facial Expressions

To extract information of facial expression, we use 287 images of facial expressions, each image using 640 by 480 pixels included face images almost in the frontal pose. Original images have been rescaled and cropped such that the eyes are roughly at the same position with a distance of 60 pixels in the final image. For edges of major facial components, an average value of the image's 2-D Gabor wavelet coefficient histogram is used. The general form of two dimensional Gabor wavelets function is given by Daugman [14]. The wave vector $\vec{k}$ of length $k \equiv \left\|\vec{k}\right\|$ defines the spatial wavelength and at the same time controls the width of the Gaussian window. The parameter σ denotes the width of the Gaussian window relative to the wavelength corresponding to k .

$$\psi_{\vec{k}}(\vec{x}) = \frac{k^2}{\sigma^2}\exp(-\frac{k^2 x^2}{2\sigma^2})[\exp(i\vec{k}\cdot\vec{x}) - \exp(-\frac{\sigma^2}{2})] \tag{1}$$

Fig. 1. Examples from the facial expression database

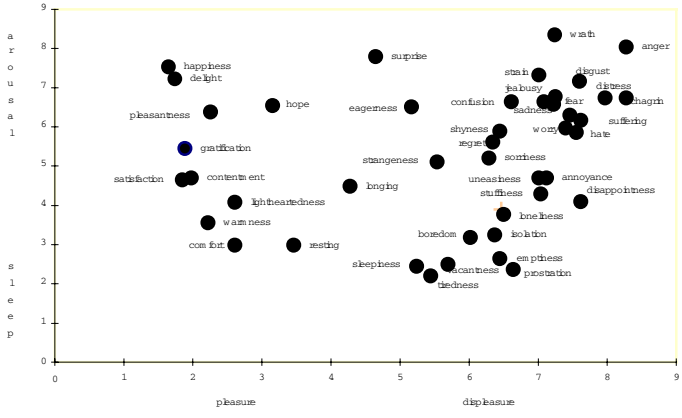
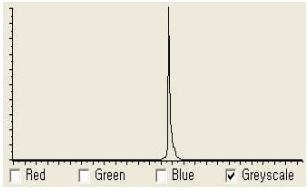


Fig. 2. Dimension Model: dimension analysis of 44 emotion words

To detect features of major face components, we use a specific frequency band, a wave number, k=0.78, and 5 distinct orientations in 22.5 ° steps between 0 and π, and chose σ=π. The complex valued $\psi_{\vec{k}}$ applied to each image combines an even and odd part. We use only the magnitudes because they represent local information of an image in a smoothly varying way. Let G be the set of Gabor function $\psi_{\vec{k}}$ to be applied to *I*. G is G_1, G_2 . The computation proceeds as follows: $\omega_1 = \sum\sum G_1 I, \quad \omega_2 = \sum\sum G_2 I, \ \varpi = \sqrt{(\omega_1^2 + \omega_2^2)}.$

Fig. 3(a) shows the result of the 2-D Gabor coefficients histogram using the magnitudes of Gabor coefficients from an expression image. This means these coefficients completely capture local facial feature points in special frequency and special orientation. Thus, we applied the average value of 2-D Gabor coefficient histogram to extract local facial feature points. The average value of Gabor coefficients histogram is controlled by optional value ±α since experimental images may be a noise. Fig. 3(b) shows the resulting image which applied an optional value to an average value of the Gabor coefficients histogram.



(a)

(b)

Fig. 3. (a) 2-D Gabor coefficient histogram. (b) Extracted edges of major face components

Extracted feature points are similar to edges of major facial components. Since Gabor vectors with neighboring pixels are highly correlated and redundant, it is sufficient to use sparse pixels on a face. We thus pick out sparse feature points based on the FCM clustering algorithm in edges extracted from the 2-D Gabor wavelet coefficient histogram. FCM algorithm applies to neutral facial images that is used as a template to extract sparse feature points from edges of major facial components on expression images.

The potentiality of fuzzy clustering algorithms can be demonstrated by their application in clustering tasks which involve a large number of feature vectors of high dimension and a large number of clusters[15]. Fuzzy C-means clustering [16] is a data clustering algorithm in which each data point belongs to a cluster to a degree specified by a membership grade. The degree of the assignment of the feature vector $\mathbf{x}_i \in X$ into various clusters is measured by the membership function $u_{ij} \in [0,1]$, which satisfy the properties $\sum_{i=1}^{c} u_{ij} = 1, \forall_j = 1,...,N$. The cost function for FCM is $J(U,c_1,...,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{N} u_{ij}^m d_{ij}^2$. c_i is the cluster center of fuzzy group *i*; $d_{ij} = \|c_i - \mathbf{x}_j\|$ is the Euclidean distance between *i*th cluster center and *j*th data point ; and $m \in [1< m ,\infty]$ is a weighting exponent. The necessary conditions for $J(U,c_1,...,c_c)$ to reach a minimum are $c_i = \sum_{j=1}^{N} u_{ij}^m X_j \Big/ \sum_{j=1}^{N} u_{ij}^m$ and $u_{ij} = 1 \Big/ \sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{2/(m-1)}$.

We determined sparse feature points using the following steps: **Step1**. Initialize the membership matrix U with random values between 0 and 1 such that the constraints in $\sum_{i=1}^{c} u_{ij} = 1$ are satisfied. **Step2.** Calculate *c* fuzzy cluster centers ($\mathbf{c}_i, i = 1,2,....,c$) using c_i. **Step3.** Compute the cost function according to $J(U,c_1,...,c_c)$, and stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold. **Step4.** Compute a new ***U*** using u_{ij}, then go to Step2. Fig. 4(a) shows a result that extracted sparse pixel points by FCM algorithm: *c*=60, *m*=2. The number of clusters is decided in the range that can reflect the same topological relationship as major face components in human vision.

After extracting the sparse feature points on neutral faces, which are used as a template to extract sparse feature points from edges on the expression images extracted

previously since each neutral face plays a standard role to decide the degree of expression change against an expression image.

To match point to point feature points on an expression face against each feature point on a neutral face, it consists of two different domains, which are called the neutral domain (N) and the expression domain (E). The expression domain contain the jets of the Gabor transformation. The Gabor jet $\vec{J}(\vec{x}_i)$ refers to the set of Gabor magnitudes obtained by sampling the image at the point $\vec{x}_i$ with sampling functions of all sizes (frequencies) and orientations. Sparse feature extraction using DLM on expression images is guided by a function S in $S(\vec{J}_i^N, \vec{J}_i^E) = \vec{J}_i^N \cdot \vec{J}_i^E \Big/ \left\|\vec{J}_i^N\right\| \left\|\vec{J}_i^E\right\|$ which determines the similarity between neutral face jet, $\vec{J}_i^N$ and expression image jet, $\vec{J}_i^E$. The entire wavelet family consists of two frequency bands, the wave number $k = \left\|\vec{k}\right\| = (\pi/4, \pi/8)$ using inverse pixels and seven different orientations from 0° to 180°, differing in 30° steps.

The linking procedure is performed under the constraint that the matching points found in the expression face have approximately the same topological relations as the preselected points in the neutral image. A match point should be chosen in the neutral face and then computed in the Euclidean distance between the preselected point in neutral face and each point in the expression image in $\vec{\Delta}_{ij}^{NE} = \vec{x}_i^N - \vec{x}_j^E$. This evaluates the quality of local topological preservation. The dynamic linking of selected points in the neutral face image to points in the expression image is formulated as an optimization problem. The cost function H in $H = \vec{\Delta}_{ij}^{NE} + \sum S(\vec{J}_i^N, \vec{J}_j^E)$ to be optimized measures the quality of proposed point matches. We chose for cost function the special form. The feature on the expression images was accepted if the cost function H satisfies two conditions at the same time : (1) Reach to the minimum value. (2) Do not exceed a maximum distance value that the matching points found in the expression face have approximately the same topological relations as the preselected points in the neutral image(see Fig. 4(b)).

(a) (b)

Fig. 4. (a) Sparse pixel points extracted with FCM algorithm on neutral face. (b) Sparse pixel points extracted with DLM on expression image

4 Facial Expression Recognition

The system for facial expression recognition uses a three-layer neural network. The first layer is the distance values from each feature point on a neutral face to each feature point on an expression face which are normalized by size from 0 to 1. The second layer is 240 hidden units and the third layer is two output nodes to recognize the two dimensions: pleasure-displeasure and arousal-sleep. Training applies error back propagation algorithm which is well known to the pattern recognition field. The activation function of hidden units uses the sigmoid function. 250 images for training and 37 images excluded from the training set for testing are used. The first test verifies with the 250 images trained already. Recognition result produced by 250 images trained previously showed 100% recognition rates. The rating result of facial expressions derived from the semantic rating of emotion words by subjects is compared with experimental results of a neural network (NN). The similarity of recognition result between human and NN is computed in $S(\vec{H},\vec{N}) = \vec{H}\cdot\vec{N} \Big/ \|\vec{H}\|\|\vec{N}\| \min\left(\|\vec{H}\|/\|\vec{N}\|, \|\vec{N}\|/\|\vec{H}\|\right)$. The dimension values of human and NN in each two dimension are given as vectors of $\overline{H}$ and $\overline{N}$.

Table 1 describes a degree of similarity of expression recognition between human and NN on two-dimensional structure of emotion. In Table 1, the result of expression recognition of NN is matched to the most nearest emotion word in 44 emotion words related to internal emotion states. The result of expression recognition of NN looks very similar to the result of expression recognition of human(see Table 1).

5 Discussion and Conclusion

This paper presents an expression recognition system based on dimension model of internal states using sparse coding. Facial expression on dimension model includes Two dimensions which are pleasure to displeasure dimension and arousal to sleep dimension. The result of expression recognition of NN looks very similar to the result of expression recognition of human. Above all, the expression images of the high level of arousal and displeasure emotion have been most effectively recognized by neural network. In a pleasure-displeasure dimension, the degree of arousal could make an effect on discriminating facial expressions like happiness, satisfaction, and comfort. The combination in displeasure dimension with the high level of arousal dimension could be well recognized by neural network. Such expressions are fear, surprise, distress, worry, and disgust. These results appear to have an effect on physical changes between neutral pattern and expression pattern in major facial components.

This study is a new approach of human's emotion processing, it is interesting to note in this context that machine vision may represent various emotions similar to human with the combination of each dimension in the internal emotion states. To future study we are planning to recognize the expressions with person independent and a wider range of emotions in much larger database than present system.

This study was supported by research funds from Chosun University, 2003.

Table 1. The result data of expression recognition between human and NN

Emotion words	Human(Mean)		Neural Network		Recognition on Neural Network	Similarity
	P – D	A – S	P –D	A – S		
happiness	1.65	7.53	3.88	3.44	lightheartedness	0.54
			4.92	4.6	boredom	0.71
			2.86	5.86	pleasantness	0.82
			1.31	5.69	gratification	0.75
			4.43	4.8	longing	0.73
satisfaction	1.85	4.65	1.49	6.07	pleasantness	0.79
			2.14	4.96	contentment	0.92
			6.32	5.9	shyness	0.52
comfort	2.61	2.98	5.0	5.7	strangeness	0.52
			3.65	3.64	lightheartedness	0.77
sadness	7.22	6.57	7.07	5.23	shyness	0.89
			3.7	6.37	hope	0.72
			6.62	7.12	surprise	0.91
tiredness	5.44	2.2	7.94	6.29	strain	0.56
			4.06	4.05	sleepiness	0.90
			4.39	4.28	longing	0.89
			4.8	5.09	strangeness	0.76
			6.39	5.65	uneasiness	0.65
worry	7.4	5.96	6.89	6.09	confusion	0.97
			7.39	6.84	strain	0.94
surprise	4.65	7.8	4.55	8.29	surprise	0.95
			4.61	7.67	surprise	0.98
			4.65	5.60	hope	0.79
disgust	7.93	6.74	6.35	3.42	isolation	0.68
			7.33	6.14	hate	0.91
			7.68	6.03	distress	0.98
			6.05	6.72	surprise	0.86
Fear	7.25	6.77	6.75	4.49	sorriness	0.80
			6.43	5.21	stuffiness	0.83
			6.68	7.97	disgust	0.94
			7.30	7.96	chagrin	0.91
			5.91	4.17	isolation	0.72
distress	7.46	6.29	7.48	7.16	disgust	0.94
			4.28	5.81	hope	0.72
			4.77	4.97	boredom	0.70
			5.60	4.11	boredom	0.71
			5.81	5.05	strangeness	0.79

References

1. Ekman, P.: Universal and cultural difference in facial expressions of emotions. In: J. K. Cole(Ed.), Nebraska symposium on motivation, Lincoln: University of Nebraska Press, (1972) 207-283
2. Lien, J.: Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity. Ph.D. Thesis, Carnegie Mellon University, (1998)
3. Oliver, N. Pentland, A., Berard, F.: LAFTER:a real-time face and lips tracker with facial expression recognition. Pattern Recognition 33 (2000) 1369-1382
4. Tian, Y.L, Kanade, T., & Cohn, J. F.: Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), (2001) 97-116

5. Cohen, I., Sebe, N., Garg, A., Chen, L. S., Huang, T. S.: Facial expression recognition from video sequence:temporal and static modeling. Computer Vision and Image Understanding , In Press (2003)
6. Yacoob, Y., Davis, L.S.: Recognizing human facial expression from long image sequences using optical flow. IEEE Trans. Pattern Anal. Machine Intell. 18(6) (1996) 636-642
7. Bartlett, M., Viola, P., Sejnowski, T., Larsen, J., Hager, J., Ekman, P.: Classfying Facial Action. In: Advances in Neural Information Processing Systems 8. D. Touretzky et al. editors, MIT Press, Cambridge, MA (1996)
8. Padgett, C., Cottrell, G.: Identifying emotion in static face images. In Proceeding of the 2nd Joint Symposium on Neural Computation, 5 (1995) 91-101
9. Essa, I. Pentland, A. : Facial Expression Recognition using Visually Extracted Facial Action Parameters. Proceedings of the International Workshop on Automatic Face and Gesture Recognition (1995) 35-40
10. Lyons, M., Akamatsu, S.:Coding facial expressions with Gabor wavelets. Proceeding of the Third International Conference on Automatic Face and Gesture Recognition, (1998) 200-205
11. von der Malsburg, C.: Nervous structure with dynamical links. Ber. Bunsenges. Phy.Chem, 89 (1985) 703-710
12. Bahn, S., Hahn, J. and Chung, C.: Facial expression database for mapping facial expression onto internal state. '97 Emotion Conference of Korea, (1997) 215-219
13. Kim, Y., Kim, J., O, S., O, K., Chung, C.: The study of dimension of internal states through word analysis about emotion. Korean Journal of the Science of Emotion and Sensibility, 1 (1998) 145-152
14. Daugman, J: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America 2 (1985) 1160-1169
15. Karayiannis, N.B., Pai, P.-I.: Fuzzy vector quantization algorithms and their application in image compression. IEEE Transactions on Image Processing, (1995)
16. Bezdek, J.C.: Fuzzy mathematics in pattern classification. Ph.D. thesis, Applied Math. Center, Cornell University, Ithaca (1973)

An Application to the Treatment of Geophysical Images through Orthogonal Projections

Sixto Romero and Francisco Moreno

Escuela Politécnica Superior La Rábida, Ctra. Palos de la Frontera s/n, 21819 Palos de la Frontera, Huelva, Spain
sixto@uhu.es, franmo@uhu.es

Abstract. The present work provides the generalization to the approach proposed by Frei and Chen to square masks of any dimension for line and edge detection in digital images. It is completed with the application of the designed algorithm to the image of a archaeological site, that to our judgement permit us to establish an correlation between the mathematical results and results of the archaeologicals explorations

1 Introduction

When we try to extract information from an image whose definition does not allow to recognize lines, edges, or isolated points, the first step to take is segmentation. This process consists in dividing the image into its parts. The result will depend on the treatment given to the analysed image.

Segmentation algorithms for monochrome images [6] are based on the two basic properties of grey-level values: discontinuity and similarity. In the first case the image partition is based on abrupt changes in grey levels and is used for the location of isolated points, lines and edges [5]. In the second case, thresholding, region growing, and region splitting and merging is used. For the detection of the three basic types of discontinuities, isolated points, lines and edges, we will use the usual way of applying a suitable mask. This technique consists in treating each pixel of the original image and creating a new image. To do this, using 3×3 masks as an example, we will change the center pixel grey level, which matches the central cell of the mask, following the equation [3]

$$R = p_1 z_1 + p_2 z_2 + \cdots + p_9 z_9 \tag{1}$$

This pixel is assigned the grey level given by R. $p_i,\ i = 1 \ldots 9$ represent the coefficients according to the mask type and $z_i,\ i = 1 \ldots 9$, represent grey levels of the pixels that make up the nine cells, according to Fig. [1]

2 Using Multimasks for Line and Edge Detection

We will focus on 3× 3 masks for line and edge detection. Let's consider the chosen pixel and the eight encompassing pixels as a nine-component vector representing the nine grey levels

$$\mathrm{z} = (z_1, z_2, z_3, \cdots, z_9)^T$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 213–220, 2004.
© Springer-Verlag Berlin Heidelberg 2004

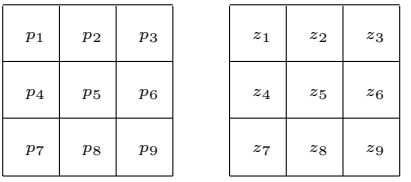


Fig. 1. A 3×3 mask

whose component z_5 represents the grey level value for that pixel and another vector, also with nine components, representing the mask coefficients

$$\mathrm{p} = (p_1, p_2, p_3, \cdots, p_9)^T$$

In a matrix form, [1] can be written

$$R = \mathrm{p}^T \mathrm{z}. \quad \mathrm{p}, \mathrm{z} \in \mathbb{R}^9$$

By properly choosing the mask coefficients so that they represent orthogonal vectors we can split up our 9-dimensional vector space into three 4-, 4- and 1-dimensional orthogonal subspaces. This approach was first proposed by Frei and Chen [8] in 1977. The first two subspaces will be called edge and line subspaces, and the last one will be called measure subspace. Frei and Chen suggested that a probability measure [4] for a pixel to belong to an edge or a line could be given by the value of the angle forming vector z, which represents each pixel, with its orthogonal projection onto each subspace. The smaller the angle, the closest the vector to its corresponding subspace and thereby the most likely that the vector belongs to the subspace. Starting from this proposal, our work will consist in calculating the projection matrix which generalizes the masks usually used by researchers. If we treat initially the masks proposed by Frei and Chen as orthogonal vectors for each subspace, and therefore basis, and if we have into account that their components sum up zero, we can choose as a basis vector for the measure subspace a vector whose components are all equal to the unit and therefore orthogonal to the remaining eight, given that we use the usual scale product. The bases for the orthogonal edge and line subspaces are given by the mask coefficients in Fig. 2, where the first four are $\mathrm{p}_1, \mathrm{p}_2, \mathrm{p}_3, \mathrm{p}_4$, suitable for edge detection; the next four $\mathrm{p}_5, \mathrm{p}_6, \mathrm{p}_7, \mathrm{p}_8$ are appropriate for line detection, and the last one, u, is added to complete a basis of the vector space $\mathbb{R}^9$.

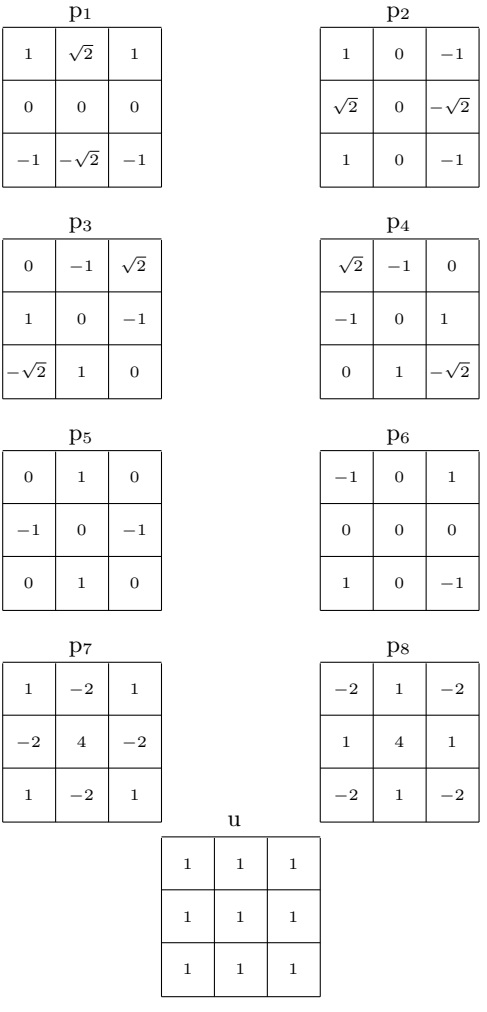


Fig. 2. Orthogonal masks suggested by Frei and Chen

2.1 Projection Matrices for 3×3 Grids

According to Frei and Chen's approach, we consider edge, line and measure subspaces formed respectively by vectors of Fig. [2], $\mathcal{B} \equiv \{p_1, p_2, p_3, p_4\}$, $\mathcal{L} \equiv \{p_5, p_6, p_7, p_8\}$ y $\mathcal{U} \equiv \{u\}$. We calculate the projection matrices on each one of them, $P_{\mathcal{B}}$, $P_{\mathcal{L}}$ y $P_{\mathcal{U}}$. These matrices will be obtained through the matrix product, for each of them, given by

$$\begin{aligned} P_{\mathcal{B}} &= B(B^T B)^{-1} B^T \\ P_{\mathcal{L}} &= L(L^T L)^{-1} L^T \\ P_{\mathcal{U}} &= U(U^T U)^{-1} U^T \end{aligned} \tag{2}$$

where the columns of these matrices represent the vectors of each subspace. For $n = 9$, the following result is obtained

$$P_{\mathcal{B}} = (b_{ij}) = \begin{cases} b_{ii} = \frac{1}{2} & \forall i \neq \frac{n+1}{2} \\ b_{i,n-i} = -\frac{1}{2} & \forall i \neq \frac{n+1}{2} \\ b_{\frac{n+1}{2},\frac{n+1}{2}} = 0 & \\ b_{ij} = 0 & \text{otherwise} \end{cases}$$

$$P_{\mathcal{L}} = (l_{ij}) = \begin{cases} b_{ii} = \frac{7}{18} & \forall i \neq \frac{n+1}{2} \\ f_{i,n-i} = -\frac{7}{18} & \forall i \neq \frac{n+1}{2} \\ f_{\frac{n+1}{2},\frac{n+1}{2}} = \frac{8}{9} & \\ f_{ij} = -\frac{1}{9} & \text{otherwise} \end{cases}$$

$$P_{\mathcal{U}} = (u_{ij}) \qquad / \qquad u_{ij} = \frac{1}{9} \quad \forall i, j = 1 \cdots n.$$

If we use the respective $3, 5, 7, \cdots$, dimensional square masks, the results are generalized giving rise to vector spaces of dimensions equal to the square of those numbers: $9, 25, \cdots$.

3 Projection Matrices on Edge and Line Subspaces

We will prove the aforementioned generalization with the two following theorems.

3.1 Projection Matrix onto the Edge Subspace

Theorem 31 *Let E, $\dim E = n$, $n = 2k + 1$, $k \in \mathbb{N}$ be an Euclidean space and let $\mathcal{B} \subset E$ be the k-dimensional vector subspace, generated by the array $\langle \boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_k \rangle$ of basis vectors of $\mathcal{B}$, whose symmetric components are opposed and their sum is zero. In these circumstances, the projection matrix onto the $\mathcal{B}$ subspace is the matrix $P_{\mathcal{B}} \in \mathcal{M}_n(\mathbb{R})$ given by the following expression*

$$P_{\mathcal{B}} = (b_{ij}) = \begin{cases} b_{ii} = \frac{1}{2} & \forall i \neq \frac{n+1}{2} \\ b_{i,n-i} = -\frac{1}{2} & \forall i \neq \frac{n+1}{2} \\ b_{\frac{n+1}{2},\frac{n+1}{2}} = 0 & \\ b_{ij} = 0 & \text{otherwise} \end{cases} \tag{3}$$

Proof. The projection matrix onto the vector subspace $P_{\mathcal{B}}$ is, according to 2, $P_{\mathcal{B}} = B(B^T B)^{-1} B^T$ where the columns of the B matrix are the vectors $\langle \boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_k \rangle$. Being a projection matrix, it will follow

$$P_{\mathcal{B}} B = B$$

and then

$$P_{\mathcal{B}} B - I_n B = \Theta \Rightarrow (P_{\mathcal{B}} - I_n) B = \Theta$$

thereby B is the matrix of eigenvectors corresponding to the eigenvalue $\lambda = 1$ of the $P_{\mathcal{B}}$ matrix. It is known from linear algebra that, according to the spectral theorem [1], every symmetric matrix is orthogonally and reciprocally diagonalizable. It is proved that the $\lambda = 1$ eigenvalue, of k algebraic multiplicity, matches the k eigenvectors which are precisely the array $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_k\}$. In order to prove this, we just need to solve the equation

$$(P_{\mathcal{B}} - I_n)\boldsymbol{b} = \boldsymbol{0}$$

the result is an undetermined system of $k+1$ equations whose solutions are the mentioned vectors with the structure suggested in theorem 31, i.e., the opposed symmetric components and the sum of them all, which is null, given that the center component is null -the number of components is odd.

3.2 Projection Matrix onto the Line Subspace

Theorem 32 *Let E, $\dim E = n$, $n = 2k+1$, $k \in \mathbb{N}$ be an Euclidean space and let $\mathcal{L} \subset E$ be the k dimensional vector subspace, generated by the array $\langle \boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_k \rangle$ of basis vectors of $\mathcal{B}$, whose symmetric components are opposed and their sum is zero. In these circumstances the projection matrix onto the $\mathcal{L}$ subspace is the matrix $P_{\mathcal{L}} \in \mathcal{M}_n(\mathbb{R})$ given by the following expression*

$$P_{\mathcal{L}} = (l_{ij}) = \begin{cases} l_{ii} = \frac{n-2}{2n} & \forall i \neq \frac{n+1}{2} \\ l_{i,n-i} = -\frac{n-2}{2n} & \forall i \neq \frac{n+1}{2} \\ l_{\frac{n+1}{n}, \frac{n+1}{n}} = \frac{n-1}{n} & \\ l_{ij} = -\frac{1}{n} & \text{otherwise} \end{cases} \tag{4}$$

Proof. The projection matrix onto the vector subspace $P_{\mathcal{L}}$ is, according to 2, where the columns of the L matrix are vectors $P_{\mathcal{L}} = L(L^T L)^{-1} L^T$. Being a projection matrix, it will follow

$$P_{\mathcal{L}} L = L$$

and then

$$P_{\mathcal{L}} L - I_n L = \Theta \Rightarrow (P_{\mathcal{L}} - I_n) L = \Theta$$

thereby L is the matrix of eigenvectors corresponding to the eigenvalue $\lambda = 1$ of the matrix $P_{\mathcal{L}}$. According to the spectral theorem [1], we know that every

symmetric matrix is orthogonally and reciprocally diagonalizable. It is proved that the $\lambda = 1$ eigenvalue, of k algebraic multiplicity, matches the k eigenvectors which are precisely the array $\{\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_k\}$. In order to prove this, we just need to solve the equation

$$(P_{\mathcal{L}} - I_n)\boldsymbol{l} = \mathbf{0}$$

the result is an undetermined system of $k+1$ equations whose solutions are the mentioned vectors with the structure suggested in theorem 32, i.e., the opposed symmetric components are equal and the sum of them all is null, therefore the center component is equal and opposed to the sum of the remaining components.

4 Algorithm for Line and Edge Detection

In the Frei and Chen's approach it is use the specifics masks and we prove, 31 and 32, that the masks aren't important, well the projections matrices only depend to the mask dimension that we can to use. The proposed algorithm can be summed up in the following steps:

Step 1. Reading of the image and calculation of the average of typical deviations for each pixel and their neighbors, according to the dimensions of the selected grid. For the computational expense not to be excessively high, a random number of the image pixels are chosen, $\pm 0.02\%$ approximately.

Step 2. Each pixel is read and its typical deviation and that of their neighbor is made, according to the dimensions of the selected grid, and this is compared to the average obtained previously and multiplied by a factor chosen at will. This way we decide whether the pixel is a candidate to be an edge pixel, a line pixel or a uniform region pixel.

Step 3. Finally, with the candidates to be edge or line pixels, and according to the approach suggested by Frei and Chen, we calculate the norms of their projections, and comparing both of them we decide whether they are edge or line pixels[2].

4.1 Application of the Algorithm to Determine Lines and Edges in a Geophysical Image

The suggested algorithm will allow to obtain lines and edges on a digital image without modifying the rest of the image. We apply masks of any dimension, odd and larger than one, and we will obtain different results depending on our interest. The Fig. 3 a) show the original image obtained with program PROSPECT [7] in graphic format standard: BMP, JPG, GIFF, ... susceptible of the studied by any commercial software implemented in MAPLE or MATLAB, for example.

5 Conclusions

We think that the more important conclusions are: is not necessary to use any mask and the original image remain the same except the lines and edges. For

streets images, highways, flat, buildings, cadasters, face photographs of persons, animal and things, it is relatively easy the edges detection. When it is considered to study a geophysical image the problem adopts a high degree of complexity. The present work shows a particular application in the one which is obtained a map from anomalies distribution electrical corresponding to the site from Cerro El Palmarón. With the different applied models have been detected important alignments related to the walls of the constructions of the period. It is provided, from the sight point of the images interpretation a structural anomalies plan that it has served as guide for the ulterior excavation. In this way the DIP represents an important tool for the historical restitution of the cited site.

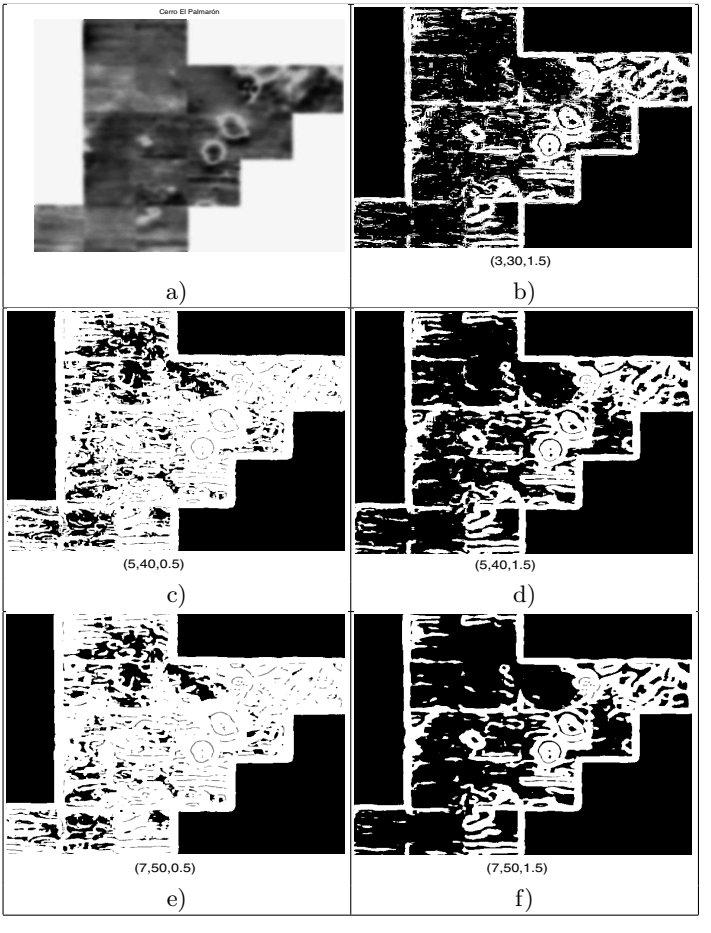


Fig. 3. a) Original image. b)c)d)e) y f) Images transformed for detections edges and lines

References

1. Noble B. , Daniel Janes W.: Applied Linear Algebra. Prentice-Hall, New York (1982)
2. Burton H. Wiejak, J.S.: Convolution with separable mask for early image processing. Computer Vision, Graphics and Image Processing **32** (1985) 279-290
3. González C.Rafael, Woods Richard E.:Digital Image Processing. Addison-Wesley, Massachusetts (1992)
4. Park D.J.: Edge detection in noisy images based on the co-ocurrence matrix. Pattern Recognition **27** (1994) 765–774
5. Gómez Lopera J.F.: An analysis of edge detection by using the jensen-shannon divergence . Journal of Mathematical Imaging and Vision **13** (2000) 35–56
6. Pratt William K.: Digital Image Processing. John Wiley & Sons, INC, New York (2001)
7. Romero S.: Modelización Matemática y Tratamiento Digital de Imágenes Geofísicas Aplicadas a la Restitución Histórica: Programa PROSPECT. IGIDL-Lisboa **1** (2000) 273–274
8. Frei W., Chen C.C..: Fast Boundary Detection: A Generalization and a New Algorithm. IEEE Trans. Computer **26** (1977) 988-998

A Derivative-Free Tracking Algorithm for Implicit Curves with Singularities

José F.M. Morgado and Abel J.P. Gomes

Dept. Computer Science and Engineering, Univ. Beira Interior
6200-001 Covilhã, Portugal
{fmorgado,agomes}@di.ubi.pt

Abstract. This paper introduces a new algorithm for rendering implicit curves. It is curvature-adaptive. But, unlike most curve algorithms, no differentiation techniques are used to compute self-intersections and other singularities (e.g. corners and cusps). Also, of theoretical interest, it uses a new numerical method for sampling curves pointwise.

1 Introduction

An implicit curve $\Gamma = \{\mathbf{p} \in \mathbb{R}^2 : f(\mathbf{p}) = 0\}$ is a level set (or zero set) of some function f from $\mathbb{R}^2$ to $\mathbb{R}$. The algorithm proposed in this paper focuses on the representation of planar implicit curves defined by real, but not necessarily analytic, functions (e.g. polynomial and transcendental functions). There are three major categories of algorithms to render implicit curves, namely:

- *Representation conversion.* Rarely, a global parameterization exists for an implicit curve. But, a local parameterization always exists in a neighborhood of a regular point of an implicit curve, i.e. a point $\mathbf{p} = (u, v)$ such that $f(\mathbf{p}) = 0$ and $\nabla f \not= 0$. This allows us to render an implicit curve by using the algorithms for parametric curves [10,2,4,6,1].
- *Space subdivision.* Recursively, it splits the ambient space into subspaces, discarding those not intersecting the curve. The subdivision terminates soon after we obtain a good approximation to the curve by a set of small subspaces (e.g. rectangles) [4,15,9]. Robust approximations can be implemented by using interval arithmetic [14], algebraic or rational techniques [8,7], and floating-point arithmetic [13].
- *Curve tracking.* It consists of sampling the curve pointwise [5,11]. This approach has its roots in the Bresenham's algorithm for rendering circles, which is essentially a continuation method in screen image space. Continuation methods are attractive because they concentrate computational processing where it is needed. However, they need a starting point on each curve component. Finding a starting point on a component can be a frustrating experience, in particular for degenerated components consisting of a single point. A way to compute these curve components is by means of the cylindrical algebraic decomposition technique from computer algebra [3].

This paper deals with the rendering of implicit curves possibly with singularities, but no derivatives are used at all.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 221–228, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Curve Sampling through Numerical Approximation

The basic idea behind the curve-tracking algorithm proposed in this paper is, given the previous and current points P, Q of the curve Γ, to determine the next point belonging to the intersection $N_Q \cap \Gamma$, where N_Q is the frontier of a small circular neighborhood centered at Q (Fig. 1(a)). The algorithm does not evaluate the intersection points analytically. Instead, any intersection point of $N_Q \cap \Gamma$ is computed by a new approximation method inspired in the false position numerical method, called *angular false position method*.

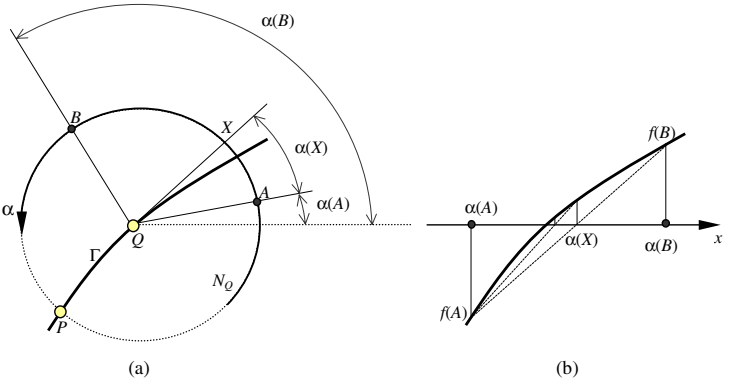


Fig. 1. The angular false position method

2.1 False Position Method: A Brief Review

The false position method is a root-finding algorithm which starts from two distinct estimates A and B for the root of $f(x) = 0$, being f a function from $\mathbb{R}$ to $\mathbb{R}$, so that $f(A) \, . \, f(B) < 0$, i.e. $f(A)$ and $f(B)$ have opposite signs or, equivalently, a root is initially bracketed in the interval $]A, B[$. The next estimate X is iteratively determined by linear interpolation given by the following formula:

$$X = B - \frac{f(B)}{f(B) - f(A)}(B - A) \tag{1}$$

This numerical method retains the prior estimate, either A or B, that together with X continue to bracket the root [12]. The formula (1) shows that the false position is an adequate sampling numerical method for generic curves, not necessarily differentiable or analytic, provided that no derivatives are used at all.

2.2 Angular False Position Method

As described above, the false position method calculates the roots of some function from $\mathbb{R}$ to $\mathbb{R}$ in the product space $\mathbb{R}^2$. But, for the curve Γ defined implicitly

by $f : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$, we are not interested in the roots of f in the product space $\mathbb{R}^3$, but in the zero set of f in $\mathbb{R}^2$, i.e. the curve itself in the domain space $\mathbb{R}^2$. For that, the curve is sampled by intersecting the zero set and a small circle N_Q centered at the current point Q (Fig. 1(a)). But, the intersection $\Gamma \cap N_Q$ occurs in the domain space of f, not in the product space. To overcome this difficulty we have first to transform the co-ordinates of both point estimates A and B on the circle onto their corresponding angle estimates $\alpha(A)$ and $\alpha(B)$ defined by a function $\alpha : \mathbb{R}^2 \to \mathbb{R}$. Then, as illustrated in (Fig. 1(b)), the false position method is used to determine an intermediate angle given by

$$\alpha(X) = \alpha(B) - \frac{f(B)}{f(B) - f(A)}[\alpha(B) - \alpha(A)] \tag{2}$$

Thus, the angular approximation method occurs in the product space of $f \circ \alpha^{-1}$, i.e. $\mathbb{R}^2$, according to the following diagram:

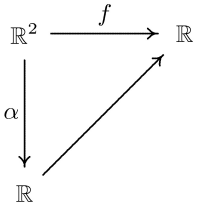


Fig. 2. Diagram of the angular false position method

3 Curve Tracking Algorithm

This algorithm confines all computations to the neighborhood N_Q to determine the intersecting points $\Gamma \cap N_Q$. The main difficulty is to correctly choose the next point X amongst those in $\Gamma \cap N_Q$. The main criterion for choosing the next point is based on the variance of curve curvature within N_Q, what is given by $\angle PQX$. Remarkably, this criterion works for any local shape, no matter whether or not the curve self-intersects, it has a cusp or a corner, it almost touches itself, or it oscillates.

3.1 Computing Neighborhood Points

The curve points $N_Q \cap \Gamma$ are numerically determined by the angular method introduced above. Instead of using the x-axis for computing the points approaching an intersection point, we use a small circular neighborhood N_Q (Fig. 3).

We could think of a uniform distribution of points on the neighborhood N_Q separated by an angle θ, and then apply the angular method to every pair of

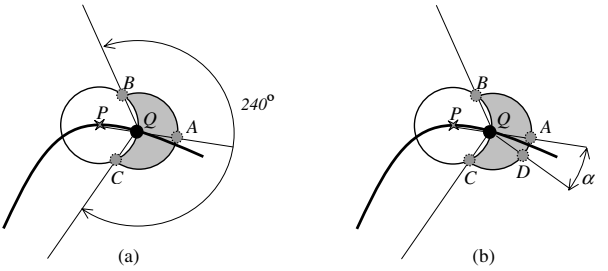


Fig. 3. Distribution of points on the neighborhood of the current point Q

points in order to determine possible intersection points. Such a distribution of points is given by $(x, y) = (x_c + r\cos\theta, y_c + r\sin\theta)$, with $\theta \in [-180°, 180°[$, being (x_c, y_c) the center of N_Q, i.e. the current point Q. To speed up our algorithm, the circle points are computed for $\theta \in [-120°, 120°]$ because the others were calculated before in the neighborhood of the previous point P (Fig. 3).

We start by determining three points, $A = Q + \overrightarrow{PQ}$, $B = Q + M_{2\pi/3}.\overrightarrow{QA}$, and $C = Q + M_{-2\pi/3}.\overrightarrow{QA}$ on N_Q, where M is the rotation matrix.

Then, one determines the point $D = Q + M_\alpha.\overrightarrow{QA}$, with $\alpha \in [5°, 10°]$, such that $f(D).f(A) < 0$, by applying the angular approximation method to the arc $\overset{\frown}{AD}$. Note that for curves with small curvature, a small α leads to very fast search for intersection or solution points. In particular, for a straight-line curve, the point A is the solution point itself. Nevertheless, we have to look for solution points along the arcs $\overset{\frown}{AB}$ and $\overset{\frown}{CD}$ in case we have more curve points on N_Q.

3.2 Cusp and Other High-Curvature Points

Most curve-tracking algorithms break down at singularities (see, for example, [5]). Nevertheless, the algorithm proposed by [11] works for curves with bifurcation points by analysis of the sign changes of the partial derivatives in a rectangle neighborhood. However, it breaks down at other singularities such as, for example, cusps, which belong to the function domain, but not to the domain of the partial derivatives. For example, it is not capable of rendering the curve $|x| + |y| - 2 = 0$ (Fig. 6(b)) with four singularities at $(0, 2)$, $(2, 0)$, $(0, -2)$ and $(-2, 0)$, where the partial derivatives do not exist.

In contrast, our algorithm needs not compute derivatives at all. This allows us to draw curves defined by both differentiable and non-differentiable functions. This reduces somehow the computation time of each point in the curve. In fact, the computation of the partial derivatives of an analytic function such as, for example, $y(9 - xy)(x + 2y - y^2)((x - 10)^2 + (y - 4)^2 - 1) = 0$ (Fig. 6(f)) may be more time-consuming than the function itself.

Cusps and corners are points at which the curvature flips. A cusp point (Fig. 4(a)), or a quasi-cusp point (Fig. 4(b)), is characterized by having a high curvature variance along the curve within N_Q. To be sure that there is a cusp

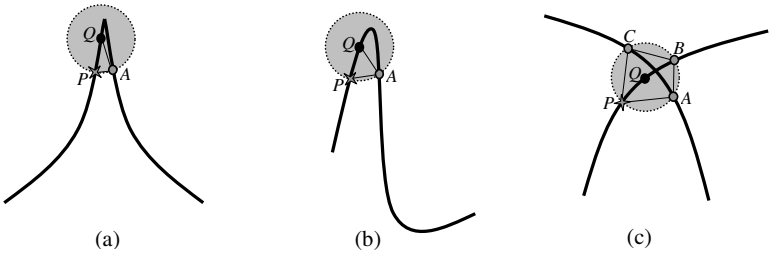


Fig. 4. (a) A cusp; (b) a quasi-cusp; (c) a self-intersection point

(or a quasi-cusp) point in N_Q, we have to check that the mediatrix of $\overline{PA}$ in N_Q intersects the curve at exactly a single point. But, first, we have to check that the angle $\angle PQA$ is relatively small. A small angle $\angle PQA$ means that A is not an appropriate point next to Q.

The strategy is then to assume that the points after the cusp (or quasi-cusp) are image points of those before it in N_Q. For example, A is the image of P. The image of Q is B by tracing a line segment parallel to $\overline{PA}$. The next point R is determined by intersecting the curve with the mediatrix of $\overline{QB}$ in N_Q. R is possibly a cusp, but if it has an image point C, we determine again the next point by computing the intersection between the curve and the mediatrix of $\overline{RC}$, stopping when the distance between the latest next point and its image is under 10^{-6}. The latest next point is nearly the cusp point (or the quasi-cusp point).

3.3 Self-Intersection Points

A self-intersection point can be seen as a generalized cusp point (Fig. 4(c)). In fact, every two adjacent curve branches incident at the crossing point can be viewed as meeting at a cusp point. As for a cusp (or a quasi-cusp), no curve point on N_Q is the next point because, with the exception of B, the points A and C form small angles with P, Q. But, B cannot be the next point either because the segment $\overline{PB}$ intersects the curve at a point. This intersecting point between P and B is a better candidate to next point than B because it is before B. It can be determined by applying the false position method between P and B.

Similar to a cusp point, we have to compute the image point D of Q by intersecting the curve with a line segment parallel to $\overline{PA}$ passing through Q in N_Q. Then, we determine the remaining image points E of D and F of E by using the segments parallel to $\overline{AB}$ and $\overline{BC}$, respectively. At this point, we can generalize the convergence process to a cusp, so that the next point R is determined by intersecting the curve with the mediatrix of $\overline{QD}$ in N_Q. This process converging to the self-intersection point stops when the distance between the latest next point and its image is under 10^{-6}. This latest point is nearly the self-intersection point.

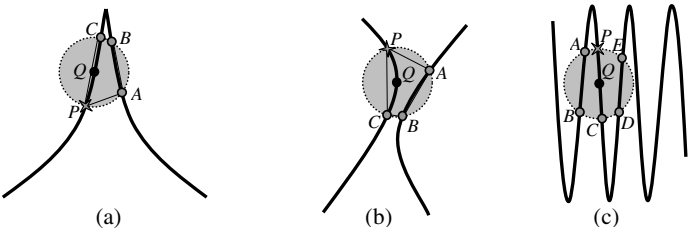


Fig. 5. Near-branch points

3.4 Near-Branch Points

Sometimes a curve almost touches itself, i.e. some of its points are very close (Fig. 5). Unlike the previous cases, one of the neighborhood-intersecting points is the point next to Q. Note that determining the next point is done without changing the neighborhood radius, even under ripples and undulations.

Let us look again at Fig. 5, where P and Q are the previous and current points, respectively. To see that the neighborhood-intersecting point C is the next point after Q, we use two criteria: *angle* (or curvature) *criterion* as above, and the *neighbor-branch criterion*. In Fig. 5(a) and (b), A cannot be the next point because the angle $\angle PQA$ is far from 180°. But, both angles $\angle PQB$ and $\angle PQC$ are close to 180°, and neither $\overline{PB}$ nor $\overline{PC}$ crosses the curve, so either B or C can be the next point. To pick up the right point, either B or C, we use the neighbor-branch criterion. This criterion is basically an elimination criterion amongst candidates to the next point, and can be described as follows:

1. Determine the midpoints of the segments $\overline{PC}$, $\overline{CB}$, and $\overline{BA}$ in Fig. 5(a)(b) (the midpoint of $\overline{PA}$ is not calculated because A is, by the angle criterion, no longer a candidate next point).
2. For each segment with midpoint M, discard its endpoints if the segment $\overline{QM'}$ intersects the curve at one or more points, being M' the projection of M on the neighborhood circle by prolonging $\overline{QM}$. This eliminates B as a candidate next point in Fig. 5(a). Note that the point B in Fig. 5(b) cannot be the next point for a different reason. In fact, the segment $\overline{PB}$ crosses the curve at a point, preventing it from being the next point.

In Fig. 5(c), the angle criterion eliminates A and E as candidate next points, whereas B and D are eliminated by the near-branch criterion. So, the point C is the next point. Note that the neighborhood radius holds constant independently of whether the curve oscillates or not.

3.5 The Algorithm

The NEXTPOINT algorithm sketched below determines the point X next to the current point Q according to criteria described above.

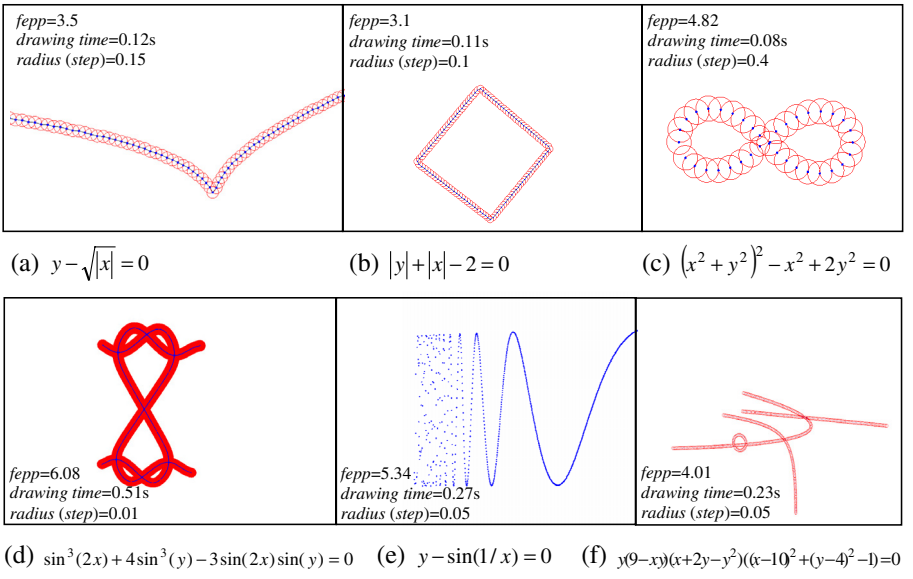


Fig. 6. Implicit plane curves

Algorithm (NEXTPOINT). *The inputs are the previous and current points P, Q, the radius r of N_Q and the function defining the curve Γ. The output is the point X next to Q.*

1. Compute the intersection points $\Gamma \cap N_Q$ by means of the angular numerical method described in Sect. 2.2.
2. **if** $(\#(\Gamma \cap N_Q) = 1)$ —a single candidate point
   (a) $X \leftarrow$ get such a single point from $\Gamma \cap N_Q$
   (b) **if** $(\angle(\overline{PQ}, \overline{PX}) \not\approx 180^{\circ})$
   - $X \leftarrow$ compute cusp or quasi-cusp through the convergence technique described in Sect. 3.2

   else —two or more candidate points
   (a) $X \leftarrow$ get such a single point from $\Gamma \cap N_Q$ by applying elimination criteria described in Sect. 3.4
   (b) **if** $(X = NULL)$ —there is a self-intersection point about Q
   - $X \leftarrow$ compute self-intersection point through the convergence technique described in Sect. 3.3
3. **return** X

This algorithm is part of the 2DCURVE algorithm that is able to render generic implicit curves in $\mathbb{R}^2$ as those in Fig. 6, which is not detailed here because of space limitations.

4 Conclusions and Experimental Results

The algorithm 2DCURVE was written in C++. Its major contributions are:

- It works for general curves defined by real functions.
- It is derivative-free. So, it does not break down at other singularities.
- It does not break down under shape oscillations and ripples either.
- It introduces a new numerical method for sampling curve points.

Fig. 6 shows interesting performance results for some curves defined by real functions in $\mathbb{R}^2$. The term 'fepp' stands for 'function evaluations per point', i.e. the average number a function is evaluated for each sampled curve point. The runtime tests were performed on a mere Windows PC equipped with a 500MHz Intel Pentium and 128MB RAM, but even so we got really fast rendering times.

References

1. Allgower, E., Gnutzmann, S.: Simplicial Pivoting for Mesh Generation of Implicitly Defined Surfaces. Comp. Aid. Geom. Des. **8** (1991) 30–46
2. Abhyankar, S., Bajaj, C.: Automatic parameterization of rational curves and surfaces III: algebraic plane curves. Technical Report CSD-TR-619, Purdue University, Computer Science Department, USA (1987)
3. Arnon, D.: Topologically reliable display of algebraic curves. Comp. Graph. **17** (1983) 219–227
4. Bloomenthal, J.: Poligonisation of implicit surfaces. Comp. Aid. Geom. Des. **5** (1988) 341–355
5. Chandler, R.: A tracking algorithm for implicitly defined curves. IEEE Comp. Graph. and Appl. **8** (1988) 83–89
6. Hobby, J.: Rasterization of nonparametric curves. ACM Trans. on Graph. **9** (1990) 262–277
7. Keyser, J., Culver, T., Manocha, D., Krishnan, S.: MAPC: a library for efficient and exact manipulation of algebraic points and curves. In Proceedings of the 15th ACM Symposium on Computational Geometry, ACM Press (1999) 360–369
8. Krishnan, S., Manocha, D.: Numeric-symbolic algorithms for evaluating one-dimensional algebraic sets. In Proceedings of the ACM Symposium on Symbolic and Algebraic Computation (1995) 59–67
9. Lopes, H., Oliveira, J., Figueiredo, L.: Robust adaptive polygonal approximation of implicit curves. In Proceedings of the SibGrapi 2001, IEEE Computer Society (2001)
10. Lorensen, W., Cline, W.: Marching Cubes: A High Resolution 3D Surface Construction Algorithm. Comp. Graph. **21** (1987) 163–169
11. Moeller, T., Yagel, R.: Efficient rasterization of implicit functions. http://citeseer.nj.nec.com/357413.html (1995)
12. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical recipes in C: the art of scientific computing. Cambridge University Press, 2nd edition, 1992.
13. Shewchuk, J.: Adaptive precision floating-point arithmetic and fast robust geometric predicates. Disc. and Comp. Geom. **18** (1997) 305–363
14. Snyder, J.: Interval arithmetic for computer graphics. In Proceedings of ACM SIGGRAPH'1992, ACM Press (1992) 121–130
15. Taubin, G.: An accurate algorithm for rasterizing algebraic curves. In Proceedings of the 2nd ACM Solid Modeling and Applications, ACM Press (1993) 221–230

Framework for Simulating the Human Behavior for Intelligent Virtual Agents. Part I: Framework Architecture

F. Luengo[1,2] and A. Iglesias[2⋆]

[1] Department of Computer Science, University of Zulia, Post Office Box #527, Maracaibo, Venezuela
fluengo@cantv.net

[2] Department of Applied Mathematics and Computational Sciences, University of Cantabria, Avda. de los Castros, s/n, E-39005, Santander, Spain
iglesias@unican.es
http://personales.unican.es/iglesias

Abstract. This paper is the first in a series of two papers (both included in this volume) describing a new framework for simulating the human behavior for intelligent virtual agents. This first paper focuses on the framework architecture and implementation issues. Firstly, we describe some requirements for such a framework to simulate realistically the human behavior. Then, the framework architecture is discussed. Finally, some strategies concerning the implementation of our framework on single and distributed CPU environments are presented.

1 Introduction

One of the most exciting fields in Computer Graphics is the simulation and animation of intelligent virtual agents (IVAs) evolving within virtual 3D worlds. This field, also known as *Artificial Life*, has received increasing attention during the last few years [1,2,3,4,5,6,12,14]. Most of this interest has been motivated by its application to the entertainment industry, from virtual and augmented reality in digital movies to video games. However, the range of potential applications also includes Architecture, Science, Education, advertising and many others.

One of the most interesting topics in the field concerns the realistic animation of the behavior of IVAs emulating the human beings. The challenge here is to provide the virtual agents with a high degree of *autonomy*, so that they can evolve freely with a minimal input from the animator. In addition, this evolution is expected to be *realistic*, in the sense that the IVAs must behave according to reality from the standpoint of a human observer.

In a previous paper [10] the authors presented a new behavioral framework able to reproduce a number of the typical features of the human behavior. The system allows the IVAs to interact among them and with the environment in a quite realistic way. A subsequent paper [8] extended the original approach by

⋆ Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 229–236, 2004.
© Springer-Verlag Berlin Heidelberg 2004

introducing some functions and parameters describing new internal, physical and mental states. The performance of that framework was also discussed in [11].

We would like to remark, however, that such a framework was exclusively designed for behavioral simulation purposes only and, consequently, it can be substantially improved in several directions. For example, neither the graphical output nor the computational efficiency did play a significant role in its design. On the other hand, it was pointed out that the use of Artificial Intelligence (AI) tools, such as neural networks and expert systems, can improve the performance of the behavioral animation schemes dramatically [7,15]. These and other extensions are the core of the present work.

This is the first in a series of two papers (both included in this volume) describing a new framework for simulating the human behavior for intelligent virtual agents. Although originally based on that introduced in [10], the current framework incorporates so many additions and improvements that it can actually be considered as a new one. Its new features concern fundamentally to the architecture and the behavioral engine. The new architecture is based on the idea of decomposing the framework into the physical and the behavioral systems and, subsequently, into their respective subsystems which carry out more specific tasks. In addition, specialized computing tools have been applied to these subsystems, so that the performance has been greatly improved. On the behavioral engine, powerful Artificial Intelligence techniques have been applied to simulate the different behavioral processes. As it will be shown later, these AI tools provide the users with a higher level of realism. Because of limitations of space, the architecture of the new framework will be described in this first paper, while the second one will focus on the application of AI tools to the behavioral engine.

The structure of this paper is as follows: in Sect. 2 we describe the main requirements of a framework to simulate the human behavior for IVAs. Then, Sect. 3 describes the architecture to fulfill those requirements. The agent's design, software tools and programming environments that have been used to implement such an architecture are also discussed in this section. Finally, Sect. 4 presents some strategies concerning the implementation of our framework on single and distributed CPU environments.

2 Framework Requirements

In this work, an *Intelligent Virtual Agent* (IVA) is the graphical representation of a virtual creature able to emulate the behavior of a living being autonomously, i.e., without the animator's intervention. Due to its inherent complexity, it is convenient to decompose our framework into different (simpler) components, which can be rather assigned to one of the following sytems:

1. the *physical system* (PS): it is responsible for the physical elements, including the 3D graphical representation of virtual agents, their motion and animation and the interaction among them and with the world's objects.
2. the *behavioral engine* (BE): it will provide the agents with emotions, feelings, thoughts, needs and beliefs (about themselves, others or the environment). Depending on their particular values, different plans will be designed by

this engine in order to accomplish the agents' goals. Although the human senses (vision, hearing, etc.) are usually associated with physical parts of our body (eyes, ears, etc.), the cognitive process itself happens at our brain, so mental routines related to perception are also included in this component. By the same reason, the different cognitive tasks related to the agent's motion control are performed at this behavioral engine[1].

Reasons for this decomposition become clear if you think about our ability to distinguish between what we are physically and mentally. In fact, we can easily assign any physical object of the 3D world (even our own body itself) to the physical system, while our emotions, beliefs, feelings or thoughts would be assigned to the behavioral engine. This separation is also extremely useful from a computational point of view. On one hand, it allows the programmer to focus on the specific module he/she is dealing with at one time. Clearly, it makes no sense to worry about the graphical motion routines when you are modifying the behavioral ones, and vice versa. On the other hand, specialized programming tools can be independently applied to each module. As a consequence, the framework's performance can be drastically optimized, provided that an adequate choice of such tools is made.

Note, however, that both systems must be strongly interconnected so that each modification in the behavioral engine (for example, if the agent is becoming tired his/her next goal might be to look for a seat to sit down) is subsequently reflected on the physical counterpart (the physical motion towards the seat) and vice versa, just as our body and brain also work as a whole. To this aim, some kind of communication between both systems must be defined. Furthermore, the better we define how these systems work and how they communicate with each other, the more effective the framework will be.

Of course, each system can be broken up into smaller subsystems, associated at its turn with more specific routines such as obstacle avoidance or path determination for the physical system, or goals or internal states for the behavioral engine. By this way, we can either work on each subsystem individually or hand out them to different people to work on. However, we should be careful with the number of levels in this sequence: indeed, too few levels will yield large codes difficult to test and debug, while too many levels will unnecesarily increase the complexity of the system.

3 Framework Architecture and Tools

3.1 Virtual Objects

The virtual agents evolve in a 3D virtual world which also comprises different kinds of objects to interact with (see Fig. 1). Basically, they can be classified into two groups: static objects and smart objects. By *smart objects* we understand those objects whose shape, location and status can be modified over time, as

[1] Note that the physical motion routines themselves still belong to the physical system. What is actually included in the behavioral engine is the simulation of the mental process that yields the orders for motion from the brain to the muscles.

opposed to the static ones. This concept, already used in previous approaches [9,13] with a different meaning, has shown to be extremely helpful to define the interactions between the virtual agents and the objects. For example, a table lamp or a radio are smart objects simply because they might be turned on/off (status set to on/off) and so are a pencil or a bottle (as they can be relocated). We point out that saying that an object is static does not mean it has null influence on the agents' actions. For instance, a tree is a static object but it should be considered for tasks such as collision avoidance and path planning.

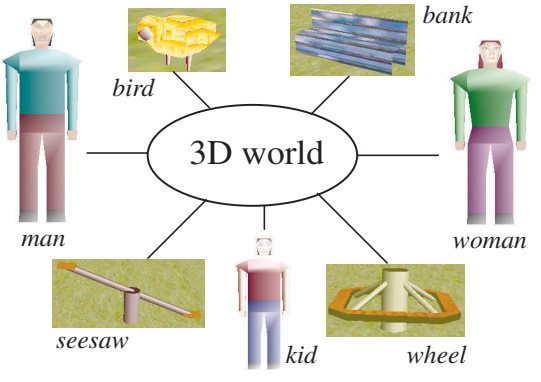


Fig. 1. The 3D world includes different kinds of virtual objects and agents

3.2 Behavioral Engine

Because the behavioral engine also includes some behavioral routines that strongly influence the graphical output (such as those for perception), we decided to split it up into the *physical control system* (PCS) and the *behavioral system* (BS), as shown in Fig. 2.

The PCS comprises two subsystems for perception and motion control tasks. The perception subsystem obtains information from the graphical environment (the virtual world) by identifying its elements (static objects, smart objects, other agents) and locations. In other words, it captures the geometry of the virtual world as it is actually done by the human beings through their senses, in which the perception subsystem is based on. On the other hand, the motion control subsystem is responsible for the conversion of the agents' plans into physical actions, as described below. At its turn, the BS (that will be described in detail in a second paper in sequence) includes several subsystems designed to perform different cognitive processes. The arrows in Fig. 2 show the information flow: the perception subsystem captures information from the virtual world which is subsequently sent to the behavioral system to be processed internally. The corresponding output is a set of orders received by the motion control subsystem,

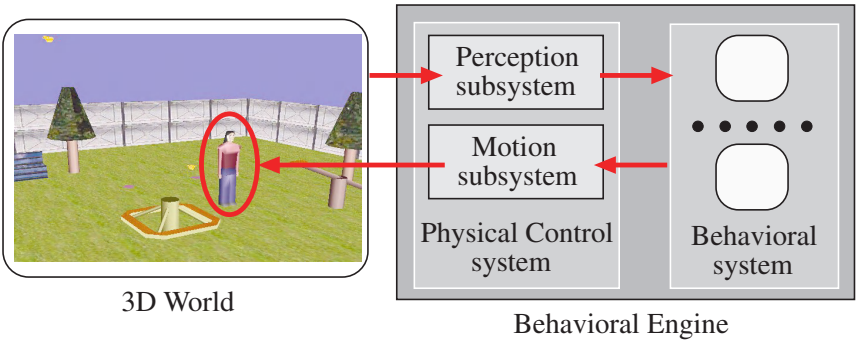


Fig. 2. Scheme of the behavioral engine of a virtual agent

which transform them into agent's physical actions animated by the physical system[2], just as the orders of our brain are sent to our muscles.

We would like to remark that this behavioral engine decomposition into the PCS and the BS is both *reasonable* and *useful*. It is reasonable because agents' reactions and decisions are mostly determined by their "personality" rather than by their physical body. Of course, the physical is also involved in "who we are", but our personality lie in another "level" of ourselves and should be analyzed separately. The usefulness comes from the fact that it is possible to reuse the BE for different virtual worlds. This leads to the concept of *adaptation*: a realistic simulation of a human being implies that the BE must be able to perform adjustments by itself in order to adapt to the changing environment. Similarly, different BE can be applied to the same virtual world. This leads to the concept of *individuality*: no two virtual agents are exactly the same as they have their own individual personality. In computational terms, this means that each virtual agent has his/her own behavioral engine, which is different from any other else.

3.3 Agents Design

As usual in Object Oriented Programming (OOP) environments, each virtual agent is represented by a class called AVA, consisting of attributes and methods. In our case, the attributes are: *AgID*, that identifies the agent, *AgSt* that accounts for the current status of the agent, and *AgVal* that stores some parameters for rendering purposes (position, direction, etc.). The methods include the *Render* method for graphical representation and those for updating the agent's attributes as a consequence of interactions with objects. Moreover, the class AVA encapsulates the attributes and methods related to the perception and the motion control subsystems. Additional methods are considered for the communication

[2] We should warn the reader about the possible confusion between "physical system" (PS) and "physical control system" (PCS). The PCS is a part of the behavioral engine, while the PS contains the routines for the graphical representation and animation of the virtual world.

from the perception subsystem to the behavioral system (*Send*) and from it to the motion control subsystem (*CallBack*). Finally, the method *Think* is used to trigger the behavioral process.

3.4 Programming Languages and Environments

Regarding the programming languages, Table 1 shows the different architecture modules of our framework as well as the software tools and programming environments used to implement such modules. The first module is the Kernel, which drives the main sequence of animation. The use of a powerful graphical library would allow the programmer to improve graphics quality dramatically with relatively little effort. By this reason, the kernel has been implemented in Open GL by using the programming environment GLUT (Open GL Utility Toolkit). The graphical representation of the virtual world (the physical system) is also a CPU demanding task. Therefore, we decided to use C++ to assure the best performance. Another reason for this choice is the excellent integration of Open GL with the C++ layer (to this purpose, we used the Visual C++ environment as programming framework). This combination of C++ with Open GL has also been used for the User Interface.

Table 1. Architecture modules of our framework and the software tools and programming environments used to implement them

Module	Software tools	Programming environment
Kernel	Open GL	GLUT
User Interface	C++ & Open GL	Visual C++ & GLUT
Physical System	C++ & Open GL	Visual C++ & GLUT
Physical Control System	C++	Visual C++
Behavioral System	C++ & Prolog	Visual C++ & Amzi! Prolog

As mentioned above, our framework consists of a physical system (PS) and a behavioral engine (BE). While the combination of C++ and Open GL works well for the physical system, the BS requires more specific tools. In particular, it has been implemented in C++ and Prolog by using the programming environment "Amzi! Prolog" (developed, at its turn, in C language). At our experience, Amzi! Prolog is an excellent tool to generate optimized code which can easily be invoked from C/C++ via Dynamic Link Libraries (DLLs), providing an optimal communication between the PCS and the BS for *standalone* applications. Furthermore, this choice provides a good solution for TCP/IP communication protocols for distributed environments, as discussed in Sect. 4.

4 Implementation on Single and Distributed CPU Environments

The framework presented in the previous sections can be developed by using only a processor or several ones. For the first case, we can either use a dynamic list of objects AVA (as shown in Fig. 3(left)) or to run each AVA in a separate process or *thread* (see Fig. 3(right)). In both cases, we must wait until all AVAs have executed to get the next animation frame. Note also that the communication between the object AVA and the behavioral system is achieved via DLLs to optimize the execution speed, avoiding other alternatives such as TCP/IP, best suited for distributed systems and networks.

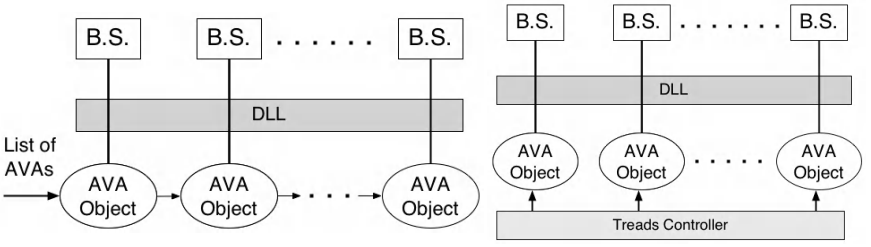


Fig. 3. Framework architectures for a single processor

Figure 4 shows the framework architecture for distributed systems. In this case, we use threads to run the different AVAs, which are connected to their corresponding BS by using sockets and TCP/IP connection. Note that parallel programming can also be applied here. For instance, we can assign each IVA behavioral system to a single processor for maximal performance.

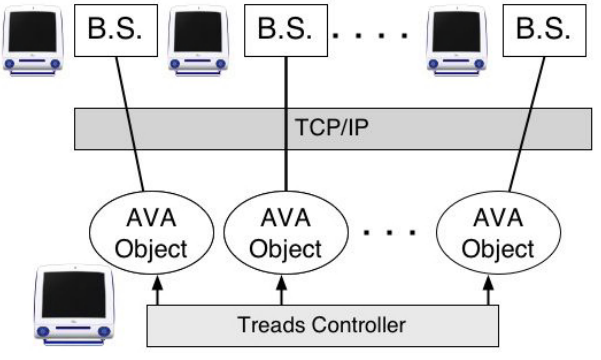


Fig. 4. Framework architecture for distributed systems

The previous single and distributed CPU architectures have been successfully implemented on PC platform (Pentium III processor). Technical details on implementation have had to be omitted because of limitations of space and will be reported elsewhere. In the second paper some interesting questions regarding the behavioral engine will be discussed.

References

1. Badler, N.I., Barsky, B., Zeltzer, D. (eds.): Making Them Move. Morgan Kaufmann, San Mateo, CA (1991)
2. Badler, N.I., Phillips, C.B., Webber, B.L.: Simulating Humans: Computer Graphics Animation and Control. Oxford University Press, Oxford (1993)
3. Blumberg, B.M., Galyean, T.A.: Multi-level direction of autonomous creatures for real-time virtual environments. Proc. of SIGGRAPH'95, ACM, New York (1995) 47-54
4. Cerezo, E., Pina, A., Seron, F.J.: Motion and behavioral modeling: state of art and new trends. The Visual Computer, **15** (1999) 124-146
5. Funge, J., Tu, X. Terzopoulos, D.: Cognitive modeling: knowledge, reasoning and planning for intelligent characters, Proceedings of SIGGRAPH'99, ACM, New York (1999) 29-38
6. Granieri, J.P., Becket, W., Reich, B.D., Crabtree, J., Badler, N.I.: Behavioral control for real-time simulated human agents, Symposium on Interactive 3D Graphics, ACM, New York (1995) 173-180
7. Grzeszczuk, R., Terzopoulos, D., Hinton, G.: NeuroAnimator: fast neural network emulation and control of physics-based models. Proceedings of SIGGRAPH'98, ACM, New York (1998) 9-20
8. Iglesias A., Luengo, F.: Behavioral Animation of Virtual Agents. Proc. of the Fourth International Conference on Computer Graphics and Artificial Intelligence, 3IA (2003) 99-114
9. Kallmann, M.E., Thalmann, D.: A behavioral interface to simulate agent-object interactions in real-time, Proceedings of Computer Animation'99, IEEE Computer Society Press, Menlo Park (1999) 138-146
10. Luengo, F., Iglesias A.: A new architecture for simulating the behavior of virtual agents. Springer-Verlag, Lecture Notes in Computer Science, **2657** (2003) 935-944
11. Luengo, F., Iglesias A.: Animating Behavior of Virtual Agents: the Virtual Park. Springer-Verlag, Lecture Notes in Computer Science, **2668** (2003) 660-669
12. Maes, P., Darrell, T., Blumberg, B. Pentland, A.: The alive system: full-body interaction with autonomous agents, Proceedings of Computer Animation'95, IEEE Computer Society Press, Menlo Park (1995) 11-18
13. Monzani, J.S., Caicedo, A., Thalmann, D.: Integrating behavioral animation techniques, Proceedings of EUROGRAPHICS'2001, Computer Graphics Forum, **20**(3) (2001) 309-318
14. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds, Proceedings of SIGGRAPH'96, ACM, New York (1996) 205-216
15. Van de Panne, M., Fiume, E.: Sensor-actuator networks, Proceedings of SIGGRAPH'93, Computer Graphics **27** (1993) 335-342

Framework for Simulating the Human Behavior for Intelligent Virtual Agents. Part II: Behavioral System

F. Luengo[1,2] and A. Iglesias[2⋆]

[1] Department of Computer Science, University of Zulia, Post Office Box #527, Maracaibo, Venezuela
fluengo@cantv.net

[2] Department of Applied Mathematics and Computational Sciences, University of Cantabria, Avda. de los Castros, s/n, E-39005, Santander, Spain
iglesias@unican.es
http://personales.unican.es/iglesias

Abstract. This paper is the second in a series of two papers (both included in this volume) describing a new framework for simulating the human behavior for intelligent virtual agents. This second paper focuses on the application of Artificial Intelligence (AI) techniques to the simulation of the human cognitive process. The paper discusses some important issues involved in this process, such as the representation and identification of objects, the information acquisition and its conversion into knowledge and the learning process. The paper also describes how some standard AI techniques (expert systems, neural networks) have been applied to tackle these problems.

1 Introduction

In the first part of this work we have reviewed some features of the architecture of a new framework for simulating the human behavior for intelligent virtual agents. In addition, we analyzed the software and programming environments used to implement such a framework, with emphasis on the graphical part. Fortunately, the huge number of software applications for 3D graphics and animation allow us to apply well-known standarized tools. The challenge is to develop a similar "machinery" for human behavior simulation.

So far, little effort was placed upon the simulation of the human cognitive processes (learning, memory, recognition, etc.) from the viewpoint of Computer Graphics. Notable exceptions are the works in [2,3,4,7,8,9,10]. In contrast, this is the primary goal of the Artificial Intelligence (AI) field. After all, most of the AI techniques (such as neural networks or expert systems) are based on the idea of reproducing the structure and behavior of the human brain. Consequently, it seems very reasonable to apply them to the simulation of the intelligent virtual agents (IVAs). This is actually the core of this paper. In particular, the paper

⋆ Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 237–244, 2004.
© Springer-Verlag Berlin Heidelberg 2004

discusses some important issues involved in this process, such as the representation and identification of objects, the information acquisition and its conversion into knowledge and the learning process. The paper also describes how some standard AI techniques (expert systems, neural networks) have been applied to tackle these problems.

2 General Scheme of the Behavioral System

The realistic simulation of the behavior of virtual agents implies that they must be able to carry out an intelligent exploration of the surrounding environment. By *intelligent* we mean that the IVAs need to walk through three main steps:

1. to identify the different objects from the virtual world (*object recognition*)
2. to obtain information from the environment (*information acquisition*)
3. this information is subsequently processed so that the agents can effectively acquire new knowledge and/or update the current one (*knowledge acquisition*).

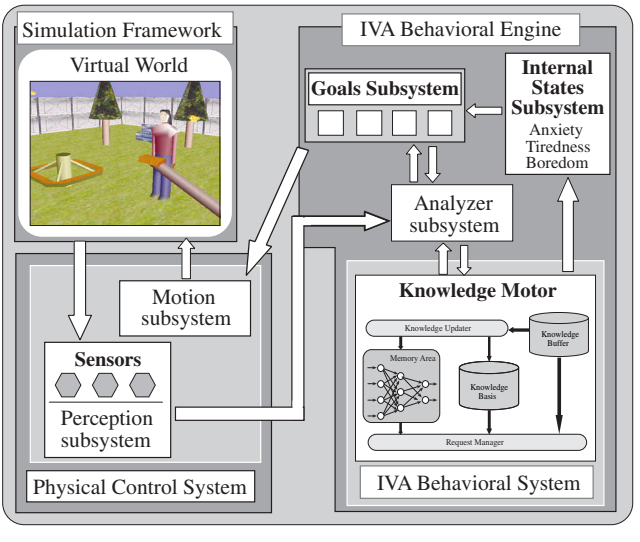


Fig. 1. General scheme for information acquisition and its conversion into knowledge

All these tasks are performed by specific subsystems and modules, as depicted in Fig. 1. The *perception subsystem* (PSB) applies routines to identify the objects and to extract information from the 3D world. Such information is subsequently sent to the *analyzer subsystem*, where the information is processed and transformed into knowledge. The *internal states subsystem* handles the information about agents personality and his/her "emotional state". With that

information, the *goal engine subsystem* updates the goals list, thus determining what the agent wants to do. Finally, the *action engine subsystem* takes a decision about the best way to achieve those goals, updates the agents status, and sends that information to the *motion subsystem* to complete the animation. The following paragraphs will analyze how these tasks have been accomplished.

2.1 Objects Representation and Identification

In order to interact with the 3D world, the IVA must be able to identify its elements, regardless their nature (smart objects, static objects, other agents) and properties (location, status, etc). These properties provide essential information, as they will determine the kind of agent-object interaction and, consequently, the future agent's actions. On the other hand, we would like this object representation to be as simple as possible. This feature is required for efficient manipulation and memory storage.

In this paper we use a representation scheme based on biological concepts such as chromosome and gene. Each object of the 3D world is represented by what we call a *chromosome*. Roughly speaking, it is a collection of multicomponent *sequences* which, at their turn, comprise single fields called *genes*, as shown in Fig. 2. For example, the chromosome in this figure consists of m sequences and n genes ($n > m$).

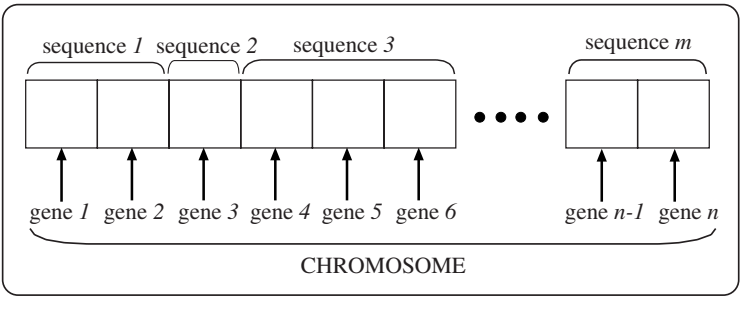


Fig. 2. Structure of the chromosome representing the objects of the 3D world

Each sequence corresponds to a certain characteristic of the object. The sequences are sorted by following a hierarchical structure, as shown in Fig. 3. In this work, we consider that the objects' chromosomes are composed of four sequences, from the most general to the most specific one: the first sequence consists of three genes that account for objects, animals and people (sequences *[1,0,0]*, *[0,1,0]* and *[0,1,1]*, respectively). The second sequence, also with three genes, adds more information about general characteristics, such as the kind of object, animal or person. In this example, the category *person* is subsequently subdivided into *kid*, *adult* and *elderly*. The third sequence consists of one gene

and is associated with the status, size or gender (for object, animal or person, respectively). Finally, the last sequence comprises five genes to store the object's ID, in order to identify an specific element within its own class.

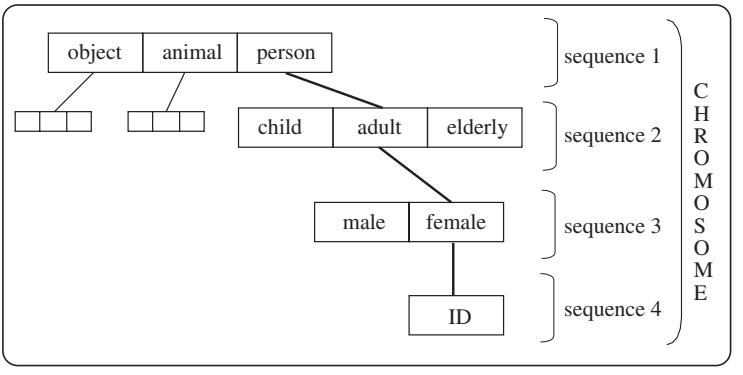


Fig. 3. Sequences of the chromosome

For example, the first woman in our environment is represented by the chromosome *[0,1,1,1,1,1,0,1,0,0,0,0,1]*. This representation is useful for identification, provided that a mathematical function to compute the distance between two arbitrary elements is defined. Given a pair of elements, the goal of such a function is to determine how close these elements are (in other words, such a function constitutes "de facto" a criterion for similarity). The *distance function* between two elements A and B at a sequence j is defined as follows:

$$dist(j, A, B) = \frac{1}{k} \sum_{i=1}^{k} |A_i^j - B_i^j| \tag{1}$$

where A_i^j denotes the ith gene at sequence j for the chromosome A, and k denotes the number of genes of such a sequence. Note that we can think of sequences in terms of levels in the tree displayed in Fig. 3. The sequence j is simply the level j down the tree at which it appears, with the top of the tree as sequence 1. We will say that two elements A and B are *similar at sequence* (or *at level*) j if $dist(j, A, B) = 0$. Further, they are *similar up to sequence* s if $dist(r, A, B) = 0$, $\forall\, r \leq s$. Note that the hierarchical structure described above imply that an arbitrary object is closer to that minimizing the distance at earlier sequences. For instance, an adult is represented by the sequence $[1, 1, 0]$ which is in-between the sequence for kids $[1, 0, 0]$ and for elder people $[0, 1, 0]$, since $dist(adult, kid) = dist(adult, elder) = \frac{1}{3}$ whereas $dist(kid, elder) = \frac{2}{3}$, meaning that an adult is closer to an elder person than a kid. Therefore, Eq. (1) provides an accurate procedure to classify objects at a glance, by simply comparing them sequentially at each level.

2.2 Information Acquisition

In this step, the *analyzer subsystem* receives the world information acquired by the PSB and then analyzes it to update the *knowledge base* accordingly. As explained in the first paper, the *perception subsystem* has been developed in a software environment different than that for the *behavioral system* (BS). Therefore, it is extremely important to define clearly a communication protocol for information exchange between both systems. In that protocol, carrying-information comprises four fields (see Fig. 4): a parameter specifying the information source (vision, hearing), the object ID or chromosome (see Sect. 2.1 for details), additional information about the location, status, etc. and a parameter called *impact index*. This last parameter is added by the analyzer to account for the impact of a new information on the agent and will be detailed later on.

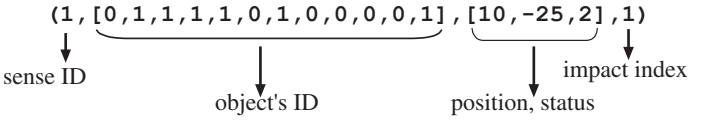


Fig. 4. Information exchange between the perception and the behavioral systems

2.3 Knowledge Acquisition

Once new information is acquired and then processed by the analyzer, it is sent to the *knowledge motor*, whose main components are displayed in Fig. 5. Firstly, the current information is temporarily stored into the *knowledge buffer*, until new information is attained. At that time, previous information is sent to the *knowledge updater*, the new one being stored into this buffer and so on.

The *knowledge base* is actually a based-on-rules expert system, containing facts and inference rules. In addition to the information provided by the updater, the facts include complex relationships among the different elements (personal relationships among agents such as friendship, relative positions of objects, etc). The inference rules, based on deductive schemes such as modus ponens, modus tollens, rule chaining, goal-oriented rule chaining and others(see, for instance, Chapter 2 of [1]), provide the system with the tools to infer new knowledge from the current one. Of course, the system's complexity is mostly determined by the number of rules and the design of the inference engine. Additional subsystems for other tasks (coherence control, action execution) have also been incorporated.

The *memory area* is a neural network that will be applied to learn from data (in our problem, the information received from the environment through the perception subsystem). A *neural network* consists basically of one or several layers of computing units, called *neurons*, connected by links. Each artificial neuron receives an input value from the input layer or the neurons in the previous layer.

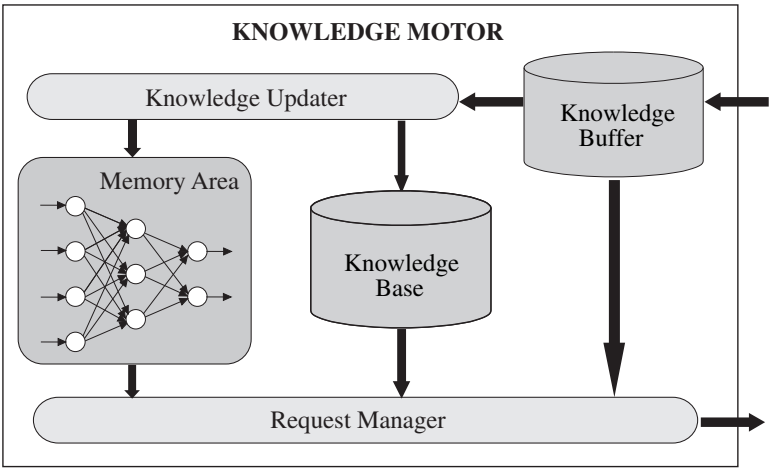


Fig. 5. Scheme of the knowledge motor and its components

Then it computes a scalar output $y = f\left(\sum w_{ik}x_k\right)$ from a linear combination of the received inputs $x_1, x_2, \ldots, x_n$ using a set of weights w_{ik} associated with each of the links and a given scalar function f (the *activation function*), which is assumed to be the same for all neurons (see [5] and [6] for details).

Among the many interesting properties of a neural network, one of primary importance is its ability to learn from the environment and to improve its performance through learning. Such an improvement takes places over time through an iterative process based on adjust the free parameters of the network (the weights). In this paper we consider the *unsupervised learning*, in which the data is presented to the network without any external information and the network must discover by itself patterns, or categories. In particular, we use an autoassociative scheme, since the inputs themselves are used as targets. In other words, the networks tries to learn the identity function, which is a problem far to be trivial as the network contains less neurons than the input/output layers, and hence, the network must perform dimensionality reduction. What the network attempts is to subdivide the chromosome space into clusters in order to associate each chromosome with a specific neuron, the nearest one in our case. To this end, we try to minimize the sum of the squared within-groups residuals, which are basically the distances of the chromosome locations to the respective group centroids. When a new chromosome is received as input, the whole structure is recomputed and the group centroids are relocated accordingly. This problem can be overcome by applying the K-means least-squares partitioning algorithm, a procedure to divide a collection of n objects into K groups. The basic algorithm consists of two main steps:

- compute cluster centroids and use them as new cluster seeds
- assign each chromosome to the nearest centroid.

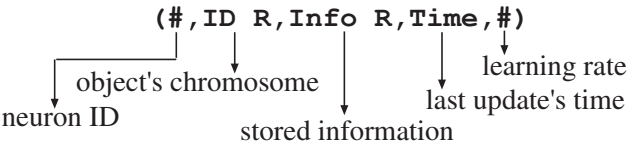


Fig. 6. Information received by the neurons

In our case, each neuron should receive the information shown in Fig. 6, namely, the neuron ID, the object's chromosome, the information to be stored by the neuron, the time at which this information is attained (which will be used for animation purposes), and the learning rate. This last parameter is introduced to describe the neuron's ability to adapt to a new information (and simultaneously, to forget the previous one). Its meaning becomes clear by simply noticing that, in our daily life, we can learn, understand and remember certain things completely, partially and sometimes not at all. In fact, certain things can never be forgotten. This "unforgettable" information is assigned to neurons whose learning rate is set to 0 so that the information is permanently stored. By this way we can deal with information which, although extremely important (i.e., with high impact index), has been received only once.

2.4 Learning Process

Let us suppose that we have a neural network with k neurons and that n data vectors $x_1, x_2, \ldots, x_n$, (with $k < n$) will eventually be perceived at different times. To update the memory area, we employ a K-means procedure for competitive networks, which are a popular type of unsupervised network architectures widely used to automatically detect clusters, or categories, within the available data. A simple competitive neural network is formed by an input and an output layer, connected by feed forward connections. Each input pattern represents a point in the configuration space (the space of inputs) where we want to obtain classes.

This type of architecture is usually trained with a *winner takes all* algorithm, so that only the weights associated with the output neuron with largest value (the winner) are updated. The procedure is based on the following strategy: at the initial stage, all the neurons are available to store new data. Therefore, the first k data vectors are sequentially assigned to these neurons, i.e., data x_i is learned by neuron i, $1 \leq i \leq k$. Simultaneously, time for neuron i is initialized to the moment at which data x_i is learned. Once the next data x_{k+1} is received, it is assigned to the neuron j such that

$$d(x_j, x_{k+1}) \leq d(x_i, x_{k+1}), \qquad \forall i = 1, \ldots, k, \quad i \neq j \tag{2}$$

When this condition is satisfied by several neurons simultaneously, the new data is assigned to that storing the oldest information. Interesting enough is the way

in which the neuron stores the new information: instead of replacing the old data by the new one, what is actually stored is a combination of both data. The basic idea behind this formulation is to overcome the limitation of having more data than neurons by allowing each neuron to store more than one data at the same time. Thus, the neuron does not exhibit a deterministic output but a probabilistic one: what is actually computed is the probability of a neuron to have a particular data at a particular time. This probability is continuously updated in order to adapt our recalls to the most recent data. This leads to the concept of *reinforcement*, based on the fact that the repetition of a particular event over time increases the probability to recall it. Of course, some particular data are associated with high-relevance events whose influence does not decrease over time (or decreases so slowly that it can be considered as a time-independent event). In those cases, the neuron must be able to store this data and maintain its probability regardless the time. The learning rate parameter introduced in Sect. 2.3 is intended to play this role.

Finally, we would like to remark that this scheme improves substantially the deterministic approaches for short-medium-long (SML)-term memory by introducing uncertainty on the agent's recalls. Combination of this scheme and fuzzy logic constitutes a better approach to the human recall process and it is currently being investigated. The conclusions of this study will be the subject of a future publication.

References

1. Castillo, E., Gutiérrez, J.M., Hadi, A.: Expert Systems and Probabilistic Network Models. Springer-Verlag, New York (1997)
2. Funge, J., Tu, X. Terzopoulos, D.: Cognitive modeling: knowledge, reasoning and planning for intelligent characters, Proceedings of SIGGRAPH'99, ACM, New York (1999) 29-38
3. Granieri, J.P., Becket, W., Reich, B.D., Crabtree, J., Badler, N.I.: Behavioral control for real-time simulated human agents, Symposium on Interactive 3D Graphics, ACM, New York (1995) 173-180
4. Grzeszczuk, R., Terzopoulos, D., Hinton, G.: NeuroAnimator: fast neural network emulation and control of physics-based models. Proceedings of SIGGRAPH'98, ACM, New York (1998) 9-20
5. Haykin, S.: Neural Networks. A Comprehensive Foundation. Macmillan Publishing, Englewood Cliffs, NJ (1994)
6. Hertz, J., Krogh, A., Palmer, R.G.: Introduction to the Theory of Neural Computation. Addison Wesley, Reading, MA (1991)
7. Monzani, J.S., Caicedo, A., Thalmann, D.: Integrating behavioral animation techniques, Proceedings of EUROGRAPHICS'2001, Computer Graphics Forum, **20**(3) (2001) 309-318
8. Ridsdale, G.: Connectionist modeling of skill dynamics. Journal of Visualization and Computer Animation, **1**(2) (1990) 6672
9. Sims, K.: Evolving virtual creatures, Proceedings of SIGGRAPH'94, ACM, New York (1994) 15-22
10. Van de Panne, M., Fiume, E.: Sensor-actuator networks, Proceedings of SIGGRAPH'93, Computer Graphics **27** (1993) 335-342

Point-Based Modeling from a Single Image

Pere-Pau Vázquez[1], Jordi Marco[1], and Mateu Sbert[2]

[1] Dept. LSI - Universitat Politècnica de Catalunya, Spain
{ppau|jmarco}@lsi.upc.es
[2] IIiA, Universitat de Girona, Spain
mateu@ima.udg.es

Abstract. The complexity of virtual environments has grown spectacularly over the recent years, mainly thanks to the use of the currently cheap high performance graphics cards. As the graphics cards improve the performance and the geometry complexity grows, many of the objects present in the scene only project to a few pixels on the screen. This represents a waste in computing effort for the transforming and clipping of maybe a lot of polygons that could be substituted by a simple point or a small set of points. Recently, efficient rendering algorithms for point models have been proposed. However, little attention has been focused on building a point-based modeler, using the advantages that such a representation can provide. In this paper we present a modeler that can generate 3D geometry from an image, completely built on points. It takes as input an image and creates a point-based representation from it. Then, a set of operators allow to modify the geometry in order to produce 3D geometry from the image. With our system it is possible to generate in short time complex geometries that would be difficult to model with a polygon-based modeler.

1 Introduction

The complexity of virtual environments has grown spectacularly over the recent years, mainly thanks to the, now affordable, high performance graphics cards. These highly complex models are made of objects that usually cover only a few, or even fractions of, pixels on the screen. Polygon-based systems are advantageous for the modeling of large objects, but, for small ones, many resources are wasted by transforming and clipping geometry which is either invisible, or very small when projected to the screen. This has led to the investigation of alternatives to pure polygon-based rendering in recent research. An interesting alternative is point-based rendering [1,2,3,4,5].

In this paper we present a modeler that takes advantage of a point-based representation to easily build 3D geometry starting from a real (or synthetic) image. It takes as input an image, builds a point-based model, and then offers a set of operators that allow the user to easily manipulate the point-based geometry and create very interesting modifications of the geometry and obtain a 3D model. Our system does not pretend to reproduce exactly a real object but offers a means to create a realistic object starting from a real image. This allows

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 245–251, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Fig. 1. The famous Dalí bread on the wall of the museum (left and right) and a 3D modification of it (center)

to create models of both small and large objects in several minutes. An example of this is shown in Fig. 1. Figure 1-left shows a photograph of one of the famous bread pieces that are part of the wall of the Salvador Dalí's museum in Figueres. Figure 1-center shows how it looks after a manipulation the user carried out in only two minutes and five seconds. In Fig. 1-right we can see the photograph as seen from the same angle.

The rest of the paper is organized as follows: Sect. 2 surveys related work. Section 3 gives an overview of our system and the user interaction method. In Sect. 4 we describe the basic operators available and show some examples. Finally, Sect. 5 concludes our work pointing to some lines of future research.

2 Previous Work

Modeling complex realistic objects is a difficult and time consuming process. Though there are many systems for modeling the real world from images, reconstruction of complex and detailed geometry is still cumbersome. Reconstructing 3D geometry from one, two, or more photos is a fundamental problem that has received extensive attention in computer vision [6,7]. Although the best results are obtained in controlled conditions, a good and complete solution has not been found yet. Another possibility is the use of expensive scanning devices, however this presents serious problems for large objects such as buildings and statues [8].

Image-Based systems replace partially or completely the geometry with images, the reconstruction techniques are similar or borrowed from Computer Vision and therefore suffer from the same drawbacks. Some examples are [9,11]. As Poulin *et al.* [10] argue, the user intervention is very important in order to improve the quality of the results of a reconstruction. They present an interactive modeling system that extracts 3D objects from photographs, using a point-based representation that allows the user to intervene and improve the obtained model. However, they not allow a *direct intervention* of the user on the geometry. In this sense, our work can be seen as complementary to theirs.

Several approaches exist on building point based representations from a synthetic geometry [2,11,12,13,14,5], where the problem of sampling is also focused.

3 Overview of the System

Our system consists in two parts, a rendering tool, which shows the results of the modifications to geometry, and a manipulation tool, that presents the initial image and allows model the geometry using the provided operators. Initially, the image is read and for each position a 3D point (with $Z = 0$) is created. In order not to lose resolution, both the rendering and the manipulation tools open windows of a size equal to the size of the image, and the point model has a point for every pixel of the image.

The **rendering tool** is allows the camera movements needed inspect the changes made to the model, that is, rotation and translation. We have decided to render the objects in orthographic view in order to avoid the perspective projection to hide details in the comparison of the generated model with the initial image. Consequently, the rendering window and the manipulation one start showing the same image. The rendering window also serves to show how the next operator will affect the resulting model (actually it shows the region that will be affected if set as modifiable) in order to aid the user to accurately predict the final changes.

The **manipulation tool** shows the image that will be modified and allows to select the regions that are going to be changed, together with the operators that will be applied. Our system allows for modifications on a single direction the Z axis. This is due to the fact that, as we are working mainly with real photographs, what there is lacking is the Z coordinate, which is what we want to somehow build or imitate. We do not want to build the 3D geometry represented in the image according to the exact measures, but, as it will be seen, our system can be extended to work with real measures, we only need to know, as a reference, the measure of some feature appearing in the image.

No special knowledge of computer graphics or 3D geometry is required to work with the modeler. The user first selects a set of regions and then sets the operator to be applied to each region. We describe now the main issues.

Region Selection: There are two kind of regions, a) *main* region, and b) *secondary regions*. All the points in the main region are transformed in the same way: their Z value is changed according to the user's definition. Secondary regions are changed according to the operator applied. A selection may be of three kinds: a) A set of (one or more) points, b) a single line, and c) a polygon. Regions consisting of single points are used for the main region as they permit very powerful manipulations to geometry, as we will see soon. In order to allow for continuous regions selections, our selection method first tests if the point we want to add to our region does currently form part of a previously defined region, if the point passes this test, it also checks if there is a previously selected region in its near environment (4 o 5 pixels around), if so, the selected point is moved next to the close region, as we assume that two regions that are very close are going to be modified together (maybe with different operators) and we do not want discontinuities (points still lying on the Z plane) on the result. We also allow to select a region that contains another region, which is very useful to define conic or spherical transformations.

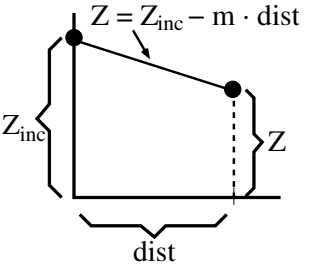


Fig. 2. Lineal operator

Operators: When the regions are already selected, we define an operator for each one (main region is always modified using the *s*tair operator). The modification consists in changing the Z value of the points inside the main region and the Z values of the points belonging to the secondary regions. The main region is assigned a constant value. Points inside secondary regions are modified according to the distance to the closest point of the main region. For a linear transformation, the new Z value of a point will be $Z = max(0, Z_{inc} - m \cdot dist)$, where Z_{inc} is the Z value assigned to the main region, as depicted in Fig. 2. Some of the regions (up to 10) can also be selected through a menu, this way it is easy to select regions formed by single points or single lines.

Hole filling: Each change in the geometry, may generate holes. Our system uses a hole filling process that is applied to the resulting model. Although we have not dealt with efficiency issues yet, and this is a very costly process, our system is able to work interactively for relatively large images. We have found that there is a delay of some seconds (4 or 5) for the whole process (modification and hole filling) in the case of quite large images (500×750 pixels) if the regions to modify are also large.

4 Operators

Differently to Poulin *et al.* [10], our operators allow to directly modify the 3D geometry. In their framework they offer tools for filling with points (the interactive modeling tool), generating more points using rules, point jittering or merging a neighborhood of points. Our philosophy is totally different, we want tools for the easy generation of shapes on the Z coordinate in order to create objects with realistic appearance. The operators implemented are: stair operator, flatten operator, lineal operator, cosine-based operator, and cape operator.

The **stair operator** performs a constant change on the Z axis. When applied to any region, each point is assigned the Z value determined by the user. In Fig. 3 we can see an example of this tool. The image on the left shows the entrance of a house, and on the right we have *pulled* the wall all around. Note that the stair function has generated a realistic 3D geometry. Only at the left part of the

Fig. 3. Stair operator

door the colours of the bricks where not copied when filling holes because the region selection at this point overlapped a little part of the wooden material. The **flatten operator** restores the points of the selected region to $Z = 0$ and eliminates the extra points that were added for hole filling during the geometry update.

The **lineal operator** serves to create slopes. The lineal operator modifies the the points of secondary regions using the following equation: $Z = max(0, Z_{inc} - m \cdot dist)$, where Z_{inc} is the Z value assigned to the main region, $dist$ is the distance to the main region and m is the slope of the linear function. Available values for m are: 0.25, 0.5, 0.75 and 1 to 5. The **conic operator**, similar to this one, creates a linear modification with a slope determined by the closer point to the main region (Z will be Z_{inc}) and the farther one (Z will be 0).

Fig. 4. Cape operator

The **cosine-based operators** allow to create roun-ded shapes. These operators allow to modify a region with a curve that is a function of the cosine of the distance, so the points in the secondary regions will have a new Z value following the formula $Z = max(0, Z_{inc} - \cos(m \cdot dist))$ where m has values from

1 to 5 and from 0.5 to 0.9 (with increments of 0.1). We also offer the functions $Z = max(0, Z_{inc} - (1/2)\cos \cdot dist)$ and $Z = max(0, Z_{inc} - (1/3)\cos \cdot dist)$. To help the user predicting the results, when one of this operations is selected, for the case of a main region consisting in a single point, the prediction of the influence area is shown as a disk in the rendering window. In this case the user may not be able to easily determine which of these functions is adequate. The operator dubbed **cape operator**, allows to create a spherical shape. It is the counterpart of the the conic operator but with round shape. In Fig. 4-right we can see a battery modeled with this operator. The user modified the model in less than 15 seconds.

5 Conclusions and Future Work

In this paper we have presented a system based on a point representation that is able to create very interesting effects in little time. Figure 5-rigth shows two examples where the main region consisted on isolated points and a linear (top) and a cosine-based (bottom) operators were applied. These manipulations only needed 90 seconds of user intervention.

Fig. 5. Different operators applied on the same image

Some other operators should be useful, for instance texture application over regions, or smoothing surfaces. A symmetry operator that could copy the same image seen from the front to the back, which is straightforward, could be interesting for symmetric objects. Some other extra functions are also possible such as using real measures from a certain part of the image if available, or the "intelligent scissors" to easily define selections that adapt to the image.

Acknowledgments. Partially supported by TIC2001-2416-C03-01 from the Spanish government, and SGR2001-00296 from Catalan government.

References

1. M. Levoy and T. Whitted. The use of points as a display primitive. Technical Report TR 85-022. CS Department, University of North Carolina at Chapel Hill, January 1985. Available from http://www-graphics.stanford.edu/papers/points/.
2. J.P. Grossman and W.J. Dally. Point sample rendering. In George Drettakis and Nelson Max editors, editors, *Rendering Techniques'98*, pages 181–192. Springer-Verlag, 1998.
3. H. Pfister, M. Zwicker, J. van Baar, and M. Gross. Surface elements as rendering primitives. In Kurt Akeley, editor, *SIGGRAPH 2000, Computer Graphics Proceedings*, Annual Conference Series, pages 335–342, Los Angeles, July 2000. ACM Press / ACM SIGGRAPH / Addison Wesley Longman.
4. S. Rusinckievicz and M. Levoy. QSplat: A multiresolution point rendering system for large meshes. In K. Akeley, editor, *SIGGRAPH 2000, Computer Graphics Proceedings*, Annual Conference Series, pages 343–352, Los Angeles, July 2000. ACM Press / ACM SIGGRAPH / Addison Wesley Longman.
5. M. Stamminger and G. Drettakis. Interactive sampling and rendering for complex and procedural geometry. In *Rendering Techniques '01, EG workshop on rendering*, pages 151–162, June 2001.
6. S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proc. of the Computer Vision and Pattern Recognition Conference*, pages 1067–1073, 1997.
7. Tomáš Werner, Tomáš Pajdla, and Martin Urban. REC3D: Toolbox for 3D Reconstruction from Uncalibrated 2D Views. Technical Report CTU-CMP-1999-4, Czech Technical University, FEL ČVUT, Karlovo náměstí 13, Praha, Czech Republic, December 1999.
8. M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk. The digital michelangelo project: 3D scanning off large statues. In K. Akeley, editor, *SIGGRAPH 2000, Computer Graphics Proceedings*, pages 131–144. ACM Press / ACM SIGGRAPH /Addison Wesley Longman, 2000.
9. P.E. Debevec, C.J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Computer Graphics Proceedings (Proc. SIGGRAPH ' 96)*, pages 11–20, 1996.
10. Pierre Poulin, Marc Stamminger, Francois Duranleau, Marie-Claude Frasson, and George Drettakis. Interactive point-based modeling of complex objects from images. In *Proceedings of Graphics Interface 2003*, pages 11–20, June 2003.
11. L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. *Proc. of SIGGRAPH 95*, pages 39–46, August 1995.
12. D. Lischinski and A. Rappoport. Image-based rendering for non-diffuse synthetic scenes. In George Drettakis and Nelson Max editors, editors, *Rendering Techniques'98*, pages 301–314, 1998.
13. S. Fleishman, D. Cohen-Or, and D. Lischinski. Automatic camera placement for image-based modeling. *Computer Graphics Forum*, 19(2):101–110, Jun 2000.
14. P.-P. Vázquez, M.Feixas, M.Sbert, and W.Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modeling. *Computer Graphics Forum*, 22(4):689–700, Dec 2003.

Introducing Physical Boundaries in Virtual Environments

Pilar Herrero and Angélica de Antonio

Facultad de Informática. Universidad Politécnica de Madrid.
Campus de Montegancedo S/N.
28.660 Boadilla del Monte. Madrid. Spain
{pherrero,angelica}@fi.upm.es

Abstract. In this paper, we present an application developed at the Universidad Politécnica de Madrid with the aim of introducing boundaries' deformations into virtual environments. These studies have been run from a geometrical point of view with the aim of analysing how boundaries can modify the shape of some of the key concepts introduced into one of the most successful awareness models in Computer Supported Collaborative Work (CSCW), known as the Spatial Model of Interaction (SMI).

1 Introduction

The Spatial Model of Interaction (SMI) [1], as its name suggests, uses the properties of space as the basis for mediating interaction. It allows objects in a virtual world to govern their interaction through some key concepts – such as medium, aura, awareness, focus, nimbus, adapters and boundaries.

In the SMI the *focus* was understood as the observing object's interest, the *nimbus* was introduced as the observed object's projection and *boundaries* were identified as those objects that divide the space into different areas or regions, controlling the interaction between all of them. An object's focus and nimbus, can be modified through boundaries.

The main concept involved in controlling interaction between objects is *awareness*. Awareness quantifies and qualifies the interaction between every pair of objects in a given medium, being manipulated via *focus* and *nimbus*.

In this paper we are going to analyse how boundaries can modify the shape of focus and nimbus and how these modifications depend not just on the kind of boundary but also on the boundary's shape. We are also going to introduce some of the algorithms developed at the Universidad Politécnica de Madrid with the aim of introducing these deformations inside several projects and applications.

2 Physical Deformation

Starting from previous implementations of these concepts [2,3] where focus has been implemented as a triangle and nimbus has been implemented as a circumference, we have made a couple of assumptions: the focus shape was a circular sector and the nimbus shape was a circumference.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 252–257, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2.1 Deforming the Object Focus

The pseudo-code associated with this deformation was:

1. Detect the boundaries with which it is possible to interact.
2. Having determined these boundaries, for each one (O_i), it is necessary to get:
   a. The set of boundary vertices, which we have called the "*Significant Points"*
   b. Having established the vertices, for each one (V_{ij}):
      i. Define segments between the centre of the focus shape and the vertex (V_{ij}).
      ii. Having established the segments, calculate the straight line defined by each of these segments (S_{ij}) and the other boundaries.
      iii. Taking into account the set of straight lines obtained in the previous point, calculate which intersect with the focus shape.
      iv. From the points of intersection calculated in the previous step, get the points that are the closest to the centre of the focus shape (*OC)* and furthest away from V_{ij}. Each of these points will be called P_{ij} and will be part of the deformed shape. If V_{ij} is not coincident with P_{ij}, then V_{ij} will also be part of the deformed shape. However, if there is any point of intersection closer to OC than V_{ij}, there will be no points of deformation related to this vertex, and the set of points obtained will be part of the set of "*Final Points*" used in the deformation.
3. From the previous step, we got all the points necessary to define the deformation of the focus shape. However, the points have to be ordered by an ordering algorithm to define the correct deformation shape.
4. It could be necessary to add some of the circular sector vertices to the list of "*Final Points*" (see section 2.1.2)
5. Divide the deformed circular sector shape into different simple figures like circular sectors and triangles to build the resulting deformed circular sector shape taking the previous points as starting points.
6. Finally, draw the deformed circular sector shape, tracing the outline of the figures that define the outline of the deformed circular sector shape (Figure 1).

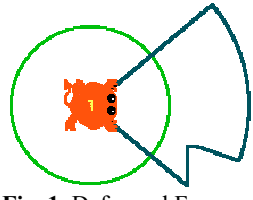


Fig. 1. Deformed Focus

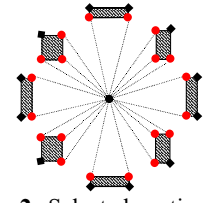


Fig. 2. Selected vertices with a circle and discarded vertices with a diamond

2.1.1 Focus Significant Points

Focus Significant Points are points that help to identify the shapes contributing to the deformed focus shape. Initially, significant points are determined by each of the boundaries that intersect with the deformed focus shape in the environment.

At first glance, one might think that these points are the vertices of these boundaries. However, this not the case, because depending on the boundary position

related to the centre of the circular sector shape, we will have to discard the vertices that are hidden by the actual boundary (see the diamond vertices in Figure 2).

Having discarded these vertices, we also have to discard, from the remaining set of vertices, the vertices that are outside the circular sector shape. Instead, we are going to consider the points of intersection between the edges coming out of these vertices and the border of the circular sector shape (see the diamond vertices in Figure 3).

2.1.2 Final Points

The best way of understanding what points have to be considered as *Final Points* is by means of an illustration using a couple of examples.

If boundaries do not interfere with each other, then each of the boundaries is contributing independently to the set of final points (Figure 4). Moreover, it is necessary to add the three vertices of the circular sector.

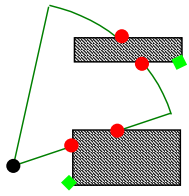


Fig. 3. Selected vertices with a circle and discarded vertices with a diamond

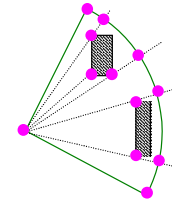


Fig. 4. Final Points

However, if a part of the object is in hiding, as is the case in the example shown in Figure 5, the final points provided by each of the boundaries have to be carefully analysed.

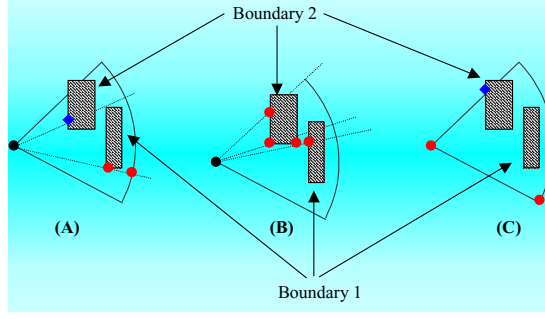


Fig. 5. Final Points with a circle

In Figure 5 (A), we can see that of all the final points that the boundary 1 should provide – 4 points if no other boundary was interfering – just 2 of these points are contributing as final points, because of the presence of boundary 2 (the diamond point means that the other two points cannot be introduced). In Figure 5 (B), we can see the final points provided by boundary 2, one of the final points provided by this boundary is on boundary 1's edge. In Figure 5 (C), we can see the final points provided by the vertex of the circular sector. From this figure, we can appreciate that the diamond

point is obstructing the incorporation of the third vertex. The set of final points is the union of all the final points provided by each of these boundaries and the circular sector figure.

2.1.3 Ordering Algorithm

We have used a basic ordering criterion and some additional criteria for some specific situations that could occur while the application is running [4].

The basic ordering criterion is that given two points P_1 and P_2, P_1 is said to be greater than P_2 related to the origin of co-ordinates OC and the origin angle α (Figure 6), and we write $P_1 >_{OC\alpha} P_2$, if $\alpha_1 \neq \alpha_2$ and $\alpha_1 > \alpha_2$, where: α is the angle defined by the lower edge of the circular sector; α_1 is the angle of the straight line defined by OC-P_1 dependent on α; α_2 is the angle of the straight line defined by OC-P_2 dependent on α.

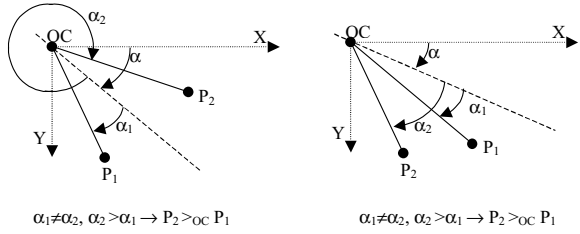


Fig. 6. Ordering Criterion

2.2 Deforming the Object Nimbus

The sequence of steps in the pseudo-code is similar for deforming both focus and nimbus, the difference lying the internal procedure for completing each of the steps, that is, how the list of significant or final points is calculated. The pseudo-code for deforming the focus shape has another important difference: an additional step (number 4), which is not necessary for deforming the nimbus shape. The deformed circular shape is showed at the Figure 7.

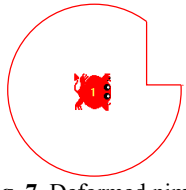


Fig. 7. Deformed nimbus

2.2.1 Significant Points

Significant points are the points that help to identify the shapes contributing to the deformed circular nimbus shape. Initially, significant points are determined by each of the boundaries that intersect with the deformed nimbus shape in the environment.

At first glance, one might think that these points are the vertex of this boundary. However, this not the case, because:

Depending on the boundary position related to the centre of the circular shape, we will have to discard the vertices that are hidden by the actual boundary (see the diamond vertices in Figure 8).

Having discarded these vertices, we also have to discard, from the remaining set of vertices, the vertices that are outside the circular shape. Instead, we are going to consider the points of intersection between the edges coming out of these vertices and the border of the circular shape (see the diamond vertices in Figure 9).

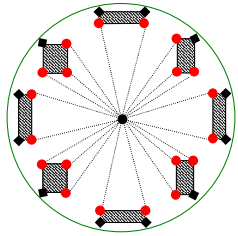


Fig. 8. Selected vertices with a circle and discarded vertices with a diamond

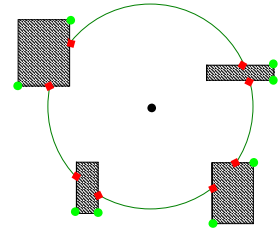


Fig. 9. Selected vertices with a diamond and discarded vertices with a circle

2.2.2 Final Points

The best way of understanding what points have to be considered as *Final Points* is by means of an illustration using a couple of examples.

If boundaries do not interfere with each other, then each of the boundaries is contributing independently to the set of final points (Figure 10).

However, if a part of the object is in hiding, as is the case in the example shown in Figure 11, the final points provided by each of the boundaries have to be carefully analysed. In Figure 11 (A), we can see that of all the final points that boundary 1 should provide – 4 points if no other boundary was interfering – just 2 of these points are contributing as final points, because of the presence of boundary 2 (the diamond point means that the other two points cannot be introduced). In Figure 11 (B), we can see the final points provided by boundary 2, one of the final points provided by this boundary is on boundary 1's edge. The set of final points is the union of all the final points provided by each of these boundaries.

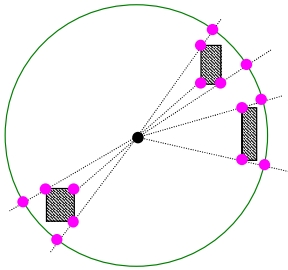


Fig. 10. Final Points

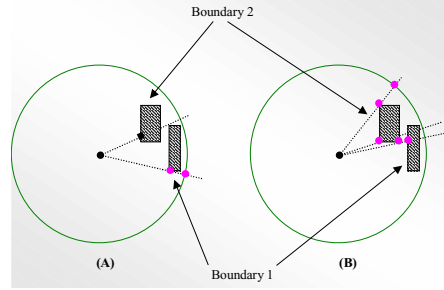


Fig. 11. Final Points with a circle

2.2.3 Ordering Algorithm

We have used a basic ordering criterion and some additional criteria for some specific situations that could occur while the application is running [4].

The basic ordering criterion is that given two points P_1 and P_2, P_1 is said to be bigger than P_2 related to the origin of co-ordinates OC (Figure 12), and we write $P_1 >_{OC} P_2$, if $\alpha_1 \neq \alpha_2$ and $\alpha_1 > \alpha_2$, where: α_1 is the angle of the straight line defined by OC-P_1; α_2 is the angle of the straight line defined by OC-P_2.

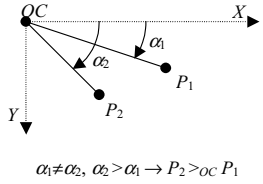


Fig. 12. Ordering criterion

3 Conclusions

We have developed an application to analyze how the presence of boundaries could modify some of the Key concepts of one of the most successful awareness models in Computer Supported Cooperative Work (CSCW), called the Spatial Model of Interaction (SMI) [1]. The SMI manages awareness in Collaborative Virtual Environments (CVEs) through these set of key concepts.

These studies have been run from a geometrical point of view with the aim of analysing how boundaries can modify the shape of two of key concepts introduced into the SMI – focus and nimbus - and how these modifications can have an effect on the awareness of interaction between each of them [4].

References

1. Benford, S., and Fahlén, L.E. *A spatial model of interaction in large virtual environments*, in Proc. Third European Conference on Computer Supported Cooperative Work (ECSCW'93), Milano, Italy. Kluwer Academic Publishers, pp. 109-124.
2. Greenhalgh, C. *Dynamic, embodied multicast groups in MASSIVE-2*, Technical Report NOTTCS-TR-96-8, Department of Computer Science, University of Nottingham, UK, 1996.
3. Greenhalgh, C., *Large Scale Collaborative Virtual Environments*, Doctoral Thesis. University of Nottingham. October 1997.
4. Herrero P. *A Human-Like Perceptual Model for Intelligent Virtual Agents* PhD Thesis. Universidad Politécnica de Madrid, June 2003.

Thin Client Access to a Visualization Environment*

Ioannis Fudos and Ioannis Kyriazis

Department of Computer Science, University of Ioannina,
GR45110 Ioannina, Greece, {fudos, kyriazis}@cs.uoi.gr

Abstract. In this paper we present a thin client system that provides Internet access to a modular visualization environment. The communication protocol for this system is designed so as to minimize the data exchanged among the server and the clients. An XML format is introduced for communicating visualization related information. Users of the system may collaborate to perform complex visualization operations and share files. The results of each operation are cached so that they may be used by the same user in a later session or by other collaborating users. Experimental performance results demonstrate the efficiency of our system when compared with commercial general purpose solutions.

1 Introduction

Graphics visualization is a demanding computational task. To process and render a complex scene of 3D objects, computationally powerful platforms are required. We have designed a client-server system that provides remote access to a visualization environment through Internet via a web browser. The system is suitable for groups of collaborating users that need to perform complex visualization related computations without having physical access to the same machine. A number of research and commercial systems have dealt with similar problems in the area of biology for MacroMolecular Modeling [6], meteorological and oceanographical purposes [9,1,2], and for general purpose world wide web applications[4,5].

In this paper we present the following technical contributions:

- a reduced XML-based communication protocol for exchanging data visualization information.
- a file caching scheme for intermediate results, which increases the performance of the system and allows user collaboration.
- real time experiments that demonstrate the efficiency of our system over commercial solutions.

* Part of this work was funded by a Greek Ministry of Education EPEAEK-HERACLETUS Grant. We would like to thank Prof. Vaclav Skala and his group for making the command line version of MVE available for this project. Also, we would like to thank Prof. Leonidas Palios for useful suggestions on early stages in the design of this system.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 258–263, 2004.
© Springer-Verlag Berlin Heidelberg 2004

As a concrete example of the above we have implemented a thin client access to MVE, a modular visualization environment [11].

The rest of this paper is organized as follows: Sect. 2 presents a short overview of the system and describes the data exchange protocol. Section 3 presents the caching scheme which is used to increase efficiency and allow user collaboration. Section 4 presents performance results.

2 Overview of the System

The system consists of three parts: The environment, which is responsible for all the computations performed during a session, the thin client, a light weight front end to the environment that allows users to access the environment remotely, and the server, which accepts requests from clients and passes them as arguments to the environment.

Our system uses the Modular Visualization Environment [11,10] to perform computations. Its operation is based on a set of independent modules, which are responsible for loading, modifying and rendering 3D graphics objects. Several modules connected to each other can form a scheme, which can be saved or executed. Each module is designed according to a common interface, regardless of the function of this module. Each module has a set of inputs, which are objects on which operations will be performed, a set of outputs which are the resulting objects, and some parameters that define its behavior during the execution of the operations. Thus modules can be represented by a data structure that describes its inputs, outputs, and parameters. When a user has created a scheme, the environment may store or execute this scheme. Execution of the scheme may be performed on a single machine or on a distributed platform.

We have used XML as many standards for interoperable graphics and visualization data are being developed in XML [8]. Also, there are portable parsers for XML that allow for porting the server part of our system easily to any platform. The client part is a plain java applet and is thus portable to any platform.

The server communicates with MVE by means of exchanging XML files. When a new client connects to the server, the server will request from MVE the list of available modules. As a response MVE will produce an XML file called modules.xml, where all currently available modules are described. An example of an XML file that contains the list of modules is shown in Fig. 1 (left). When a client wishes to execute a scheme, the server will provide a file to the visualization environment with an XML description of the scheme (scheme.xml). An example of a simple scheme described in an XML file is shown in Fig. 1 (right). At startup, the client establishes a connection to the server and requests the currently available list of modules. After receiving the modules list, the user can develop a scheme. When ready, the client sends the scheme to the server for execution. The scheme is executed, and the results are sent to the client for visualization. Besides executing the scheme, the client may request to save the scheme on the server's shared folders, or to load a scheme from the server. During a session, a client has to communicate with the server only to post a request for

```
<?xml version="1.0" encoding="iso-8859-1"
                    standalone="no"?>
<!DOCTYPE MODULES SYSTEM "modules.dtd">
<MODULES TIME="Tue Jun 26 18:48:54.471 2001">
   <MODULE NAME="Decimation" TYPE="2"
           DLL_NAME="Modules\ Decim_module.dll"
           DESCR="Triangle Mesh Decimation Module">
      <PARAMETERS>
         <PARAM NAME="Reduction" TYPE="NUM"
                DESCR="Percent of reduction"
                MANDATORY="FALSE"/>
      </PARAMETERS>
      <INPUT TYPE="2" DESCR=""/>
      <OUTPUT TYPE="2" DESCR="Output : Reduced list of
              vertices, reduced list of triangles"/>
   </MODULE>
   <MODULE NAME="TriangleSaver" TYPE="3"
           DLL_NAME="Modules\ Triangle_Modules.dll"
           DESCR="Save Triangle Data">
      <PARAMETERS>
         <PARAM NAME="FileName" TYPE="STR"
                DESCR="STL File to save triangles"
                MANDATORY="TRUE"/>
         <PARAM NAME="Binary" TYPE="BOOL"
                DESCR="Binary file ?"
                MANDATORY="FALSE"/>
      </PARAMETERS>
      <INPUT TYPE="2" DESCR="Triangles to Save"/>
   </MODULE>
   <MODULE NAME="IsoExtractor" TYPE="2"
           DLL_NAME="Modules\ Volume_Modules.dll"
           DESCR="Compute iso-surface from Volumetric
                  data (IsoExtractor)">
      <PARAMETERS>
        Not available
      </PARAMETERS>
      <INPUT TYPE="5" DESCR="Volumetric Data Input"/>
      <OUTPUT TYPE="2" DESCR="Triangle Data Output"/>
   </MODULE>
</MODULES>
```

```
<?xml version="1.0" encoding="iso-8859-1"
                    standalone="no"?>
<!DOCTYPE SCHEME SYSTEM "scheme.dtd">
<SCHEME>
   <OPTIONS PARALLEL="FALSE"/>
   <MODULES>
      <MODULE ID="TriLoad1"
              MOD_TYPE="TriangleLoader"
              PARAMETERS="Cow.tri"/>
      <MODULE ID="Dec1" MOD_TYPE="Decimation"
              PARAMETERS="80"/>
      <MODULE ID="TriSave1"
              MOD_TYPE="TriangleSaver"
              PARAMETERS="Cow2.stl;FALSE"/>
      <MODULE ID="TriSave2"
              MOD_TYPE="TriangleSaver"
              PARAMETERS="Cow3.stl;FALSE"/>
   </MODULES>
   <CONNECTIONS>
      <CONNECTION FROM="TriLoad1" OUTPUT="0"
                  TO="Dec1" INPUT="0"/>
      <CONNECTION FROM="TriLoad1" OUTPUT="0"
                  TO="TriSave2" INPUT="0"/>
      <CONNECTION FROM="Dec1" OUTPUT="0"
                  TO="TriSave1" INPUT="0"/>
   </CONNECTIONS>
</SCHEME>
```

Fig. 1. (left) The modules.xml file; (right) the scheme.xml file

executing, saving or loading a scheme. No communication with the server is required while developing a scheme. As shown in Fig. 2 (left), the information communicated among the server and the clients is minimized, as the messages exchanged are few and short in length. This makes our system appropriate for clients running on machines with slow network connections.

The server is the part of the system that connects the clients to the MVE, and provides them with appropriate information whenever requested, whether this involves calling the environment or not. As the server is multithreaded it can serve many clients at the same time. A client may request from the server to send the list of currently available modules, to execute a scheme and render the results, to save a scheme, or to load a previously saved scheme. The server also includes a login manager, which is responsible for user authentication and identification, and handles the ownership and permissions of the schemes. This way, a client may share a scheme with other users or groups, and set access rights for its schemes. Besides the login manager, the server includes a cache manager, which is responsible for caching the results of an execution, hashing the schemes to organize the cached results, and searching the cached results to retrieve a previously executed scheme. The client is designed as lightweight as possible, and is able to execute on any java-enabled web-browser. No computations are performed on the client machine, only some basic interaction with the server is carried out. The execution of an operation is performed by the server. The client performs only the visualization of the results. For user's convenience, the GUI of the client is similar to the MVE. Fig. 2 shows messages exchanged during a session, and a snapshot of the Internet client.

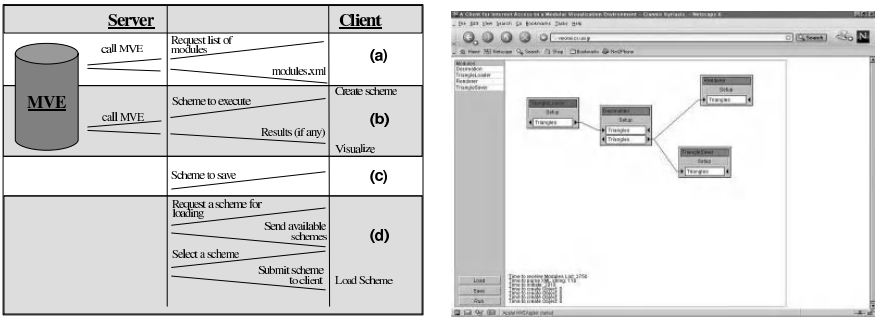


Fig. 2. (left) The messages exchanged during a session: (a) request the list of Modules, (b) execute a Scheme, (c) save a Scheme and (d) load a Scheme; (right) the Internet Client

3 File Sharing, Caching, and User Collaboration

When a client stores or loads a scheme, this scheme is actually stored on the server's site. These folders may contain other users' schemes. This way, users may collaborate by sharing their schemes. They also may form groups to share their schemes only with the members of the same group. A login process is used to identify and authenticate the client at startup. Then, when a scheme is stored, it has an owner. Owners may choose to share their scheme with members of their group, or to make it available to everybody. Like in a unix file system (UFS), the files have an owner and a group, and the owner sets the mode of the file for himself/herself, his/her group and others (read, write, execute for each such category).

To reduce the workload of the server further, we cache results of a scheme execution, so if a scheme has already been executed in a previous session, even from a different user, it will not have to be executed again. Caching only the final result of an execution would be useful only if the scheme for execution matches the previously executed scheme exactly. Even if one parameter had a different value, the cached results would be useless, as they would produce different results. This is why we cache the intermediate results as well, so that even partially matching schemes may use these results. Since the result of a module execution is the same when the input and the parameters are identical, even if the rest of the scheme differs we cache the result of each module seperately. To locate a cached result, we use a hash table. The hash function [7,12] uses the {module, input, parameter} set as input, and returns an entry in the hash table for the output file. Figure 3 shows the structure of the hash function and the hash table.

The module ID, along with its parameters and inputs are hashed as a whole, and the hash output is stored along with the output of the execution. If there is more than one output, each of them is stored separately, as they may be used as different inputs for some other module. To distinguish among the different

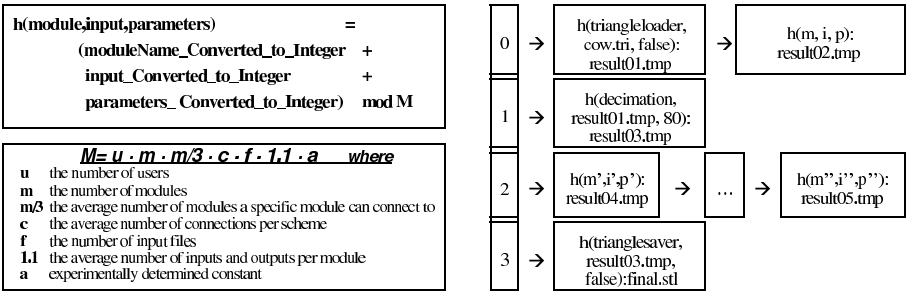


Fig. 3. Details on the hash function and the hash table

outputs of the same hash bucket, we store the {module, input, parameter} set as well. Since the input of each module was the output result of another module execution, there is a result file for this input already, so we can use this file instead of the input. The length of the hash table depends on the number of users, the number of different modules, the average number of connections per scheme, and the number of input files. If the table becomes very large, we remove the least used entries. If M is the length of the hash table, we should not allow more than 3M entries to be cached, as this increases the time to search a cached result. We have determined experimentally an efficient size which is illustrated in Fig. 3. In order to locate the cached results in the database, the server must first hash the clients scheme, to find the hash entries that may contain the cached results. The modules that participate in a scheme are hashed from the first to the last, and a hash entry is returned for each output. It is the server who searches for cached results, as it is the one that has the necessary information available. The client just sends the scheme to the server. The search is done backwards so that we can find a matching result as soon as possible. If we find a matching output, we use it as input to execute the rest of the scheme. The new results of the execution are cached as well.

4 Performance Evaluation

We tested our prototype system and evaluated its performance under various client, server, and network configurations.

In the first experiment, we compared the response time of our client-server system with a popular commercial tool that provides access to the desktop of a remote platform [3]. We measured the time it takes for the environment to start, and the times to load, save and execute a specific scheme. As shown in Fig. 4 (left), the response time for our client is relatively small, compared to the commercial tool. Our client performs well even over slow network connections, as the messages exchanged between the client and the server are few and short.

In the second experiment, we measured the performance of our client in various configurations concerning the state of the server and the sites where the

	56 kbps modem		100 Mbps LAN	
	Commercial	Our Client	Commercial	Our Client
Time to start	13 sec	2130 ms	2 sec	1650 ms
Time to execute	6 sec	975 ms	4 sec	845 ms
Save Scheme	2 sec	10 ms	About 1 sec	10 ms
Load Scheme	2 sec	280 ms	About 1 sec	130 ms

	LAN Client	Dialup Client		
Time to:	Normal	Normal	Network Traffic	Workload on server
receive Modules List	1750	2210	4030	6373
parse XML string	60	80	100	100
Startup	1910	2932	4216	8201
create module	0	0	0	0
save scheme	0	0	50	0
exec scheme	270	350	585	1010
receive file's list	0	130	370	3405
load large scheme	60	160	380	780

Fig. 4. Results of the experiments

server and the client run, such as increased network traffic, low system resources, and many clients connected on the server. The response times measured are the times to save, load, and execute a scheme, as well as to receive the list of available modules, and to startup the environment. As shown in Fig. 4 (right), the time to receive the list of modules is the main reason of delaying the initiation of the client. The time to receive the modules is relatively long, because it takes the server considerable time to generate the list.

References

1. Ferret, data visualization and analysis. http://ferret.wrc.noaa.gov/Ferret/.
2. Geovista center, collaborative visualization. http://www.geovista.psu.edu/research/ collaborativevisualization/.
3. Symantec pcanywhere. http://www.symantec.com/pcanywhere/Consumer/.
4. L. Beca, G. Cheng, G.C. Fox, T. Jurga, K. Olszewski, M. Podgorny, P. Sokolowski, and K. Walczak. Tango, a collaborative environment for the world wide web. http://trurl.npac.syr.edu/tango/papers/tangowp.html.
5. L. Beca, G. Cheng, G.C. Fox, T. Jurga, K. Olszewski, M. Podgorny, P. Sokolowski, and K. Walczak. Web technologies for collaborative visualization and simulatio. http://trurl.npac.syr.edu/tango/papers/tango_siam.html.
6. M. Bhandarkar, G. Budescu, W.F. Humphrey, J.A. Izaguirre, S. Izrailev, L.V. Kalt, D. Kosztin, F. Molnar, J.C. Phillips, and K. Schulten. Biocore: A collaboratory for structural biology. In *Proceedings of the SCS International Conference on Web-Based Modeling and Simulation*, pages 242–251, 1999.
7. J.L. Carter and M.N. Wegman. Universal classes of hash functions. *Journal on Computing Systems and Science*, 18(2),79:143–154.
8. J. Ferraiolo, F. Jun, and D. Jackson. Scalable vector graphics. Technical Report TR-11, SVG, 2002.
9. S. Franklin, J. Davison, and D.E. Harrison. Web visualization and extraction of gridded climate data with the ferret program. http://www.pmel.noaa.gov/ferret/ferret_climate_server.html.
10. M. Rousal and V. Skala. Modular visualization environment - mve. In *Proceedings of International Conference ECI 2000*, pages 245–250, 2000.
11. V. Skala. The mve and complete programming documentation and user's manual. http://herakles.zcu.cz.
12. R. Sprugnoli. Perfect hashing functions: A single probe retrieving methods for static sets. *CACM*, 20(11),77:841–850.

Interactive Visualization of Relativistic Effects with the Hardware Acceleration

Radoslaw Mantiuk, Karina Murawko-Wisniewska, and Dorota Zdrojewska

Technical University of Szczecin, Zolnierska 49, 71-210 Szczecin, Poland,
phone +48(91)4495661,
rmantiuk@wi.ps.pl, http://zgk.wi.ps.pl

Abstract. This paper describes an interactive rendering technique for the special relativistic visualization of fast moving objects. The method takes into consideration relativistic aberration of light and Doppler effect. Programmable graphics hardware is used to achieve interactive visualization. Standard texture mapping pipeline is modified on per-pixels basis and relativistic effects are rendered. The basis of optical relativistic phenomena and a survey of relativistic real time rendering techniques are also presented in the paper.

Keywords: rendering of relativistic effects, vertex and pixel shaders, interactive image synthesis, computer graphics.

1 Introduction

Mankind is limited to very low velocities compared to the speed of light. Therefore, computer simulations are the only means of seeing the world from the perspective of fast moving vehicle. Properties of space, time, and light in relativistic physics are totally different from those in classical, Newtonian physics. They are contrary to human experience. Until now, people could only imagine how such effects like relativistic aberration, Doppler shift, Terrell rotation look like. Computer visualization of relativistic effects could help the intuition of people and make the relativistic phenomena more understandable. Interactive visualization of relativistic phenomena brings additional value. The observer can change speed or direction of motion and see the results immediately.

Real time rendering methods need to be used to achieve interactive visualization of relativistic effects. Moreover the computation should be aided by the 3D graphics hardware to generate smooth animation. This paper describes such interactive rendering technique. The method is based on environment texture mapping technique. Standard texture mapping pipeline is modified on per-pixels basis and relativistic effects are rendered. Shader programs, running on graphics board, are used to speed up calculations.

The next section describes the previous works on interactive visualization of relativistic effects. In Sect. 3, the basis of relativistic physic are presented. Section 4 gives details of our relativistic rendering algorithm based on hardware acceleration. In Sect. 5 implementation is described and results are shown. The paper ends with conclusion and an outlook on future work.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 264–269, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Previous Works

The issue of visual perception in special relativity was ignored for a long time. Hsiung and Dunn [1] were the first to use visualization techniques for image shading of fast moving objects. They proposed the extension of ray tracing for generation of relativistic effects. The first real time approach for relativistic rendering was presented by Hsiung et al. in [2]. In the last few years two different approaches, which allow rendering relativistic phenomena in real time were discussed.

The first one extends the normal rendering pipeline by an additional transformation of vertices and is an extension of Hsiung T-buffer. This approach, called polygon rendering technique, was investigated by Gekelman et al. in [3]. In [4] and [5] Gekelman's method was generalized to any object built with polygon mesh. The goal of polygon rendering technique is to create a new copies of 3D objects by relativistic transformation of existing objects. Objects' vertices are considered as point lights. The light from these light sources runs to observer not along the straight lines and causes relativistic deformations. The main drawback of polygon rendering technique is the necessity of polygon tessellation. Relativistic transformation causes bending of polygon edges and polygons need to be tessellated to achieve acceptable image quality.

The second method of relativistic effects rendering is based on texture mapping. This method was proposed by Weiskopf in [6] and developed in [14]. The main goal of the method is to use extended texture mapping technique to transform standard image of the environment to the coordinate system of very fast moving observer. Objects are deformed by the relativistic aberration of light and finite light speed effects. These deformations can be achieved by mapping of an environment texture on the sphere surrounding the observer. During the mapping the relativistic calculation are done and colour and position of texels in environment map are changed. Interior of the sphere with mapped texture is projected on the plane to achieved final image with relativistic phenomena. In [7] above method was extended to allow visualization of Doppler effect (colour shift) and radiance fluctuation (brightness shift). The texture mapping technique does not require any changes in 3D scene structure, only pixel transformations (changes of pixel colour and position) are done. The main drawback of the technique is limitation to the walkthrough environments (the environment textures need to be recalculated when the position of objects or lights is changed). Also the achieved rendering speed (5-7 fps) doesn't allow for smooth animation.

The relativistic phenomena visualization method presented in our paper is an extension of Weiskopf method. We propose the modification of texture mapping technique and the usage of reprogrammable hardware to speed up rendering.

3 Relativistic Transformations

Take two coordinate systems S and S' moving relatively to each other with velocity v. When velocity v is close to the speed of light an observer in system S

experiences events differently than the observer in system S'. These differences, explained by Einstein in his Theory of Special Relativity, can be computed using Lorentz transformation [9]. Lorentz transformation of point (P, t_p) measured in system S to system S' is given by

$$P' = P + [\frac{(\gamma - 1)}{\beta_l^2} * P \cdot \beta + \gamma * c * t_p] * \beta,\ t'_p = \gamma * t_p + \gamma * \frac{\beta \cdot P}{c} \qquad (1)$$

where c - speed of light, $\beta = v/c$, $\beta_l = |\beta|$, $\gamma = \frac{1}{\sqrt{1-\beta_l}}$.

The direction of a light beam depends on the velocity of the light source relative to the observer, this effect is called the optic aberration of light (relativistic aberration). Consider a particle moving with velocity v (along the x-axis) through the S coordinate system [13], it emits the light beam at an angle f' to the x'-axis (measured in S' system, in which particle is at rest), in the S system the light beam is pushed to the new angle f. The relation between the angles (in spherical coordinates (Θ, ϕ)) is given by

$$\cos\Theta' = \frac{\cos\Theta - \beta}{1 - \beta * \cos\Theta},\ \phi' = \phi. \qquad (2)$$

The shift of light wave's frequency (v) during the relative motion of a light source and an observer is called the Doppler effect [11]. When the source and the observer are approaching each other colors are shifted toward the high end of the visible spectrum (the blue shift). The Doppler shift of light waves in vacuum which arrive with an angle f with respect to direction of travel is:

$$\nu' = \nu * \gamma * (1 - R_d \cdot \beta). \qquad (3)$$

More detailed description of special relativity can be found in [9][12][10].

4 Real Time Visualization of Relativistic Aberration and Doppler Effects

The algorithm for relativistic image synthesis is designed upon the Weiskopf's texture-based relativistic rendering method [6]. Hardware acceleration is accomplished by modifying the method in order to use the vertex and pixel shaders (an extension of modern graphic cards).

In preprocessing a set of images representing the environment must be generated. The single image is not sufficient, because the aberration of light changes the field of view (it could widen the field of view beyond the image). The images allow looking from the observation point in any direction, similar as in environment mapping techniques (we use cube mapping technique). The new set of textures must be prepared when the scene objects' or lights' position change and in other specific situations (when the information in textures is not enough for calculations).

In the first step of algorithm the texture coordinates are transformed with the relativistic aberration equation, in our system it is achieved with vertex shader. Then the prepared textures are mapped on the sphere, which surrounds the observer. The last step is the Doppler effect calculation, the change of textures' pixel colour is done by pixel shader. For calculating of colour shift in RGB the algorithm from ray tracer Backlight [8] was used. It allows calculating every component separately by using a simplified modeling of spectrum. The RGB components are shifted according to equation 3.

There is also possibility for interactive change of speed. The elements of the scene (scene objects and light sources) are at rest relative to each other. The relative motion of the scene and the observer is limited to the direction along the z-axis, but can be expanded by using additional rotations of the coordinate system.

5 Implementation and Results

Our implementation of the relativistic visualization algorithm is based on hardware acceleration in form of vertex and pixel shaders. The algorithm is implemented in Direct3D 9.0 with NVIDIA's Cg language.

The most intuitive and fastest method of projecting the environment onto the sphere is the use of cube texture. Its' faces represent different parts of the scene viewed from the six orientations of the camera positioned at the coordinate system origin.

In our implementation, cube mapping and aberration of light are calculated in the vertex shader program (for each vertex of the sphere surrounding the observer). All essential data is gathered in vertex shaders registers. Vertex position and normal are passed from the vertex buffer into input registers (POSITION, NORMAL). β coefficient, necessary to calculate the light aberration, is placed in one of the constant registers. Vertices coordinates are transformed to spherical coordinates and relativistic light aberration is calculated (according to equation 2). The results are passed in output registers (POSITION, COLOR0 and TEXCOORD0).

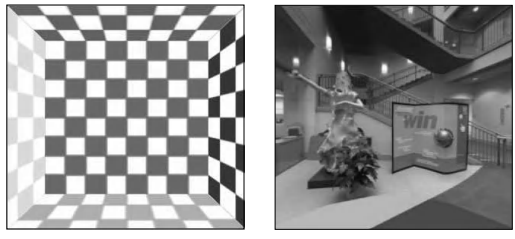


Fig. 1. Test scenes: left - cube with the checkered textures, right - inside the hall (the texture from DirectX samples)

Pixel shader does not have POSITION register, so we use COLOR0 register to store coordinates of each vertex of the sphere (calculated by vertex shader). These coordinates are used in pixel shader to compute Doppler effect.Pixel shader program computes final environment texture coordinates, and then gets colour of the corresponding texel. Afterwards the Doppler effect is applied - colour of each pixel is changed with the equation 3, and passed in the output COLOR register. After colour computation, primitives are drawn on the screen.

Figure 1 presents two example scenes to which relativistic effects are going to be applied. The camera is motionless.

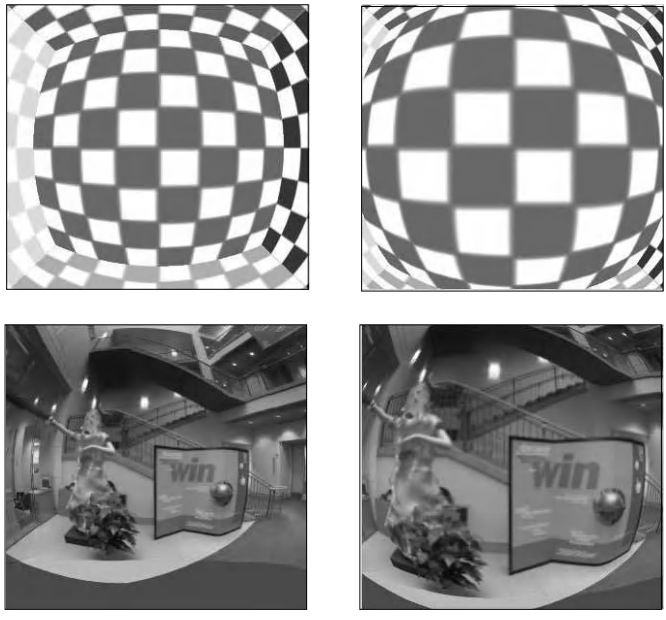


Fig. 2. Relativistic light abberation: top/left - velocity v=0.6c, top/right - velocity v=0.9c, bottom/left - velocity v=0.6c, bottom/right - velocity v=0.9c

Relativistic light aberration effect applied to example scenes can be seen in the fig.2. The camera is moving into the scene with 0.6c or 0.9c velocity.

As seen in the pictures above, change of the movement velocity causes impression of disturbance in the scene geometry. When the camera is moving with a large speed, close to the speed of light, straight lines seem to be curved.

The Doppler effect causes shift of colour into high frequencies when the observer is moving into the scene. It appears in change of colour into blue.

The rendering speeds 30 fps without Doppler effect and 15 fps with Doppler effect were achieved (for images of resolution 320x320 pixels). The colour shift algorithm plays a dominant role for the rendering performance.

6 Conclusion and Future Work

In this paper a hardware aided texture mapping approach to relativistic visualization has been presented. The physical basis is the relativistic aberration of light and Doppler effect. Interactive visualization of relativistic phenomena was achieved. The most important parts of the rendering algorithm are calculated in shaders, programs activated in hardware. Our method doesn't require any special hardware, it can be activated in a standard graphics card which supports vertex and pixel shaders.

In future work, the implementation of searchlight effect is considered. Also we plan to increase the performance of the rendering. Faster algorithms of colour shifting and new features of shaders will be explored. The implementation of hybrid method of rendering which uses texture mapping and Image Based Rendering approach together is also considered. It would allow for real time visualization of such relativistic effects as Terrell rotation.

References

1. Hsiung P.K., Dunn R.H.P.: Visualizing relativistic effects in spacetime. In Proceedings of Supercomputing '89 Conference, pp. 597-606, 1989.
2. Hsiung P.K., Thibadeau R.H., Wu M.: T-buffer: Fast visualization of relativistic effects in spacetime. Computer Graphics, 24(2), pp. 83-88, March 1990.
3. Gekelman W., Maggs J., Xu L.: Real-time relativity. Computers in Physics, pp. 372-385, 1991.
4. Rau R.T., Weiskopf D., Ruder H.: Special Relativity in Virtual Reality. Mathematical Visualization, H.-C. Hege, K. Polthier, Springer, pp. 269-279, 1998.
5. Weiskopf Daniel: An Immersive Virtual Environment for Special Relativity. WSCG 2000 Conference, pp. 337-344, 2000.
6. Weiskopf Daniel: A Texture Mapping Approach for the Visualization of Special Relativity. IEEE Visualization '99 Late Breaking Hot Topics Proceedings, A. Varshney, C.M. Wittenbrink, H. Hagen, ACM Press, pp. 41-44, 1999.
7. Weiskopf Daniel: Fast Visualization of Special Relativistic Effects on Geometry and Illumination. W. de Leeuw, R. van Liere, Data Visualization 2000 (Proceedings of the EG/IEEE TCVG Symposium on Visualization), Springer 2000, Wien, pp 219-228, 2000.
8. Backlight documentation: Http://www.anu.edu.au/Physics/Searle/
9. Bobrowski Czeslaw: Fizyka - krotki kurs. Wydawnictwa Naukowo-Techniczne, Warszawa, 3rd edition, 1993 (in Polish).
10. Orear Jay: Fizyka. Vol. 1. Wydawnictwa Naukowo-Techniczne, Warszawa, 5th edition, 1998 (in Polish).
11. Savage C.M., Searle A.C.: Visualising special relativity. Http://www.anu.edu.au, 1997.
12. Schutz Bernard F.: Wstep do ogolnej teorii wzglednosci. PWN, Warszawa 2002, ISBN 83-01-11666-8 (in Polish).
13. Chang M-C., Lai F., Chen W-C.: Image Shading Taking into Account Relativistic Effects. ACM Transactions on Graphics 15, 4 (Oct. 1996), 265-300.
14. Weiskopf D., Kraus U., and Ruder H.: Searchlight and Doppler effects in the visualization of special relativity: A corrected derivation of the transformation of radiance. ACM Transactions on Graphics, 18(3), July 1999.

Design of Interactive Environment for Numerically Intensive Parallel Linear Algebra Calculations*

Piotr Luszczek[1] and Jack Dongarra[1,2]

[1] Innovative Computing Laboratory, Computer Science Department, University of Tennessee Knoxville
[2] Computational Science and Mathematics Division, Oak Ridge National Laboratory

Abstract. We focus our attention in this article on how to provide parallel numerical linear algebra capabilities to Problem Solving Environments. Instead of describing a particular implementation, we present an exploration of the design space and consequences of particular design choices. We also show tests of a prototype implementation of our ideas with emphasis on the performance perceived by the end user.

1 Introduction

Numerical linear algebra may well be regarded as the most basic and thus essential component of problem solving environments (PSE) for numerical calculations. In this article, we intend not to focus on the user tool for accessing the parallel numerical capabilities we propose, but rather, on exploration of the design space available for such PSEs. To the user tool we refer as a *host environment*. The challenge is, we believe, in seamlessly integrating parallel computing capabilities with these environments.

The applicability of our arguments exceeds by far the scope of pure numerical linear algebra on dense matrices. Appropriate design of basic objects and their manipulations invites easy introduction of additional features such as sparse and eigenvalue solvers.

2 Related Work

Exhaustive survey of interactive environments for scientific computing deserves an article on its own. Therefore, we give only references to what we believe are the most relevant efforts that are related to numerical linear algebra. Python is an object-oriented programming language but it invites very much interactive

* This work is partially supported by the DOE LACSI – Subcontract #R71700J-29200099 from Rice University and by the NSF NPACI – P.O. 10181408-002 from University of California Board of Regents via Prime Contract #ASC-96-19020.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 270–277, 2004.
© Springer-Verlag Berlin Heidelberg 2004

style of development and experimentation [1]. Consequently, there exist numerous libraries that extend Python's numerical capabilities, the most popular include Numeric[2], Numarray[1], SciPy[2], MatPy[3], and ScientificPython[4]. Just for completeness' sake, we should also mention a similar framework for Perl called *The Perl Data Language*[5] with its shell for interactive work called `perldl`. Commonly known environments for interactive numerical calculations are Matlab [3], Octave[6], Scilab [4], Interactive Data Language [5], and Rlab[7]. Also, there exist environments that focus on symbolic manipulations with numerical capabilities, they are surveyed elsewhere [6], here we only mention a few: Mathematica [7], Maple [8], Macsyma [9], and Maxima [10]. Finally, there exist relatively many parallel extensions to Matlab[8] despite some scepticism dating back to 1995 [11]. Out of these extensions, Matlab*P [12,13,14] seems to be the most intensively developed, reaching its third major release version at the time of this writing.

LAPACK for Clusters (LFC) [15] is one of the projects of the Self-Adapting Numerical Software (SANS) framework [16]. It is intended to meet the challenge of developing next generation software by automated management of complex computing environments while delivering to the end user the full power of flexible compositions of the available algorithmic alternatives. LFC, in particular, automates the process of resource discovery and selection, data distribution, and execution of parallel numerical kernels for linear algebra calculations. As such, we believe, it is suitable for the interactive environment we describe in this article.

3 Network Model

We consider primarily a typical two-tier client-server architecture without overloading the server with extra functionality that is left for the client. In such scenario there exists clear separation of capabilities – the server only needs to provide high performance computing capabilities.

Similar reasoning is behind placing the object logic on the client rather than the server (which only holds, presumably large, object data). It simplifies the design of the server and makes it possible to use it on a wider variety of platforms. The client, on the other hand, may leverage existing software technologies for remote management of computational objects.

4 Object-Oriented Features

While designing our system, the first decision to make of is to choose either 0-based (first matrix entry is row 0 and column 0) or 1-based indexing scheme.

[1] `http://www.stsci.edu/resources/software_hardware/numarray/`
[2] `http://www.scipy.org/`
[3] `http://matpy.sourceforge.net/`
[4] `http://starship.python.net/~hinsen/ScientificPython/`
[5] `http://pdl.perl.org/`
[6] `http://www.octave.org/`
[7] `http://rlab.sourceforge.net/`
[8] `http://supertech.lcs.mit.edu/~cly/survey.html`

There exist large amount of code in production use that requires us to implement both. The problem cannot be easily solved by following the convention of the host environment. Such a solution does not allow for code migration between two host environments that use conflicting indexing schemes, therefore, we allow for both.

A related problem is how the end of a range is specified. This may be illustrated with an array declaration (of size `N`) in Fortran: "`REAL A(N)`" and in C: "`float A[N];`". While both declarations use `N` as the upper bound specifier, Fortran uses it inclusively (the allowed indices are `1`, `2`, ..., `N`) and C uses it exclusively (allowed indices are `0`, `1`, ..., `N-1`). Similarly, Matlab uses inclusive convention and Python uses exclusive one. Since there is no single scheme used across different host environments we need to provide for both.

An important decision to make is to decide whether matrix objects should operate with copy or view semantics. The most common situation when this decision has to be made is during submatrix operations. Consider an m by n matrix A partitioned as follows: $A = [A_1\ A_2]$, where A_1 and A_2 are m by n_1 and m by n_2 matrices, respectively, with $n_1 + n_2 = n$. A common way to refer to A_1 is `A[:,:n1]`. The question is whether such a reference should create a copy of the appropriate portion of A or, instead, only produce an alias (a view). There exist situations where either the former or the latter solution is preferable. Different systems solve this problem differently: Matlab and Python use the copy semantics while Fortran 90 uses the view semantics. Most likely, end users will opt for copy semantics, while developers will prefer the view semantics. Therefore, we choose to allow both in our system.

The flexibility of multiple data types comes at the price of resolving issues with mixed-type operations. Automatic variable casting is a feature of almost any programming language in wide spread use. An expression like `x + y` is handled correctly even if `x` and `y` are variables of different numerical type. The most common behavior in such a case is to *promote* (type-cast to the larger type) one of the values and then perform calculations. The promotion rule works well for statically typed languages but most PSEs use some form of dynamic typing and therefore it is harder to ensure correct type for the result. The two major issues to consider are the memory allocation (promotion could potentially require a few times more space to be used) and tensor-rank change (an outer product of two vectors produces a matrix: $A = xx^T$ – a different data type all together). Various solutions may be more appropriate in different some situations. Hence we opt for providing means for ensuring appropriate kind of automatic casting.

The type of PSE that we are describing, deals with tensors of different ranks: 0 – numerical values, 1 – vectors, and 2 – matrices. Such environments add a unique aspect to the type-casting problem described above: reduction of tensor rank. Consider a general case of matrix-matrix multiply: $C = AB$, where: A is m by k, B is k by n, and C is m by n. If either m or n is 1 then the multiply reduces the tensor rank by 1. If k is 1 then the reduction is by 2. However, the type of the result cannot be changed even if potential tensor rank reduction occurs: if a matrix algorithm (such as an iterative method or a dense linear solver) is formulated in terms of submatrices (so called *block algorithm*) then it is expected to work even if the submatrices degenerate to single values (*block*

size is 1). There is no general way of detecting when type change should follow a tensor rank reduction. Therefore, we choose not to perform the type change by default (with type change being optional) since this facilitates interactive work.

5 Host Environment Integration

Networking capabilities are the most essential for our system. Out of the host environments that we initially target, networking is fully supported in Python. Maple, Mathematica, Matlab, and Octave require an extension written in a native language – this creates a portability problem. Luckily, most of the aforementioned environments support Java so this is the way to write just one code and use it in many environments. Finally, since Octave does not support Java as of this writing, an extension can be written using system calls such as `system()`.

Support of name spaces is an important but not essential feature that we would like to use. Python offers more sophisticated way of dealing with this problem – it has a hierarchical module system comparable to that of ISO C++ and Java. Matlab comes close to it by implementing functions only relevant in the context of one particular class of objects (they are commonly referred to as object methods but in Matlab have invocation syntax just like regular functions). Mathematica implements contexts and packages to cope with name resolution. For all other environments we need to use the prefixing technique.

Object-orientation is an important feature as it allows, among others, for a simple statement like `a+b` to be interpreted differently depending on what `a` and `b` are. Most of the host environments that we know are only object-based. Matlab is somewhat more advanced as it allows for creation of new objects and operator overloading but does not have object destructors. This is an important capability in the presence of overloaded operators since they tend to produce anonymous temporary objects which cannot be reclaimed even manually. This problem can be somewhat alleviated by using Java from within Matlab. Python is an object-oriented language which makes it suitable for our system. In other environments we need to resort to function syntax – it takes a lot from expressiveness but still allows to use the functionality that we offer.

6 Parallel Execution

The first issue to resolve in parallel processing is the fact that vectors and matrices most often have different requirements for data layout: vector computations are likely to benefit from 1D (one dimensional) layout, while for matrices, 2D distribution is preferable. One way to automate the decision process for novice users is to be distributing vectors in 1D fashion and matrices in 2D. In a case when a matrix and vector are to be used together, the vector needs to be made conformant to the matrix' layout to perform the operation efficiently. Such a solution involves relatively small communication penalty. For more advanced users, full control of data distribution is the preferable way.

Another aspect is execution synchronization between the client and the server. The term *lazy evaluation* is used to refer to one of the possible scenarios [17].

Simply put, it means that only every other remote request is blocking the client until the server's completion. Generalization of this way of communication is referred to as asynchronous mode. Such a mode, in our opinion, is not good for an interactive environment since it splits the call process into two phases: submission and completion requests. It is not the way existing sequential environments operate – their behavior is equivalent to a synchronous mode (each request is blocked on the client side until the server fulfills the request). A midway solution is transactional processing: the user starts a transaction, then all the computational requests are submitted, and then the call finalizing the transaction is made which blocks until all the computational requests are served. It allows the server to order the computational steps for better performance.

7 Miscellaneous Issues

An important aspect of any numerical system is compliance with the IEEE 794 standard [18]. While the standard is commonly accepted by many hardware vendors, it is still rare to find fully compliant product. We are bound here by what is the typical behavior of the host environment and what is available on the server. Some environments have a way of dealing with non-conformant hardware or system libraries, e.g. in Python, floating-point exceptions are caught by a Unix signal handler.

There exist a few options for data storage and transfer that we consider useful. Certainly, users will have some data sets stored locally on their client machines. These local data need to be transferred to the server for manipulation. During calculation, the best place for data would be the server while at the end, the results need to be transfered back to the client (in case the server does not provide reliable storage capabilities). In the meantime, the data is prone to be lost due to hardware or software crashes so at some point fault-tolerance should be considered. Another scenario is downloading data from an external source. A very helpful extension is support for scientific data formats.

Security is an important asset of a software piece that provides server-like capabilities. In this area, we only intend to leverage existing solutions with initial focus on the port-forwarding feature of `ssh(1)`. It seems relevant in the presence of firewalls and NATs (Network Address Translation) that prevent connections to all but few selected ports.

When it comes to changing the behavior of a computational environment; two main configuration styles need to be considered: global and local. The global type includes: configuration files (*dot-files* in Unix), environment variables, command line options, and global program variables. In a sense, all of them provide a similar functionality with different timing and scoping. However, since a PSE may be regarded as a language, it is important to maintain its semantic consistency. Therefore, global configuration is a valid solution when there is only one default setting mandated as standard and other choices are only optional. Relevant local configuration types include: object attributes, shadow objects or explicit syntax. The first two are somewhat similar as shadow objects are just aliases of their originals with some of the attributes changed. For example if `A` is a square matrix, `A.I` (a shadow object of `A`) could indicate inverse of `A` but

using `A.I` would not immediately produce a numerical inverse of `A` but rather, LU decomposition would be used instead. Compared to object attributes, shadow objects are more explicit. From clarity standpoint, object attributes are not as good as explicit syntax (e.g. function call) but are far more succinct and more suitable for interactive environments.

8 Implementation

At the moment, the basic infrastructure of our design has been implemented and successfully applied to a dense matrix factorization and iterative solution method in Matlab and Python environments. Our preliminary tests show that the overhead of remote execution can be offset when problem sizes become prohibitive for a sequential environment and it is indeed possible to reap the benefits of parallel computation.

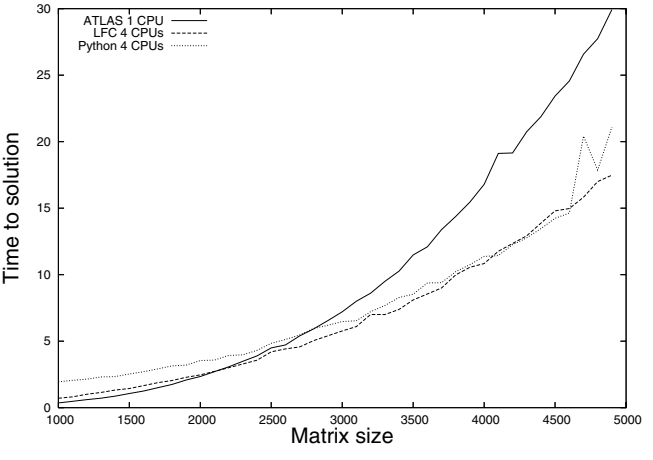


Fig. 1. Comparison of time to solution of a system of linear equations of varying size with different methods

In our tests, we used two dual Xeon 2.4 GHz computers connected with switched Gigabit Ethernet. MPICH 1.2.4 was used as the MPI implementation. Figure 1 shows the timing results for our tests that were performed on a non-dedicated system. The objective was to solve in double precision floating-point arithmetic a system of linear equations by means of LU factorization. Three scenarios were used to obtain a solution: sequential computation, parallel computation, remotely controlled parallel computation.

For the first scenario, ATLAS [19,20] library was used on a single CPU. In particular, the functional equivalent of LAPACK's [21] `DGESV()` routine was used that performs LU decomposition *in-situ*. The second scenario utilized 4 nodes

that performed computations with the LFC's equivalent of ScaLAPACK's [22] `PDGESV()` routine. Again, no data copying was involved. The third scenario used the same hardware and software as the second one but the execution initiation and timing was done on a remote computer running Python interpreter. The round-trip time between the client and one of the nodes of the computational server grid (as measured by the `ping` program) was about 82 milliseconds – a value representing a 16-hop connection (as measured by the `tracepath` program) through wireless access point and an ADSL line. In this scenario, a copy was made of the system matrix to store its LU factors computed by `PDGESV()`: $x = A^{-1}b$ was written as `x = A.I * b` but the inverse of A was not calculated explicitly but rather the LU decomposition of a copy of A was used. It's a trade-off between convenience and optimality (the optimal notation being for example "`pgesv(A, x, b)`") and we intended for our tests to reveal how much this convenience costs.

Figure 1 reveals two important matrix sizes: the size for which parallel execution is faster than sequential execution (3000 in our case) and the size for which the matrix copy overhead is negligible (4000 in our case). The graph shows counter-intuitive effect of copy-free solve being slower than the solve with copy overhead – this is to be expected on a non-dedicated system and is more likely to occur the longer the time to solution is. Worth noting for matrices larger than 4500 is the unexpected increase of time to solution for the remote execution. Very likely explanation is a sudden surge in the load of the network that connects the client and server.

9 Future Work

Our implementation might exhibit itself as an OGSA-compliant service. Such a service would not be running on the server but rather on a proxy capable of OGSA interaction. The proxy would interact with the actual computational server through a simplified protocol – like NetSolve's three-tier approach [23]. A direction to pursue is creation of compilation system so that it is possible to translate existing scripts to a stand-alone executable. Such capability provides opportunity to have a client-server environment for experimentation and debugging while the compiled executable could be used on systems with only batch queue access where setting up a server is not possible.

References

1. Venners, B.: Programming at Python speed: A conversation with Guido van Rossum (2003) Available at `http://www.artima.com/intv/speed.html`.
2. Dubois, P., Hinsen, K., Hugunin, J.: Numerical Python. Computers in Physics **10** (1996)
3. Mathworks Inc.: MATLAB 6 User's Guide. (2001)
4. Gomez, C., ed.: Engineering and Scientific Computing with Scilab. Birkhäuser, Boston (1999)
5. Gumley, L.: Practical IDL Programming. First edn. Morgan Kaufmann Publishers (2001)

6. Schrüfer, E.: EXCALC – a package for calculations in modern differential geometry. In Shirkov, D., Rostovtsev, V., Gerdt, V., eds.: Proc. IV Int. Conf. Comp. Algebra in Physical Research, Dubna, U.S.S.R., World Scientific, Singapore, 1990, 71–80
7. Wolfram, S.: Mathematica: A System for Doing Mathematics by Computer. Addison-Wesley, Reading, Mass. (1988)
8. Char, B., et al.: Maple V, Language Reference Manual. Springer (1991)
9. Rand, R.: Computer algebra in applied mathematics: an introduction to MACSYMA. Number 94 in Research notes in mathematics. Pitman Publishing Ltd., London, UK (1984)
10. de Souza, P., Fateman, R., Moses, J., Yapp, C.: The Maxima book. 2003
11. Moler, C.: Why there isn't parallel Matlab. Mathworks Newsletter (1995).
12. Choy, L., Edelman, A.: MATLAB*P 2.0: A unified parallel MATLAB. Technical report, Massachusetts Institute of Technology (2003)
13. Choy, L.: MATLAB*P 2.0: Interactive supercomputing made practical. Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (2002)
14. Husbands, P.: Interactive Supercomputing. PhD thesis, Department of Electrical Engineering and Comp. Science, Massachusetts Institute of Technology (1999)
15. Chen, Z., Dongarra, J., Luszczek, P., Roche, K.: Self-adapting software for numerical linear algebra and LAPACK for clusters. Parallel Computing **29** (2003) 1723–1743
16. Dongarra, J., Eijkhout, V.: Self adapting numerical algorithms for next generation applications. International Journal of High Performance Computing Applications **17** (2003) 125–132 ISSN 1094-3420.
17. Norris, B.: An environment for interactive parallel numerical computing. Technical Report UIUCDCS-R-99-2123, University of Illinois, Urbana, Illinois (1999)
18. IEEE 754: Standard for binary floating point arithmetic. Technical report, Institute of Electrical and Electronics Engineers (1985)
19. Whaley, R., Petitet, A., Dongarra, J.: Automated empirical optimizations of software and the ATLAS project. Parallel Computing **27** (2001) 3–35
20. Dongarra, J., Whaley, C.: Automatically tuned linear algebra software (ATLAS). In: Proceedings of SC'98 Conference, IEEE (1998)
21. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK User's Guide. Third edn. Society for Industrial and Applied Mathematics, Philadelphia (1999)
22. Blackford, L., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J., Dhillon, I., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D., Whaley, R.: ScaLAPACK Users' Guide. Society for Industrial and Applied Mathematics, Philadelphia (1997)
23. Agrawal, S., Dongarra, J., Seymour, K., Vadhiyar, S.: NetSolve: Past, present, and future – a look at a grid enabled server. In Berman, F., Fox, G., Hey, A., eds.: Grid Computing: Making the Global Infrastructure a Reality. Wiley Publisher (2003)

Computer Algebra for Real-Time Dynamics of Robots with Large Numbers of Joints

Ramutis Bansevicius[1], Algimantas Cepulkauskas[2], Regina Kulvietiene[2], and Genadijus Kulvietis[2]

[1] Kaunas University of Technology, Donelaicio 73, Kaunas 3006, Lithuania
bansevicius@cr.ktu.lt
[2] Vilnius Gediminas Technical University, Sauletekio 11, Vilnius 2040, Lithuania
{algimantas.cepulkauskas,regina.kulvietiene,
genadijus.kulvietis}@gama.vtu.lt

Abstract. This paper deals with the description of a theoretical background of systematic computer algebra methods for analyzing the real-time dynamics of robots with a large numbers of joints. Many numerical methods based on different principles of mechanics were developed to obtain the equations that model the dynamic behavior of robots. In this paper, the efficiency of computer algebra application was compared with the most popular methods of forming the dynamic equations of robots in real time. To this end, the computer algebra system VIBRAN was used. A real-time dynamic model in closed form of the robots with large numbers of joints has been developed, using the computer algebra technique with the following automatic program code generation.

1 Introduction

The application of general control theory to complex mechanical systems, such as robots, aircrafts, etc., represents an extremely difficult problem because of prominent nonlinearity and complexity of mathematical models of these systems. If industrial robots have large numbers of joints, the application of such a theory and development of new control algorithms are unavoidable in order to achieve a high positioning speed and accuracy. In on-line control, the calculation of model equations must be repeated very often, preferably at the sampling frequency that is no lower than 50Hz. However, the problem of forming the dynamic equations of robots in real time by means of today's computers is rather difficult and complex. It appears necessary to develop computer methods of mathematical modelling for at least two reasons. One of them is that it is impossible to immediately choose the most convenient configuration when designing robots. The term configuration should be interpreted as the structure (i.e., kinematic scheme) and parameters (i.e., dimensions, masses, etc.). Thus, it is necessary to analyze a number of different robot configurations and choose the one, most appropriate to the future purpose of the device. Knowing how complex a task it is to write a mathematical model by hand, the need for an algorithm that would enable a computer to perform the task seems quite logical. The other

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 278–285, 2004.
© Springer-Verlag Berlin Heidelberg 2004

reason is the need in multiple applications for real-time control of robots. The development of computer methods, such that perform real-time calculations of robot dynamics, is a direct contribution to the synthesis of control algorithms for practical purposes. Particularly this problem is much more complex for the robots with a large number of joints [7], [12], [13].

In the last three decades, numerous investigators have used different principles of dynamics in order to obtain the equations that model the dynamic behavior of robot arms. The first formulations to be developed were based on a closed form representation of the equations, and the Lagrange-Euler (L-E) equations were preferentially used for this purpose. These formulations were found to be inefficient due to the high number of algebraic operations involved. A solution to this problem was found with the use of relationships present in the dynamic equations. The Newton-Euler (N-E) equations were found to be the most appropriate dynamic principle for this type of formulation and they have been used to develop the most efficient formulations known so far. Other formulations, based on the Kane equations, have yielded algorithms whose computational complexity is similar to that found in formulations based on the N-E equations. The use of dynamic principles different from those employed in the formulations based on L-E, N-E or Kane equations was minor and, furthermore, has produced formulations of high computational complexity. Currently it is believed that the use of diverse dynamic principles will lead to similar formulations of equivalent computational complexity. This has been partially proved by applying the appropriate relationships to the L-E equations in order to obtain an equivalent formulation to that given by the N-E equations, although a greater effort is required in order to reach the final equations [14]. It is for this reason that most of the formulations that produce efficient algorithms have been developed from the N-E equations. Featherstone and Orin [6] make a detailed review of these methods and algorithms derived.

The Gibbs-Appell (G-A) equations are one of the principles that has been used the least for solving the dynamic problem of manipulating robots. The simple form of these equations deal with mechanical systems subjected to holonomic and non-holonomic type of constraints is also emphasized in the specialized technical literature. Surprisingly, a bibliographical review of the literature on this area reveals a limited use of the G-A equations in modern dynamics. A few years ago, the supposed relationship of the G-A equations and Kane's dynamic equations caused a great number of works and comments on the matter [14]. In the field of robotics, Popov proposed a method, later developed by Vukobratovic [14], in which the G-A equations were used to develop a closed form representation of high computational complexity. This method was used by Desoyer and Lugner [11], [14] to solve, by means of the recursive formulation $O(n^2)$ (n is the number of the degree-of-freedom), an inverse dynamic problem, using the Jacobian matrix of the manipulator, with the view of avoiding the explicit development of partial derivatives. Another approach was suggested by Vereshchagin [14] who proposed manipulator motion equations from Gauss' principle and Gibbs' function. This approach was used by Rudas and Toth [11] to solve

the inverse dynamic problem of robots. Recently, Mata et al. [10] have presented a formulation of order *O(n)* that solves the inverse dynamic problem and establishes recursive relations that involve a reduced number of algebraic operations. The algorithms that model the dynamic behavior of manipulators are divided into two types: algorithms that solve the inverse dynamic problem and those that give a solution to the forward dynamic problem. In the former, the forces exerted by the actuators are obtained algebraically for certain configurations of the manipulator (position, velocity and acceleration). On the other hand, the forward dynamic problem computes the acceleration of joints of the manipulator once the forces, exerted by the actuators, are put. This problem is part of the process that must be followed in order to simulate the dynamic behavior of the manipulator. This process is completed after it has calculated the velocity and position of the joints by means of the process of numerical integration in which the acceleration of the joints and the initial configuration are data input to the problem.

The first efficient recursive algorithm for solving the inverse dynamic problem was proposed by Luh et al. [9]. This algorithm, based on the N-E equations, has been improved repeatedly in the course of years [2], [6]. Other authors have developed efficient recursive algorithms to solve the inverse dynamic problem, based on other principles of dynamics. As examples of these, we have the work of Hollerbach [14] that uses the L-E equations; and those of Kane and Levinson [14], and Angeles et al. [1], which use the Kane equations. The complexity of the above mentioned numerical algorithms will be compared with computer algebra realization. Some efforts to apply symbolic calculations in the dynamics of robots were made [11], [14], but due to tremendous final closed form equations these efforts were unsuccessful.

Simulations by means of numerical methods are powerful tools for investigations in mechanics but they do have drawbacks, e.g., finite precision, errors generated when evaluating expressions. The computerized symbolic manipulation is a very attractive means to reliably perform analytic calculations even with complex formulas and expressions. But frequently a semi-analytic approach, combining the features of analytical and numerical computations, is the most desirable synthesis. This allows the analytic work to be pushed further before numerical computations start.

For numerical-symbolic computation of the real-time dynamics of robots with large numbers of joints the computer algebra system VIBRAN [5], [8] was used [11]. The computer algebra system VIBRAN is a FORTRAN preprocessor for analytical computation with polynomials, rational functions and trigonometric series. Special VIBRAN's procedure can generate an optimized FORTRAN code from the obtained analytical expressions, which can be directly used in the programs for a further numerical analysis.

2 Real-Time Dynamics of Robot

The real-time dynamic model of a robot was constructed using the Uicker-Kahn method [11], [14], based on the L-E equations, that is very convenient for computer algebra implementation [3], [11]. This method enables the calculation of all the matrices of the dynamic robot model: the inertial matrix, the matrix of Coriolis and centrifugal effects and the gravity vector. The dynamic equations of an n-degree-of-freedom manipulator, derived using this method, are of the following form:

$$P_i = \sum_{j=i}^{n} \left\{ \sum_{k=1}^{j} \left[tr\left(\frac{\partial W_j}{\partial q_i} J_j \frac{\partial W_j^T}{\partial q_k} \right) \right] \ddot{q}_k + \right.$$
$$\left. + \sum_{k=1}^{j} \sum_{l=1}^{j} \left[tr\left(\frac{\partial W_j}{\partial q_i} J_j \frac{\partial^2 W_j^T}{\partial q_k \partial q_l} \right) \dot{q}_k \dot{q}_l \right] - m_j \overrightarrow{g}^T \frac{\partial W_j}{\partial q_i} \tilde{r}_{j0} \right\}, \qquad (1)$$

where P_i is a driving torque acting at the i-th joint; q_i is a generalized joint coordinate corresponding to the i-th degree of freedom; W_i is the transformation matrix between the i-th local coordinate system and the reference system; J_i is the inertia matrix of the i-th link with the respect to local coordinate system; m_i is the mass of the link i; $\tilde{r}_{i0}$ is the distance vector between the center of mass of the link i and the origin of the reference coordinate system, expressed in the local coordinate system of the i-th link; $\overrightarrow{g}$ is the gravity vector.

The matrix W_i may be expressed as

$$W_i = A_0^1 A_1^2 ... A_{i-1}^i,$$

where A_{k-1}^k is a (4×4) transformation matrix between two local coordinate systems.

Equation (1) may be expressed in the matrix form

$$P = H(q)\ddot{q} + \dot{q}^T C(q)\dot{q} + g(q), \qquad (2)$$

where P is the vector of driving torques; $H(q)$ is the inertial matrix of the system; $C(q)$ is the $n \times n \times n$ matrix of Coriolis and centrifugal effects; $g(q)$ is the vector of gravity effects.

Fig. 1 illustrates a flexible robot with a large number of joints [3], [4]. The robot consists of cylindrical piezoceramic transducers and spheres. Here the resonant oscillations of every piezoelectric transducer are controlled by a microprocessor, switching on and off the high-frequency and high-voltage signal from the signal generator. The phase and duration of every pulse, applied to the electrodes of transducers, are synchronized with the rotation of an unbalanced rotor, mounted in the gripper of the robot.

The external torque vector, placed in the gripper and rotating in the plane perpendicular to the gripper direction, is expressed in the form

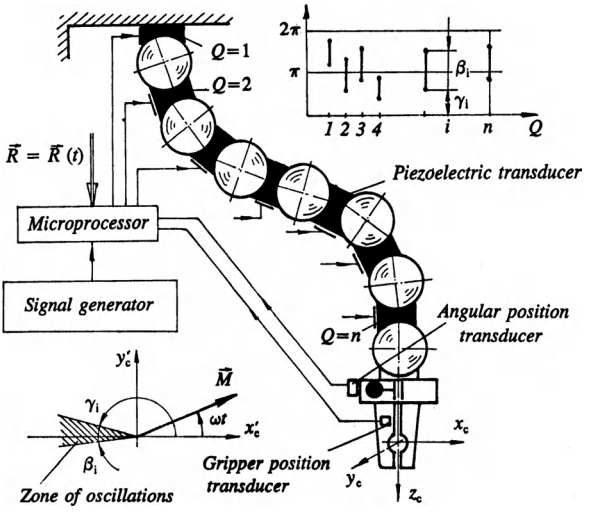


Fig. 1. The scheme of a robot with a large number of joints

$$\overrightarrow{F} = \left\{ \begin{array}{c} F_x = m_0 r\omega^2 \cos(\omega t) \\ F_y = m_0 r\omega^2 \sin(\omega t) \\ 0 \end{array} \right\}, \tag{3}$$

where m_0 is the mass of unbalance; r is a radius; ω is the angular velocity.

The recursive algorithm consists of two steps for each local coordinate. Therefore, the first step is the calculation of active forces and the second one is the definition of active torques. This algorithm may be expressed in the form

$$\begin{aligned} \overrightarrow{F}_i &= \tilde{A}_i^{i+1} \overrightarrow{F}_{i+1} \\ \overrightarrow{M}_i &= \tilde{A}_i^{i+1} \overrightarrow{M}_{i+1} + \overrightarrow{h}_{i,i-1} \times \overrightarrow{F}_{i+1}, \end{aligned} \tag{4}$$

where $\overrightarrow{F}_n = \overrightarrow{F}$, see formula (3). Expressions (4) are calculated starting from

$$i = n-1 \quad to \quad i = 1.$$

The generalized torque for the i-th joint may be obtained in the form

$$Q_i = \overrightarrow{M}_i \overrightarrow{z}_{i0}, \tag{5}$$

where $\overrightarrow{z}_{i0}$ is the unit vector of the respective axis.

3 Computer Algebra Implementation

In the algorithm for automatic generation of the analytical model, it will be assumed that the parameters of a robot (length, mass, inertia, etc.) are known

and will be treated as constants. Joint coordinates as well as their derivatives will be treated as independent variables, i.e., as symbols. Using the computer algebra technique, the Uicker–Kahn method is very convenient, because it enables us to obtain the equations of motion in closed form and may be applied in solving either the direct or the inverse problem of dynamics.

Fig. 2 illustrates a fragment of the VIBRAN program that implements the Uicker–Kahn method. In this program the sparse matrix technology was used to achieve the best performance. To have a possibility to compare various results and algorithms, only two joints of the proposed robot will be considered.

```
POLINOM A(16),B(20),C(20)
RACIONAL D,E,U
INTEGER*2 NA(18),NB(22),NC(22)
DATA G/0.,0.,-9.80621,0./
.....
100 RSMP(U,E,D,N)
ADDA(U,D)
100 RSMP(U,E,D,N)
```

Fig. 2. A fragment of the VIBRAN program

This program calculates all the elements of matrices *H(q), C(q), g(q).* These matrices were calculated for the discussed flexible robot with the 6-th-degree-of-freedom. The kinematic parameters of this robot in Denavit–Hartenberg's notation [3], [11], [14] are presented in the table below.

N	q_i	α_i	a_i	d_i
1	q_1	0	0	0
2	q_2	90°	0	0
3	q_3	0	0.04	0
4	q_4	-90°	0	0
5	q_5	-90°	0	0
6	q_6	0	0	0.04

For simplicity, a substitution was made to avoid numerical trigonometric calculation of the function

$$S_i = \sin q_i, \quad C_i = \cos q_i.$$

The fragment of analytical calculations of flexible robot matrices performed by the VIBRAN program is presented in Fig. 3.

In total 153 elements were calculated and about 15% of them were equal to zero.

A special VIBRAN procedure [5] , [8] generates two FORTRAN subroutines from the obtained analytical expressions of robot matrices. The code of the first

generated subroutine contains a dictionary of monomials included into the expressions of robot's matrices. This dictionary of monomials is sorted in ascending order of monomial multiindices to reduce the number of floating point multiplications. The code of the second generated subroutine contains the calculation of common members included in all the expressions and all the elements of robot's matrices. The generated subroutines can be immediately compiled and used for real-time operation, simulation or control synthesis.

```
H11   = .8326E-4+.1296E-3*C3**2-.9964E-4*C3**2*C4**2*C5**+.9964E
  -4*C3*S3*S4*C4*C5**2-
.................
G3 = -.113752E-6*S5*C4*C3+.113752E-6*S5*S4*S3+.14121E-5*C4*S3*S6
.........
G6 = .14121E-5*S3*S4*C6-.14121E-5*C4*C3*C6-.14121E-5*S3*C4*C5*S6
  -.14121E-5*C3*S4*C5*S6
```

Fig. 3. Analytical expressions of robot's matrices

The number of floating point product operations required to construct the dynamic model by the Uicker–Kahn method numerically depends on n^4 (n is the number of the degree-of-freedom) and, by contrast, the recursive methods based on the N-E or G-A equations have a linear dependency on the number of the degree-of-freedom. Some differences appear using the computer algebra technique. The Uicker–Kahn method produces closed-form differential equations and only recursive equations can be obtained from other well-known algorithms which means that only the numerical implementation is possible and this method suits only for inverse dynamics. The code presented in Fig. **??** contains only 371 floating point product. The computational complexity of the proposed approach is comparable with that of the most efficient algorithms known so far, as shown in the table below.

Authors	Principle	Products $(n+6)$	Number of operations
Luh et al. [9]	N-E	$150n - 48$	852
Angeles et al. [1]	Kane	$105n - 109$	521
Balafoutis and Patel [2]	N-E	$93n - 69$	489
Mata et al. [10]	G-A	$96n - 101$	475
This work	L-E	Closed form	371

Generalized torques were calculated in the same manner . These torques are needed to complete the control scheme of the robot. Another VIBRAN program calculates the acting forces and torques, using formula (4), and generalized torques, using formula (5).

4 Conclusions

The proposed mixed numerical-analytical implementation of the Uicker–Kahn method drastically reduces the number of floating point operations, particularly for robots with a large number of joints. The use of the computer algebra technique enables us to obtain the equations of motion in closed form. It can be applied in solving both the direct and the inverse problem of dynamics as well as in real-time dynamics modelling for intelligent control scheme realization.

References

1. Angeles, J., Ma, O., Rojas, A.: An algorithm for the inverse dynamics of n-axis general manipulators using Kane's equations. Comp. Math. Appl. 17 (12) (1989) 1545–1561.
2. Balafoutis, C.A., Patel, R.V.: Dynamic Analysis of Robot Manipulators: A Cartesian Tensor Approach, Kluwer Academic Press, Boston (1991).
3. Barauskas, R.; Bansevicius; R. Kulvietis; G. Ragulskis K.. 1988. Vibromotors for Precision Microrobots. Hemisphere Publishing Corp., USA.
4. Bansevicius R., Parkin R., Jebb, A., Knight, J.: Piezomechanics as a Sub-System of Mechatronics: Present State of the Art, Problems, Future Developments. IEEE Transactions on Industrial Electronics, vol. 43, (1) (1996) 23–30.
5. Cepulkauskas, A., Kulvietiene, R., Kulvietis G.: Computer Algebra for Analyzing the Vibrations of Nonlinear Structures. Lecture Notes in Computer Science, Vol. 2657. Springer-Verlag, Berlin Heidelberg New York (2003) 747–753.
6. Featherstone, R., .Orin, D. E.: Robot dynamics: equations and algorithms. Proceedings of the 2000 IEEE International Conference on Robotics and Automation, San Francisco (2000) 826–834.
7. Knani J.: Dynamic modelling of flexible robotic mechanisms and adaptive robust control of trajectory computer simulation. Applied Mathematical Modelling , Vol. 26. (12) (2002) 1113–1124.
8. Kulvietiene, R., Kulvietis, G.: Analytical Computation Using Microcomputers. LUSTI, Vilnius (1989).
9. Luh, J.Y.S., Walker, M.W., Paul, R.P.: On-line computational scheme for mechanical manipulators. J. Dyn. Syst. Meas. Control 102 (1980).
10. Mata, V., Provenzano, S., Valero, F., Cuadrado, J., I.: Serial-robot dynamics algorithms for moderately large numbers of joints. Mechanism and Machine Theory, 37 (2002) 739–755.
11. Rovetta, A., Kulvietis, G.: Lo sviluppo di software per il controllo dinamico di robot industriali. Dipartimento di Meccanica, Politecnico di Milano, Milano (1986).
12. Surdhar, J., S., White, A., S.: A parallel fuzzy-controlled flexible manipulator using optical tip feedback. Robotics and Computer-Integrated Manufacturing, Vol. 19 (3) (2003) 273–282.
13. Tso, S., K., Yang, T., W., Xu, W., L., Sun, Z., Q.: Vibration control for a flexible-link robot arm with deflection feedback. International Journal of Nonlinear Mechanics, 38 (2003) 51–62.
14. Vukobratovic, K., M., Kircanski M., N.: Real-time Dynamics of Manipulation Robots, Springer-Verlag, Berlin Heidelberg New York (1985).

Development of SyNRAC—Formula Description and New Functions

Hitoshi Yanami[1,2] and Hirokazu Anai[1,2]

[1] Information Technology Core Laboratories, Fujitsu Laboratories Ltd.
Kamikodanaka 4-1-1, Nakahara-ku, Kawasaki 211-8588, Japan
yanami@flab.fujitsu.co.jp, anai@jp.fujitsu.com

[2] CREST, Japan Science and Technology Agency
Kawaguchi Center Building, 4-1-8, Honcho, Kawaguchi 332-0012, Japan

Abstract. In this paper we present newly developed functions in Maple-package SyNRAC, for solving real algebraic constraints derived from various engineering problems. The current version of SyNRAC provides quantifier elimination (QE) for the quadratic case and an environment dealing with first-order formulas over the reals (including new simplifiers of formulas) on Maple.

1 Introduction

We presented Maple-package SyNRAC for solving real algebraic constraints in 2003 [1]. SyNRAC stands for a Symbolic-Numeric toolbox for Real Algebraic Constraints and is aimed to be a comprehensive toolbox composed of a collection of symbolic, numeric, and symbolic-numeric solvers for real algebraic constraints derived from various engineering problems.

In this paper we show the current status of development of SyNRAC. In the previous version of SyNRAC [1] the following algorithms were available

- a special QE by the Sturm-Habicht sequence for sign definite condition,
- a special QE by virtual substitution for linear formulas,
- some naive simplifications of quantifier-free formulas.

Besides, the current version of SyNRAC provides the following:

- an environment dealing with first-order formulas over the reals,
- a special QE by virtual substitution for quadratic formulas,
- some new standard simplifiers of formulas.

Since we firstly presented SyNRAC, we have introduced some new operational symbols and fixed a notation system for expressing formulas. We are now developing our tool under the basis of the new environment. The QE algorithms previously equipped have also been reimplemented after the latest setting. These new features greatly extend the applicability and tractability of SyNRAC for solving real algebraic constraints in engineering. The current notation for first-order logic over the reals is much easier to read than the previous one. This helps users describe mathematical formulas for various types of real algebraic constraints. A

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 286–294, 2004.
© Springer-Verlag Berlin Heidelberg 2004

special QE method for quadratic formulas widens the application areas of SyNRAC in actual problems (see [2]). The simplifiers can reduce the size of a given formula. This contributes not only to improve recognition of formulas but also to remarkably improve the efficiency of special QE procedures based on virtual substitution.

Furthermore, using SyNRAC as a kernel, we are now pushing the further development of design tools based on computer algebra (in particular, QE) in various application fields: One successful attempt is the development of a toolbox for *parametric robust control design* on MATLAB [3] based on the authors' previous works concerning QE-based robust control design [4,5,6,7].

2 A New Environment for First-Order Formulas over the Reals

When we say a real algebraic constraint, what we have in mind is a first-order formula over the reals. We describe what type of formulas we are dealing with and how they are expressed in SyNRAC.

An *atomic formula* is an equality or inequality $f(x_1, \ldots, x_n)\ \rho\ g(x_1, \ldots, x_n)$, where f and g are polynomials in a finite number of indeterminates over $\mathbb{Q}$ and ρ is one of the relational operators $\{=, \neq, \leq, <\}$. A *formula* is a string obtained by appropriately arranging atomic formulas, logical operators, or existential/universal quantifiers. Here is an example of existential formulas with respect to x, y, and z

$$\exists x \exists y \exists z\ (f_1 \wedge f_2 \wedge (h_1 \vee h_2) \wedge f_3) \Longrightarrow \neg(g_1 \wedge g_2)\ ,$$

where f_i, g_i, and h_i are atomic formulas.

To express formulas in SyNRAC, we need to prepare and fix notational symbols for $\exists$, $\forall$, $\wedge$, $\vee$, $\neg$, and so forth. In the earlier stages of implementation, we were using relational and logical operators bundled in Maple. As we proceeded, it turned out that some of the Maple's operators are unsuitable for our purpose. Let us show a simple example. Let x be just an indeterminate. The `evalb` command, which evaluates a relation in Boolean context, in Maple returns `false` when $x = 0$ is input. This behavior does not meet our expectation, because we want to remain $x = 0$ unchanged unless x is assigned a value.

To avoid such reactions, we have introduced a user-defined operator `&=`[1] and replaced it for the Maple's equality symbol '='. To maintain consistency, the other relational operators are redefined by adding "`&`" at the forefront of the respective commands (see Table 1). Some of them are just an alias for the Maple's corresponding command. Logical operators and quantifier symbols have also been redefined in the same way as in Tables 2 and 3. In SyNRAC, the atomic formula $x^2 - 2y^2 - 3z^2 \leq xy - 6yz - z + 7$ is expressed in

[1] A Maple user can form a neutral operator symbol by using `&name` (the ampersand character "`&`" followed by one or more characters).

Table 1. The relational operators in SyNRAC

Operator	$=$	$\neq$	$\leq$	$<$	$\geq$	$>$
Notation	`&=`	`&<>`	`&<=`	`&<`	`&>=`	`&>`

Table 2. The logical operators in SyNRAC

Operator	$\wedge$	$\vee$	$\neg$	$\Longrightarrow$	$\Longleftarrow$	$\Longleftrightarrow$
Notation	`&and`	`&or`	`&not`	`&impl`	`&repl`	`&equiv`

Table 3. The quantifiers in SyNRAC

Operator	$\exists x_1 \cdots \exists x_n \varphi$	$\forall x_1 \cdots \forall x_n \varphi$
Notation	`&Ex`$([x_1, \ldots, x_n], \varphi)$	`&All`$([x_1, \ldots, x_n], \varphi)$

`(x^2-2*y^2-3*z^2) &<= (x*y-6*y*z-z+7)` .[2]

The example formula above is expressed in the following:

`&Ex([x,y,z],` (f_1 `&and` f_2 `&and (`h_1 `&or` h_2`)` `&and` f_3`)` `&impl &not(`g_1 `&and` g_2`))` .

The operators `&and` and `&or` can also be used as a prefix operator, taking a list of operands as an argument. The expression `&and([`f_1`,` f_2`, ...,` f_n`])` is equivalent in SyNRAC to f_1 `&and` f_2 `&and` $\cdots$ `&and` f_n.

According to these notational rules, QE algorithms has been (re)implemented in SyNRAC. In addition, several basic utility functions on formulas are provided in SyNRAC, for example, functions for counting the number of atomic formulas, extracting atomic formulas from a formula as a list, and so on. Moreover, some computations for the disjunctive normal form[3] are also available.

3 Solving Quadratic Algebraic Constraints over the Reals

Here we briefly explain a special QE by virtual substitution of parametric test points that is applicable to formulas in which the quantified variables appear at most quadratically (see [8] for details). We call a formula whose atomic subformulas are at most quadratic (linear) with respect to its quantified variables a *quadratic* (*linear*) formula, respectively.

Let

$$\psi(p_1, \ldots, p_m) \equiv Q_1 x_1 \cdots Q_n x_n \varphi(p_1, \ldots, p_m, x_1, \ldots, x_n)$$

[2] The polynomials both sides should be enclosed within parentheses since the user-defined operator `&name` in Maple has higher priority than the basic arithmetic operators. In the examples in the present paper, we leave them out when too convoluted.

[3] A formula is called a *disjunctive normal form* if it is a disjunction (a sequence of $\vee$'s) consisting of one or more disjuncts, each of which is a conjunction (a sequence of $\wedge$'s) of one or more atomic formulas.

be a linear or quadratic formula, where $Q_i \in \{\forall, \exists\}$ and φ is a quantifier-free formula. By using the equivalence $\forall x\varphi(x) \Longleftrightarrow \neg(\exists x \neg\varphi(x))$, we can change the formula into an equivalent one of the form $(\neg)\exists x_1 \cdots (\neg)\exists x_n\ (\neg)\varphi$. The negation '$\neg$' that precedes a quantifier-free formula can be easily eliminated (use De Morgan's law and rewrite the atomic subformulas), which is not essential part of QE. Therefore we may focus our attention on an *existential formula,* i.e., a formula of the form $\exists x_1 \cdots \exists x_n \varphi(p_1, \ldots, p_m, x_1, \ldots, x_n)$. Furthermore, it is sufficient to show how to eliminate $\exists x$ in $\exists x\varphi$, since all the quantifiers in the formula can be eliminated by removing one by one from the innermost one.

Now our main purpose is to eliminate the quantified variable $\exists x$ in

$$\psi'(p_1, \ldots, p_m) \equiv \exists x\ \varphi(p_1, \ldots, p_m, x)$$

with $\varphi(p_1, \ldots, p_m, x)$ quantifier-free and quadratic, and obtain an equivalent quantifier-free formula $\psi'(p_1, \ldots, p_m)$. For fixed real values $q_1, \ldots, q_m$ for the parameters $p_1, \ldots, p_m$, all polynomials appearing in $\varphi(x)$ are linear or quadratic. Therefore, the set $M = \{r \in \mathbb{R} | \varphi(q_1, \ldots, q_m, r)\}$ of real values r for x satisfying φ is a finite union of closed, open, or half-open intervals over $\mathbb{R}$. The endpoints of these intervals are among $\pm\infty$ and the real zeros of atomic formulas in φ. Then candidate terms, say, $t_1, \ldots, t_k$, for those zeros can be constructed by the solution formulas for linear or quadratic equations.

If φ does not contain any strict inequalities, all the intervals composing M are either unbounded or closed. In the closed case such an interval contains its real endpoint. So M is nonempty if and only if the substitution of $\pm\infty$ or of one of the candidate solutions t_j for x satisfies φ. Let S be the candidate set $S = \{t_1, \ldots, t_k, \pm\infty\}$. Such a set is called an *elimination set* for $\exists x\varphi$. We obtain a quantifier-free formula equivalent to $\exists x\varphi$ by substituting all candidates in S into φ *disjunctively:*

$$\exists x\varphi \Longleftrightarrow \bigvee_{t \in S} \varphi(x//t)\ .$$

We note that there is a procedure assigning the expression $\varphi(x/t)$ obtained from φ by substituting t for x an equivalent formula [8]. We denote the resulting formula by $\varphi(x//t)$.

If φ contains strict inequalities, we need to add to S other candidates of the form $s \pm \epsilon$, where s is a candidate solution for some left-hand polynomial in a strict inequality and ϵ is a positive infinitesimal.

For improving the efficiency of this method, the following two points are crucial: (i) refining the elimination set S by a scrupulous selection of a smaller number of candidates in S; (ii) integrating with sophisticated simplifications of quantifier-free formulas. SyNRAC now employs three types of elimination sets proposed in [9]. Simplifications in SyNRAC are discussed in the next section.

Moreover, (heuristic) techniques for decreasing the degree during elimination are important for raising the applicability of quadratic QE, because after one quantifier is eliminated for a quadratic case the degree of other quantified variables may increase. Only simple degree-decreasing functions are implemented in the current version of SyNRAC.

4 Simplification

In the present paper, the term simplification is used for simplification of quantifier-free formulas. When a quantifier is eliminated in a given first-order formula with a special QE procedure, its quantifier-free part usually gets larger. During a QE algorithm, formulas under manipulation tend to get extremely large, deeply nested and highly redundant. That is why simplification procedures, which equivalently change a quantifier-free formula into more concise one, are important. Utilizing simplification algorithms combined with a special QE algorithm contributes to improve not only readability of the resulting formula but efficiency of the computation.

As for simplification, Maple, on which we implement our toolbox SyNRAC, can simplify certain formulas. By using Maple's `evalb` command for the inequality $3 < 5$, the value `true` are obtained. But it does not work for, say, '$x < 3$ `and` $x < 5$'; the `evalb` command does nothing and just returns '$x < 3$ `and` $x < 5$', not '$x < 5$'. Dolzmann and Sturm [10] summarize the rule for simplifying such formulas, to be precise, the formula '$f\ \rho_1\ 0$ `and`/`or` $g\ \rho_2\ 0$', where f and g differ only by a constant c, and ρ_1 and ρ_2 are an (in)equality. They called these laws *ordering theoretical smart simplification* when $c = 0$, i.e., $f = g$ and *additive smart simplification* when $c \neq 0$, respectively.

Automatic formula simplifiers are implemented in REDLOG[4] and QEPCAD[5] (see [13,10] for possible simplifications). Several simplification rules including ordering theoretical and additive smart simplification are implemented in SyNRAC, which greatly increase the efficiency of our QE commands. These rules dramatically work especially when the number of quantified variables are large.

5 Commands in SyNRAC

In this section we show some computational examples to illustrate how commands in SyNRAC are used.[6]

First, you need to load the packages:

```
> read "synrac";   with(combinat);
```

You can use `qe_sdc` to solve the formula $\forall x > 0,\ f(x) > 0$, called the *sign definite condition* (SDC). The first argument of `qe_sdc` is polynomial f and the second is the variable to be eliminated. The next example shows how to use the command to solve the problem $\forall x > 0,\ a_2x^2 + a_1x + a_0 > 0$,

```
> qe_sdc(a2*x^2+a1*x+a0, x);

                  ( -a0 &< 0 &and  a1 &< 0 &and -4*a0+a1^2 &< 0 ) &or
                  ( -a0 &< 0 &and -a1 &< 0 &and -4*a0+a1^2 &< 0 ) &or
                  ( -a0 &< 0 &and -a1 &< 0 &and  4*a0-a1^2 &< 0 )
  time = 0.02, bytes = 123614
```

[4] REDLOG is a QE package based on virtual substitution on REDUCE.

[5] QEPCAD is a general QE package that is applicable to all first-order formulas based on cylindrical algebraic decomposition (CAD) [11,12].

[6] All computations were executed on a Pentium III 1 GHz processor.

By using `qe_lin` command, you can solve the existential linear QE problem. This command takes two arguments; the former is a list of quantified variables and the latter a quantifier-free formula. In the following example, `qe_lin` eliminates the two quantified variables in $\exists x \exists y(y > 2x + 3 \wedge x > 0 \wedge y < s)$ and returns a condition with regard to s.

```
> qe_lin(&Ex([x,y], y&>2*x+3 &and x&>0 &and y&<s));

                                  -1/2*s &< -3/2
  time = 0.03, bytes = 144686
```

The `qe_quad` command can deal with quadratic QE problems. You can solve the quadratic QE problem $\exists x \exists y(x^2 - 4x - 5 \leq y \wedge 3 \leq x \wedge y \leq -5s + 6)$ as follows:

```
> qe_quad(&Ex([x,y], &and[(x^2-4*x-5)&<=y, 3&<=x, y&<=(-5*s+6)]));

                                  -14+5*s &<= 0
  time = 0.03 sec, bytes = 233514
```

The two examples below show that if a decision problem is given, i.e., the input contains no free variables, each command returns the `true` or `false` value:

```
> qe_sdc(x^5-x^2+3*x-9,x);

                                    false
  time = 1.11, bytes = 8774262

> qe_lin(&Ex([x,y], y&<2*x+2 and y&<=-3*x+12 and y&>(1/3)*x+5));

                                     true
                    A sample point:  [x, y], [52/25, 144/25]
  time = 0.03, bytes = 155078
```

A sample point is one that makes the formula true.
By calling the `qfsimple` command, you can simplify quantifier-free formulas with ordering theoretical and additive smart simplification.

```
> qfsimple((x&<5 &and x&<c &and x&>=10) &or (x&<=3 &and x&<=5 &and x&>=-5
            &and x&<>3) &or (x&>7 &and x&<=d));

         (-3+x &<= 0 &and -5-x &<= 0) &or (-x &< -7 &and -d+x &<= 0)
  time = 0.00, bytes = 44974
```

The `substsimple` command simplifies quantifier-free formulas by making use of simple atomic equations. This command repeats the following two procedures: (i) solving the linear atomic equations with only one variable in each conjunctive formula and substituting its solution for the variable as far as its influence goes; (ii) calling the `qfsimple` command and simplifying the resulting formula. These are redone until such linear equations run out.
In the next example, z in the input formula is firstly substituted by $3/2$ except in the 4th atomic one, and then by using the 1st equation in the resulting formula, x is replaced by $3/5$ in three places.

```
> substsimple(5*x&=2*z &and 9&>=3*y-x &and x+4*y+z&>0 &and 2*z-3&=0
              &and 5*x+2*y&<=z+3);
```

```
          x &= 3/5 &and -40*y &< 21 &and z &= 3/2 &and -3+4y &<= 0
time = 0.00, bytes = 97406
```

6 Examples

We show two example problems from mathematical programming and solve them with SyNRAC.

Example 1 First consider the following convex quadratic programming:
minimize $x_1^2 + x_1x_2 + 2x_2^2$,
subject to $x_1 + 4x_2 \geq 16$, $3x_1 + 2x_2 \geq 18$, $x_1 \geq 0$, $x_2 \geq 0$.

To obtain a description of the first-order formula, we add an unqualified variable z and express the problem in

$$\exists x_1 \exists x_2 (z - (x_1^2 + x_1x_2 + 2x_2^2) \geq 0 \wedge x_1 + 4x_2 \geq 16 \wedge 3x_1 + 2x_2 \geq 18 \wedge x_1 \geq 0 \wedge x_2 \geq 0) .$$

Eliminating the quantified variables x_1 and x_2, we can obtain a condition on z, from which we would obtain the range of the objective function. Quantifier elimination procedure in SyNRAC outputs the condition below in 1.78 sec:

```
&or([46 - z &<= 0, &and([567 - 16 z &<= 0,
    &or([(46 - z) &= 0, &and([46 - z &<= 0, -162 + z &<= 0]),
    &and([-466 + z &<= 0, 2659 - 40 z &<= 0]), 2659 - 40 z &<= 0])]),
    &and([46 - z &<= 0, -256 + z &<= 0])])
```

A little computation tells us that this formula is equivalent to $z \geq 46$. Thus the minimum of the objective function $x_1^2 + x_1x_2 + 2x_2^2$ equals 46.

Example 2 Next we consider the following nonconvex programming:
minimize $x_1 + 3x_2$,
subject to $x_1^2 + x_2 - 4x_1 - 3 \geq 0$, $x_1^2 + 2x_2 - 12x_1 + 32 \geq 0$, $x_1 \geq 0$, $x_2 \geq 0$.

As in the first example, we rewrite the problem by using a slack variable z into

$$\exists x_1 \exists x_2 (z - (x_1 + 3x_2) \geq 0 \wedge x_1^2 + x_2 - 4x_1 - 3 \geq 0 \wedge x_1^2 + 2x_2 - 12x_1 + 32 \geq 0 \wedge x_1 \geq 0 \wedge x_2 \geq 0) .$$

Quantifier elimination procedure as well as simplification after QE outputs the condition below in 6.12 sec:

```
-155 + 25*42^(1/2) - z &<= 0
```

Thus the minimum of the objective function $x_1 + 3x_2$ is $-155 + 25\sqrt{42}$, or approximately 7.02.

7 Conclusion

We presented a newly developed functions in Maple-package SyNRAC. The current version of SyNRAC, in particular, provides quantifier elimination for quadratic case and some standard simplifiers of formulas over the new environment for

first-order formulas over the reals on Maple. The new features greatly extend the applicability and tractability of SyNRAC for solving real algebraic constraints in engineering. We are continually improving the efficiency of implemented algorithms and are going to implement other algorithms (including symbolic-numeric algorithms) for solving real algebraic constraints into SyNRAC.

Now we note that based on SyNRAC the development of a toolbox for parametric robust control design on MATLAB is ongoing.

We are aware that there is still a considerable way for SyNRAC to be a sophisticated symbolic-numeric tool. Hence we will keep progressing to bridge the gap. Our goal is to develop innovative symbolic-numeric methods and to build novel design tools via SyNRAC for various fields in engineering.

Acknowledgements. The authors would like to thank Volker Weispfenning for his invaluable advice.

References

1. Anai, H., Yanami, H.: SyNRAC: A Maple-package for solving real algebraic constraints. In: Proceedings of International Workshop on Computer Algebra Systems and their Applications (CASA) 2003 (Saint Petersburg, Russian Federation), P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2657, Springer (2003) 828–837
2. Dolzmann, A., Sturm, T., Weispfenning, V.: Real quantifier elimination in practice. In Matzat, B.H., Greuel, G.M., Hiss, G., eds.: Algorithmic Algebra and Number Theory. Springer, Berlin (1998) 221–247
3. Sakabe, K., Yanami, H., Anai, H., Hara, S.: A MATLAB toolbox for parametric robust control system design based on symbolic computation. In: Bulletin (Kokyuroku) of RIMS (Research Institute for Mathematical Sciences, Kyoto Univ.) Workshop on Computer Algebra—Algorithms, Implementations and Applications 2003 (15-18 December 2003), (To appear)
4. Anai, H., Hara, S.: Fixed-structure robust controller synthesis based on sign definite condition by a special quantifier elimination. In: Proceedings of American Control Conference 2000. (2000) 1312–1316
5. Anai, H., Hara, S.: Linear programming approach to robust controller design by a quantifier elimination. In: Proceedings of SICE Annual Conference 2002 (Osaka, Japan). (2002) 863–869
6. Anai, H., Hara, S.: A parameter space approach for fixed-order robust controller synthesis by symbolic computation. In: Proceedings of IFAC World Congress on Automatic Control b'02. (2002)
7. Anai, H., Yanami, H., Hara, S.: SyNRAC: a maple-package for solving real algebraic constraints toward a robust parametric control toolbox. In: Proceedings of SICE Annual Conference 2003 (Fukui, Japan). (2003) 1716–1721
8. Weispfenning, V.: Quantifier elimination for real algebra—the quadratic case and beyond. Applicable Algebra in Engineering Communication and Computing **8** (1997) 85–101
9. Loos, R., Weispfenning, V.: Applying linear quantifier elimination. The Computer Journal **36** (1993) 450–462 Special issue on computational quantifier elimination.
10. Dolzmann, A., Sturm, T.: Simplification of quantifier-free formulae over ordered fields. Journal of Symbolic Computation **24** (1997) 209–231

11. Collins, G.E.: Quantifier elimination for the elementary theory of real closed fields by cylindrical algebraic decomposition. In Brakhage, H., ed.: Automata Theory and Formal Languages. 2nd GI Conference. Volume 33 of Lecture Notes in Computer Science., Gesellschaft für Informatik, Springer-Verlag, Berlin, Heidelberg, New York (1975) 134–183
12. Collins, G.E., Hong, H.: Partial cylindrical algebraic decomposition for quantifier elimination. Journal of Symbolic Computation **12** (1991) 299–328
13. González-Vega, L.: A combinatorial algorithm solving some quantifier elimination problems. In Caviness, B., Johnson, J., eds.: Quantifier Elimination and Cylindrical Algebraic Decomposition. Texts and monographs in symbolic computation. Springer-Verlag (1998) 365–375

DisCAS: A Distributed-Parallel Computer Algebra System*

Yongwei Wu[1], Guangwen Yang[1], Weimin Zheng[1], and Dongdai Lin[2]

[1] Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, China
[2] State Key Laboratory of Information Security, Institute of Software,
Chinese Academy of Sciences, Beijing, 100080, China.

Abstract. The DisCAS system employs and interacts with multiple ELIMINOs distributed over the Internet to achieve a distributed-parallel computing environment. ELIMINO is a computer algebra system developed to support Wu's method for computing characteristic sets of polynomials and for other related operations. GridPPI, an MPI-like interface for grid computing, could couple multiple computing tools distributed over grid to run complex computing problems. DisCAS combines grid technology, GridPPI and ELIMINOs to deliver high performance computing to Internet users. The overall ELIMINO, GridPPI, and grid technology, as well as the DisCAS architecture are presented. The way to access and apply DisCAS and related works are also discussed at last.

1 Introduction

ELIMINO [7] is a new computer algebra system being developed at the Key Laboratory of Mathematics Mechanization, Chinese Academy of Sciences. Capabilities of ELIMINO include manipulation of multi-precision numbers and polynomials, computation of characteristic sets in Wu's method [9], polynomial equation solving, geometric theorem proving etc. As a universal system for a broad class of problems, ELIMINO is very computation intensive.

Polynomial characteristic sets are especially very computation intensive. Even medium-sized characteristic set problems can take a very long time to solve. Consequently,it is reasonable and promising to use many ELIMINOs distributed over Internet to improve the computing performance. DisCAS, a distributed-parallel computer algebra system, aims to use multiple ELIMINOs over Internet to speed up the computation.

The grid [1,2] technology uses high-speed networks to integrate heterogeneous computers distributed over a network to form a virtual supercomputer. Grid computing is an important and current research area and it promises to supply supercomputing powers by combining a grid of networked workstations. By using grid technology, *Globus Toolkit* (GT) [3], multiple ELIMINOs distributed over a

* This Work is supported by NSFC (60373004,60373005) and China Postdoctoral Foundation

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 295–302, 2004.
© Springer-Verlag Berlin Heidelberg 2004

grid can provide high performance symbolic computing services for users as an integrated system: DisCAS.

GridPPI[17] is a coarse-grained distributed parallel programming interface (PPI) for grid computing. As a MPI-like programming model[12], GridPPI provides a group of generic and abstract function prototypes with well-specified semantics. It supports high-level dynamic parallel tasking over grid too. Through GridPPI, users could couple multiple ELIMINOs distributed over multiple heterogeneous machines to run practical complex computing applications.

By adopting GT and GridPPI, DisCAS achieves the following specific results with minimal effort.

- Delivering the powerful distributed-parallel symbolic computation to Internet users.
- Demonstrating grid computing as a way to speed up computer algebra systems.
- Parallelizing GCD, factorization and characteristic-sets based computations over the Internet.
- By implementing GridPPI for DisCAS, providing one MPI-like programming model for users.
- Achieving interoperability with other GridPPI compliant systems, such as grid operation system *TsingHua University Grid* (THUG)[18].

We begin with an overview of the ELIMINO system, and the GridPPI specification. We then introduce the *Open Grid Service Architecture* (OGSA) and GT. These pave the way for describing the architecture and implementation of DisCAS. Use of DisCAS and related works are then discussed.

2 ELIMINO

IELIMINO is a new computer-mathematics research system developed at the Key Laboratory of Mathematics Mechanization, Institute of Systems Science, Chinese Academy of Sciences, as part of the "Mathematics Mechanization and its Applications" project. A long-standing goal at MMRC is to automate Wu's method independent of existing computer algebra systems. In ELIMINO, many different kinds of mathematical objects and data structures are provided. As an interactive system, ELIMINO is designed to focus on the implementation of Wu's method for researchers to perform sophisticated mathematical computations. It has very general capabilities for treating numbers, polynomials and characteristic sets.

To facilitate mathematical research, ELIMINO is kept open and flexible. The architecture of ELIMINO consists of three parts (see Figure 1):

- **Kernel part** is the soul of the system, it contains implementation of number system, polynomial manipulation system, characteristic sets method. The kernel part can be viewed as a powerful algebraic compute engine.

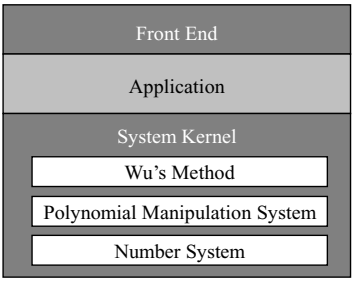


Fig. 1. ELIMINO system Architecture

- **Applications** are packages or programs developed using the ELIMINO library. Examples include the polynomial system solver and the geometry theorem prover. A package may be built-in or loaded into ELIMINO on demand.
- **Front-end** is the interface between the system and users. The front end handles the interaction between the user and the system.

3 GridPPI Specification

GridPPI[17] is extension of *Open Mathematical Engine Interface*OMEI[5,6]. It aims to be an *application programming interface* (API) general enough to work for most grid computing environments. It specifies a set of function prototypes together with their syntax and semantics to give a MPI-like programming level interface for computing engines. These function prototypes supports all operations that are necessary for secure access and coordinated use of multiple computing tools, including service discovering and selecting, task submitting and reporting, communication between subtasks, atom task executing and status report, etc..

GridPPI supports task-level dynamic parallel tasking over grid too. Through GridPPI, users could couple multiple computing tools distributed over multiple heterogeneous machines to run practical complex computing applications. As an attempt in standardizing programming interface for grid computing, GridPPI achieve several objectives:

– **Achieving Cooperative Use of Multiple Compute Engines**
GT framework can make heterogeneous machines internet accessible. Compute engines over these machines can be called through GridPPI easily, and more, can be cooperatively used.
– **Providing a MPI-like Programming Model**
MPI[12] programming model is easy accepted for most high performance computing requirers. GridPPI provides one distributed-parallel computing interface available over the Internet.

- **Application Portability**
  An application or user interface developed using any GridPPI-compatible interface would be portable among different grid systems, as long as those systems have GridPPI drivers available.
- **Integration of Different Grid Systems**
  Since an application can access multiple grid system by loading multiple GridPPI drivers, an integrated grid system with more powerful and combined capabilities can be accomplished under GridPPI programming model.

4 Open Grid Service Architecture and GT3

The grid [1,2] is a virtual supercompter consisting of heterogeneous computers (nodes) distributed over a network. Grid computing is a research area about how to combine networked workstations and harness their computation powers.

The Open Grid Service Architecture (OGSA) [2] uses key grid technologies and Web services mechanism [11] to create an integrated, distributed system framework. It specifies a uniform exposed service semantics (the Grid service), defines standard mechanisms for creating, naming, and discovering transient Grid service instances, provides location transparency and multiple protocol bindings for service instances, and thus supports integration with underlying native platform facilities. GT3 (Globus Toolkit 3) is a reference implementation of the Open Grid Service Infrastructure (OGSI)[14]. It provides a development environment including programming models for exposing and accessing grid service implementations.

The GT3 provides a uniform Java programming model for programmers to build and deploy their own grid services. Figure 2 shows the architecture of the globus platform and the way users access the grid service. To a globus platform, computing and data resources of a single node are considered *grid services*. A grid service is a network service that provides a set of well-defined interfaces that follow specific conventions [2].

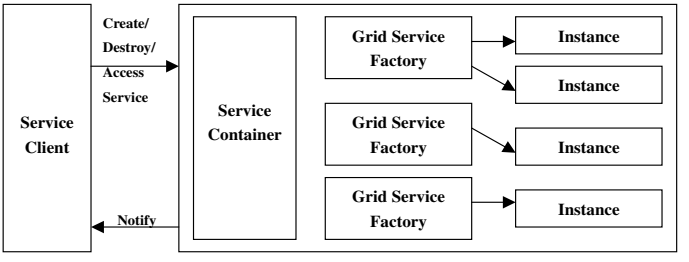


Fig. 2. Architecture of GT3

The GT3 *Service Container* (Figure 2) listens for incoming service requests. For a create-service request, the service container first performs security checks.

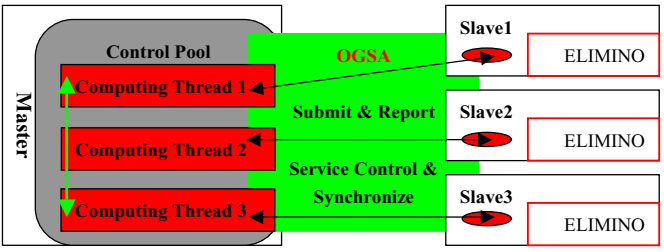


Fig. 3. DisCAS Architecture

It then calls the *Grid Service Factory* class to create a new service instance. A *Uniform Resource Identifier* URI for this service instance is returned to the requester. This URI is known as a *Grid Service Handle* (GSH). With the GSH, the service client can use and control the service instance. At the end of computations, the service instance can be destroyed. Each node that provides grid services has a service container that manages all grid services in that node. A grid service factory acts as a service resource provider. It manages all service instances of a specialized grid service.

5 DisCAS Architecture and Implementation

Through OGSA, multiple ELIMINO engines distributed over a grid can provide powerful computing services for Internet users as a virtual supercomputer. Figure 3 shows the architecture of DisCAS. It is a master-slave arrangement. The master program, on the client side, instantiates and controls multiple slave ELIMINO servers, each with a front end. The master runs a *control pool* of threads. Each control thread is in charge of the interaction with one particular remote ELIMINO engine. The control pool loads the task class and allocates server resources for the required tasks. For each task, the control pool first creates a service instance (a slave ELIMINO) in the allocated server node and then creates a new control thread for the task. Because the actual task is executed in the ELIMINO server, the responsibility of the control thread is to supply service control and synchronization, communication between threads.

Developers can easily create and access the ELIMINO computing services following the GridPPI compliant API Just like writing an MPI [12] program, developers simply write a Java class that describes the task for each ELIMINO server and send this class to the control pool. The control pool accesses the remote ELIMINO servers through *remote drivers*. The remote drivers in turn access computing service through a grid service locator. Figure 4 shows the control flow of a computing thread in the pool.

As shown in Figure 4, an *ELIMINO server* is an ELIMINO deployed as a grid service through a local driver. This service is mapped to a GSH (service locator

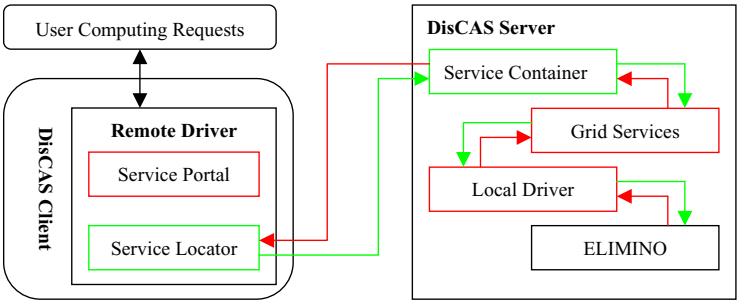


Fig. 4. Control Flow of One Computing Thread

in Figure 4) held by the DisCAS master through the GT3 service container, DisCAS can deliver the high performance computing power to remote users.

6 Use of DisCAS

Figure 5 shows the user interface we developed to access DisCAS. First, the `Nodes` box displays all the available grid nodes that can provide computational services. You can also add or delete nodes from this interface. The computation tasks can be seen in the `Task Lists` box. This list is editable by adding or deleting tasks. Once the node list and task list have been set up, you can click on the `Execute` button and the tasks will be assigned and submitted to grid nodes for computing. The ongoing status and results sent back from grid nodes will be displayed in the `Result` box.

Another way is Java programming through GridPPI. This way is much more flexible and practical for users. Developers could use the implementation of GridPPI we provided to couple multiple ELIMINOs over the Internet to complete the complex symbolic computation.

7 Related Works

By all means there have been various attempts to provide a distribute-parallel computing environment through grid technology. Many such efforts have been collected and catalogued by the grid application research group of the Global Grid Forum. Among the most famous and similar with our DisCAS are PSE and IAMC. PSE also provides an API for grid-based computing. Users could couple multiple computing nodes, potentially of different architectures, to run chemical computational problems through PSE[13]. IAMC[8,4] aims to make mathematical computations accessible easily and widely over Internet.

PSE (Problem Solving Environments)[13] inherits some interesting solutions exploited in Charlotte [16] and NetSolve[15]. It is designed to provide all the computational facilities needed to solve a target class of problems over a grid.

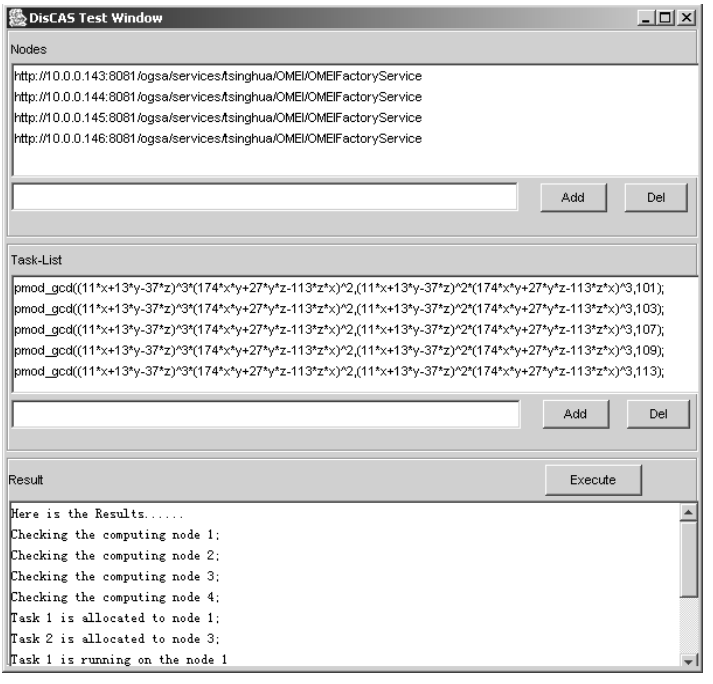


Fig. 5. DisCAS User Interface

As one chemical application grid, PSE also supply a completely transparent support to the user who does not have to care about the localization and the allocation of computing resources.

Internet Accessible Mathematical Computation (IAMC)[8] is a research project at the Institute of Computational Mathematics (ICM) at Kent State University. The goal of IAMC is to make mathematical computations Internet accessible easily and widely. By loading multiple OMEI [6] drivers, it can also access multiple compute engines. IAMC is an interactive computing environment over Internet. It gets the users' single computing request from the IAMC client and sends the request to one compute engine one by one.

8 Conclusion and Future Work

By using grid technology, DisCAS integrates multiple ELIMINOs distributed over the Internet to provide high performance computing services for remote users. It provides parallel GCD, factorization and characteristic-set based computations. By implementing GridPPI, DisCAS provides one MPI-like programming model for users. At the same time, DisCAS could achieve interoperability with other GridPPI compliant systems, such as grid operation system THUG.

The design and implementation of DisCAS is not final. THUG provides one testbed for our DisCAS. On-going work on DisCAS include design refinements, an efficient grid task manager, reliability and performance test.

References

1. I. Foster, C. Kesselman, S. Tuecke, *The Anatomy of the Grid: Enabling Scalable Virtual Organization*, International J. Supercomputer Applications, 15(3), (2001)
2. I. Foster, C. Kesselman, *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*, J. Nick, S. Tuecke, (2002)
3. I. Foster, C. Kesselman, *Globus: A Metacomputing Infrastructure Toolkit*, International J. Supercomputer Application, (1997), 11(2), 115-128
4. LIAO, W. and WANG, P. S. *Building IAMC: A Layered Approach*, Proc. International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'00), (2000), 1509-1516.
5. LIAO, W. and WANG, P. S. *Specification of OMEI: Open Mathematical Engine Interface*, ICM Technical Report, (2001) http://icm.mcs.kent.edu/reports/index.html.
6. LIAO W., LIN D. and WANG P. S. *OMEI: Open Mathematical Engine Interface*, Proceedings of ASCM'2001, pp 83-91, Matsuyama, Japan, (2001)
7. LIN D., LIU J. and LIU Z. *Mathematical Research Software: ELIMINO.* Proceedings of ASCM'98. Lanzhou Univ., China, (1998), 107-116
8. WANG, P. S. *Design and Protocol for Internet Accessible Mathematical Computation.* In Proc. ISSAC'99, ACM Press, (1999), 291-298.
9. WU, W. T. *Basic Principle of Mechanical Theorem Proving in Elementary Geometries,* J. Syst. Sci. Math. Sci. 4, (1984), 207-235
10. WU, Y., LIAO, W., LIN, D., WANG, P. S., *Local and Remote User Interface for ELIMINO through OMEI.* Proceedings of International Congress on Mathematical Software (ICMS 2002). World Scientific Press. Aug. (2002)
11. Graham, S., Simeonov, S., Boubez, T, Daniels, G., Davis, D., Nakamura, Y. and Neyama, R. *Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI.* Sams, (2001)
12. W Gropp, E. Lusk, *User's Guide for MPICH, a Portable implementation of mpI*, Argonne National Laboratory, University of Chicago,(1996)
13. Baraglia, R., Laforenza, D., Lagana, A., *A Web-based Metacomputing Problem-Solving Environment for Complex Applications*, Proceedings of Grid Computing 2000, (2000), 111–122
14. Tuecke, S., Czajkowski, K., Foster, I. , et.al.: *Open Grid Services Infrastructure (OGSI) Version 1.0*, Global Grid Forum Draft Recommendation. (2003).
15. Casanova H., Donfarra, J., *NetSolve: A network Server for Solving Computational Science Problems*, Intl. Journal of Supercomputing Application and High Performance Computing, **11(3)** (1998)
16. Baratloo, A., Karaul, M., *Charlotte: Metacomputing on the Web, Special Issue on Metacomputing*, Future Generation Computer Systems, (2001) 559-570.
17. Guangwen Yang, Yongwei Wu, Qing Wang, Weiming Zheng, *GridPPI: Task-level Parallel Programming Interface for Grid Computing*, Accepted by International Journal of Grid and Utility Computing, (2003)
18. Dazheng Huang, Fei Xie, Guangwen Yang, *T.G.: a Market-oriented Computing System with Fine-grained Parallelism*, 9th Workshop on Job Scheduling Strategies for Parallel Processing Seattle, Washington, (2002)

A Mathematica Package for Solving and Displaying Inequalities

R. Ipanaqué[1] and A. Iglesias[2⋆]

[1] Mathematics Department, National University of Piura, Perú
robertchero@hotmail.com
[2] Department of Applied Mathematics and Computational Sciences, University of Cantabria, Avda. de los Castros, s/n, E-39005, Santander, Spain
iglesias@unican.es
http://personales.unican.es/iglesias

Abstract. Solving inequalities is a very important topic in computational algebra. In fact, the most important computer algebra systems include sophisticated tools for solving different kinds of inequalities in both symbolic and graphical ways. This paper presents a new Mathematica package, `InequationPlot`, for displaying the two-dimensional solution sets of several inequalities. In particular, it extends Mathematica's capabilities by providing graphical solutions to many inequalities (such as those involving trigonometric, exponential and logarithmic functions) that cannot be solved by using the standard Mathematica commands and packages. The package also deals with inequalities involving complex variables by displaying the corresponding solutions on the complex plane. The output obtained is consistent with Mathematica's notation and results and the package provides a similar output for those cases already solved by Mathematica. The performance of the package is discussed by means of several illustrative and interesting examples.

1 Introduction

Solving inequalities is a very important topic in Mathematics, with outstanding applications in many problems of theoretical and applied science. Inequalities play a key role simply because many problems cannot be completely and accurately described by only using equalities. However, since there is not a general methodolody for solving inequalities, their symbolic computation is still a challenging problem in computational algebra. Depending on the kind of the functions involved, there are many "specialized" methods such as those based on cylindrical algebraic decomposition, Gröebner basis, quantifier elimination, etc. In addition, some "ad hoc" methods are to be implemented. The interested reader is referred to [1,3,4,5,7] for an introduction to the field. See also [8] for a recent survey on geometric inequalities.

In spite of the difficulties mentioned above, the most important computer algebra systems (Mathematica, Maple, etc.) include sophisticated tools for solving

⋆ Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 303–310, 2004.
© Springer-Verlag Berlin Heidelberg 2004

different kinds of inequalities in both symbolic and graphical ways. However, the problem is far from being solved in all its generality and, consequently, there is still a long way to walk in order to tackle this issue.

This paper presents a new Mathematica package, `InequationPlot`, for displaying the two-dimensional solution sets of of several inequalities. In particular, it extends Mathematica's capabilities by providing graphical solutions to many inequalities (such as those involving trigonometric, exponential and logarithmic functions) that cannot be solved by using the standard Mathematica commands and packages. The package also deals with inequalities involving complex variables by displaying the corresponding solutions on the complex plane. The output obtained is consistent with Mathematica's notation and results and the package provides a similar output for those cases already solved by Mathematica. The performance of the package is discussed throughout the paper by means of several illustrative and interesting examples.

The structure of this paper is as follows: Section 2 describes the main standard Mathematica tools for solving inequalities. Then, Section 3 introduces the new Mathematica package, `InequationPlot`, and describes the commands implemented within. The performance of the package is also discussed in this section by using some illustrative examples. In addition, we show some inequalities that are solvable with our package while they cannot be solved by applying the standard Mathematica kernel and packages. Finally, Section 4 closes with the main conclusions of this paper and some further remarks.

2 Standard Mathematica Tools for Solving Inequalities

Mathematica incorporates a number of sophisticated commands and packages for solving different kinds of inequalities. For example, the Mathematica command `Reduce[`*eqns,vars, elims*`]` simplifies equations trying to eliminate the variables *elims*. In this process, the command generates both equations and inequalities combined with the operators `&&` (and) and `||` (or). In addition, Mathematica 3.0 includes a package, `Algebra‘AlgebraicInequalities‘`, for solving systems of strong polynomial inequalities [9]. In particular, the command `SemialgebraicComponents[`$\{e_1, ..., e_n\}, \{x_1, ..., x_m\}$`]` gives at least one point in each connected component of the open semialgebraic set defined by inequalities $\{e_1, ..., e_n\}$, where both sides of each e_i are polynomials in variables $\{x_1, ..., x_m\}$ with rational coefficients. This command is based on the cylindrical algebraic decomposition algorithm (see, for instance, [2] and [6] for a description). In addition, the Mathematica 3.0 package `Algebra‘InequalitySolve‘` provides the solution set of an expression containing logical connectives and univariate polynomial equations and inequalities in a single or several variables. Other recent commands to deal with this problem incorporated in Mathematica version 4.0 are described in [10]. That paper also discusses the main algorithms used in this task.

The visualization of the solution sets of given systems of inequalities has also been the subject of further research. For example, the add-on Mathematica

4.1 package `Graphics‘InequalityGraphics‘` provides commands to visualize logical combinations of polynomial and algebraic inequalities in two- and three-dimensional real spaces (commands `InequalityPlot` and `InequalityPlot3D` respectively). Some additional developments to find real solutions of systems of real equations and inequalities have also been recently incorporated into Mathematica version 5.0. For example, the command `FindInstance` can be used to find a real solution of a system of equations and inequalities. Even users can directly apply the `CylindricalDecomposition` command to compute solutions of certain kinds of inequalities.

However, there are several families of inequalities that cannot be solved by using the standard Mathematica commands described above. This limitation can easily be understood by analyzing the methods applied to solve those inequalities, mostly based on the cylindrical algebraic decomposition (CAD). In this decomposition, each $S \subset \mathbb{R}^n$ is represented as a finite union of disjoint cells. The concept of cell applied here can be defined as follows:

1. a cell in $\mathbb{R}$ is either a point or an open interval
2. a cell in $\mathbb{R}^{n+1}$ is either defined by the set

$$\{(x, y)/x \in C, f(x) < y < g(x)\} \tag{1}$$

or the set

$$\{(x, y)/x \in C, y = f(x)\} \tag{2}$$

where $x = (x_1, x_2, \ldots, x_n)$, C is a cell in $\mathbb{R}^n$ and f and g are either
a) continuous functions on C such that for some polynomials F and G we have $F(x, f(x)) = 0$ and $G(x, g(x)) = 0$ or
b) $\pm\infty$
and $f(x) < g(x)$, $\forall x \in C$.

Given a finite set F of polynomials in n variables, the CAD algorithm calculates an F-invariant[1] cylindrical algebraic decomposition of $\mathbb{R}^n$. This algorithm is particularly well suited for obtaining the solution set of a system of polynomial equations and inequalities for n variables. For example:

```
In[1]:= CylindricalDecomposition[x^2 + y^2 +z^2 < 1,{x,y,z}]
```

$$Out[1] := \{-1 < x < 1 \ \&\& \ -\sqrt{1-x^2} < y < \sqrt{1-x^2} \ \&\& \\ -\sqrt{1-x^2-y^2} < z < \sqrt{1-x^2-y^2}\}$$

Unfortunately, as will be shown later, inequalities involving trigonometric, exponential or logarithmic functions cannot be decomposed in cells and, therefore, the CAD method can no longer be applied to those cases. The package `InequationPlot`, described in the next section, overcomes this limitation and allows the user to solve a large family of real and complex inequality systems and equations and display their two-dimensional solution sets.

[1] A cylindrical algebraic decomposition of a set $S \subset \mathbb{R}^n$ is said to be *F-invariant* if each of the polynomials from F has a constant sign on each cell of that decomposition.

3 The Package `InequationPlot`: Some Illustrative Examples

Inequalities involving trigonometric functions cannot be solved by applying the CAD algorithm described in Section 2. For example, let us try to display the solution sets of each of the inequalities

$$sin(x+y) > \frac{1}{2} \tag{3}$$

and

$$sin(2x) + cos(3y) < 1 \tag{4}$$

on the set $[-8,8] \times [-8,8]$ by using the standard Mathematica commands. In this case, we must use the command `InequalityPlot` of the Mathematica 4.1 package:

```
In[2]:=<<Graphics`InequalityGraphics`
```

Unfortunately, since the region defined by inequality (3) on the prescribed domain cannot be broken down into cylinders, the CAD algorithm fails to give the solution:

```
In[3]:=InequalityPlot[Sin[x + y] > 1/2, {x, -8, 8}, {y, -8, 8}]
```

Out[3] := *InequalityPlot* :: *region* :
The region defined by $sin(x+y) > 1/2 \bigwedge -8 <= x <= 8 \bigwedge -8 <= y <= 8$ *could not be broken down into cylinders.*

The previous inequalities can be solved, however, by loading the package developed by the authors:

```
In[4]:=<<InequationPlot`
```

which includes the command

`InequationPlot[`$ineqs, \{x, xmin, xmax\}, \{y, ymin, ymax\}, opts$`]`

for displaying the two-dimensional region of the set of points satisfying the inequalities *ineqs* of real numbers inside the square $[xmin, xmax] \times [ymin, ymax]$, which is actually the problem we are going to solve. For example, inequalities (3)-(4) can be solved as follows:

```
In[5]:=InequationPlot[#, {x, -8, 8}, {y, -8, 8}, AspectRatio ->
         Automatic]& /@ {Sin[x+y]>1/2, Sin[2 x]+Cos[3 y]<1}
```

Out[5]:= See Figure 1

Similarly, Fig. 2 displays the solution sets of the inequalities $F(x) + F(y) = 1$ and $F(x^2) + F(y^2) = 1$ (where F stands for the floor function) on the squares $[-4,4] \times [-4,4]$ and $[-2,2] \times [-2,2]$, respectively. We would like to remark that the Mathematica command `InequalityPlot` does not provide any solution for these inequalities either.

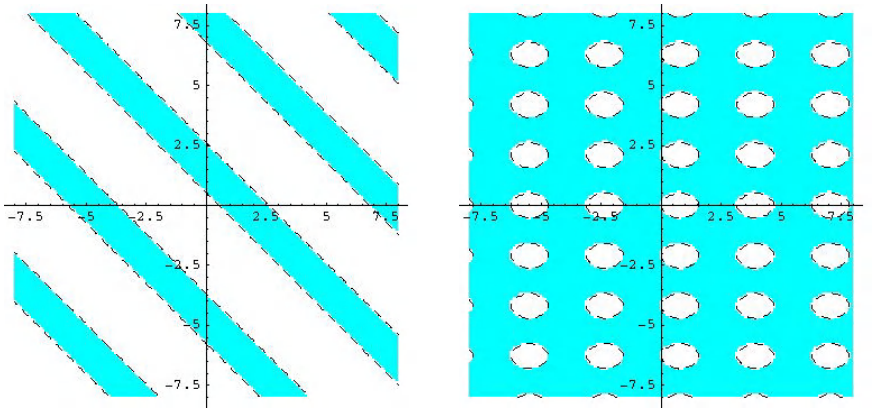


Fig. 1. Some examples of inequality solutions on the square $[-8, 8] \times [-8, 8]$: (left) $sin(x + y) > \frac{1}{2}$; (right) $sin(2x) + cos(3y) < 1$

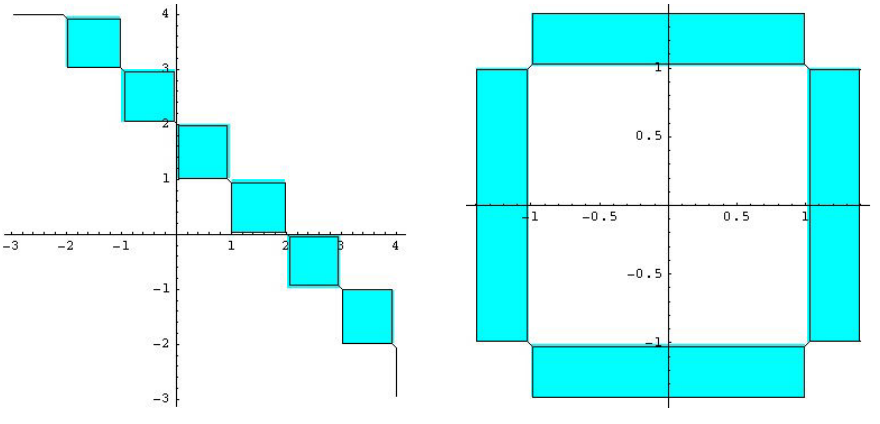


Fig. 2. Some examples of inequality solutions: (left) $floor(x) + floor(y) = 1$ on the square $[-4, 4] \times [-4, 4]$; (right) $floor(x^2) + floor(y^2) = 1$ on the square $[-2, 2] \times [-2, 2]$

The previous command, `InequationPlot`, can be generalized to inequalities involving complex numbers. The new command

`ComplexInequationPlot[`$ineqs, \{z, \{Rezmin, Rezmax\}, \{Imzmin, Imzmax\}\}$, $opts$`]`

displays the solution sets of the inequalities *ineqs* of complex numbers inside the square in the complex plane given by $[Rezmin, Rezmax] \times [Imzmin, Imzmax]$.

In this case, the functions appearing within the inequalities need to be real-valued functions of a complex argument, e.g. *Abs*, *Re* and *Im*. For example:

```
In[6]:=ComplexInequationPlot[#,{z,{-2, 3},{-3, 3}}, AspectRatio->
    Automatic]& /@ {1<Abs[z^2-z+1]<4, 1<Abs[z^2-2z]/Abs[z^2+3]<4}
```

Out[6]:= See Figure 3

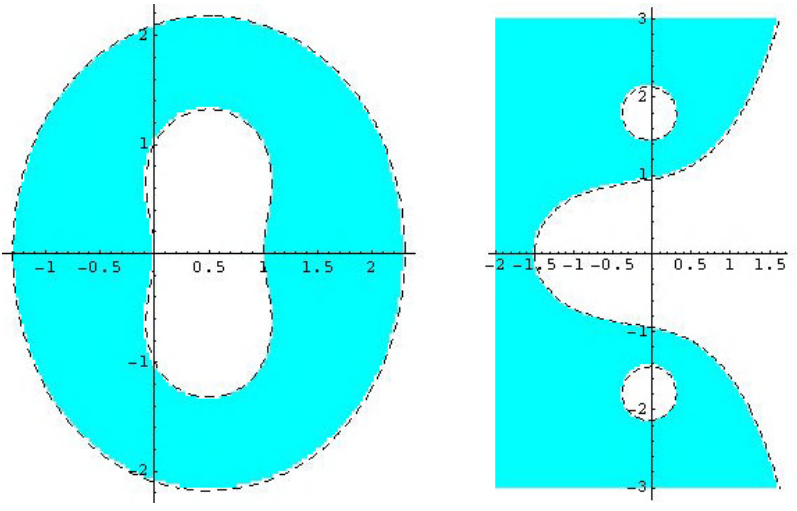


Fig. 3. Some examples of inequality solutions for $z \in \mathbb{C}$ such that $Re(z) \in [-2, 3]$ and $Im(z) \in [-3, 3]$: (left) $1 < ||z^2 - z + 1|| < 4$; (right) $1 < \dfrac{||z^2 - 2z||}{||z^2 + 3||} < 4$

We would like to point out that the output obtained from the package is consistent with Mathematica's notation and results and the package provides a similar output for those cases already solved by Mathematica. Figure 4 illustrates this fact: on the left, the solutions provided by the standard Mathematica command `InequalityPlot` to the inequalities $x^2 + y^2 \leq 8 \bigwedge y^2 \leq 2x$ (top) and $x^2 - y^2 \geq 1 \bigwedge (x^2 - 1)^3 y^2 \leq 1$ (bottom) on $[0, 3] \times [-3, 3]$ and $[-3, 3] \times [-2, 2]$ respectively are displayed. On the right, the solutions obtained by using the new command `InequationPlot` are shown. As the reader can see, the new solution sets match perfectly those obtained from the standard Mathematica commands.

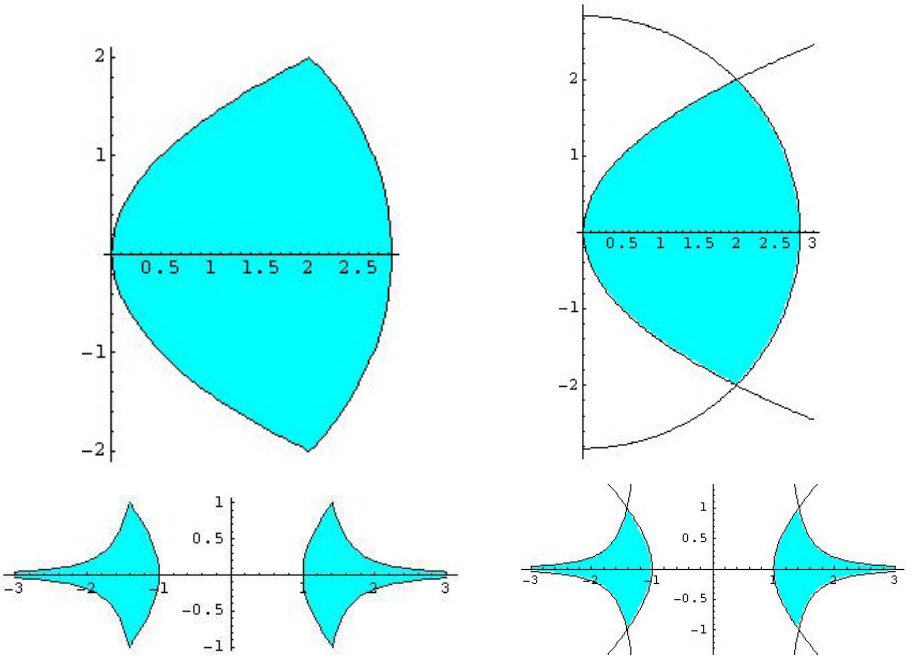


Fig. 4. Solution sets for the inequality systems: (top) $x^2 + y^2 \leq 8 \bigwedge y^2 \leq 2x$; (bottom) $x^2 - y^2 \geq 1 \bigwedge (x^2 - 1)^3 y^2 \leq 1$. The solutions have been obtained by applying: (left) the standard Mathematica command `InequalityPlot`; (right) the new command `InequationPlot`.

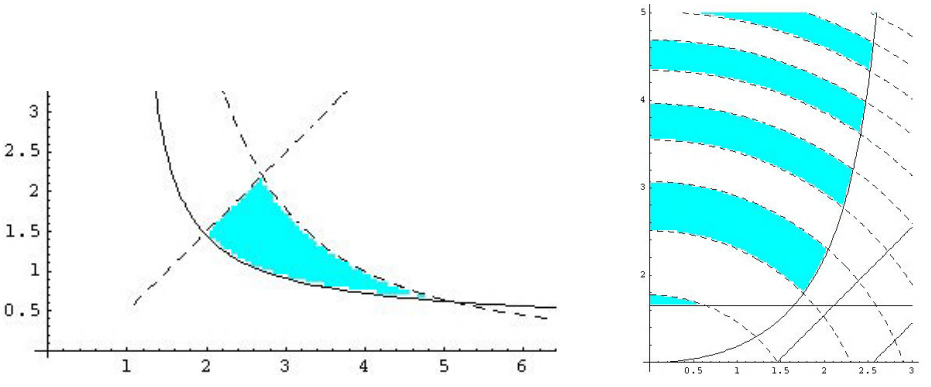


Fig. 5. (left, right) Solution sets for the inequality systems given by Eqns. (5) and (6) respectively

The last example aims to show how complicated the inequality systems can be: in addition to include exponential, logarithmic and trigonometric functions, combinations and even compositions of these (and other) functions can also be considered. In Figure 5 the solutions sets of the inequality systems:

$$e^y \geq 1 \bigwedge log(x)y \geq 1 \bigwedge x\sqrt{y} < 4 \bigwedge x - y > \frac{1}{2} \tag{5}$$

$$log(y) \geq \frac{1}{2} \bigwedge sin(x)y \geq x \bigwedge cos(e^{x-y}) \geq 0 \bigwedge sin(x^2 + y^2) > 0 \tag{6}$$

on $[1, 10] \times [0, 10]$ and $[0, 3] \times [1, 5]$ respectively are displayed.

4 Conclusions and Further Remarks

In this paper a new Mathematica package, `InequationPlot`, to solve real and complex inequalities and display their associated two-dimensional solution sets is introduced. The package extends Mathematica's capabilities by providing graphical solutions to many inequalities (such as those involving trigonometric, exponential and logarithmic functions) that cannot be solved by using the standard Mathematica commands and packages. The performance of the package has been illustrated by means of several interesting examples. In all the cases, the output obtained is consistent with Mathematica's notation and results and the package provides a similar output for those cases already solved by Mathematica. Perhaps the weakest feature of the package is the computation time, which is substantially higher than that usually required by Mathematica for solving the common cases. Further work will be developed in order to decrease this computation time and also to extend our results to the three-dimensional case. The obtained results will be reported elsewhere.

References

1. Beckenbach, E.F., Bellman, R.E.: An Introduction to Inequalities. Random House, New York (1961)
2. Brown, C.W.: Simple cylindrical algebraic decomposition construction and its applications. Journal of Symbolic Computation, **31** (2001) 521-547
3. Caviness, B.F., Johnson, J.R.: Quantifier Elimination and Cylindrical Algebraic Decomposition. Springer-Verlag, New York (1998)
4. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities (Second Edition). Cambridge University Press, Cambridge (1952)
5. Kazarinoff, N.D.: Geometric Inequalities. Random House, New York (1961)
6. McCallum, S.: Solving polynomial strict inequalities using cylindrical algebraic decomposition. The Computer Journal, **36**(5) (1993) 432-438
7. Mitrinovic, D.S.: Analytic Inequalities. Springer-Verlag, New York (1970)
8. Mitrinovic, D.S., Pecaric, J.E., Volenec, V.: Recent Advances in Geometric Inequalities. Kluwer, Dordrecht (1989)
9. Strzebonski, A.: An algorithm for systems of strong polynomial inequalities. The Mathematica Journal, **4**(4) (1994) 74-77
10. Strzebonski, A.: Solving algebraic inequalities. The Mathematica Journal, **7** (2000) 525-541

Choleski-Banachiewicz Approach to Systems with Non-positive Definite Matrices with *MATHEMATICA*®

Ryszard A. Walentyński

Silesian University of Technology, Faculty of Civil Engineering
ul. Akademicka 5, PL44-101 Gliwice, Poland
rwal@kateko.bud.polsl.gliwice.pl

Abstract. The paper presents the discussion on extension of potential application of the Choleski-Banachiewicz algorithm to the system of linear equations with non-positive definite matrices. It is shown that the method is also stable in case of systems with negative definite matrices and sometimes may be also successful if the matrix is neither positive nor negative definite. The algorithm handles systems with complex symmetric (not Hermitian) matrices, too. This fact has deep physical sense and engineering applications since systems with negative definite matrices are common in tasks of dynamics and post buckling analysis in civil and mechanical engineering. Possibility of utilization of Choleski-Banachiewicz algorithm to such problems can be very practical. The entire analysis has been carried out within *MATHEMATICA*® environment.

1 Introduction

A linear system of equations with a symmetrical matrix can be efficiently solved using the so-called Cholesky decomposition, Weisstein [11]. This routine was not invented by Cholesky (Cholewski, Choleski [8]). The method was presented for the first time by Dolittle [5]. His work was forgotten and the algorithm was reinvented by several mathematicians including Cholesky whose contribution was published by Benoit [2] and by Banachiewicz [1]. Due to many inventors the algorithm has a lot of names: Cholewski m., Cholesky-Croot m., Croot m., Dolittle-Choleski m. and so on. In my country — Poland the routine is called Choleski-Banachiewicz method. However, the procedure can be called "a square root m." to avoid problems with personal names. A brief discussion on the method including its historical context has been presented by Janczura [6].

It is believed that the Choleski-Banachiewicz method is limited to the positive definite matrices, Hermitian in case of complex matrix and symmetrical in case of real matrix, Weisstein [11]. The aim of the paper is to show that this opinion is true as far as it refers to symmetry.

MATHEMATICA® version 5, which is an advanced computer algebra system, has a built-in procedure called `CholeskyDecomposition[]`, Wolfram [12, 14]. This routine is limited to Hermitian and positive definite matrices. In other cases it fails.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 311–318, 2004.
© Springer-Verlag Berlin Heidelberg 2004

I built my own procedure called **CholeskiBanachiewicz[]** within *MATHEMATICA*® environment. This routine not only decomposes the matrix, but also solves the system of linear equations. Due to its unique features it is used to solve linear systems of equations within Refined Least Squares Method, [10]. Testing the Choleski-Banachiewicz procedure I found that it is not limited to the system with positive definite matrices. It was already mentioned in [9]. Using symbolic computation tools I will try to explain the problem a bit wider in this contribution. It can be shown that the presented procedure works well not only with systems with symmetric non-positive real matrices but also with symmetric complex matrices (symmetric complex matrix is not Hermitian).

The described feature of the Choleski-Banachiewicz method can have important application in computational mechanics and probably in other branches of physics. Therefore a physical sense of the problem and its engineering utilization are discussed.

2 Algorithm and Its Implementation

The algorithm is described in many books and therefore will not be repeated here. The routine has been built according to the recipe given by Burden and Faires [4]. Implementation of the algorithm is straightforward within *MATHEMATICA*®. The code is as long as a procedure. This is one of the advantages of this symbolic computation system.

```
In[1]:= CholeskiBanachiewicz[aa_List, bb_List, prec_] :=
          Module[{a = aa, b = bb, i, j, n = Length[bb]},
            Do[{a[[i, i]] = Sqrt[N[a[[i, i]], prec] - Sum[a[[i, k]]^2, {k, 1, i-1}]],
              b[[i]] = (b[[i]] - Sum[a[[i, j]] b[[j]], {j, 1, i-1}])/a[[i, i]],
              Do[a[[j, i]] = (a[[j, i]] - Sum[a[[j, k]] a[[i, k]], {k, 1, i-1}])/a[[i, i]],
                {j, i + 1, n}],
                  Do[a[[j, i]] = 0, {j, i - 1}]}, {i, n}];
            Do[b[[i]] = (b[[i]] - Sum[a[[j, i]] b[[j]], {j, i+1, n}])/a[[i, i]], {i, n, 1, -1}];
            {a, b}]
```

The arguments of the procedure are: matrix $\boldsymbol{A}$, vector $\boldsymbol{b}$ and desired working precision. If two first arguments are exact numbers, symbolic or have precision higher than machine precision – the computations can be done with precision up to the infinity. It is important for ill-conditioned tasks. The output of the routine are decomposed matrix $\boldsymbol{L}$ and the vector $\boldsymbol{x}$.

3 Dealing with Non-positive Matrices

As it was mentioned in the introduction, we will show that the algorithm successfully deals with not positive definite matrices. We will discuss two examples of the systems with real and complex matrices.

3.1 Real Matrix

The presented example, for the sake of conciseness, is limited to system of matrices 4×4. One can evaluate it with bigger dimensions of matrices.

```
In[2]:= dim = 4;
```

Let us build an arbitrary matrix $\boldsymbol{m}$ with random elements.

```
In[3]:= m = Table[Random[Integer, {-100, 100}], {i, dim}, {j, dim}];
```

On this basis we can produce a symmetrical matrix $\boldsymbol{m}_1$.

```
In[4]:= MatrixForm[m1 = m + Transpose[m]]
```

$$\text{Out[4]=}\begin{pmatrix} 96 & -2 & -80 & 58 \\ -2 & -142 & -40 & 66 \\ -80 & -40 & 162 & 29 \\ 58 & 66 & 29 & 22 \end{pmatrix}$$

If we try to decompose the matrix $\boldsymbol{m}_1$ with the built-in procedure:

```
In[5]:= CholeskyDecomposition[m1]
```

the attempt finishes with a message announcing: *The matrix is not sufficiently positive definite to complete the Cholesky decomposition to reasonable accuracy*. It is especially odd and conservative behavior since the matrix is exact (integer). This behavior is probably caused by the fact that the matrix is non-positive definite since some of its eigenvalues are negative numbers. Nevertheless the pitfall should be not connected with accuracy.

```
In[6]:= eig = Sort[Eigenvalues[N[m1]]]
Out[6]= {-176.436, -4.60622, 99.5497, 219.493}
```

Let us build a new matrix function $\boldsymbol{m}_2$ of the parameter λ.

```
In[7]:= m2[λ_] := m1 - IdentityMatrix[dim] λ
```

The parameter λ modifies elements of the matrix diagonal. It is easy to check that the matrix $\boldsymbol{m}_2$ is positive definite for any λ smaller than minimal eigenvalue of the original matrix $\boldsymbol{m}_1$:

```
In[8]:= Sort[Eigenvalues[N[m2[-400]]]]
Out[8]= {223.564, 395.394, 499.55, 619.493}
```

and negative definite for any value of the parameter λ grater than maximal eigenvalue of the matrix $\boldsymbol{m}_1$.

```
In[9]:= Sort[Eigenvalues[N[m2[400]]]]
Out[9]= {-576.436, -404.606, -300.45, -180.507}
```

Let us create a vector $\boldsymbol{b}$, which will be a vector of free elements of the system of linear equations.

```
In[10]:= b = Table[Random[Integer, {-100, 100}], {i, dim}]
Out[10]= {16, -59, -62, 5}
```

The l_2-norm of this vector $\|\boldsymbol{b}\|_2$ is equal to:

```
In[11]:= normb2 = Norm[b, 2]
Out[11]= √7606
```

Since both matrix $\boldsymbol{m}_1$ and vector $\boldsymbol{b}$ are exact numbers we can solve the system exactly, to enforce it the third argument that represents working precision is set to ∞. This command returns the decomposed matrix $\boldsymbol{L}$, only. Solution of the system is omitted by setting the argument **[[1]]** in the command.

```
In[12]:= MatrixForm[CholeskiBanachiewicz[m1, b, ∞][[1]]]
```

$$\text{Out[12]}= \begin{pmatrix} 4\sqrt{6} & 0 & 0 & 0 \\ -\frac{1}{2\sqrt{6}} & \frac{1}{2}\,i\sqrt{\frac{3409}{6}} & 0 & 0 \\ -10\sqrt{\frac{2}{3}} & 250\,i\sqrt{\frac{2}{10227}} & \sqrt{\frac{366658}{3409}} & 0 \\ \frac{29}{2\sqrt{6}} & -\frac{1613\,i}{2\sqrt{20454}} & \frac{196421}{\sqrt{1249937122}} & i\sqrt{\frac{4439473}{366658}} \end{pmatrix}$$

The decomposed matrix is complex. It is not a problem for contemporary computer algebra systems and numerical systems like MATLAB since all of them can handle complex numbers.

In spite of the fact that the input matrix is non-positive definite one we can check that $\boldsymbol{L} \cdot \boldsymbol{L}^T \equiv \boldsymbol{m}_1$.

```
In[13]:= %.Transpose[%] === m1
Out[13]= True
```

We can also test if really *the matrix is not sufficiently positive definite to complete the Cholesky decomposition to reasonable accuracy*. Let us decompose the matrix $\boldsymbol{m}_1$ with PC machine precision that is equal to 16.

```
In[14]:= MatrixForm[CholeskiBanachiewicz[m1, b, 16][[1]]]
```

$$\text{Out[14]}= \begin{pmatrix} 9.79796 & 0 & 0 & 0 \\ -0.204124 & 11.9181\,i & 0 & 0 \\ -8.16497 & 3.49608\,i & 10.3709 & 0 \\ 5.9196 & -5.63917\,i & 5.55576 & 3.47965\,i \end{pmatrix}$$

This test shows that the round-off error is not very big.

```
In[15]:= MatrixForm[%.Transpose[%] - m1]
```

$$\text{Out[15]=}\begin{pmatrix} 0.\times10^{-14} & -0.\times10^{-16} & -0.\times10^{-15} & 0.\times10^{-15} \\ -0.\times10^{-16} & -0.\times10^{-14} & 0.\times10^{-15} & -0.\times10^{-15} \\ -0.\times10^{-15} & 0.\times10^{-15} & -0.\times10^{-14} & 0.\times10^{-14} \\ 0.\times10^{-15} & -0.\times10^{-15} & 0.\times10^{-14} & 0.\times10^{-14} \end{pmatrix}$$

The actual aim of my Choleski-Banachiewicz procedure is not the decomposition of the matrix itself but solution of the system.

The function $e(\lambda)$ finds a normalized relative error of the system solution with Choleski-Banachiewicz method of the system $\boldsymbol{m}_2(\lambda)\cdot\boldsymbol{x} == \boldsymbol{b}$:

$$e(\lambda) = \frac{\|\boldsymbol{m}_2(\lambda)\cdot\boldsymbol{x} - \boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2}, \tag{1}$$

what is implemented with:

```
In[16]:= normchol[x_] :=
  Norm[m2[x].Chop[CholeskiBanachiewicz[m2[x], b, 16][[2]]] - b]
                                    normb2
```

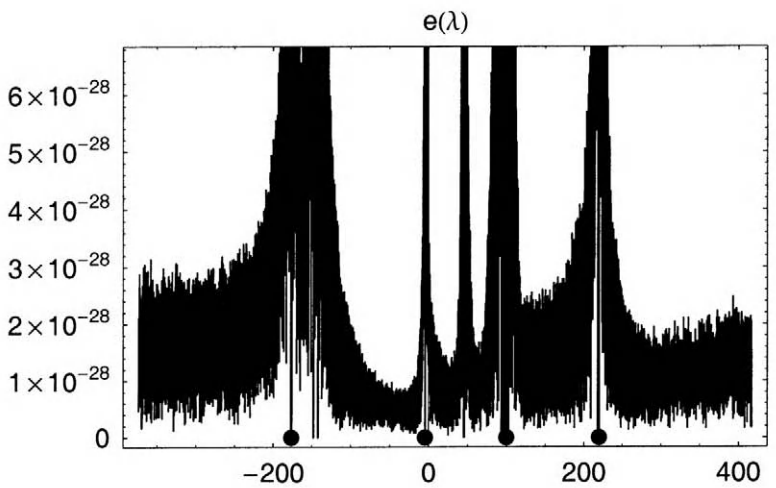


Fig. 1. The relative error of the solution with Choleski-Banachiewicz method $e(\lambda)$ with regard to the parameter λ – description in text

Figure 1 shows the diagram of the function $e(\lambda)$ in a lot of points from the interval $\lambda \in \langle -374, 417\rangle$. Four big points in the bottom of the diagram represents position of the $\boldsymbol{m}_1$ matrix eigenvalues. It can be found that the precision of the solution becomes poor near these points. The matrix $\boldsymbol{m}_2(\lambda)$ for λ equal to eigenvalues of $\boldsymbol{m}_1$ becomes singular and in their neighborhoods badly conditioned.

3.2 Complex Matrix

Let us consider a symmetric complex matrix $\boldsymbol{A}$:

$$\boldsymbol{A} := \begin{pmatrix} 1 & -2+i & -i \\ -2+i & 19 & -3 \\ -i & -3 & 3 \end{pmatrix}. \tag{2}$$

Attempt of use of built-in function **CholeskyDecomposition[]** fails with a message: *The matrix is not Hermitian or real and symmetric.*

Let us attempt to solve a system with this matrix with **CholeskiBanachiewicz[]** procedure.

```
In[17]:= MatrixForm[
           (sol = CholeskiBanachiewicz[{{1, -2 + i, -i}, {-2 + i, 19, -3}, {-i, -3, 3}},
             {1, 1, 1}, ∞]) [[2]]]
```

The result is saved in variable **sol** and the output presents the decomposed matrix $\boldsymbol{L}$:

$$\text{Out[17]=} \begin{pmatrix} 1 & 0 & 0 \\ -2+i & 2\sqrt{4+i} & 0 \\ -i & -\dfrac{2+i}{\sqrt{4+i}} & \sqrt{\dfrac{52}{17} - \dfrac{13\,i}{17}} \end{pmatrix}$$

It can be checked that $\boldsymbol{L} \cdot \boldsymbol{L}^T \equiv \boldsymbol{A}$.

```
In[18]:= MatrixForm[%.Transpose[%]]
Out[18]= {{1, -2 + i, -i}, {-2 + i, 19, -3}, {-i, -3, 3}}
```

The solution with **CholeskiBanachiewicz[]** procedure is exact since both matrix $\boldsymbol{A}$ and free vector are exact numbers and working precision has been set to ∞.

```
In[19]:= sol[[2]]
Out[19]= {15/13 + 4 i/13, 7/26 + i/26, 1/2 + 11 i/26}
```

Let us check if numerical computations with machine precision will finish with satisfactory precision. The result of Choleski-Banachiewicz procedure:

```
In[20]:= CholeskiBanachiewicz[{{1, -2 + i, -i}, {-2 + i, 19, -3}, {-i, -3, 3}},
           {1, 1, 1}, 16][[2]]
Out[20]= {1.15385 + 0.307692 i, 0.269231 + 0.0384615 i, 0.5 + 0.423077 i}
```

One can easily check that a built-in function **LinearSolve[]** returns the same result.

4 Physical Sense and Engineering Applications

Let us consider a mechanical, dynamical, discrete system described with the following equation of motion, Langer [7].

$$\boldsymbol{B} \cdot \ddot{\boldsymbol{q}} + \boldsymbol{C} \cdot \dot{\boldsymbol{q}} + \boldsymbol{K} \cdot \boldsymbol{q} = \boldsymbol{P}, \tag{3}$$

where $\boldsymbol{B}$ is a matrix of inertia, $\boldsymbol{C}$ is a matrix of viscotic dumping, $\boldsymbol{K}$ is a stiffness matrix, $\boldsymbol{P}$ is a vector of external load and $\boldsymbol{q}$ is a vector of generalized Lagrangian displacements.

If the system is loaded with stationary periodic load $\boldsymbol{P} := \boldsymbol{P}_s \sin(2\pi\, pt)$, where p is a frequency of the load, $[p] = Hz$, and dumping is negligible and we can set $\boldsymbol{C} := 0$ in the equation of motion (3) we can derive the following system of linear algebraic equations.

$$\left(\boldsymbol{K} - 4\pi^2 p^2 \boldsymbol{B}\right)\boldsymbol{q}_s = \boldsymbol{P}_s . \tag{4}$$

The matrix $\boldsymbol{K}$ is positive definite but the matrix $\left(\boldsymbol{K} - 4\pi^2 p^2 \boldsymbol{B}\right)$ becomes non-positive definite if frequency p becomes bigger than the first frequency of free vibration of the system.

There are two rational engineering approach to design of the dynamical systems, both of them assume that the frequency of external load to be far from the resonance (eigenvalues) zone. The first approach consists in such design of the dynamical system that the free vibration frequencies become higher than frequency of excitations. It results in very stiff and heavy (expensive) structures. Another approach consists in design of very flexible structure with frequencies of free vibrations much smaller than the frequency of excitation. The disadvantage of such approach is that the dynamical system has to go through "resonance zone" during acceleration. Such phenomenon can be observed in many facilities like domestic washing machines, computer processor fans and turbogenerators in power stations. This problem may be overcome by decreasing the time of acceleration during turning on and off. On the other hand the mathematical model of the dynamical system excited with frequencies close to free vibrations has to consider dumping. In this case equation (4) is not valid. Some more complicated formulas have to be used, Langer [7]. A shape of upper envelope of the diagram of the normalized error in Figure 1 has some analogy to the diagram of the dynamical coefficient used in engineering dynamical analysis.

Systems with negative definite matrices can be stable not only in tasks of civil engineering; example is Bulatović [3]. It seems to be good news that such systems can be effectively solved with Choleski-Banachiewicz method.

5 Conclusions

Choleski-Banachiewicz procedure can be applied not only to positive definite matrices but also to negative definite ones and sometimes to matrices which are neither positive nor positive definite. It has been also shown that the procedure can be successfully employed to systems with symmetrical complex matrices.

Possible application of the method to negative definite matrices can be very useful in analysis of dynamical systems in civil and mechanical engineering. Another field can be post-buckling behavior of structures.

The analysis was carried out thanks to the computer algebra system *MATHEMATICA*®. Despite of conservative behavior of the built-in procedure, the system enabled me to challenge the problem and extend the potential of the procedure.

References

1. Banachiewicz, T.: Methode de Resolution Numerique des Equations Lineaires, Du Calcul des Determinants et des Inverses et de Reduction des Formes Quadratiques. Bull. Ac. Pol. Sci. A. (1938) 393–404
2. Benoit, A.: Note Surface ton Methode, Proce de Du Commendent Cholesky. Bull. Geod. Toulouse (1924) 67–77
3. Bulatović, R.M.: On Stability Criteria for Gyroscopic Systems with Negative Definite Stiffness. Facta Universitatis, series Mechanica, Automatic Control and Robotics **2**(10) (2000) 1081–1087 http://facta.junis.ni.ac.yu/facta/macar/macar2000/macar2000-07.pdf
4. Burden, R.L., Faires, J.D.: Numerical Analysis. 5th edn. PWS Publishing Company, Boston (1993)
5. Dolittle, M.H.: Method Employed in the Solution of Normal Equations and the Adjustment of Transquation. U. S. Survey Report (1878) 115–120
6. Janczura, A.: Solving of Linear Equations with Bounded Right-Hand Side. Mechanika i Komputer, Warsaw **8** (1990) 235–250
7. Langer, J.: Dynamics of Bar Structures. In Structural Mechanics with Elements of Computer Approach, volume 2, chapter 7. Arkady, Warsaw (1984) 76–156
8. Pattison, D.H.: Summary: Pronunciation of 'Cholesky'. S-News Mailing List Archives Division of Biostatistics at Washington University School of Medicine, Saint Louis (2000) http://www.biostat.wustl.edu/archives/html/s-news/2000-10/msg00172.html
9. Walentyński, R.A.: Refined Least Squares Method for Shells Boundary-Value Problems. In: Tazawa, Y. et al. (eds.): Symbolic Computation: New Horizons. Proceedings of the Fourth International *MATHEMATICA*® Symposium. Tokyo Denki University Press, Tokyo (2001) 511–518 (extended version on the proceedings CD-ROM)
10. Walentyński, R.A.: Application of Computer Algebra in Symbolic Computations and Boundary-Value Problems of the Theory of Shells. Silesian University of Technology Press, Gliwice, Civil Engineering **1587**(100) (2003)
11. Weisstein, E.: Cholesky Decomposition. In: World of Mathematics, a Wolfram Web Resource. CRC Press LLC, Wolfram Research, Inc., Boca Raton Champaign (2003) http://mathworld.wolfram.com/CholeskyDecomposition.html
12. Wolfram Research: CholeskyDecomposition. In: Built-in Functions, chapter Lists and Matrices - Matrix Operations. Wolfram Media, Champaign (2003) http://documents.wolfram.com/v5/Built-inFunctions/ListsAndMatrices/MatrixOperations/CholeskyDecomposition.html
13. Wolfram Research: MatrixManipulation. In: Standard Add-on Packages, chapter Linear Algebra. Wolfram Media, Champaign (2003) http://documents.wolfram.com/v5/Add-onsLinks/ StandardPackages/LinearAlgebra/MatrixManipulation.html
14. Wolfram, S.: Advanced Matrix Operations. In: The *MATHEMATICA*® Book – Online, chapter 3.7.10. 5th edn. Wolfram Media, Champaign (2003) http://documents.wolfram.com/v5/TheMathematicaBook/AdvancedMathematicsInMathematica/LinearAlgebra/3.7.10.html

A Closed Form Solution of the Run-Time of a Sliding Bead along a Freely Hanging Slinky

Haiduke Sarafian

The Pennsylvania State University
York, PA 17403, USA
has2@psu.edu

Abstract. The author has applied Lagrangian formalism to explore the kinematics of a bead sliding along a frictionless, freely hanging vertical Slinky. For instance, we derived a closed analytic equation for the run-time of the bead as a function of the traversed coil number. We have applied *Mathematica* to animate the 3-dimensional motion of the bead. The derived run-time is incorporated within the animation to clock the bead's actual motion. With the help of *Mathematica* we have solved the inverse run-time equation and have expressed the traversed coil number as a function of the run-time. The latter is applied to further the analysis of the problem conducive to analytic time-dependent equations for the bead's vertical position, its falling speed and its falling acceleration, and its angular velocity about the symmetry axis of the Slinky. It is also justified that a Slinky is a device capable of converting the gravitational potential energy of a sliding bead into pure rotational energy.

1 Introduction

A Slinky is a massive, soft spring-like object and has curious dynamic and static features. By hanging masses to a freely, vertically suspended Slinky and setting it in motion, the authors of [1] have investigated some of its dynamic features. The Slinky's static characteristics have been studied in [2]. In both references, analytically, it is proven how the Slinky's own weight contributes to the uneven spacing of the adjacent coils along a vertically hung Slinky. The proven equations match the intuitive expectations – the spacing between the adjacent coils for the coils closer to the support are wider than the ones at the bottom. Furthermore, the lower coils are less slanted verses the ones closer to the support. In fact, the bottom coils are almost horizontal.

To incorporate these known characteristic features and to broaden the scope of the Slinky's related issues, we studied a kinematic problem. We considered the effects of the uneven spacing of the Slinky's coils to calculate the run-time of a sliding bead under the gravity pull. We have shown, although the calculation of the run-time of a sliding bead along a theoretical, massless evenly-spaced vertically hung spring is trivial, it is not so for a real Slinky. We were able to solve the Slinky problem exactly and derived an analytic closed form equation to express the run-time of a sliding bead as a function of the traversed coil number.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 319–326, 2004.
© Springer-Verlag Berlin Heidelberg 2004

We have applied *Mathematica* [3] to animate the 3-dimensional motion of a bead. We considered a typical Slinky and have applied its geometrical lengths to clock the run-time of a sliding bead. The numeric values of the run-time were embedded in the 3-d animation so that one can visually correspond the actual movement of the bead to its run-time and the traversed coil number. Because of the length limitation of the article the *Mathematica* code is not included, the oral presentation will feature the animation.
To further our analysis with the help of *Mathematica* we have solved the inverse run-time equation, we expressed the traversed coil number as a function of the run-time. The detail of the procedure and the specific use of *Mathematica* in achieving this goal is given in section 4. The latter is applied to further the analysis of the problem conducive to analytic time-dependent equations for the bead's vertical position, its falling speed and its falling acceleration, and its angular velocity about the symmetry axis of the Slinky. For comprehensive understanding, the derived equations are plotted v.s. time.

2 The Physics and the Analysis of the Problem

We denote the number of Slinky's coils by N, the radius of the circular coils by R, its un-stretched axial length when laid on a level table by L_0 and its stretched length when freely suspended vertically by L. Figure 1 depicts one such Slinky. We have applied *Mathematica's* ParamtericPlot3D command to display the Slinky. The origin of a right-handed Cartesian coordinate system is set at the bottom of the Slinky with the z-axis pointing to the top of the page. The first coil, $n = 0$, is at the $z = 0$ and the top coil, $N = 35$, is at the support. The height of the individual coil is measured from the bottom, the data is shown in Fig. 2. The size of the data points indicates the accuracy of the measurements along the vertical axis.

According to [1,2] the height of the nth coil is given by; $z_n = an + bn^2$ with $a = \frac{L_0}{N}$ and $b = \frac{L-L_0}{N^2}$. In Fig. 2, z_n, is shown by the solid line – it perfectly fits the data.

The position vector of a bead in the aforementioned Cartesian coordinate system is $\boldsymbol{r} = \{x(t), y(t), z(t)\}$ and can be written as $\boldsymbol{r} = \{R\cos(2\pi n^{'}), R\sin(2\pi n'), z_{n'}\}$ where $n^{'}$, is the number of the traversed coils and equals $n^{'} = N - n$. The kinetic energy $T = \frac{1}{2}m\nu^2$ and the potential energy $V = mgz_{n'}$ of a freely released bead of mass m in terms of the traversed coil number $n^{'}$ are:

$$T = \frac{1}{2}m\dot{n^{'}}^2\{c^2 + [a + 2b(N - n^{'})]^2\} \tag{1}$$

$$V = mg[a(N - n^{'}) + b(N - n^{'})^2] \tag{2}$$

where $\dot{n^{'}} = \frac{d}{dt}n^{'}$, $c = 2\pi R$ and g is the gravity.

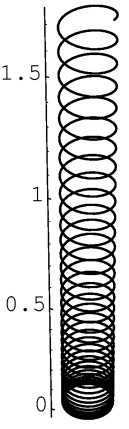


Fig. 1. The display of a typical Slinky with specs of $N = 35$, $R = 5.0cm$, $L_0 = 7.0cm$ and $L = 172.0cm$

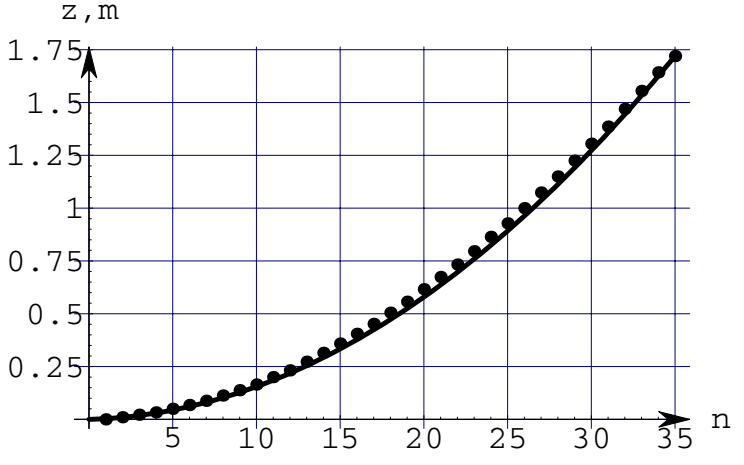


Fig. 2. The dots are the data and the solid line is $z_n = an + bn^2$

The Euler-Lagrange equation, $\frac{d}{dt}\frac{\partial L}{\partial \dot{n'}} = \frac{\partial L}{\partial n'}$ for Lagrangian $L = T - V$ with T and V subject to (1) and (2) is,

$$\ddot{n'}\{c^2 + [a + 2b(N - n')]^2\} - 2b\dot{n'}^2[a + 2b(N - n')] - g[a + 2b(N - n')] = 0. \quad (3)$$

To solve (3) we introduce $\xi = a + 2b(N - n')$. In terms of ξ, (3) becomes,

$$\ddot{\xi}(c^2 + \xi^2) + \dot{\xi}^2\xi + \nu^2\xi = 0 \quad (4)$$

Here, $\nu^2 = 2bg$ and assumes the speedsquared dimension. To solve (4) we set, $\eta = \dot{\xi}^2$. In terms of η and ξ, (4) yields

$$\frac{\dot{\eta}}{2\dot{\xi}}(c^2 + \xi^2) + \xi(\eta + \nu^2) = 0 \tag{5}$$

By separating the variables of (5) and integrating both sides of the resulting equation with appropriate limits we arrive at

$$\int_0^{\eta} \frac{d\eta}{\eta + \nu^2} = -\int_{\xi_0}^{\xi} \frac{2\xi d\xi}{c^2 + \xi^2} \tag{6}$$

here, $\xi_0 = a + 2bN$. The integration of (6) yields,

$$\eta = \nu^2 \frac{{\xi_0}^2 - \xi^2}{c^2 + \xi^2} \tag{7}$$

In (7) we replace η with $(\frac{d\xi}{dt})^2$, and by rearranging the terms and integrating the result, we arrive at

$$t = \frac{1}{\nu}\int_{\xi_0}^{\xi} \sqrt{\frac{c^2 + \xi^2}{{\xi_0}^2 - \xi^2}}\, d\xi \tag{8}$$

this yields

$$t = \frac{1}{\nu}\sqrt{c^2 + {\xi_0}^2}\, E(\arccos(\frac{a + 2bn}{\xi_0}), \frac{\xi_0}{\sqrt{c^2 + {\xi_0}^2}}) \tag{9}$$

$E(\delta, r)$ is the Elliptic integral of the second kind [4]. Equation (9)is the run-time; i.e. it is the time a bead starting from the top coil takes to traverse to the *nth* coil.

We have noticed that (8) is a convoluted analog of the kinematics of one dimensional uniform motion. I.e. (8) can be viewed as $\ell = \nu t$, with ℓ being the Slinky's characteristic length given by

$$\ell = \int_{\xi_0}^{\xi} w(c, \xi_0, \xi) d\xi \tag{10}$$

In other words, the bead slides along the characteristic length, ℓ, given by the weighted differential length $d\xi$. The weight, w, is defined by

$$w(c, \xi_0, \xi) = \sqrt{\frac{c^2 + \xi^2}{{\xi_0}^2 - \xi^2}} \tag{11}$$

As mentioned earlier, the bead's characteristic speed ν is $\nu = \sqrt{2gb}$. Intuitively, one would expect the run-time of a bead, for a skinny Slinky to be the same as the run-time of a freely falling bead released at the same height. To verify this, in (9) we set $R = 0$, this yields,

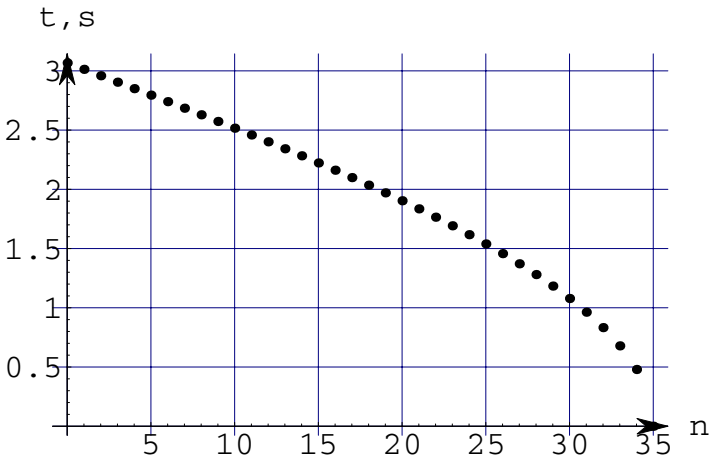


Fig. 3. The run-time of a bead v.s. the coil number n. The specs of the Slinky are the ones used in Fig. 1

$$t = \frac{1}{\nu}\xi_0 E(\arccos(\frac{a+2bn}{\xi_0}), 1) \tag{12}$$

On the other hand, since. $L_0 \ll L$, (12) simplifies further,

$$t \simeq \sqrt{\frac{2L}{g}} E(\arccos(\frac{n}{N}), 1) \tag{13}$$

For $n = 0$, (13) yields the run-time of a bead traversing the entire length of the stretched Slinky, L. The E in (13) for $n = 0$ yields, $E(\frac{\pi}{2}, 1) = 1$ and yields the shortest run-time,

$$t \simeq \sqrt{\frac{2L}{g}} \tag{14}$$

Equation (14) is identical to the time of flight of a free falling object, $L = \frac{1}{2}gt^2$, released at height L. Equation (14) for $L = 172.0cm$, the length of a hanging Slinky gives $t = 0.592s$ and matches the numeric value of (9) for small values of R, e.g. $R = 0.005cm$. It is instructive to display the run-time t, given by (9) v.s. the coil number n. For the Slinky on hand, this is shown in Fig 3.

3 Corollary Topics of Interest

It is curious to find out at any given time how high the bead is from the bottom of the Slinky, how fast the bead is falling and its falling acceleration. To address

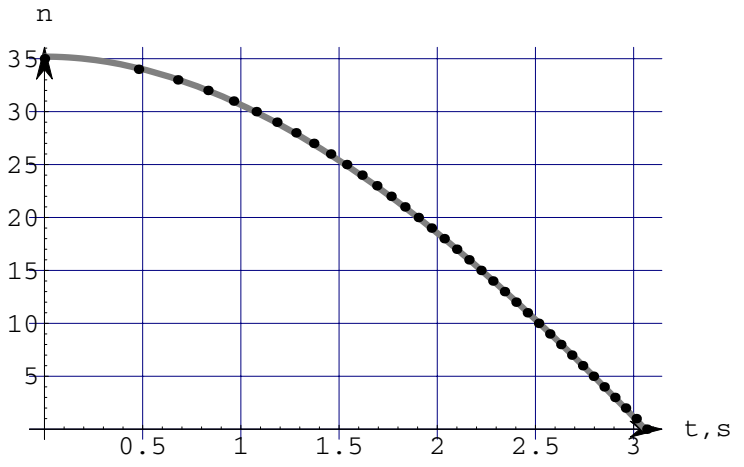


Fig. 4. The plots of the coil number n v.s. t. The solid line is the fitted polynomial

these questions, there is a need to solve (9) for t, this is problematic. However, we pursued the goal graphically. First, by interchanging the horizontal and vertical axes in Fig. 3, we re-plot the data points. We then apply *Mathematica's* Fit command to fit the data with a suitable polynomial. Figure 4 displays the output.

For the Slinky on hand the fitted function is an incomplete third order polynomial with the given fitted coefficients $n(t) = 35.190 - 4.925t^2 + 0.381t^3$. By substituting $n(t)$ in z_n we evaluate the vertical height of the bead, $z(t)$, its falling speed, $\dot{z}(t)$ and its falling acceleration $\ddot{z}(t)$. These quantities are plotted in Fig. 5.

The ordinate of Fig. 5, is calibrated in MKS units, and hence, m, $\frac{m}{s}$, $\frac{m}{s^2}$ are to be used to read the height, the velocity and the acceleration.
According to Fig. 5, the bead starts off with an initial zero velocity and of about $1\frac{m}{s^2}$ acceleration. In $1.6s$, it reaches its maximum, $0.83\frac{m}{s}$ velocity and acquires zero acceleration. In according to Fig. 3, $t = 1.6s$, corresponds to $n = 24$, that is the bead reaches its maximum velocity after traversing 11 coils. From this point on, it takes another $1.4s$ to decelerate to the bottom.
We also noted, a Slinky is a device capable of converting the gravitational potential energy to a pure rotational energy. This is because the bottom coils are horizontal and the bead upon reaching the bottom is to circulate about the Slinky's symmetry axis with no further fall.
We quantify our claim by applying the conservation of energy principle to the two ends of the Slinky; the potential energy at the top and the rotational kinetic energy at the bottom. That is $(PE)^i = (KE)^f_{rot}$, this gives $mgL = \frac{1}{2}I\omega^2_{max}$, here, $I = mR^2$ is the moment of inertia of the bead about the Slink's symmetry axis and ω_{max} is its maximum angular velocity. Solving this equation for ω_{max} yields $\omega_{max} = \frac{1}{R}\sqrt{2gL}$. For the specs of the Slinky on hand this yields

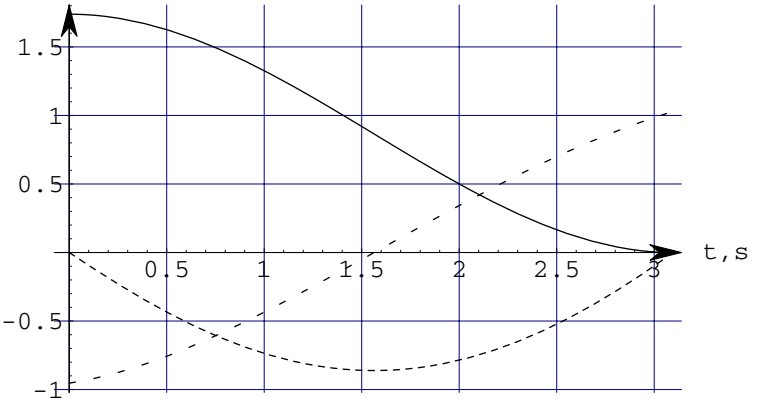


Fig. 5. The plot of $z(t)$(solid-line),the falling velocity,$\dot{z}(t)$(short dashed-line) and the falling acceleration $\ddot{z}(t)$ (long dashed-line)

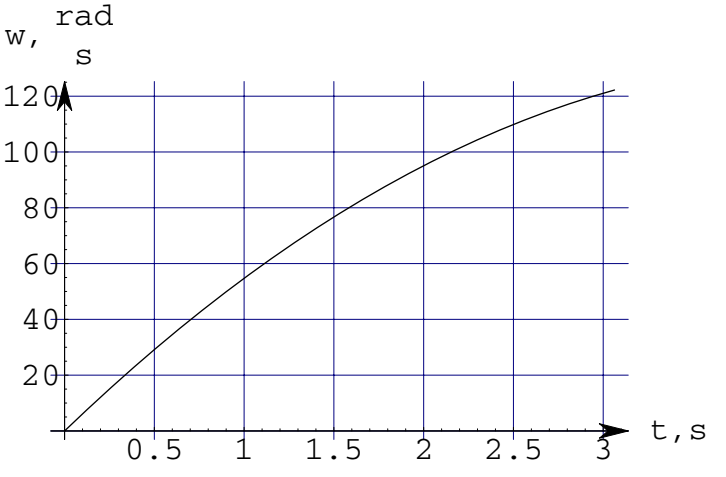


Fig. 6. Plot of angular velocity, ω, v.s. t. The Slinky's specs are the ones used in Fig. 1

$\omega_{max} = 116\frac{rad}{s}$. On the other hand, the angular velocity of the bead is $\omega(t) = \frac{d}{dt}\{2\pi[N - n(t)]\}$. The quantity in the braces is the traversed azimuthal angle about the symmetry axis of the Slinky and $n(t)$ is the aforementioned fitted polynomial.
Figure 6 displays $\omega(t)$ v.s. time. The maximum value of $\omega_{max}(t)$ at the end of the run is $120\frac{rad}{s}$, this is in good agreement with the predicted, $116\frac{rad}{s}$.

References

1. Thomas C. Heard, Neal D. Newby, Jr. : Behavior of a soft spring. Am. J. Phys. **45** (1977) 1102–1106
2. French, A. P. : The Suspended Slinky - A Problem in Static Equilibrium. The Physics Teacher **32**(1994) 244–245
3. Wolfram, S. The *Mathematica* book. New 4th edn. Cambridge Press (1999).
4. Gradshteyn, I. S., Ryzhik, I. M. : Table of Integrals, Series and Products, 2nd edn. Academic Press, p.276 (1980)

Analytical Theory of Motion of a Mars Orbiter

J.F. San Juan[1], S. Serrano[2], and A. Abad[2]

[1] Universidad de La Rioja, 26004 Logroño.
[2] Universidad de Zaragoza, 50009 Zaragoza, Spain.

Abstract. The design of spatial missions to Mars requires the development of analytical theories in order to put artificial satellites in orbit around Mars.
In this paper, we present a complete third order analytical model of a satellite perturbed by the zonal $J_2, \ldots, J_6$ harmonics of the Mars potential. Two Lie transformations, the elimination of the Parallax and the elimination of the Perigee, and the Krylov–Bogoliubov–Mitropolsky method are applied to obtain a complete integration of the model. The algebraic expressions of the generators, the Hamiltonians and the integrals, together with a software code to compute the ephemeris of the satellite, are automatically obtained using our computer algebra system ATESAT.

1 Introduction

Mars is one of the main challenges for the aerospace community. Since 1965, when the Mariner IV made the first successful flyby of Mars, man has wanted to reach the planet. The study of the Martian topography, gravity surface composition, atmospheric structure, etc., depends on the orbital missions around Mars. Besides, the possibility of future manned missions to Mars forces to select brief parking orbits with a precise analytical determination of the position of the satellite [9].

The interest of the Centre National D'Etudes Spatiales (CNES, France) in this kind of orbiters, together with the experience of the Space Mechanics Group (GME) of the University of Zaragoza in obtaining analytical theories of Earth artificial satellites by using symbolic computational tools, drove us to collaborate in order to obtain an analytical theory of a Martian artificial satellite with an error less than one kilometer in one month. To this aim, we used the same methods as those applied to Earth satellites.

The Lie–Deprit method [7] based on Lie canonical transformations was used to develop efficient analytical theories of the satellite problem. To obtain our theory we applied two Lie transformations: the elimination of the Parallax and the elimination of the Perigee. After these eliminations, the simplified Hamiltonian was reduced to one degree of freedom. To obtain a theory without singularities in eccentricity and inclination and without series expansions in eccentricity, we applied the Krylov-Bogoliubov-Mitropolsky (KBM) method instead of the classical Delaunay normalization. Both, the Lie-Deprit method and the KBM technique are very well suited methods for symbolic computation.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 327–334, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In general, during the computing of an analytical theory in the artificial satellite problem one has to cope with long series expansion, handling a huge amount of terms and particular objects like the so-called Poisson series[1]. This makes general purpose algebraic manipulators ineffective and forces to create specific computer algebra systems to handle these type of objects in an efficient way. Our software ATESAT ([2,3,13]) uses the Poisson Series Processor PSPC [1,14] to generate automatically analytical theories and their corresponding ephemeris programs for satellite orbits.

With ATESAT we developed the complete third order theory of the motion of an artificial satellite perturbed by the first sixth order zonal harmonics of the potential of Mars. With this theory we obtained the required precision; in fact the error for a wide range of initial conditions is less than 400 meters per month.

2 Dynamical Model

The Hamiltonian formalism that describes the orbit of an artificial satellite around Mars is written in Whittaker's or polar-nodal variables[2] as

$$\mathcal{H} = \sum_{n\geq 0} \frac{\epsilon^n}{n!} \mathcal{H}_n, \tag{1}$$

with

$$\begin{aligned} \mathcal{H}_0 &= \frac{1}{2}\left(R^2 + \frac{\Theta^2}{r^2}\right) - \frac{\mu}{r}, & \mathcal{H}_1 &= \frac{\mu}{r}\left(\frac{\alpha}{r}\right)^2 P_2(s_i \sin\theta), \\ \mathcal{H}_2 &= 2\,\frac{\mu}{r}\sum_{n\geq 3}^{6}\left(\frac{\alpha}{r}\right)^n \frac{J_n}{J_2^2}\, P_n(s_i \sin\theta), & \mathcal{H}_k &= 0, \quad k \geq 3, \end{aligned} \tag{2}$$

where P_n is the Legendre polynomial of degree n, μ is the gravitational constant, α is the equatorial radius, J_n are the zonal harmonic coefficients, $s_i = \sin i = \sqrt{1-(N^2/\Theta^2)}$, is a function of the momenta N and Θ, and the small parameter ϵ is the Martian constant of oblateness.

The variable ν is cyclic in the zonal problem, and therefore we have a dynamical system with two degrees of freedom.

[1] A Poisson series is a multivariate Fourier series, whose coefficients are multivariate Laurent series

$$\sum_{i_0,\ldots,i_{n-1},j_0,\ldots,j_{m-1}} C_{i_0,\ldots,i_{n-1}}^{j_0,\ldots,j_{m-1}} x_0^{i_0}\ldots x_{n-1}^{i_{n-1}} \begin{pmatrix}\sin\\ \cos\end{pmatrix}(j_0 y_0 + \ldots + j_{m-1} y_{m-1}).$$

[2] $(r, \theta, \nu, R, \Theta, N)$, where r is the radial distance from the planet's mass center to the satellite, θ is the argument of latitude, and ν is the argument of the ascending node. The variables R, Θ and N are the conjugate momenta to the coordinates r, θ and ν, respectively.

3 Canonical Lie Transformations

A Lie transformation [7,10,11] can be defined as an infinitesimal contact transformation $\varphi : (\boldsymbol{y}, \boldsymbol{Y}, \epsilon) \to (\boldsymbol{x}, \boldsymbol{X})$, such that $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{Y}; \epsilon), \boldsymbol{X}(\boldsymbol{y}, \boldsymbol{Y}; \epsilon)$ satisfy the differential equations

$$\frac{d\boldsymbol{x}}{d\epsilon} = \nabla_{\boldsymbol{X}} W(\boldsymbol{x}, \boldsymbol{X}, \epsilon), \quad \frac{d\boldsymbol{X}}{d\epsilon} = -\nabla_{\boldsymbol{x}} W(\boldsymbol{x}, \boldsymbol{X}, \epsilon), \tag{3}$$

with the initial conditions $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{Y}, \epsilon = 0) = \boldsymbol{y}, \boldsymbol{X}(\boldsymbol{y}, \boldsymbol{Y}, \epsilon = 0) = \boldsymbol{Y}$, where $\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Y} \in \mathbb{R}^m$, and $W = \sum_{i \geq 0} (\epsilon^i / i!)\, W_{i+1}(\boldsymbol{x}, \boldsymbol{X})$ is the generator of the transformation.

In the particular case of the Hamiltonian formalism, a Lie transformation converts a Hamiltonian $\mathcal{H}(\boldsymbol{x}, \boldsymbol{X}; \epsilon) = \sum_{i \geq 0} (\epsilon^i / i!) \mathcal{H}_{i,0}(\boldsymbol{x}, \boldsymbol{X})$ into a new one $\mathcal{K}(\boldsymbol{y}, \boldsymbol{Y}; \epsilon) = \sum_{i \geq 0} (\epsilon^i / i!) \mathcal{H}_{0,i}(\boldsymbol{y}, \boldsymbol{Y})$ by means of the relations

$$\mathcal{H}_{p,q} = \mathcal{H}_{p+1,q-1} + \sum_{k=0}^{p} \binom{p}{k} \left(\mathcal{H}_{p-k,q-1}; W_{k+1}\right), \tag{4}$$

where $(_;_)$ stands for the Poisson bracket. Equation (4) is called *Lie triangle*. Note that with this approach, there appear many intermediate terms $\mathcal{H}_{p,q}$ which must be computed and stored.

The Lie-Deprit method [7] tries to find the generator of a Lie transformation to turn the Hamiltonian into a new one satisfying some prefixed conditions.

This method looks for the generator order by order. Once the order $(n-1)$ is solved we know the expressions of W_i, $i = 0, \ldots, n-1$ and $\mathcal{H}_{p,q}$, $p + q \leq n - 1$. By joining the equations in (4) for $p + q = n$, we find the *homological equation*

$$\mathcal{L}_0 W_n = (\mathcal{H}_0, W_n) = \mathcal{H}_{0,n} - \tilde{\mathcal{H}}_{n,0}, \tag{5}$$

where $\tilde{\mathcal{H}}_{n,0}$ can be computed from the previous orders by means of the Lie triangle. To solve the order n, we must follow three steps:

1. Compute $\tilde{\mathcal{H}}_{n,0}$ from the expressions of order $(n-1)$.
2. Choose $\mathcal{H}_{0,n}$.
3. Find an integral W_n of the homological equation.

3.1 Elimination of the Parallax

The first step of this theory looks for a Lie transformation called elimination of the Parallax. This canonical transformation [8] reduces the complexity, not the number of degrees of freedom, of the Hamiltonian (1). Besides, the elimination of the Parallax algorithm allows us to compute the expression in a close form of the eccentricity, and therefore to obtain general purpose theories valid for any kind of elliptic eccentricity.

The homological equation (5) is expressed, in Whittaker variables, as follow

$$\mathcal{L}_0 \mathcal{W}_n = R \frac{\partial \mathcal{W}_n}{\partial r} - \left(\frac{\mu}{r^2} - \frac{\Theta^2}{r^3}\right) \frac{\partial \mathcal{W}_n}{\partial R} + \frac{\Theta}{r^2} \frac{\partial \mathcal{W}_n}{\partial \theta} = \mathcal{H}_{0,n} - \tilde{\mathcal{H}}_{n,0}. \tag{6}$$

Instead of looking for an integral of the previous equation, we will apply the property

$$\mathcal{L}_0\left(\sum_{j\geq 0}\left(C_j\sin j\theta+S_j\cos j\theta\right)\right)=\frac{\Theta}{r^2}\sum_{j\geq 1} j\left(C_j\cos j\theta-S_j\sin j\theta\right),$$

valid for any function of the algebra $\mathcal{F}=\{F=\sum_{j\geq 0}\left(C_j\cos j\,\theta+S_j\sin j\,\theta\right),\ C_j,S_j\in\ker(\mathcal{L}_0)\}$. Particularly, using the $C=e\cos g$, $S=e\sin g$ and $p=a(1-e^2)$ functions of the orbital elements a, e and g, the Hamiltonian (1) can be expressed as a function of $\mathcal{F}$ if we take into account the relations $1/r=1/p+C/p\cos\theta+S/p\sin\theta$, $R=C\Theta/p\sin\theta-S\Theta/p\cos\theta$, since $C,S,p,\Theta\in\ker(\mathcal{L}_0)$.

After applying the previous change of variables, the equation (6) becomes

$$\mathcal{L}_0\left(\mathcal{W}_n\right)+\frac{\Theta}{r^2}C_0=\frac{\Theta}{r^2}F$$

and then we apply the three steps of the method described in the previous section

1. Computing $\quad F=(r^2/\Theta)\,\tilde{\mathcal{H}}_{n,0}=\sum_{j\geq 0}\left(C_j\cos j\theta+S_j\sin j\theta\right).$
2. Choosing $\quad \mathcal{H}_{0,n}=(\Theta/r^2)\,C_0.$
3. Finding $\quad \mathcal{W}_n=\sum_{j\geq 1}\left[(C_j/j)\sin j\theta-(S_j/j)\cos j\theta\right].$

Note that after computing $\tilde{\mathcal{H}}_{n,0}$ by means of the Lie triangle, obtaining $\mathcal{H}_{0,n}$ and $\mathcal{W}_n$ becomes a simple symbolic exercise of coefficients reordering.

After the Parallax elimination, the transformed Hamiltonian and the generator of third order, expressed as a Poisson Series, have 176 and 1837 terms, respectively. The number of terms of the direct and inverse expression of this transformation are given in the following table

Parallax	r	θ	ν	R	Θ	N
Direct	2056	5000	4738	1988	3057	1
Inverse	2052	4981	4610	1989	3034	1

3.2 Elimination of the Perigee

Apparently, the elimination of the Parallax removes the argument of latitude θ, however, this variable appears implicitly in the state functions C and S, which are functions of the perigee g. In order to reduce the problem, we can eliminate the perigee by applying a new Lie transformation whose generator has two terms

$$\mathcal{W}_n=\mathcal{W}_n^*(C,S,\Theta,\theta)+\tilde{\mathcal{W}}_n(C,S,\Theta,_),$$

$\mathcal{W}_n^*$, which depends on θ, and $\tilde{\mathcal{W}}_n$, which does not depend on θ. The Lie operator $\mathcal{L}$ applied to this generator

$$\mathcal{L}_0\mathcal{W}_n=R\frac{\partial\mathcal{W}_n}{\partial r}-\left(\frac{\mu}{r^2}-\frac{\Theta}{r^3}\right)\frac{\partial\mathcal{W}_n}{\partial R}+\frac{\Theta}{r^2}\frac{\partial\mathcal{W}_n}{\partial\theta}=\frac{\Theta}{r^2}\frac{\partial\mathcal{W}_n^*}{\partial\theta},$$

only depends on $\mathcal{W}_n^*$.

With such a property, the solution $\mathcal{W}_n^* = \int r^2(\tilde{\mathcal{H}}_{n,0} - \mathcal{H}_{0,n})/\Theta\, d\theta$ of the homological equation only gives one part of the expression of the generator.

The homological equation (5) is obtained by detaching from the n-th line of the Lie triangle the known terms of order $n-1$ of the unknown terms. Then, there are more unknown terms depending on $\tilde{\mathcal{W}}_{n-1}$, and the homological equation becomes $\mathcal{L}_0\mathcal{W}_n^* = \tilde{\mathcal{H}}_{n,0} + \mathcal{H}_{0,n} + n(\mathcal{H}_{1,0}; \tilde{\mathcal{W}}_{n-1})$.

By computing $\tilde{\mathcal{H}}_{n,0}$ and splitting it into the part $\tilde{\mathcal{H}}_{n,0}^{\theta}$ that depends on θ and the part $\tilde{\mathcal{H}}_{n,0}^{*}$ that does not depends on θ, we can choose the new Hamiltonian of order $n > 1$ as the expression

$$\mathcal{H}_{0,n} = \langle \tilde{\mathcal{H}}_{n,0}^{*} \rangle_g = \frac{1}{2\pi} \int_0^{2\pi} \tilde{\mathcal{H}}_{n,0}^{*}(C, S)\, dg,$$

that does not depends on g.

Eventually, taking into account the expression

$$(\mathcal{H}_{1,0}; \tilde{\mathcal{W}}_{n-1}) = \mathcal{F}_1(\theta) - \frac{3\mu\alpha^2}{2\Theta^3 r^2}(4 - 5s_i^2)\frac{\partial \tilde{\mathcal{W}}_{n-1}}{\partial g},$$

we can use it to obtain $\tilde{\mathcal{W}}_{n-1}$

$$\tilde{\mathcal{W}}_{n-1} = \frac{2\Theta^3 r^2}{3n\mu\alpha^2} \frac{1}{(4 - 5s_i^2)} \int \left(\mathcal{H}_{0,n} - \langle \tilde{\mathcal{H}}_{n,0}^{*} \rangle_g \right) dg,$$

and $\mathcal{W}_n^*$

$$\mathcal{W}_n^* = \int \frac{r^2}{\Theta} \left(\tilde{\mathcal{H}}_{n,0}^{\theta} + n\mathcal{F}_1(\theta) \right) d\theta.$$

This algorithm called elimination of the Perigee [5] is applied to remove the argument of the perigee g from the perturbation. This elimination is not a normalization in the sense that the transformed Hamiltonian does not belong to the kernel of the Lie derivative $\mathcal{L}_0$ associated with $\mathcal{H}_0$, rather it reduces by one the number of degrees of freedom.

The Hamiltonian after the two previous transformations can be expressed as

$$\mathcal{H} = \frac{1}{2}\left(R^2 + \frac{\Theta^2}{r^2} \right) - \frac{\mu}{r} + \sum_{n \geq 1} \frac{\epsilon^n}{n!} \mathcal{M}_n^{i,j} \frac{R^i}{r^j}, \tag{7}$$

where $\mathcal{M}_n^{i,j}$ are functions of the constants α and μ and the momenta Θ and N. This transformed Hamiltonian has 317 terms, and the generator has 4521 terms. The number of terms of the direct and inverse expression of this transformation, which is considerably greater than in the previous transformation, is given in the following table

Perigee	r	θ	ν	R	Θ	N
Direct	83410	155376	81893	80156	66322	1
Inverse	83294	154769	81435	79915	66322	1

4 The Satellite as a Perturbed Harmonic Oscillator

The variables θ and ν are cyclic in the Hamiltonian (7), then the momenta Θ and N are constant and we will consider only the two first Hamilton's equations

$$\frac{dr}{dt} = \frac{\partial \mathcal{H}}{\partial R}, \quad \frac{dR}{dt} = -\frac{\partial \mathcal{H}}{\partial r}, \tag{8}$$

in order to find the temporal evolution of r and R. After finding the solution of previous differential equations, the problem lies in the computation of two quadratures to obtain θ and ν.

By differentiating again the first equation in (8), combining both to obtain a second order differential equation, and changing the variables r and dr/dt and the time t for two new variables u and v, and a new time s defined by

$$u = \frac{1}{r} - \frac{\mu}{\Theta^2}, \qquad r^2 \frac{ds}{dt} = \Theta, \qquad v = \frac{du}{ds}, \tag{9}$$

we obtain the equation of a perturbed harmonic oscillator

$$\frac{d^2u}{ds^2} + u = \sum_{n\geq 1} \frac{\epsilon^n}{n!} \mathcal{K}_n^{i,j} u^i v^j, \tag{10}$$

where $\mathcal{K}_n^{i,j}$ are polynomials in the constants μ, Θ and $\mathcal{M}_m^{i,j}$.

4.1 The Krylov–Bogoliubov-Mitropolski (KBM) Method

The solution of the differential equation (10) of the perturbed harmonic oscillator can be written as an asymptotic expansion $u = \sum_{n\geq 0} (\epsilon^n/n!)\, u_n(\delta, f)$, where $u_0 = \delta \cos f$, $u_n(\delta, f)$ are 2π–periodic functions in f, and the variation of δ and f with respect to the time is given by

$$\frac{d\delta}{ds} = \sum_{n\geq 0} \frac{\epsilon^n}{n!} A_n(\delta), \quad \frac{df}{ds} = \sum_{n\geq 0} \frac{\epsilon^n}{n!} B_n(\delta), \quad A_0 = 0, \quad B_0 = \omega. \tag{11}$$

The KBM method, [6,12], gives an iterative way to find the terms A_n, B_n and u_n. Assuming that the order $(n-1)$ has been solved, we find a function U_n that depends on all expressions until order $(n-1)$. Then, by calling $c_j(F)$ and $s_j(F)$ respectively the coefficients of $\cos jf$ and $\sin jf$ in the Fourier expansion of $F(f)$, the expressions of A_n, B_n are $A_n = -s_1(U_n)/(2\omega)$, $B_n = -c_1(U_n)/(2\omega\delta)$, and the function u_n is given by its Fourier expansion whose coefficients are $c_0(u_n) = c_0(U_n)/\omega^2$, $c_j(u_n) = c_j(U_n)/\omega_j$, $s_j(u_n) = s_j(U_n)/\omega_j$, $j \geq 2$, with $\omega_j = \omega^2(1-j^2)$.

Applying the KBM method to the equation (10) we find the variation of δ and f with respect to the new time s by means of the expressions

$$\frac{d\delta}{ds} = 0\,, \quad \frac{df}{ds} = n_f = 1 + \sum_{n\geq 1} \frac{\epsilon^n}{n!} \Psi(\delta, \mathcal{K}_n^{i,j}). \tag{12}$$

Equation (12) shows that δ has a constant value. n_f is also a constant since the values of $\mathcal{K}_n^{i,j}$ are constant. However, we will not integrate equation (12) to obtain the relation between f and s, rather, we will obtain a generalized Kepler equation in the next section. This latter equation will give us a direct relation between f and t.

Besides we obtain the expressions of u and v. These expressions together with the change of variables (9), allow us to find the expression of R

$$R = \frac{\mu e}{\Theta} \sin f - \epsilon \mathcal{K}_1^{1,0} \frac{\mu e}{\Theta} \sin f + \frac{\epsilon^2}{2!} \sum_{k=1}^{5} \mathcal{K}_k^R \sin kf, \tag{13}$$

and the expression of $1/r$, that once inverted gives

$$r = \frac{p}{1 + e\cos f} - \epsilon \frac{p^2 \mathcal{K}_1^{0,0}}{(1 + e\cos f)^2} + \frac{\epsilon^2}{2!} \left[\frac{2p^3 \left(\mathcal{K}_1^{0,0}\right)^2}{(1 + e\cos f)^3} \right. \tag{14}$$

$$\left. - \frac{p^2}{(1 + e\cos f)^2} \left(\mathcal{K}_0^r + \mathcal{K}_2^r \cos 2f + \mathcal{K}_3^r \cos 3f + \mathcal{K}_4^r \cos 4f + \mathcal{K}_5^r \cos 5f\right) \right],$$

where $\mathcal{K}_k^R$ and $\mathcal{K}_k^r$ are functions of constants, and we use the generalized semi–latus rectum, eccentricity and semi–mayor axis. These expressions are used to obtain the variables θ and ν and the momenta Θ and N.

5 Generalized Kepler Equation

The variation of the generalized true anomaly, f, with respect to time, t, is obtained after considering the definition of s given by (9) from which we get

$$n_f \, \Theta \, dt = r^2 \, df, \tag{15}$$

in which we substitute r by the expression (14) in terms of f.

Then, we apply a change of variables as in the case of the transformation of the true anomaly into the eccentric anomaly in the non–perturbed two body problem $p/(1 + e\cos f) = a(1 - e\cos E)$, $df = \sqrt{1 - e^2}(1 - e\cos E)dE$. This change of variables is usually used in orbital mechanics to transform integrals with powers of $(1 + e\cos f)$ in the denominator into terms with powers of $(1 - e\cos E)$ in the numerator. By doing so, we can integrate these expressions. In our problem, the existence of factors $\cos nf, \sin nf$ in the numerator of the perturbation terms implies that, even after applying the change, some terms still have powers of $(1 - e\cos E)$ in the denominator.

A practical way to avoid this formal problem is to apply first the change of f into w defined by $w = 1 + e\cos f$, $\cos f = (w - 1)/e$. By using the properties of Chebyshev's polynomials, T_n, U_n, we can express each term of (15) as a positive or negative power of w, multiplied or not by $\sin f$. Then, to integrate (15) we

just have to consider integrals of the form $\int w^n\,df$, $\int w^n \sin f\,df$, that can be easily performed [4].

Finally, we obtain the equation $n(t-T) = E - e\sin E + \sum_{i>0}(\epsilon^i/i!)\mathcal{K}_i(E,f)$, where T represents the value of t when $f = E = 0$, equation known as the generalized Kepler's equation. Note that for orders greater than one, some terms in f and E are mixed up in this generalized Kepler equation.

Acknowledgements. We are very grateful to I. Tijera for her suggestions. This work has been supported in part by the Ministerio de Educación y Ciencia (DGICYT BFM2002-03157 and # BFM2003-02137) and the Department of Space Mathematics of Centre National d'Etudes Spatiales (France).

References

1. Abad, A. and San Juan, J. F.: PSPC: A Poisson Series Processor coded in C. *Dynamics and Astrometry of Natural and Artificial Celestial Bodies.* Kurzynska et al eds. Poznam, Poland, (1993), 383–389.
2. Abad, A. and San Juan, J. F.: ATESAT: software tool for obtaining automatically ephemeris from analytical simplifications. *Conseil de L'Europe. Cahiers du Centre Européen de Géodynamique et de Séismologie.* A. Elipe and P. Paquet eds. Luxembourg. **10**, (1995), 93–98.
3. Abad, A., Elipe, A., Palacián, J. and San Juan, J. F.: ATESAT: A Symbolic Processor for Artificial Satellite Theory. *Mathematics and Computers in Simulation.* **45**, (1998), 497–510.
4. Abad, A., San Juan, J. F., and Gavín, A.: Short term evolution of artificial satellites.*Celestial Mechanics and Dynamical Systems.* **79**, (2001), 277–296.
5. Alfriend, K. T. and Coffey, S. L.: Elimination of the Perigee in Satellite Problem, *Celestial Mechanics*, **32**, (1984), 163–172.
6. Bogoliubov, N. N. and Mitropolsky, Y. A.: *Asymptotic Method in the Theory of Nonlinear Oscillations*, Gordon and Breach. New York, (1961).
7. Deprit, A.: Canonical Transformations Depending on a Small Parameter, *Celestial Mechanics*, **1**, (1969), 12–30.
8. Deprit, A.: The Elimination of the Parallax in Satellite Theory. *Celestial Mechanics*, **24**, (1981), 111-153.
9. Desai, P. N., Braun, R. D. and Powell, R. W.: Aspects of Parking Orbit Selection in a Manned Mars Mission, NASA TP-3256, (1992).
10. Henrard, J.: On a perturbation theory using Lie Transform, *Celestial Mechanics*, **3**, (1970), 107–120.
11. Kamel, A. A.: Perturbation methods in the theory of nonlinear oscillations, *Celestial Mechanics*, **3**, (1970), 90–106.
12. Krylov, N. and Bogoliubov, N. N.: *Introduction to Nonlinear Mechanics.* Princeton University Press, Princeton N.Y., (1947).
13. San Juan, J. F.: ATESAT: Automatization of theories and ephemeris in the artificial satellite problem, Tech. rep. CT/TI/MS/MN/94-250, CNES, France, (1994).
14. San Juan, J. F.: Manipulación algebraica de series de Poisson. Aplicación a la teoría del satélite artificial. Ph. D. Dissertation, Univ. of Zaragoza, (1996).

Computing Theta-Stable Parabolic Subalgebras Using LiE

Alfred G. Noël

The University of Massachusetts, Department of Mathematics, Boston, MA 02125-3393, USA
Massachusetts Institute of Technology, Department of Mathematics, Cambridge, MA 02139-4307, USA

Abstract. This paper describes an algorithm for computing representatives of conjugacy classes of θ-stable parabolic subalgebras of a semisimple complex Lie algebra $g_{\mathbb{C}}$ relative to any of its non-compact real forms g of inner type. These subalgebras are important for studying representations of g.

1 Introduction

The notion of a θ-stable parabolic subalgebra of $g_{\mathbb{C}}$ was introduced by Vogan [**1**] in the 1970's in order to study representations of semisimple Lie groups. Since then such subalgebras have been used by several authors to understand certain questions related to the theory of nilpotent orbits. In many instances, it is desirable to compute representatives of certain classes of such subalgebras under the action of a given Lie group. In this paper we propose an algorithm for computing such representatives under the action of a connected complex Lie group $K_{\mathbb{C}}$. The algorithm was implemented in the computer algebra system LiE [**2**] and was used to show that a theorem of Peter E. Tapa for classical real Lie groups does not extend to exceptional Lie groups [**3**]. It is also being used to study *polarization* in the exceptional Lie groups [**4**].

Let g be a real semisimple Lie algebra with adjoint group G and $g_{\mathbb{C}}$ its complexification. Also let $g = k \oplus p$ be the Cartan decomposition of g where k is a Lie algebra and p, a vector space. Finally, let θ be the corresponding Cartan involution of g. Then $g_{\mathbb{C}} = k_{\mathbb{C}} \oplus p_{\mathbb{C}}$ where $k_{\mathbb{C}}$ and $p_{\mathbb{C}}$ are obtained by complexifying k and p respectively. Denote by $K_{\mathbb{C}}$ the connected subgroup of the adjoint group $G_{\mathbb{C}}$ of $g_{\mathbb{C}}$, with Lie algebra $k_{\mathbb{C}}$. Then $k_{\mathbb{C}}$ and $p_{\mathbb{C}}$ are the $+1$-eigenspace and the -1-eigenspace of the extension of θ on $g_{\mathbb{C}}$ respectively. We shall call such an extension θ also. The subgroup $K_{\mathbb{C}}$ preserves $p_{\mathbb{C}}$ under the adjoint action.

Let $h_{\mathbb{C}}$ be a Cartan subalgebra and $\Phi = \Delta(g_{\mathbb{C}}, h_{\mathbb{C}})$ the root system of $g_{\mathbb{C}}$ determined by $h_{\mathbb{C}}$. A **Borel subalgebra** of $g_{\mathbb{C}}$ is a subalgebra $b = h_{\mathbb{C}} \oplus n$ where $n = \bigoplus_{\alpha \in \Phi^+} g_{\mathbb{C}}^{\alpha}$ for some positive system Φ^+ within Φ and $g_{\mathbb{C}}^{\alpha}$ denotes the root space of α. Any subalgebra q of $g_{\mathbb{C}}$ containing a Borel subalgebra is called a **parabolic subalgebra** of $g_{\mathbb{C}}$. If $q = q \cap k_{\mathbb{C}} \oplus q \cap p_{\mathbb{C}}$ then we shall say that q is a θ-stable parabolic subalgebra of $g_{\mathbb{C}}$. Any parabolic subalgebra decomposes as $q = l + u$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 335–342, 2004.
© Springer-Verlag Berlin Heidelberg 2004

where l is a Levi subalgebra of $g_{\mathbb{C}}$ and u, the nilradical of q, is a vector space consisting of nilpotent elements of $g_{\mathbb{C}}$.

We shall give an algorithm to compute representatives of all the $K_{\mathbb{C}}$-conjugacy classes of θ-stable parabolic subalgebras of $g_{\mathbb{C}}$ when g is of inner type, that is, when $rank(g) = rank(k)$, and $g_{\mathbb{C}}$ simple. Slight modifications of the algorithm are necessary to handle the cases where g is not of inner type. More will be said on this at the end.

2 Algorithm Description and Proof of Correctness

The algorithm is divided into two main components:

i. Computation of the $K_{\mathbb{C}}$-conjugacy classes of systems of simple roots. This task will be performed by the function Compute_Chambers().

ii. Computation of representatives of conjugacy classes of θ-stable parabolic subalgebras. The function Compute_Parabolics() will perform this computation.

The algorithm will be written in "pidgin" LiE, that is, we shall use a lot of built-in functions from LiE. Readers should consult **[2]** to find information on such functions. The built-in functions will be written in italics in the definition of the algorithm. We should point out that the algorithm can be implemented on any computer algebraic systems containing some Lie-theoretical capabilities. All variables will be of type integer. This is the only type that LiE accommodates. To see a very brief evaluation of LiE consult our ICCS 2003 paper **[5]**. Finally, we shall use the notation $\|S\|$ for the cardinality of a set S and comments will be placed between square brackets.

Description of the algorithm.

Input:

G: semisimple type of $G_{\mathbb{C}}$

K: semisimple type of $K_{\mathbb{C}}$

R_{nc}: set of non compact roots of $G_{\mathbb{C}}$

Output:

P: one-dimensional array of integers containing the indices of the roots in each parabolic subalgebra.

P_addr: two-dimensional array of integers containing the start and the end of each parabolic subalgebra in P.

This variable plays an important role when processing the parabolic subalgebras.

num_parabolics: number of representatives computed

Begin [Main]

n = n_pos_roots(G);

l = Lie_rank(G);

[Compute the number of $K_{\mathbb{C}}$-conjugacy classes of root systems]

n_ch = ‖W(G)‖/‖W(K)‖; [W(G) and W(K) are the Weyl groups of G and K]

P = $null$(n_ch$*2^l, l$); [Create and initialize P]

P_addr = $null$(n_ch$*2^l$, 2);

```
    dim_par = null(n_ch*2^l);
    [ Initialize lists and queues ]
    ch = null (n_ch*l,l); qu = null ((n_ch*l),l);
    cur_ch =null(l,l); new_ch = null(l,l);
    [ Initialize counters ]
    l_count = 1; q_indx = l; l_indx = l;
    d_index = 0;
    Compute_Chambers();
    Compute_Parabolics();
End [ Main]
```

Compute_Chambers();

begin

[This algorithm computes the $K_{\mathbb{C}}$ conjugacy classes of simple roots systems as follows: Starting with a Vogan system in the usual bourbaki system it looks for other non $K_{\mathbb{C}}$-conjugate systems by performing reflection along non compact imaginary roots]

```
    for i = 1 to l do ch[i] = pos_roots(G)[i] ; qu[i] = pos_roots(G)[i]; od;
    [ Main Loop ]
    while l_count < n_ ch do
    [ reflect through all non compact roots in current chamber ]
    [ be sure that both queue and list are properly maintained ]
    for i =1 to l do cur_ch[i] = q[i];od; [ dequeue ]
    for  i = 1 to l do

    if  cur_ch[i] ∈ R_nc then
    [ this is a noncompact root reflect through it ]
    for k = 1 to l do
    new_ch[k] = cur_ch[k] - Cartan(cur_ch[k], cur_ch[i], G)
    *cur_ch[i]; od;
    [ check for duplicate ]
    for j =1 to l do if new_ ch[j] != ch[(k-1)*l + j] then
    uniq = 0; break; fi;  od;
    if uniq == 1 then break; fi; od;
    if  uniq ==0 then for k = 1 to l do ch[l_indx + k] = new_ch[k];
    qu[q_indx +k] = new_ch[k]; od; l_indx = l_indx + l;
    q_indx = q_indx + l;
    fi; fi; od;

    for i = 1 to (q_indx- l) do qu[i] = qu[i+l]; od; [ rebuild queue ]
    q_indx = q_indx - l;
    od; [Main Loop]
end [ Compute_Chambers() ]
```

Compute_Parabolics();

[This algorithms computes a list which contains all the representatives of the $K_{\mathbb{C}}$-conjugacy classes of θ-stable parabolic subalgebras of $g_{\mathbb{C}}$. Since each chamber given by the previous algorithm is equivalent to a θ-stable Borel subalgebra the algorithm builds standard parabolic subalgebras in each chamber by using the subsets of the chamber.]

begin

chamb = null (l,l); i = 1; par_index = 0;
for j = i **to** (i+l-1) **do** count= count+1; chamb[count]= ch[j];**od**;
[initialize the nilradical of the Borel subalgebra generated by chamb]
u = null (*n_pos_roots*(G), l);
for u_indx = 1 **to** *n_pos_roots*(G) **do**
for r_indx = 1 **to** l **do** u[u_indx] = u[u_indx]+ pos_roots(G)[u_indx,r_indx]*
chamb[r_indx]; **od**; **od**;
[find all subsets of the simple roots in the class
and build the resulting parabolic $q = l \oplus u$]
cc[ii] = ii; **od**;
null(kk,l);
[retrieve the subset of simple roots defining the Levi subalgebra]
chamb[cc[ii]];**od**;
l_matrix= null (kk,l); [Cartan matrix for the Levi subalgebra]
for i_l = 1 **to** kk **do for** j_l =1 **to** kk **do**
l_matrix[i_l,j_l] = *Cartan*(G)(lev_roots[i_l],lev_roots[j_l]); **od**; **od**;
g = *Cartan_type*(G)(lev_roots); m_cartan = *Cartan*(G)(g);
if l_matrix $\neq$ m_cartan **then**
good_lev = lev_roots; nnn = kk; pi =null(nnn+2); p = null(nnn+1);
d = null(nnn+1); previous = null(nnn); current = null(nnn);
for iii = 2 **to** nnn+1 **do** pi[iii] = iii; p[iii] = iii; d[iii] = -1; **od**;
d[1] = 0; m = nnn+2; pi[1] = m; pi[m] = m; counter = 0; **for** c = 2 **to** nnn+1
do current[c-1] = pi[c] - 1; **od**;
lev_roots[current[i_l]] **od**; **to** kk **do** l_matrix[i_l,j_l] =
Cartan(G)(good_lev[i_l],good_lev[j_l]); **od**; **od**;
if l_matrix == m_cartan **then** lev_roots =good_lev ; break; fi;
[right order found]
m = nnn+1; **while** pi[p[m]+d[m]] >m **do** d[m] = -d[m]; m = m-1;
if m ==1 **then** break; **fi**; **od**;
bucket = pi[p[m]]; pi[p[m]] = pi[p[m]+d[m]]; pi[p[m]+d[m]] = bucket;
previous = current;
od; **fi**;
[end permutation, continue to process subsets]
jj=kk;
while cc[jj]== (l-kk +jj) **do** jj = jj-1; **if** jj == 0
then break **fi**; **od**;
if jj !=0 **then** cc[jj] = cc[jj] +1 **fi**;
for ii = jj+1 **to** kk **do if** ii == 1 **then** cc[ii] = 1 **else**
cc[ii] = cc[ii-1] +1; **fi** ;**od**;

```
    n_pieces = n_comp(g);
    nilp_u = null(n,l); levi_index = 0; nilp_u_index = 0; lev_ptr = 0;
    sg = g[ii]; l_p = Append ( pos_roots(sg),  (-pos_roots(sg)));
    for jk = 1 to 2 n_pos_roots(sg) do
    levi_index = levi_index +1;
    for ll=1 to Lie_rank(sg) do
    levi_subalg[levi_index] = levi_subalg[levi_index] +
    l_p[jk,ll]*lev_roots[lev_ptr + ll]; od;od;
    lev_ptr = lev_ptr + Lie_rank(sg); od;
    for ik = 1 to n do trouver = 0;
    break;fi; od;
    if trouver == 0  then nilp_u_index = nilp_u_index + 1;
    nilp_u[nilp_u_index] = u[ik]; fi; od;
    [ check for duplicate subalgebras and build the list]
    found = 0; u_qq = null(nilp_u_index,l);
    for ik = 1 to nilp_u_index do u_qq[ik] = nilp_u[ik]; od;
    q = sort(Append (levi_subalg,u_qq));  [ q = l ⊕ u ]
    dimq = levi_index+nilp_u_index;
    data = Append (levi_subalg u_qq);
    if n_parabolics == 0 then
    n_parabolics = n_parabolics + 1; P_addr[1,1] = 1; P_addr[1,2] = dimq;
    for ik =1 to dimq do par_index = par_index +1;
    P[par_index] = data[ik]; od;
    d_index = d_index +1; dim_par[d_index] = dimq+l;
    else found = 0; [ Check for duplicates ]
    P[P_addr [ik,1] +jk -1] od; break; fi; fi; ood;
    dimq do par_index = par_index +1;
    P[par_index] = data[ik]; od;
    d_index = d_index +1; dim_par[d_index] = dimq+l;
    fi; fi; od; od; od;
end  [ Compute_Parabolics() ]
```

Remark. At the end of Compute_Parabolics() the list P will contain representatives of all classes of parabolic subalgebras except those of the Borel subalgebras. However, the Borel subalgebras are completely determined by the roots stored in the variable **chamb** which defines the Cartan subalgebra equivalent to the Levi subalgebra in this case. The variable **u** contains the appropriate positive roots and is in fact the nilradical of the Borel representative.

Proof of correctness

Theorem. *The above algorithm is correct.*

Proof. Maintaining the above notations, it is known that the parabolic subalgebras q containing a Borel subalgebra b of $g_{\mathbb{C}}$ are parametrized by the set of subsets of Δ the set of simple roots that defines b (See **[7]** Proposition 5.90 for a proof). Let Φ be the root system generated by Δ and let Γ be a subset of Δ. Define q_Γ to be the subalgebra of $g_{\mathbb{C}}$ generated by $h_{\mathbb{C}}$ and all of the root

spaces $g_{\mathbb{C}}^{\alpha}$ such that $\alpha \in \Delta$ or $-\alpha \in \Gamma$. Let $\langle \Gamma \rangle$ denote the subroot system of Φ generated by Γ and put $\langle \Gamma \rangle^+ = \langle \Gamma \rangle \cap \Phi^+$. Define

$$l = h_{\mathbb{C}} \oplus \bigoplus_{\alpha \in \langle \Gamma \rangle} g_{\mathbb{C}}^{\alpha} \qquad u = \bigoplus_{\alpha \in \Phi^+ \setminus \langle \Gamma \rangle^+} g_{\mathbb{C}}^{\alpha}$$

Then $q_\Gamma = l \oplus u$ is a parabolic subalgebra containg b and is said to be a standard parabolic subalgebra. Moreover every parabolic subalgebra of $g_{\mathbb{C}}$ is conjugate to a standard parabolic subalgebra of $g_{\mathbb{C}}$. Since we assume that g is of inner type we conclude that all parabolic subalgebras are θ-stable. The above argument is valid for each $k_{\mathbb{C}}$-conjugacy class of Borel subalgebras. Hence, the algorithm generates a list containing representatives of all the $k_{\mathbb{C}}$-conjugacy classes of θ-stable parabolic subalgebras of $g_{\mathbb{C}}$.

In order to finish the proof we need to show that the computation $\langle \Gamma \rangle$ is correct. This is done in Compute_Parabolics() by obtaining subsets of Δ and permuting the set of roots in such subsets when necessary. We generate the permutations using minimal change order as described in [**6**]. To compute subsets of Δ we use an implementation of Algorithm 5.8 in [**6**] also. The proofs of correctness of both algorithmic schemes are found in [**6**]. Hence, the theorem follows.

□

Complexity

The complexity of the algorithm depends on that of the built-in functions. In LiE such functions are well designed and seem to perform in an optimal manner. Since LiE is not a multipurpose software package, the designers were able to use clever and faster algorithms to enhance performance. The reader should realize that most of the work in this algorithm is done by Compute_Parabolics() which computes subsets and permutations of elements of Δ which is of size l the rank of $g_{\mathbb{C}}$. The analysis of both schemes is done in [**6**] and it is not too difficult to see that the permutation scheme will take $\mathcal{O}(k!)$ to permute k elements of Δ and the determination of the subsets of size k is proportional to the number of combinations of subsets of k elements of Δ that is $\mathcal{O}(\binom{l}{k})$. Hence both schemes perform quasi-optimally. Of course, this is a worst case analysis. For $l \leq 8$ the algorithm performs very well on average. We are not in a position to give an average case analysis at this time. However the reader should know that we were able to compute the representatives of the classes of theta-stable parabolic subalgebras for all the exceptional non compact simple Lie groups relative to all their real forms of inner type. This is not trivial [**4**]. The computations were carried on an IMac G4 with speed 1GHz and 1Gb SDRAM of memory.

3 Some Applications of Representation Theory

A representation of a group is a mathematical map which associates a matrix to each element of the group. Matrices are very concrete objects that facilitate difficult computations which would be impossible otherwise. This was recognized

after the discovery of quantum mechanics. Hence, given a group, if all or a lot of its representations are available then the investigator has a better chance of finding one which fits the problem at hand. Finding all the representations of a given reductive real Lie group is one the most important unsolved problems in Mathematics.

An other incentive to study Representation theory comes from Number theory. Here, we should point out that the field of Number theory, although considered one of the purest branches of mathematical thought, turns out to have very important and concrete applications in our industrial world. One such application is the design and development of almost unbreakable codes in cryptography allowing the possibility of making transactions on the internet using credit cards.

Representation theory is used in quantum chemistry, quantum computing, construction of telephone networks, radar and antenna design, robotics, coding theory, computer vision and many other branches of science and engineering. Readers who are interested in real world applications of Representation theory should visit the following website:

http : *//web.usna.navy.mil/wdj/repn_thry_appl.htm*

The work presented in this paper is part of a program whose aim is to compute new representations of reductive real Lie groups. See [**8**] for more details.

4 Conclusion

In this paper we proposed and used and algorithm which produces a list containing representatives of all the $K_{\mathbb{C}}$-conjugacy classes of theta-stable parabolic subalgebras of a complex simple Lie algebra $g_{\mathbb{C}}$ relative any of its real non-compact forms g of inner type. We proved the correctness of the algorithm and gave a worst case analysis of its complexity. We also mentioned that the average performance of the algorithm is quite good because we were able to use it to compute data from all the exceptional simple Lie groups. However we still have more work to do. First, we need to extend the algorithm to the cases where g is not of inner type. This can be done as follows: the group $K_{\mathbb{C}}$ should be replaced by $G^{\theta}_{\mathbb{C}}$ the subgroup of $G_{\mathbb{C}}$ that fixes $k_{\mathbb{C}}$ and the computation of the theta stable parabolic will be more complicated. We should be able to do this soon. In order to manage space more effectively we need to develop a formula for the number of $K_{\mathbb{C}}$-conjugacy classes of theta-stable parabolic subalgebras. We have not been able to find such a formula in the literature. From our conversation with experts there are reasons to believe that the formula is not known. One way to circumvent this issue is to use the Weyl group of $K_{\mathbb{C}}$. This solution does not scale well because the Weyl group grows fast as the rank of g increases and traversing it becomes a challenging problem. We are currently developing some new strategies to solve this problem.

There are also some software issues. LiE provides essentially two types of data structures, the **vector**, a one-dimensional array of type integer, and the **matrix**, a two-dimensional array of type integer and does not allow dynamic allocation. These two factors complicate the handling of large data sets. Since the LiE source

code is available we plan to solve these problems in the future. We believe that in general the mathematical algorithms in LiE are well designed. However we would welcome some serious work on a good user-interface.

Many mathematicians are currently using Computer Algebra Systems in their research not only as simulation tools but also as a way of generating important counterexamples and conjectures. As these systems become more and more powerful we should expect a stronger cooperation between mathematicians, system designers and computer scientists.

Acknowledgment. The author wishes to thank the referees for their helpful and insightful comments.

References

1. Vogan D. jr:*The algebraic structure of the representation of semisimple Lie groups I*, Annals of Math. **109** (1979), 1-60
2. Van Leeuwen M. A. A., Cohen A. M., Lisser B.:*LiE A package for Lie Group Computations*, Computer Algebra Nederland, Amsterdam The Netherlands (1992)
3. Noël A. G.: *Appendix to "Richardson Orbits for Real Classical Groups" by Peter E. Trapa (Counterexamples in F_4)* , to appear in Journal of Algebra
4. Noël A. G.: *Some remarks on Richardson Orbits in Complex Symmetric Spaces*, (preprint)
5. Noël A. G.: *Computing maximal tori using LiE and Mathematica*, Lectures Notes in Computer Science, Springer-Verlag. **2657** (2003) 728-736
6. Reingold E. M., Nievergelt J., Deo N.: *Combinatorial Algorithms Theory and Practice*, Prentice-Hall (1977)
7. Knapp A. W.:Lie Groups Beyond and introduction second edition, Birkhaüser Progress in Mathematics **140** (2002)
8. Peter E. Trapa.: *Richardson Orbits for Real Classical Groups* , to appear in Journal of Algebra

Graphical and Computational Representation of Groups

Alain Bretto and Luc Gillibert

Université de Caen, GREYC CNRS UMR-6072,Campus II, Bd Marechal Juin BP 5186, 14032 Caen cedex, France. {alain.bretto,lgillibe}@info.unicaen.fr

Abstract. An important part of the computer science is focused on the links that can be established between group theory and graph theory. CAYLEY graphs can establish such a link but meet some limitations. This paper introduces a new type of graph associated to a group: the G-graphs. We present an implementation of the algorithm constructing these new graphs. We establish a library of the most common G-graphs, using GAP and the `SmallGroups` library. We give some experimental results with GAP and we show that many classical graphs are G-graphs.

1 Introduction

The group theory, especially the finite group theory, is one of the main parts of modern mathematics. Groups are objects designed for the study of symmetries and symmetric structures, and therefore many sciences have to deal with them. Graphs can be interesting tools for the study of groups, a popular representation of groups by graphs being the CAYLEY graphs, an extended research has been achieved in this direction [1]. The regularity and the underlying algebraic structure of CAYLEY graphs make them good candidates for applications such as optimizations on parallel architectures, or for the study of interconnection networks [4]. But these properties are also a limitation: many interesting graphs are not CAYLEY graphs.

The purpose of this paper is to introduce a new type of graph – called G-graphs – constructed from a group and to present an algorithm to construct them. This algorithm is used for establishing some experimental results and for finding which graphs are G-graphs and which graphs are not. In fact, G-graphs, like CAYLEY graphs, have both nice and highly-regular properties. Consequently, these graphs can be used in any areas of science where CAYLEY graphs occur. Moreover many usual graphs, as the cube, the hypercube, the cuboctahedral graph, the Heawood's graph and lots of others, are G-graphs. We prove that some generic and infinite families of graphs, such as the complete bipartite graphs, are G-graphs. We establish a catalogue of the most common G-graphs, and for each of these graphs we exhibit the corresponding group, using the GAP's `SmallGroups` library. We also show that some non-vertex-transitive graphs, such as the Gray graph and the Ljubljana graph, are also G-graphs. In contrast, notice that CAYLEY graphs are always vertex-transitive.

The G-graphs are very informative about the groups from which thy are constructed: (1) they can be used for studying subgroups, via the correspondence

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 343–350, 2004.
© Springer-Verlag Berlin Heidelberg 2004

between induced subgraphs and subgroups, and (2) the orders of the elements of a group can be read in the corresponding graph. In [2], it was shown that two isomorphic groups give two isomorphic graphs and that two abelian groups are isomorphic if and only if their associated graphs are themselves isomorphic. Thus, G-graphs can establish a link between the graph isomorphism problem and the abelian group isomorphism problem. But what happens for non abelian groups? We give some answers with an experimental simulation on all the groups of small order, again using GAP and the `SmallGroups` library.

2 Basic Definitions

We define a graph $\Gamma = (V; E; \epsilon)$ as follows::

- V is the set of vertices and E is the set of edges.
- ϵ is a map from E to $P_2(V)$, where $P_2(V)$ is the set of subsets of V having 1 or 2 elements.

In this paper graphs are finite, i.e., sets V and E have finite cardinalities. For each edge a, we denote $\epsilon(a) = [x; y]$ if $\epsilon(a) = \{x, y\}$ with $x \neq y$ or $\epsilon(a) = \{x\} = \{y\}$. If $x = y$, a is called *loop*. The set $a \in E, \epsilon(a) = [x; y]\}$ is called *multiedge* or *p-edge*, where p is the cardinality of the set. We define the degree of x by $d(x) = |\{a \in E, x \in \epsilon(a)\}|$.
In this paper, groups are also finite. We denote the unit element by e. Let G be a group, and let $S = \{s_1,\ s_2, \ldots,\ s_k\}$ be a nonempty subset of G. S is a *set of generators* of G if any element $\theta \in G$ can be written as a product $\theta = s_{i_1} s_{i_2} s_{i_3} \ldots s_{i_t}$ with $i_1, i_2, \ldots i_t \in \{1,\ 2, \ldots,\ k\}$. We say that G is *generated* by $S = \{s_1,\ s_2, \ldots,\ s_k\}$ and we write $G = \langle s_1, s_2, \ldots, s_k\rangle$. Let H be a subgroup of G, we denote Hx instead of $H\{x\}$. The set Hx is called *right coset* of H in G. A subset T_H of G is said to be a *right transversal* for H if $\{Hx, x \in T_H\}$ is precisely the set of all cosets of H in G.

3 Graph Group Process

Let (G, S) be a group with a set of generators $S = \{s_1,\ s_2,\ s_3 \ldots\ s_k\}$, $k \geq 1$. For any $s \in S$, we consider the left action of the subgroup $H = \langle s\rangle$ on G. Thus, we have a partition $G = \bigsqcup_{x \in T_s} \langle s\rangle x$, where T_s is a right transversal of $\langle s\rangle$. The cardinality of $\langle s\rangle$ is $o(s)$ where $o(s)$ is the order of the element s. Let us consider the cycles

$$(s)x = (x, sx, s^2x, \ldots, s^{o(s)-1}x)$$

of the permutation g_s: $x \longmapsto sx$. Notice that $\langle s\rangle x$ is the support of the cycle $(s)x$. Also ust one cycle of g_s contains the unit element e, namely $(s)e = (e, s, s^2, \ldots, s^{o(s)-1})$. We now define a new graph denoted $\Phi(G; S) = (V; E; \epsilon)$ as follows:

- The vertices of $\Phi(G;S)$ are the cycles of g_s, $s \in S$, i.e., $V = \sqcup_{s\in S} V_s$ with $V_s = \{(s)x, x \in T_s\}$.
- For all $(s)x, (t)y \in V$, $\{\langle s\rangle x, \langle t\rangle y\}$ is a p-edge if $card(\langle s\rangle x \cap \langle t\rangle y) = p$, $p \geq 1$.

Thus, $\Phi(G;S)$ is a k-partite graph and any vertex has a $o(s)$-loop. We denote $\tilde{\Phi}(G;S)$ the graph $\Phi(G;S)$ without loop. By construction, one edge stands for one element of G. One can remark that one element of G labels several edges. Both graphs $\Phi(G;S)$ and $\tilde{\Phi}(G;S)$ are called *graph from group* or *G-graph* and we say that the graph is *generated* by the groups $(G;S)$. Finally, if $S = G$, the G-graph is called a *canonic graph*.

3.1 Algorithmic Procedure

The following procedure constructs a graph from the list of the cycles of the group:

```
Group_to_graph_G(L)
for all s in L
    Add s to S
    for all s' in L
        for all x in s
            for all y in s'
                if x=y then Add (s,s') to A
```

An implementation of this procedure has been written in C++: we call it `Gro2gra`. The complexity of our implementation is $O(n^2 \times s^2)$ where n is the order of the group G and s is the cardinal of the family S. An other procedure constructs the vertices, that is the list of the cycles from the group G and the family S. The implementation of this procedure requires a high-level language that can manipulate groups: we use GAP Release 4.3 (Groups, Algorithms, and Programming) [5]. The following algorithm uses two functions:

1. `c_cycles`: computes a list of lists of lists, in fact the list of the lists of the cycles of each element s.
2. `fx`: writes the results of `c_cycles` in a file.

Only the procedure `c_cycles` is interesting:

```
InstallGlobalFunction (c_cycles, function(G, ga)
local ls1,ls2,gs,k,x,oa,a,res,G2;
res:=[]; G2:=List(G);
for a in ga do
      gs:=[]; oa:=Order(a)-1;
      ls2:=Set([]);
      for x in G do
            if not(x in ls2) then
                  ls1:=[];
                  for k in [0..oa] do;
                        Add(ls1, Position(G2, (a^k)*x));
```

```
                        AddSet(ls2, (a^k)*x);
                od; Add(gs, ls1);
            fi;
      od; Add(res, gs);
od;
return res; end);
```

For each s, $T_s = \{x_1, x_2, \ldots, x_j\}$, the right transversal of $\langle s \rangle$, is computed during the construction of the cycles $(s)x_i$. For this purpose, all the elements y of the cycle $\langle s \rangle x_i$ are added to the set `ls2`, then the procedure chooses an element x_{i+1} in G that does not appears in `ls2`, computes $(s)x_{i+1}$ and adds the elements of $\langle s \rangle x_{i+1}$ in `ls2`. The operation is repeated until all the elements of G are in `ls2`. Then, the set `ls2` is emptied; a new $s \in S$ is chosen and the operation is repeated. The second function, `fx`, is only here for the human's interface.

3.2 Example

In order to compute the cycles of the graph $\Phi(C_2 \times C_2; S)$ with $S = C_2 \times C_2$, we only have to call the function:

```
fx(AbelianGroup([2,2]),AbelianGroup([2,2]),"c2c2");
```

The procedure `fx` creates the following cycles in a file `c2c2`:

```
(1)(2)(3)(4)(1 2)(3 4)(1 3)(2 4)(1 4)(2 3)
```

Then, the program `Gro2gra` creates the following edges:

```
"(1)"--"(1 2)" "(1)"--"(1 3)" "(1)"--"(1 4)" "(2)"--"(1 2)"
"(2)"--"(2 4)" "(2)"--"(2 3)" "(3)"--"(3 4)" "(3)"--"(1 3)"
"(3)"--"(2 3)" "(4)"--"(3 4)" "(4)"--"(2 4)" "(4)"--"(1 4)"
"(1 2)"--"(1 3)" "(1 2)"--"(2 4)" "(1 2)"--"(1 4)"
"(1 2)"--"(2 3)" "(3 4)"--"(1 3)" "(3 4)"--"(2 4)"
"(3 4)"--"(1 4)" "(3 4)"--"(2 3)" "(1 3)"--"(1 4)"
"(1 3)"--"(2 3)" "(2 4)"--"(1 4)" "(2 4)"--"(2 3)"
```

These graph is shown in Fig. 1.

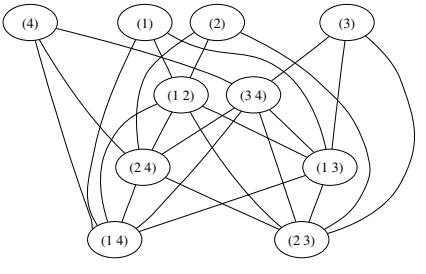


Fig. 1. $C_2 \times C_2$

4 Experimental Results

4.1 The Dihedral Group, the Generalized Quaternion Group, and the Product of Two Cyclic Groups

Let the dihedral group D_{2n} be the group of presentation:

$$\langle r, s \mid r^n = e, s^2 = e, sr = r^{n-1}s \rangle$$

Proposition 1: For $S = \{r, s\}$, the graph D_{2n}, $\tilde{\Phi}(D_{2n}; S)$ of the dihedral group is the complete bipartite graph $K_{2,n}$.
See Fig. 2 for an example.

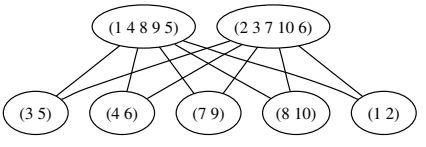


Fig. 2. $\tilde{\Phi}(D_{10}; \{a, b\})$

Let the generalized quaternion group Q_n be the group of presentation:

$$\langle a, b \mid a^{2n} = e, b^2 = a^n, ab = ba^{2n-1} \rangle$$

Proposition 2: For $S = \{a, b\}$, the graph Q_n, $\tilde{\Phi}(Q_n; S)$ of the generalized quaternion group is the complete double-edged bipartite graph $K^2_{2,n}$.
See Fig. 3 for an example.

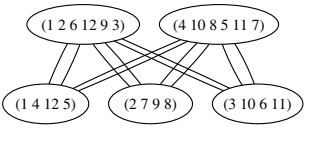


Fig. 3. $\tilde{\Phi}(Q_3; \{a, b\})$

Let $C_n \times C_k$ be the product of two cyclic groups. Such a product is generated by two elements, a and b, with $a^n = b^k = e$. More precisely, $C_n \times C_k$ is the group of presentation:

$$\langle a, b \mid a^n = e, b^k = e, ab = ba \rangle$$

Proposition 3: For $S = \{a, b\}$, the graph $C_n \times C_k$, $\tilde{\Phi}(C_n \times C_k; S)$ of the product of two cyclic groups, is the complete bipartite graph $K_{n,k}$.
See Fig. 4 for an example.

4.2 How to Recognize a *G*-Graph

Given a G-graph Γ, an interesting problem is how to find a group G and a family S such that $\tilde{\Phi}(G; S)$ isomorphic to Γ. If both G and S exist, we say that

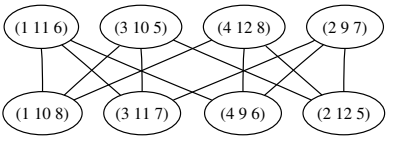


Fig. 4. $\tilde{\Phi}(C_3 \times C_3; \{a, b\})$

Γ is a G-graph. Here, we use the `SmallGroups` library from GAP. This library gives us access to all groups of certain small orders. The groups are sorted by their orders and they are listed up to isomorphism. Currently, the library contains the groups of order at most 2000 except 1024 (423 164 062 groups). In this section, we prove that many usual graphs are G-graph and we exhibit their corresponding groups.

The cube – Let us consider the skeleton of a cube. It is a graph with 8 vertices and 12 edges. All vertices are of degree 3 and the graph is bipartite. Suppose the cube is a G-graph $\tilde{\Phi}(G; S)$. Then the corresponding group G is of order 12 and is generated by a family S of cardinality 2, because the graph is bipartite. The alternate group with 12 elements, A_4, subgroup of S_4, is generated by the two cycles $(1, 2, 3)$ and $(1, 3, 4)$. Let S be the family $\{(1, 2, 3), (1, 3, 4)\}$. If we compute the graph $\tilde{\Phi}(A_4; S)$ with our algorithm, then we find the graph depicted in Fig 5.

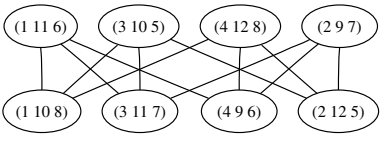


Fig. 5. $\tilde{\Phi}(A_4; S)$

It is easy to check that this graph is isomorphic to the cube. Thus, the cube is a G-graph as expected .

The hypercube – Let us consider the skeleton of an hypercube of dimension 4. It is a graph with 16 vertices and 32 edges. All vertices are of degree 4 and the graph is bipartite. Suppose the hypercube is a G-graph $\tilde{\Phi}(G; S)$. Then the corresponding group G is of order 32 generated by a family S of cardinal 2, because the graph is bipartite. The order of the elements of the family S must be 4 because the vertex degree is 4. If we look at the library `SmallGroups` we find 51 groups of order 32. Only 7 groups of order 32 can be generated by two elements of order 4: the groups number 2, 6, 10, 11, 13, 14 and 20. If we compute the corresponding graphs with our algorithm we find that `SmallGroup(32,6)` matchs (see Fig. 6).

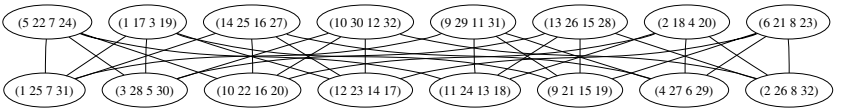


Fig. 6. $\tilde{\Phi}(\texttt{SmallGroup}(32,6); S)$

Others G-graphs – We give here some examples of G-graphs. The corresponding groups are indicated between parenthesis:

1. Bipartite complete graphs ($G = C_n \times C_k$, $S = \{(1,0)(0,1)\}$)
2. The 3-prism ($G = C_3 \times C_3$, $S = \{(1,0)(0,1)\}$)
3. The cuboctahedral graph ($G = C_2 \times C_2 \times C_2$, $S = \{(1,0,0),(0,1,0),(0,0,1)\}$)
4. The square (G is the Klein's group, $G = \{e, a, b, ab\}$, and $S = \{a, b\}$)
5. The cube ($G = A_4$, $S = \{(1,2,3),(1,3,4)\}$)
6. The hypercube ($G =$`SmallGroup(32,6)`, $S = \{f1, f1 * f2\}$)
7. The generalized Petersen's graph $P_{8,3}$
   ($G =$`SmallGroup(24,3)`, $S = \{f1, f1 * f2\}$)
8. The 2×2 grid on a torus ($G = Q_2$, $S = \{a, b\}$)
9. The 3×3 grid on a torus ($G = D_6$, $S = \{s \in G, Ordre(S) = 2\}$)
10. The 4×4 grid on a torus ($G =$`SmallGroup(32,6)`, $S = \{f1, f1 * f2\}$)
11. The Heawood's graph ($\langle a, b \mid a^7 = b^3 = e, ab = baa\rangle$, $S = \{b, ba\}$)
12. The Pappus's graph
    ($G = \langle a, b, c \mid a^3 = b^3 = c^3 = e, ab = ba, ac = ca, bc = cba\rangle$, $S = \{b, c\}$)
13. The Mobius-Kantor's graph ($G =$`SmallGroup(24,3)`, $S = \{f1, f1 * f2\}$)
14. The Gray graph ($G =$`SmallGroup(81,7)`, $S = \{f1, f2\}$)
15. The Ljubljana graph ($G =$`SmallGroup(168,43)`, $S = \{f1, f1 * f2 * f4\}$)

4.3 Couples of Non-isomorphic Groups Giving Isomorphic Graphs

One of the main goals of the G-graphs was originally the study of the graph isomorphism problem. A result in [2] says that two isomorphic groups give two isomorphic graphs. Another result in the same paper says that two abelian isomorphic groups are isomorphic if and only if their associated graphs are isomorphic. But what happens with non abelian groups? With our implementation of the algorithm `Gro2gra` and the `SmallGroups` library of GAP, it is possible to check automatically all the couples of non-isomorphic groups up to the order 100 in only a few days of computation.
Only couples of groups having the same number of elements of the same order can give isomorphic graphs. Such couples are be called "suspicious" couples in the table bellow. All isomorphisms are tested with Nauty [7]. Only orders with a least one suspicious couple are listed.

It is easy to see that only a few percent of suspicious couples give isomorphic graphs. The conclusion is that G-graphs are informative enough in the majority of the cases to allow the identification of the group by the graph. Finally we can notice that the couples of groups giving isomorphic graphs share the same properties:

They are non-simple, non-perfect, solvable and super-solvable.

Order	Number of groups	Number of suspicious couples	Non-isomorphic groups giving isomorphic graphs
16	14	7	1
27	5	2	2
32	51	66	3
48	52	20	3
54	15	6	2
64	267	1425	24
72	50	1	0
80	52	23	7
81	15	15	13
96	231	345	12
100	16	1	1

References

1. L. Babai. *Automorphism groups, isomorphism, reconstruction.* Chapter 27 of Handbook of combinatorics, 1994.
2. Alain Bretto and Alain Faisant, *A new graph from a group*, To appear in Compte rendu de l'academie des sciences, Paris, 2003.
3. John F. Humphreys, *"A course in Group Theory"*, Oxford University Press, 1997.
4. G. Cooperman and L. Finkelstein and N. Sarawagi. *Applications of Cayley Graphs.* Appl. Algebra and Error-Correcting Codes. Springer Verlag. Lecture Notes in Computer Sciences, Vol. 508 1991, 367–378.
5. The GAP Team, (06 May 2002), *"GAP - Reference Manual"*, Release 4.3, http://www.gap-system.org.
6. Joseph Lauri and Raffaele Scapellato, *Topics in Graphs Automorphisms and Reconstruction*, London Mathematical Society Student Texts, 2003.
7. Brendan D. McKay, Computer Science Department, Australian National University, (1981), *"Practical graph isomorphism"*, Congressus Numerantium 30, p. 45-87.

First Order ODEs: Mathematica and Symbolic-Numerical Methods

Ciro D'Apice, Giuliano Gargiulo, and Manzo Rosanna

University of Salerno, Department of Information Engineering and Applied Mathematics, via Ponte don Melillo,
84084 Fisciano (SA), Italy
{dapice,gargiulo,manzo}@diima.unisa.it

Abstract. The use of information technology in addition to traditional lectures affords a means to develop student intuition and curiosity, reaching in the same time a deep knowledge of the subject of study. The aim of this work is to show the didactic use of a Computer Algebra System to illustrate and compare different symbolic-numerical methods for solving first order ordinary differential equations (ODEs). In particular, we apply, relate and compare the built-in functions of Mathematica, the method of integration by series, the Picard process and the linearization method in solving some first order ODEs. This approach allows students not only to master the basic methods for solving ODEs, but also to be naturally led to theoretical deepening of such areas as power series, stability and convergence theory, elements of functional analysis or the local-global relationship via linearization.

1 Symbolic Methods in Solving ODEs

1.1 Introduction

Consider a differential equation of the first order in normal form

$$y' = f[x, y] \tag{1}$$

and one of its solutions, that is a function u:J R, where J is an interval which satisfies :

$$u'[x] == f[x, u[x]] \quad \forall x \in J\ . \tag{2}$$

If f is continuous (C^0), the right-hand member of (2) is continuous, since u is C^1 (continuously differentiable). If f is more smooth, for example, C^1 the right-hand member of (2) (being composition of C^1 functions) is C^1 and therefore u is C^2. For induction it can be easily proved that, if f is C^k (k-time continuously differentiable), u is k+1-time continuously differentiable, if f is infinitely continuously differentiable, also u is infinitely continuously differentiable.
In these cases, the derivatives at x_0 of a Cauchy problem solution

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 351–357, 2004.
© Springer-Verlag Berlin Heidelberg 2004

$$\begin{aligned} y' &= f[x, y] \\ y[x_0] &= y_0 \end{aligned} \tag{3}$$

can be computed starting from the derivatives of f at (x_0 ,y_0). In fact, obviously it turns out that

$$\begin{aligned} u[x_0] &= y_0 \\ u'[x_0] &= f[x_0, y_0] \end{aligned} \tag{4}$$

Moreover, computing derivatives of (2) - it is possible if f is enough regular – we obtain:

```
eq = u'[x] == f[x, u[x]];
relat = {x -> x0, u[x0] → y0, u'[x0] → f[x0, y0]};
newrelat = D[eq, x] //. relat
relat = Join[relat, {u''[x0] -> newrelat[[2]]}];
```

$$u''[x0] == f[x0, y0]\, f^{(0,1)}[x0, y0] + f^{(1,0)}[x0, y0]$$

```
newrelat = D[eq, {x, 2}] //. relat
relat = Join[relat, {u'''[x0] -> newrelat[[2]]}];
```

$$\begin{aligned} u^{(3)}[x0] == {} & f^{(0,1)}[x0, y0]\,(f[x0, y0]\, f^{(0,1)}[x0, y0] + f^{(1,0)}[x0, y0]) + \\ & f[x0, y0]\, f^{(1,1)}[x0, y0] + \\ & f[x0, y0]\,(f[x0, y0]\, f^{(0,2)}[x0, y0] + f^{(1,1)}[x0, y0]) + f^{(2,0)}[x0, y0] \end{aligned}$$

It is clear that if we go on in this way all the derivatives of u at x_0 can be computed (when the derivatives of f exist and are known).

1.2 ODE Integration by Series

Alternatively - assuming for simplicity that x_0=0 - we suppose that f admits Taylor development respect to y near x, for all x in a neighbourhood of x_0:

```
uSer = SeriesData[x, 0, Table[a[k], {k, 0, 3}], 0, 4] /. a[0] → y0
```

$$y0 + a[1]\,x + a[2]\,x^2 + a[3]\,x^3 + O[x]^4$$

So also u has a Taylor development:

```
fSer = SeriesData[y, y0, Table[b[k, x], {k, 0, 3}], 0, 3]
```

$$b[0, x] + b[1, x]\,(y - y0) + b[2, x]\,(y - y0)^2 + O[y - y0]^3$$

We require that the composite of the two series (i.e., the series of the composite function f[x, u [x]]) is equal to the derived series of u (i.e., the series of derivates) …

```
eqcoeffSer = LogicalExpand[(CoefficientList[#1, x] &) /@ eqSer]
```

$$\begin{aligned} & b[0, 0] == a[1] \;\&\&\; a[1]\, b[1, 0] + b^{(0,1)}[0, 0] == 2\, a[2] \;\&\&\; \\ & a[2]\, b[1, 0] + a[1]^2\, b[2, 0] + \\ & \quad a[1]\, b^{(0,1)}[1, 0] + \frac{1}{2}\, b^{(0,2)}[0, 0] == 3\, a[3] \end{aligned}$$

```
Solve[eqcoeffSer, Table[a[k], {k, 3}]] // Flatten
```

$$\left\{a[3] \to \frac{1}{6}\left(b[0,0]\,b[1,0]^2 + 2\,b[0,0]^2\,b[2,0] + b[1,0]\,b^{(0,1)}[0,0] + 2\,b[0,0]\,b^{(0,1)}[1,0] + b^{(0,2)}[0,0]\right),\ a[2] \to \frac{1}{2}\left(b[0,0]\,b[1,0] + b^{(0,1)}[0,0]\right),\ a[1] \to b[0,0]\right\}$$

Obviously, the coefficients of the series can be expressed in terms of the derivatives of u and f, so this approach is equivalent to the previous…

```
Series[f[x, y], {x, 0, 2}]
```

$$f[0,y] + f^{(1,0)}[0,y]\,x + \frac{1}{2} f^{(2,0)}[0,y]\,x^2 + O[x]^3$$

```
Series[u[x], {x, 0, 2}]
```

$$u[0] + u'[0]\,x + \frac{1}{2} u''[0]\,x^2 + O[x]^3$$

1.3 Picard Method and Linearization

In the Picard iterative process, we search the fixed point of a transformation among abstract spaces (Banach spaces). We can expect that if we approximate this transformation with a simpler one (for example a linear one) the fixed point we find in this way is on one side the approximation of the searched fixed point (solution of the ODE) and moreover on the other side it is simpler to compute. The transformation is:

$$T[y][x] = y_0 + \int_{x_0}^{x} f[t, y[t]]dt \tag{5}$$

The best local linear approximation is the differential that in our case can be written in the following form:

$$\begin{aligned} T'[y,\varphi][x] &= \int_{x_0}^{x} \partial_y f[t, y[t]]\varphi(t)dt \\ dT'[y,\varphi][x] &= T[y][x] + T'[y, \varphi - y][x] = \\ & y_0 + \int_{x_0}^{x} f[t, y[t]]dt + \int_{x_0}^{x} \partial_y f[t, y[t]](\varphi[t] - y[t])dt \end{aligned} \tag{6}$$

to be seen as a function of φ , with y fixed). It is easy to see that the fixed point of dT (computed in y) is a solution of the following problem:

$$\varphi[x] = y_0 + \int_{x_0}^{x} f[t, y[t]]dt + \int_{x_0}^{x} \partial_y f[t, y[t]](\varphi(t) - y[t])dt$$

The latter is a linear (non homogeneous) ODE and therefore Mathematica is able to solve it symbolically. The solution of (6) or (7) is an approximation of the solution of the initial ODE.

$$\begin{aligned} &\varphi[x_0] = y_0 \\ &\varphi'[x] = f[x, y[x]] + \partial_x f[x, y[x]](\varphi[x] - y[x]) \end{aligned} \tag{7}$$

It can be used as an approximation or as initial point (near, and so with fast convergence rate) in the Picard process applied to the original ODE. Also, it can be used as an abstract version of Newton-Raphson method. In particular, the search of the fixed points of the transformation (5) is equivalent to the search of the fixed points of the associated Newton-Raphson transformation:

$$NR(T)[y] = (y - T[y]) - (Id - T'[y])^{-1}(y - T[y]) \tag{8}$$

where Id is the identical transformation and $(\ .\ )^{-1}$ is the transformation inverse. Since the inversion corresponds to the solution of a linear ODE we can apply again, even if only in part, the power of symbolic calculus of Mathematica to NR(T). We remark that analogous results are valid if T' and dT are computed with a fixed approximation $y=y_1$, instead of a variable one.

2 The Particular Case y'==y^2

Dsolve

```
solution =
 y[x] /. Flatten[DSolve[{y'[x] == y[x]^2, y[0] == 1}, y[x], x]]
```

$$\frac{1}{1-x}$$

Picard method

Let us define the following integral function. We can observe that as it follows from the definition, when we give in input only the function H, it assumes for default 0 as initial point and x as final point.

```
H[function_, variable_: x, intialpoint_: 0] :=
 Module[{y},
  y[variable] /.
   Flatten[DSolve[{y'[variable] == function^2, y[intialpoint] == 1},
     y[variable], variable]]]
```

```
Clear[step]; step[1] = Nest[H, 1, 1]
```

$$1 + x$$

```
Table[step[n] = Nest[H, 1, n], {n, 2, 5}]
```

$$\left\{\frac{1}{3}(3 + 3x + 3x^2 + x^3),\right.$$
$$\frac{1}{63}(63 + 63x + 63x^2 + 63x^3 + 42x^4 + 21x^5 + 7x^6 + x^7),$$
$$\frac{1}{238140}(238140 + 238140x + 238140x^2 + 238140x^3 + 238140x^4 +$$
$$206388x^5 + 158760x^6 + 109620x^7 + 67095x^8 + 36120x^9 +$$
$$\left.16632x^{10} + 6300x^{11} + 1890x^{12} + 420x^{13} + 60x^{14} + 4x^{15}),\ \ldots\right\}$$

```
Normal[Series[1/(1 - x), {x, 0, 5}]]
```

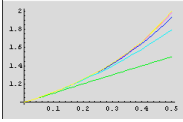
$$1 + x + x^2 + x^3 + x^4 + x^5$$

```
Normal[Series[step[5], {x, 0, 5}]]
```

$$1 + x + x^2 + x^3 + x^4 + x^5$$

By comparison of the result obtained with Picard iterative process and the exact solution we can see that the difference between the two terms decreases. We plot the approximations of the solutions at the first 5 steps and the exact solution to evaluate their difference.

```
Plot[Evaluate[Join[{solution}, Table[step[n], {n, 5}]]],
  {x, .0001, .5}, PlotStyle → Evaluate[Table[Hue[k/6], {k, 5}]]];
```



We are thus led to conjecture that the exact solution is $1+x+x^2+x^3+x^4+x^5+\ldots$, i.e. $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$.

Linearization

```
T[function_, variable_: x, initialpoint_: 0] :=
  ∫_initialpoint^variable function^2 dvariable + 1
```

```
L[function_] :=
 y[x] /.
  Flatten[DSolve[{y'[x] - 2 y[x] == D[function, x],
     y[0] == (function /. x -> 0)}, y[x], x]]
```

```
P[f_] := f - L[f - T[f]]
```

```
Timing[p1 = P[1]]
```

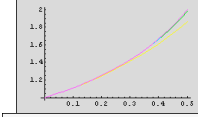
$$\left\{0.04\ \text{Second},\ 1 + \frac{1}{2}\left(-1 + e^{2x}\right)\right\}$$

```
Timing[p2 = P[P[1]]]
```

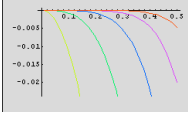
$$\left\{0.28\ \text{Second},\ 1 + \frac{1}{2}\left(-1 + e^{2x}\right) + \frac{1}{8}\left(-1 + e^{4x} - 4\,e^{2x}\,x\right)\right\}$$

Timing[p_3=P[p_3]]…
Timing[p_4=P[p_4]]…

```
Plot[{p1, p2, p3, p4, solution}, {x, .0001, .5},
  PlotStyle → Evaluate[Table[Hue[k/6], {k, 5}]]];
```



```
Plot[{p1 - solution, p2 - solution, p3 - solution, p4 - solution},
  {x, .0001, .5}
, PlotStyle -> Evaluate[Table[Hue[k / 5], {k, 6}]]];
```



By comparison of the plots it turns out that the last method seems to approximate the solutions with a greater rate than the Picard process.

Integration by series

```
y[x_] := SeriesData[x, 0, Table[a[i], {i, 0, 6}]]
```

```
y[x]
```

$$a[0] + a[1]\,x + a[2]\,x^2 + a[3]\,x^3 + a[4]\,x^4 + a[5]\,x^5 + a[6]\,x^6 + O[x]^7$$

```
systODE = LogicalExpand[∂x y[x] == y[x]^2]
```

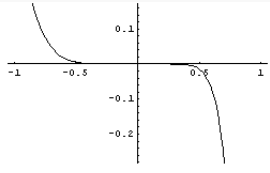
$$a[0]^2 - a[1] == 0 \;\&\&\; 2\,a[0]\,a[1] - 2\,a[2] == 0 \;\&\&$$
$$a[1]^2 + 2\,a[0]\,a[2] - 3\,a[3] == 0 \;\&\&$$
$$2\,a[1]\,a[2] + 2\,a[0]\,a[3] - 4\,a[4] == 0 \;\&\&$$
$$a[2]^2 + 2\,a[1]\,a[3] + 2\,a[0]\,a[4] - 5\,a[5] == 0 \;\&\&$$
$$2\,a[2]\,a[3] + 2\,a[1]\,a[4] + 2\,a[0]\,a[5] - 6\,a[6] == 0$$

```
coeff = Flatten[Solve[{systODE, a[0] == 1}, Table[a[i], {i, 0, 6}]]]
```

$$\{a[6] \to 1,\ a[5] \to 1,\ a[4] \to 1,\ a[3] \to 1,\ a[2] \to 1,\ a[1] \to 1,\ a[0] \to 1\}$$

Now we plot the difference between the solution approximated by "series" seriesol=Normal [y[x]] /. coeff ($1+x^2+x^3+x^4+x^5+x^6$) and the symbolic solution .

```
Plot[Evaluate[seriesol] - solution, {x, -1, 1}];
```



References

1. Albano, G., D'Apice, C., Tomasiello, S.: Simulating Harmonic Oscillator and electrical circuits: a didactical proposal. iJMEST, Vol.33, No 2, (2002) 157-185.
2. D'Apice, C., De Simone, T., Manzo, R., Tibullo, V.: MOSFET: a Virtual Laboratory with *Mathematica*. Proceedings ICTMT6 Volos, Greece, (2003) 252-257.
3. Kampas F.,Lockheed M.: Iterative Solution of Highly Nonlinear Differential Equations Using *Mathematica* WorldWide Mathematica Conf. (1998).
4. Mezzino, M.: Discovering Solutions to Ordinary Differential Equations. WorldWide Mathematica Conf. (1998).
5. Olness, F.: Integrating Mathematica in the Undergraduate Science Curriculum: Teaching Computer Literacy with Mathematica. WorldWide *Mathematica* Conf. (1998).

Evaluation of the Fundamental Physical Constants in *Mathematica*

Andrey S. Siver

Institute of High Energy Physics, Protvino, Russia

Abstract. We present PAREVAL package consisting of a number of *Mathematica* modules used to re-evaluate basic CODATA:1998 recommended fundamental physical constants (FPC). Package can be found at http://sirius.ihep.su/∼siver. Values of the basic FPC-1998 with the positive defined correlation matrix and a number of functions for the FPC usage can be found therein. Among other PAREVAL functions there are functions for energy levels of hydrogen and deuterium, electron and muon magnetic moment anomaly and muonium ground-state hyperfine splitting.

1 Introduction

Hopes for the discovery of new physics at present time are frequently connected with high-precision experiments combined with corresponding high-precision calculations. These calculations usually involve values of the FPC and should also use their uncertainties and correlation matrix. But none of the available resources provides the last one (correlation coefficients presented on the official site [3] are incorrect because they are rounded to three decimal figures).

Design of PAREVAL package has been mostly motivated by two facts:

- Methodological doubt about validity of CODATA recommended FPC-1998 [1]. This doubt arises due to the non-positive semi definiteness of the correlation matrix of the input experimental data presented in [1]. More information can be found in [4];
- Absence of powerful IT resources for scientific activities in FPC studies. Critical notes about several resources can be found in [4]

Our FPC-1998 re-evaluation has been generally based on review [1]. We have checked the values of the basic FPC and got their correlation matrix. The rest (derived) FPC can be expressed as a functions of the basic ones and thus be calculated.

2 Package Structure

PAREVAL consists of a number of *Mathematica* [2] modules which can be ranged as followings:

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 358–363, 2004.
© Springer-Verlag Berlin Heidelberg 2004

1. Modules for FPC usage. Contains data and functions for the FPC usage in high-precision calculations;
2. Modules with physical formulae collection. Contains functions designed to calculate energy levels of hydrogen and deuterium, electron and muon magnetic moment anomaly, muonium ground-state hyperfine splitting and some other functions;
3. Modules for experimental data processing. Contains functions transforming experimental data between input and internal forms;
4. Module for parameters evaluation. Contains *Mathematica* functions for evaluation of parameters based on least-squares method;
5. Modules for results presentation. Contains functions used to present results of parameters evaluation in LaTeX system.

The package has general public license (GPL) [5] and can be found at http://sirius.ihep.su/∼siver. Values of the basic FPC-1998 with the positive defined correlation matrix, a number of useful functions for the FPC usage, *Mathematica* notebook in which calculations have been carried out can be found therein.

Most important modules are in a nutshell described below.

2.1 Modules for FPC Usage

The first module contains several *Mathematica* variables which can be used in calculations. The second one contains a function for the calculation of the FPC uncertainties propagation with FPC covariance matrix.

Module	Function	Input	Output	Example	Description of the symbols
fpc-usage-1	prop	f(z)	u[f(z)]	Fig.1	Calculates propagation of the uncertainties of the FPC for f(z) according to the law: $u[f(z_1, ..., z_n)] = \sqrt{\sum_{i,j=1}^{N} \frac{\partial f}{\partial z_i} cov[z_i, z_j] \frac{\partial f}{\partial z_j}}$
fpc-usage-1	info	const	name, unit, symbTeX	Fig.1	name — 'const's name; unit — 'const's unit; symbTeX — 'const's TeX symbol

2.2 Modules with Physical Formulae Collection

These modules contain some realization of several physical functions used in FPC-1998 evaluations.

Function name	Mathem. symbol	Description
E1tot	$E_X(n, \mathrm{L}, j)$	Energy levels of hydrogen and deuterium[1]
a_e	a_e	Electron magnetic moment anomaly [1]
a_μ	a_μ	Muon magnetic moment anomaly[1]
$\Delta\nu_{Mu}$	$\Delta\nu_{Mu}$	Muonium ground-state hyperfine splitting [1]
ν	ν	Function conjugated to muonium ground-state hyperfine splitting([1], p.387)

Module	Func.	Input	Example	Description of the symbols
cod-15-3	E1tot	X, n, L, j	Fig. 2	X=(1 or 2), X=1 for hydrogen and X=2 for deuterium; n — principal quantum number; L — nonrelativistic orbital angular momentum; j — angular momentum quantum number;
formulae-2	a_e	alpha, dae		alpha — fine structure constant; dae — value of theoretical uncertainty for a_e (see [1], p.476).
formulae-2	a_μ	alpha, damu		alpha — fine structure constant; damu - value of theoretical uncertainty for a_μ (see [1], p.479).
formulae-2	$\Delta\nu_{Mu}$	mE, mM, aMU, alpha, R	Fig.3	alpha — fine structure constant; mE — electron mass; mM — muon mass; R — Rydberg constant. aMU — Muon magnetic moment anomaly

```
In[7]:= dir1 = "d:\\siver\\4mathematica\\par-0.9.2\\";

In[8]:= << (dir1 <> "reFPC-1998-2.m"); << (dir1 <> "fpc-usage-1.m");

In[9]:= info[R]

Out[9]= {Rydberg constant, m^-1, R_{\infty}}

In[10]:= info[hPl]

Out[10]= {Planck constant, J s, h}

In[11]:= info[alpha]

Out[11]= {fine-structure constant, 1, \alpha}

In[12]:= m$e = R / alpha^2 / c * 2 * hPl /. c → 299792458;
         val = m$e /. subsFPC (* === get value === *)

Out[13]= 9.10938187149259 × 10^-31

In[14]:= unc = prop[m$e] (* === get uncertainty === *)

Out[14]= 7.20642221827284 × 10^-38

In[15]:= unc / val (* === get relative uncertainty === *)

Out[15]= 7.91098926352514 × 10^-8
```

Fig. 1. Example of the usage of modules for calculations with the FPC

```
In[12]:= Collect[
           Simplify[
            Expand[
             (E1tot[HH, 2, 0, 1/2] - E1tot[HH, 1, 0, 1/2]) /.
                {Z → 1, m_N → m_p, m_l → m_p, RR[1] → RP} /. m_e → 9.109381874910758 * 10^(-31) /.
              m_p → 1.672621582915101 * 10^(-27) /. m_μ -> 1.883531087707727 * 10^(-28) /.
             c → 299792458], α > 0 && α < 1/2], α]

Out[12]= -2.22578 × 10^-17 α^2 - (1.40384 × 10^-55 RP^2 α^4)/h^2 +
         1/8 (8.90312 × 10^-17 + 497. h - 6.55057 × 10^-13 Sqrt[1 - α^2] + 4.63132 × 10^-13 Sqrt[1 + Sqrt[1 - α^2]]) +
         1/8 α^5 (5.76497 × 10^-13 + (3.27423 × 10^-75 RP^3)/h^3 + 4.85732 × 10^-13 Log[α]) +
         1/8 α^8 (2.50237 × 10^-14 - 3.1776 × 10^-16 Log[α] + 5.83615 × 10^-13 Log[α]^2) +
         1/8 α^7 (7.04894 × 10^-12 + 1.89846 × 10^-12 Log[α] + 7.28499 × 10^-13 Log[α]^2) +
         1/8 α^6 (-1.78255 × 10^-12 + (2.12019 × 10^-54 RP^2)/h^2 + (1.12307 × 10^-54 RP^2 Log[(1.71496×10^-21 RP α)/h])/h^2)
```

Fig. 2. Lamb shift. Example of the usage of the 'E1tot' function. Unfortunately, it's impossible to show the expression for classical Lamb shift in a reasonable form obtained as *Mathematica* output. So we make extra substitutions for m_e, m_p, m_μ and c using the values of FPC-1998.

3 Further Work. Discussion

A long time ago it was realized that evaluation of the FPC gave two important results: (i) values of the FPC and (ii) test for modern theories on agreement with each other and with experiments [1].

The first task can be solved with the help of PAREVAL package. But in order to solve the second one a system for monitoring the values of FPC should be created. This system should include:

- Collection of methods for adjustment of parameters of the theories;
- A database of all measured experimental data which can be used in parameters evaluation;
- A database of all self-consistent modern theoretical formulae, relevant to the experimental database. Besides, the database of the formulae should be connected to calculation media;
- Collection of methods that could test statistical hypotheses and seek possible systematical errors or uncertainties of calculation methods and programming "bugs";
- Subsystem for the presentation of results;

```
In[38]:= temp = Simplify[Expand[Δν_Mu[m_e, m_μ, a_μ, α, R] /. subsX /. dMu → 0],
           α > 0 && α < 1]
```

$$
\begin{aligned}
\text{Out[38]}= \; & 175 - \frac{8\,c\,R\,\alpha^4\,(13 + 36\,\text{Log}[2])\,m_e^2}{9\,(m_e + m_\mu)^2} + \frac{1}{135\,(m_e + m_\mu)^5} \\
& \left(4\,c\,R\,\alpha^3\,m_e^2\,m_\mu \left(650\,\alpha\,m_\mu^2 - 1440\,\alpha\,\text{Log}[2]\,m_\mu^2 - 360\,\alpha\,\text{Log}[\alpha]\,m_\mu^2 + \right.\right. \\
& \frac{540\,\text{Log}\left[\frac{m_\mu}{m_e}\right] m_\mu^2\,(m_e + m_\mu)}{\pi\,(m_e - m_\mu)} + 30\,\alpha\,(m_e + m_\mu)^2 + \frac{700\,\alpha\,(m_e + m_\mu)^2}{\pi^2} + \\
& \frac{7758\,\alpha^2\,(m_e + m_\mu)^2}{\pi^3} - \frac{10260\,\alpha^2\,(m_e + m_\mu)^2}{\pi} + \frac{840\,\alpha^2\,\text{Log}[\alpha]^2\,(m_e + m_\mu)^2}{\pi} + \\
& \frac{195\,\alpha\,\text{Log}\left[\frac{m_\mu}{m_e}\right]\,(m_e + m_\mu)^2}{\pi^2} + \frac{540\,\alpha^2\,\text{Log}[\alpha]\,\text{Log}\left[\frac{m_\mu}{m_e}\right]\,(m_e + m_\mu)^2}{\pi} - \\
& \frac{360\,\alpha\,\text{Log}\left[\frac{m_\mu}{m_e}\right]^2\,(m_e + m_\mu)^2}{\pi^2} + \frac{240\,\alpha^2\,\text{Log}\left[\frac{m_\mu}{m_e}\right]^2\,(m_e + m_\mu)^2}{\pi^3} - \\
& \left.\left. \frac{240\,\alpha^2\,\text{Log}\left[\frac{m_\mu}{m_e}\right]^3\,(m_e + m_\mu)^2}{\pi^3} + \frac{1890\,\alpha\,(m_e + m_\mu)^2\,\text{Zeta}[3]}{\pi^2} \right) \right) - \\
& \frac{1}{607500\,\pi^3\,(m_e + m_\mu)^3} \left(c\,R\,\alpha^2\,(1 + a_\mu)\,m_e\,m_\mu^2 \right. \\
& \left(358500\,\pi^4\alpha^3 - 135000\,\pi^3\,(24 + 51\,\alpha^4 + 2\,\alpha^2\,(-11 + \text{Log}[64])) + \right. \\
& 4500\,\pi^2\,\alpha^3\,(1124 - 3840\,\text{Log}[2] + 15\,\pi\,\alpha\,(-547 + 240\,\text{Log}[2]))\,\text{Log}[\alpha] + \\
& 4320000\,\pi\,\alpha^3\,(2\,\pi + \alpha)\,\text{Log}[\alpha]^2 + 8\,\pi^2\,\alpha\,(-202500 + 4860000\,\alpha^3 + \alpha^2 \\
& \quad (-15709037 + 13410000\,\text{Log}[2] + 562500\,\text{Log}[2]^2 - 466875\,\text{Zeta}[3])) + \\
& 22500\,\pi\,\alpha^2\,(-197 + 12384\,\alpha^2 - 108\,\text{Zeta}[3]) - 625\,\alpha^3\,\left(28259 + 7200 \right. \\
& \left.\left.\left. \text{Log}[2]^4 + 172800\,\text{PolyLog}\left[4, \frac{1}{2}\right] + 40032\,\text{Zeta}[3] - 46440\,\text{Zeta}[5] \right) \right) \right)
\end{aligned}
$$

Fig. 3. Example of the usage of the '$\Delta\nu_{Mu}$' function - muonium ground-state hyperfine splitting

– Subsystem for automatical or semi-automatical search for scientific information (old, modern and which just have appeared), which can be helpful to researcher.

As far as we know, none of such systems actually exists. PAREVAL package can be considered as a prototype of such system.

At last we would like to note that our working experience in *Mathematica* tells us that this computer algebra system is powerful enough to realize a system for monitoring values of fundamental physical constants.

References

1. P. J. Mohr and B. N. Taylor, "CODATA recommended values of the fundamental physical constants: 1998", Rev. Mod. Phys. **72** (2000) 351.
2. *Mathematica*, http://www.wolfram.com
3. Fundamental Physical Constants from NIST, http://physics.nist.gov/cuu/Constants/
4. Siver A.S., Ezhela V.V., "On the CODATA recommended values of the fundamental physical constants: V3.2(1998)&V4.0(2002)", IHEP Preprint 2003-34, Protvino, 2003; arXiv:physics/0401064
5. For General Public License (GPL) note see http://www.gnu.org

Symbolic Polynomial Interpolation Using Mathematica

Ali Yazici[1], Irfan Altas[2], and Tanil Ergenc[3]

[1] Computer Engineering Department, Atilim University, Ankara - Turkey
aliyazici@atilim.edu.tr
[2] School of Information Studies, Wagga Wagga - Australia
ialtas@csu.edu.au
[3] Mathematics Department, Middle East Technical University, Ankara - Turkey
tanil@metu.edu.tr

Abstract. This paper discusses teaching polynomial interpolation with the help of Mathematica. The symbolic power of Mathematica is utilized to prove a theorem for the error term in Lagrange interpolating formula. Derivation of the Lagrange formula is provided symbolically and numerically. Runge phenomenon is also illustrated. A simple and efficient symbolic derivation of cubic splines is also provided.

1 Introduction

Students use software tools such as Mathematica or Maple to test mathematical ideas and access the Internet to develop an awareness of the wider learning environment. Kaput [1] has suggested that the mathematical thinking ability to recognize translation from one representation of a function to another can be assisted by the use of computers. The use of a symbolic algebra package in combination with the Internet can develop new strategies and a deeper understanding of many mathematical concepts.

A numerical treatment of Hermite interpolation is studied in [2] using Mathematica. In this paper, we demonstrate the symbolic use of Mathematica in teaching interpolation concepts in a numerical computing course offered to sophomore Engineering students at Atilim University. In addition to 3 hours/week theoretical lectures, practical laboratory sessions are held (2 hours/week) for a group of 20 students to support the theory.

In Section 2 polynomial interpolation and Lagrange interpolating polynomials are discussed. Mathematica instructions are used to derive polynomials and the error formula for the linear case in Sections 2.2 and 2.3 respectively. Runge's phenomenon is demonstrated numerically with equally spaced nodes in Section 2.4. Use of the Chebyshev nodes as interpolation points are displayed in Section 2.5. Finally, Section 3 is devoted to the derivation of piecewise cubic spline interpolation symbolically.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 364–369, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Polynomial Interpolation

Let $x_o, x_1, ..., x_n$ be a set of $n+1$ distinct real or complex numbers on $[a, b]$, and let $y_o, y_1, ..., y_n$ be associated function values. Then the polynomial $p(x)$ is said to interpolate $f(x)$ at these points if $p(x_k) = f(x_k)$, $k = 0, 1, ..., n$.

2.1 Lagrange Form of the Interpolating Polynomial

Lagrange form of interpolating polynomial [3] is based on the polynomials

$$L_{n,j}(x) = \prod_{i=0}^{j} \frac{(x - x_i)}{(x_j - x_i)}, \quad i \neq j, \quad j = 0, 1, \ldots n \tag{1}$$

where x_i's are interpolation nodes in $[a, b]$. For each j, $L_{n,j}(x)$ is a polynomial of degree n and has the property

$$L_{n,j}(x_i) = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

We will write $L_{n,j}(x)$ simply as $L_j(x)$ when there is no confusion as to its degree. It is easy to see that polynomial $p(x) = \sum_{k=0}^{n} L_k(x) f(x_k)$ has degree n and satisfies the interpolation condition $p(x_j) = f(x_j), j = 0, 1, \ldots n$.

If $f^{n+1}(x)$ is continuous on $[a, b]$, then interpolating polynomial $p(x)$approximates $f(x)$ for each x in $[a, b]$ with an error

$$e(x) = f(x) - p(x) = \frac{f^{n+1}(c)}{(n+1)!} W(x) \tag{3}$$

where $W(x) = (x - x_o)(x - x_1) \ldots (x - x_n)$ and c is a number in (a, b).

The function $W(x)$ plays an important role in determining the size of the error bound. Among all possible choice for distinct x_j's $j = 0, 1, \ldots, n$ in $[a, b] = [-1, 1]$, maximum of $W(x)$ is minimized if x_i's are the roots of the $(n + 1)st$ degree Chebyshev polynomial.

2.2 Session: Lagrange Interpolation with Equally Spaced Points

A Mathematica session is set up to demonstrate the theoretical and practical aspects of interpolation in a step wise manner. In this paper only a part of the experiments are discussed. Complex coding and programming are avoided for educational purposes at the cost of computational efficiency. Firstly, the error formula (3) given above will be proven for the linear case using Mathematica.

The Mathematica instructions to prove the result is given below:

- Define the point set X for x_o and x_1, and compute length of X = m = n+1
  `In[1]:=` $X = \{xo, x1\}$, `In[2]:=` $m = Length[X]$ `Out[2]= 2`

- Define a general function F and define L_j's symbolically in product form
  `In[3]:=` $F[x_] := f[x]$
  `In[4]:=` $L[j_, x_] := Product[If[i == j, 1, (x - X[[i]])/(X[[j]] - X[[i]]), \{i, 1, m\}]$
- Compute, say, $L[2, x]$ and display the interpolating polynomial $p_1[x]$
  `In[5]:=` $L[2, x]$ `Out[5]=` $\frac{x-x_o}{x_1-x_o}$
  `In[6]:=` $p_1[x] := F[x_0]L[1, x] + F[x_1]L[2, x]$
- Define a special function g[t]. Here, x, x_o, and x_1 are constants with respect to t, and g[t] is zero at these points. Also, assume that $x_0 < x < x_1$
  `In[7]:=` $g[t_] := f[t] - p_1[t] - e1[x]\frac{(t-x_0)(t-x_1)}{(x-x_0)(x-x_1)}$
- Compute the first derivative of g with respect to t
  `In[8]:=` $g\prime[t]$
  `Out[8]=` $-\frac{(t-x_0))e1[x]}{(x-x_0)(x-x_1)} - \frac{(t-x_1)e1[x]}{(x-x_0)(x-x_1)} - \frac{f[x_o]}{x_o-x_1} - \frac{f[x_1}{-x_o+x_1} + f\prime[t] - p_1\prime[t]$
- Compute the second derivative of g with respect to t at c_1
  `In[9]:=` $g\prime\prime[z]$ `Out[9]=` $-\frac{2e1[x]}{(x-x_o)(x-x_1)} + f\prime\prime[z]$
- Applying Rolle's theorem to g[t] on $[x_o, x]$ to find a value d_o in (x_0, x) so that $g\prime[d_o] = 0$. A second application of Rolle's theorem to g[t] on $[x, x_1]$ produces a value d_1 in (x, x_1) so that $g\prime[d_1] = 0$. Observe that $g\prime[t]$ is zero at d_o, and d_1. Therefore, again by Rolle's theorem (applied to $g\prime[t]$ on $[d_o, d_1]$) we find the value c_1 for which $g\prime\prime[c_1] = 0$ which ends the proof.
  `In[10]:=` $Solve[g\prime\prime[c_1] == 0, e1[x]]$
  `Out[10]=` $\{\{e1[x] \to \frac{1}{2}(x - x_o)(x - x_1)f\prime\prime[c_1]\}\}$

This experiment encourages the students to utilize Mathematica for some constructive proofs of simple theorems and for deriving required identites in the method.

2.3 Session: Deriving Lagrange Interpolation Symbolically with Mathematica

In this section, Lagrange interpolating polynomial p2 will be derived symbolically for n=2 using 3 interpolation points.

- Define the point set X for the three (m=n+1) points x_o, x_1, and x_2
  `In[1]:=` $X = \{xo, x1, x2\}$ `Out[1]=` $\{xo, x1, x2\}$
  `In[2]:=` `m=Length[X]` `Out[2]=` `m=3`
- Define a general function F and L_j's symbolically in product form
  `In[3]:=` $F[x_] := f[x]$
  `In[4]:=` $L[j_, x_] := Product[If[i == j, 1, (x - X[[i]])/(X[[j]] - X[[i]]), \{i, 1, m\}]$
- Display L[1,x] symbolically
  `In[5]:=` `L[1,x]` `Out[5]=` $\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$
- Form the quadratic interpolating polynomial p2 symbolically
  `In[6]:=` $p2[x_, X_] = Sum[F[X[[i]]]L[i, x], \{i, 1, m\}]$
  `Out[6]=` $\frac{(x-x_1)(x-x_2)f[x0]}{(x_0-x_1)(x_0-x_2)} + \frac{(x-x_0)(x-x_2)f[x1]}{(-x_0+x_1)(x_1-x_2)} + \frac{(x-x_0)(x-x_1)f[x2]}{(-x_0+x_2)(-x_1+x_2)}$

Out[6] is the quadratic interpolating polynomial as an approximation to f.

2.4 Session: Runge's Function Illustrated

The polynomial interpolation problem has a unique solution. However, use of equally spaced interpolation points with polynomials of high degree may cause divergence. This is known as Runge's phenomenon and illustrated below.

Consider the function $F(x) = \frac{1}{1+25x^2}$ over the interval [-1,1]. The interpolating polynomials $p2[x]$, and $p8[x]$ with 3, and 9 equally spaced points respectively, in $[-1, 1]$ are formed.

- Define the function and the interpolation points
  `In[1]:=` $F[x_] := \frac{1}{1+25x^2}$, `In[2]:=` $xo = -1; x1 = 0; x2 = 1$
  `In[3]:=` $X = \{xo, x1, x2\}$ `Out[3]=` $\{-1, 0, 1\}$
  `In[4]:=` $m = Length[X]$ `Out[4]=` 3
- Form L_j's symbolically in product form and compute $p2[x]$.
  `In[5]:=` $L[j_, x_] := Product[If[i == j, 1, (x - X[[i]])/(X[[j]] - X[[i]]), \{i, 1, m\}]$ `In[6]:=` $p2[x_, X_] = Simplify[Sum[F[X[[k]]]L[k, x], \{k, 1, m\}]$
  `Out[6]=` $1 - \frac{25x^2}{26}$
- Display $p8[x]$. Intermediate steps are similar and omitted.
  `Out[7]=` $1. - 1.77636\text{x}10^{-15}x - 13.20303x^2 - 2.13163\text{x}10^{-14}x^3 + 61.36721x^4 + 7.10543\text{x}10^{-14}x^5 - 102.81501x^6 + 4.973799\text{x}10^{-14}x^7 + 53.68930x^8$

Now, students can obtain a plot of these polynomials to observe the divergence with higher degree polynomials.

2.5 Session: Lagrange Interpolation with Chebyshev Points

A remedy to Runge's phenomenon is to employ unequally spaced Chebyshev nodes as interpolation points as discussed above. Now, Lagrange interpolating polynomials will be formed at the Chebyshev nodes in a similar fashion.

- Compute the quadratic chebp2 at the Chebyshev nodes, c_k.
  `In[10]:=` $m = 3$
  `In[11]:=` $For[k = 1, k \leq m, k++, c[k] = -Cos[Pi(2k - 1)/(2m)]]$;
  `In[12]:=` $X = Table[c[k], k, 1, m]$ `Out[12]=` $\{-0.866025, 0., 0.866025\}$
  `In[13]:=` $L[j_, x_] := Product[If[i == j, 1, (x - X[[i]])/(X[[j]] - X[[i]]), \{i, 1, m\}]$
  `In[14]:=` $chebp2[x_, X_] = Simplify[Sum[F[X[[k]]]L[k, x], \{k, 1, m\}]]$
  `Out[14]=` $1 + 0.x - 1.26582x^2$
- chebp8 is computed in a similar fashion.
  `Out[15]=` $1. + 0.x - 9.51343^2 - 8.88178\text{x}10^{-16}x^3 + 31.3482x^4 - 1.95399\text{x}10^{-14}x^5 - 40.3504x^6 - 6.21725\text{x}10^{-15}x^7 + 17.6203x^8$
- Finally, a plot the graph of F, p4, and chebp8 is given to illustrate the Runge's phenomenon (Fig.1) and use of Chebyshev nodes
  `In[17]:=` $Plot[F[x], p4[x, X], chebp8[x, X], x, -1, 1,$
  $PlotRange- > All, AxesLabel- > x, y, Ticks- > Automatic,$
  $AxesOrigin- > 0, 0]$

Observe that, Chebyshev nodes produces a good approximation and Lagrange method with equally spaced points seems to diverge.

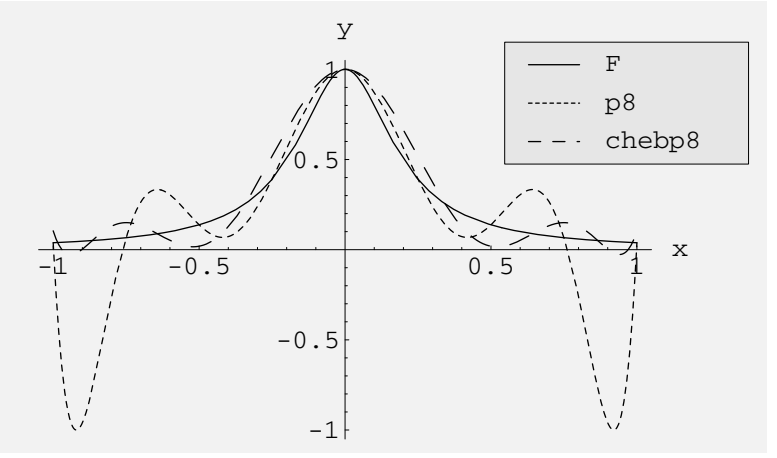


Fig. 1. Plot of Runge function together with p8, and chebp8

3 Cubic Splines

3.1 Background

As shown above, high degree interpolating polynomials may produce divergent approximations. To alleviate these difficulties, piecewise polynomial interpolation is provided [4], [5].

In the sequel, a natural cubic spline over two subintervals using 3 data points will be derived symbolically using Mathematica.

3.2 Session: Deriving Cubic Splines with Mathematica

- Consider a set of three points $(t_i, y_i), i = 1, 2, 3$. The required natural cubic spline is defined by two separate cubic polynomials p1 and p2 in $[t_1, t_2]$ and $[t_2, t_3]$.
  `In[1]:=` $p1[t_] := a + bt + ct^2 + dt^3$, `In[2]:=` $p2[t_] := e + ft + gt^2 + ht^3$
- A total of 8 parameters (a,b,c,d,e,f,g, and h) are to be determined. Using the interpolation condition at the end points, we obtain 4 equations
  `In[3]:=` $eq1 = y1 == (a + bt[1] + ct[1]^2 + dt[1]^3)$
  `In[4]:=` $eq2 = y2 == (a + bt[2] + ct[2]^2 + dt[2]^3)$
  `In[5]:=` $eq3 = y3 == (e + ft[2] + gt[2]^2 + ht[2]^3)$
  `In[6]:=` $eq4 = y3 == (e + ft[3] + gt[3]^2 + ht[3]^3)$
- Requiring the first derivative to be continuous at t_2 gives an additional equation as follows:
  `In[7]:=` $eq5 = p1\prime[t[2]] == p2\prime[t[2]]$ `Out[7]=` $b + 2ct[2] + 3dt[2]^2 == f + 2gt[2] + 3ht[2]^2$
- Requiring the second derivative to be continuous at t_2 gives
  `In[8]:=` $eq6 = p1\prime\prime[t[2]] == p2\prime\prime[t[2]]$ `Out[8]=` $2c + 12d == 2g + 12h$

- Finally, by the conditions of natural cubic spline
  `In[9]:=` $eq7 = p1''[t[1]] == 0$ `Out[9]=` $2c + 6d$
  `In[10]:=` $eq8 = p2''[t[3]] == 0$ `Out[10]=` $2g + 18h$
- Cubic splines can be easily determined by solving the linear system of 8 equations in 8 unknowns using Mathematica's Solve function
  `In[11]:=` $sol = Simplify[Solve[eq1, eq2, eq3, eq4, eq5, eq6, eq7, eq8, \{a, b, c, d, e, f, g, h\}]]$
  `Out[11]=`$\{\{a- > 2y1-y2, b- > \frac{1}{2}(-y1+y3), c- > -\frac{3}{4}(y1-2y2+y3), d- > \frac{1}{4}(y1-2y2+y3), e- > 6y1-9y2+4y3, f- > \frac{1}{2}(-13y1+24y2-11y3), g- > \frac{9}{4}(y1-2y2+y3), h- > \frac{1}{4}(-y1+2y2-y3))\}\}$

4 Conclusions

In this paper, use of Mathematica software is demonstrated for teaching polynomial interpolation problem in an educational setting. It has been shown that, symbolic packages are quite effective in deriving the required formula, and even proving some simple constructive theorems as shown to prove the error term of the Lagrange interpolation for the linear case.

Our experience reveals that students learned the concepts better and deeper. At the same time, this approach and the symbolic power of Mathematica provided an interactive research environment for the students to develop new skills and ideas.

With this approach, students may grasp important and difficult concepts easily in a laboratory environment.

The symbolic power provided by Mathematica, has provided a platform to discuss the fundamental and difficult issues related to the interpolation problem and the cubic splines.

The authors are involved in the design of an interactive tool to put all the ideas together in an integrated way to support teaching of numerical methods.

References

1. Kaput, J.: Technology and Mathematics Education, in Handbooks of Research on Mathematics Teaching and Learning (Ed. Grouws, D.A.), MacMillan, New York (1992) 515-556.
2. Reiter, C.A.: Exploring Hermite Interpolation with Mathematica, Primus, 2, 2(1992) 173-182.
3. Mathews, J.H.: Numerical Methods For Computer Science, and Mathematics, Prentice-Hall International (1987).
4. De Boor, C.: A Practical Guide to Splines, Springer Verlag, (1978).
5. Heath, M.T.: Scientific Computing: An Introductory Survey, McGraw-Hill International Editions (1997).

Constant Weight Codes with Package CodingTheory.m in Mathematica

Igor Gashkov

Karlstad University, Department of Engineering Sciences, Physics and Mathematics 65188
Karlstad Sweden
Igor.Gachkov@kau.se

Abstract. The author offers the further development of package CodingTheory.m [1] in the direction of research of properties and parameters of Constant weight codes (lower and upper bounds) based on works [2], [3] and also using the Table of Constant Weight Binary Codes (online version, Neil J. A. Sloane: Home Page http://www.research.att.com/~njas/codes/Andw/) and the table of upper bounds on *A(n, d, w)* (which in many cases also gives lower bounds) maintained by Erik Agrell, Alexander Vardy and Kenneth Zeger, and which is an electronic supplement to their paper Upper bounds for constant-weight codes, http://www.s2.chalmers.se/~agrell/bounds/cw.html. The offered package allows to carry out the comparative analysis of parameters of new codes with classical upper bounds such as Johnson bound, Linear programming (Ph Delsarte) bound …, and also with already available classes of codes. As an example we consider a some construction of codes as union of two codes with parameters *(n, 2a, a + b)* and *(m, 2b, a + b),* that in some cases gives codes the best than codes obtained with use Juxtaposing .

1 Introduction

An *(n, d, w)* constant weight binary code is a binary code of length *n*, code distance *d* in which all code words have the same number, *w*, of "ones." We will denote the maximal possible size of (*n, d, w*) constant weight code by *A (n, d, w)*. The most important and interesting problem is finding the largest possible size *A (n, d, w)* of a (*n, d, w*) constant weight code (hereafter called optimal codes). The results of code searching used to be put in tables of optimal codes. The first lower bound appeared in 1977 in the book of MacWilliams and Sloane ([4], pp.684-691). A table of binary constant weight codes of length $n \le 28$ with explicit constructions for most of the 600 codes was presented in the encyclopedic work of Brouwer, Shearer, Sloane, and Smith [2]. Today Neil J. A. Sloane presents his table of constant weight codes online and performs continual updates. There is a companion table of upper bounds on *A(n, d, w)* (which in many cases also gives lower bounds) maintained by Erik Agrell, Alexander Vardy and Kenneth Zeger, and which is an electronic supplement to their paper Upper bounds for constant-weight codes. The fragment of Sloane's table of constant weight codes with parameters $20 \le n \le 70$, $d = 18$ and $10 \le w \le 16$ looks as

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 370–375, 2004.
© Springer-Verlag Berlin Heidelberg 2004

follows (the point means that the appropriate code is optimal code and index give the key till example “j” means Juxtaposing (see below))

TABLE d=18
Lower bounds on A(n,18,w)

n,w	10	11	12	13	14	15	16
20	2.	2	1	1	1	?	?
…	…	…	…	…	…	…	…
26	2.	2.	2.	2.	2	?	?
27	$3.^{j}$	$3.^{j}$	$3.^{j}$	$3.^{j}$	3	?	?
28	3.	3.	3.	$4.^{j}$	$4.^{j}$	?	?
...	...	...	...	...	...	...	...
69	$18.^{z18}$	?	?	?	?	?	?
70	$21.^{z18}$	?	?	?	?	?	?
n,w	**10**	**11**	**12**	**13**	**14**	15	16

2 The Package “Constant Weight Codes”

The package ” Constant Weight Codes” is a file written in MATHEMATICA and will be read into MATHEMATICA with the commands.

```
In[1] := <<ConstantWeightCodes`
```

The package consists of two parts: one part with Table of Lower bounds on *A(n, d, w)* based on the Sloane`s table with explanations

```
In[2] := LowerboundsCWC[[26,12,10]]
Out[2]=33
```

We can find information about this lower bound using command

```
In[3] := LowerboundsCWC[[26,12,10,information]]
Out[3]= K. J. Nurmela, M. K. Kaikkonen and P. R. J.
Östergård, New constant weight codes..., IEEE Trans.
Info. Theory 43 (1997), 1623-1630.
```

and the second part contains commands describing the classical upper bounds of constant weight codes. The information on the list of possible upper bounds can be received using the command ?*Bound*.

```
In[4] := ?*Bound*
Out[4]=
BergerBound JohnsonBound SecondJohnsonBound LPBound …
```

The complete information about a command is received by using the command
? *Name*.

```
In[5] : = ? LPBound
Out[5]= LPBound [n,d,w] give Linear programming upper
bound  ( Ph. Delsarte ) on the size for constant weight
code n-length of code, d- code distance, d is even
number and w- weight of the code.
```

We can see also as far as the code (see In[2], In[3], lower bound) differs from known upper bounds

```
In[6] : = FreimanBound[26,12,10] BergerBound[26,12,10]
Out[6]=   195                    189
```

If the number of code vectors on a code coincides with one of the upper bound this give, that the code is optimal code

```
In[7] : = OptimalCWC[[58,18,10]] OptimalCWC[[70,18,10]]
Out[7]=    ?                     21
```

We also have an opportunity to make changes in a package, in case of occurrence of new codes with parameters it is better known

```
In[8] : = LowerboundCWC[[58,18,10]]=12;
OptimalCWC[[58,18,10]]=12;

LowerboundCWC[[58,18,10,information]]= "Gashkov I. Some
optimal constant weight code, The Karlstad workshop on
Applied mathematics; September 7-11, 2003";

DeleteFile["LowerboundCWC","OptimalCWC"];Save[Lowerboun
dCWC", LowerboundCWC];Save["OptimalCWC", OptimalCWC];
```

Now we will show how, using MATHEMATICA, we can, base on Sloane`s table (see above) update this table. Trivial values we can obtain from well knows theorem 1. [2]

Theorem 1.

a) If d is odd, $A(n, d, w) = A(n, d+1, w)$.

b) $A(n, d, w) = A(n, d, n - w)$

c) $A(n, d, w) = 1$ if $2w < d$

d) If $d = 2w$ then $A(n, d, w) = [n/w]$, where $[n/w]$ is greatest integer less then or equal n/w.

e) $A(n, 2, w) = \binom{n}{w}$.

By juxtaposing two codes (the method of construction a new codes by placing them side by side) we can obtain lower bound $A(n1 + n2, d1 + d2, w1 + w2) \geq min\{A(n1, d1, w1), A(n2, d2, w2)\}$.Codes obtain with use juxtaposing we shall denote through $m(n1, d1, w1) + (n2, d2, w2)$ where m is $min\{A(n1, d1, w1), A(n2, d2, w2)\}$ .The command Juxtaposing[…] changes the table of codes supplementing with its codes obtained above by a stated method, in a case if the "j" code has the greater or equal

number of code vectors, than already available in the table. As example we chose the parameters of the table $20 \le n \le 70$, $d = 18$ and $10 \le w \le 16$.

```
In[9]:= LowerboundsCWC = Juxtaposing[LowerboundsCWC];
```

We chose the parameters of the table 20 ≤ n ≤ 70, d = 18 and 10 ≤ w ≤16.

```
In[10]:= n1=20;n2=20;w1=10;w2=16;d=18;
Table[Take[LowerboundsCWC[[n,d]],{w1,w2}],{n,n1,n2}]
Out[10]=
{{2(2,2,1)+(18,16,9)   2(2,2,1)+(18,16,10)   1   ...   1},
 {2(2,2,1)+(19,16,9)   2(2,2,1)+(19,16,10)   ...   ...   1},
 {      ...                    ...            ...   ...   ...},
 {   18 Optimal        42 (24,6,4)+(45,12,7)   ...   ...   ...}}
```

We can give the next generation of this table of lower bounds, using the following theorem.

Theorem 2. If $A(n1, 2a, a + b) = M1$ and $A(n2, 2b, a + b) = M2$ then $A(n1 + n2, 2(a + b), \min\{n1, n2\}) \ge \min\{M1, M2\}+1$

Proof. Let $n1 \ge n2$. We placing two codes $(n1, 2a, a + b)$ and $(n2, 2b, n2 - a - b)$ side-by-side. Using Juxtaposing we obtain code with parameters $(n1+ n2, 2(a + b), n2)$. But we can add to code one binary vector more $v = (0,0,\dots 0, 1,1,\dots 1)$, where the first $n1$ coordinates are equal 0, the last $n2$ equal 1. We can calculate distance between any vectors belonging to code $(n1+ n2, 2(a + b), n2)$ and vector v. We can calculate this distance as the sum of two distances: the first distance is the distance between an any vector of a code $(n1, 2a, a + b)$ and a zero vector, the second is the distance between an any vector of a code $(n2, 2b, n2 - a - b)$ and a "one " vector (vector consisting of 1).

Example: Codes with parameters $(6, 4, 3)$ and $(4, 2, 3)$ consist of the same number of vectors.

Code (6, 4, 3)	Code (4, 2, 3)	Code (10, 6, 4)
V1= (1, 1, 1, 0, 0, 0)	(1, 1, 1, 0)= W1	(1, 1, 1, 0, 0, 0, 0, 0, 0, 1)
V2= (0, 0, 1, 1, 1, 0)	(1, 1, 0, 1)=W2	(0, 0, 1, 1, 1, 0, 0, 0, 1, 0)
V3= (0, 1, 0, 1, 0, 1)	(0, 1, 1, 1)=W3	(0, 1, 0, 1, 0, 1, 1, 0, 0, 0)
V4= (1, 0, 0, 0, 1, 1)	(1, 0, 1, 1)=W4	(1, 0, 0, 0, 1, 1, 0, 1, 0, 0)
		(0, 0, 0, 0, 0,0, 1, 1, 1, 1)

We shall denote such codes through $m\ (n1, 2a, a + b) \cup (n2, 2b, a + b)$ where m is $\min\{M1, M2\}+1$. We can transform the table of constant weight codes taking into

account the theorem 2 and we obtain a new codes

```
(13,6,8)∪(20,10,8)->A(33,16,13)≥18
(13,6,8)∪(21,10,8)->A(34,16,13)≥19
(16,8,9)∪(19,10,9)->A(35,18,16)≥17
(13,6,9)∪(23,12,9)->A(36,18,13)≥11
```

$(13,6,9)\cup(24,12,9)\to A(37,18,13)\geq 14$

$(8,2,6)\cup(31,10,6)\to A(39,12,8)\geq 29$

$(14,6,9)\cup(25,12,9)\to A(39,18,14)\geq 26$

$(14,6,9)\cup(26,12,9)\to A(40,18,14)\geq 27$

$(14,6,9)\cup(27,12,9)\to A(41,18,14)\geq 29$

$(11,2,8)\cup(57,14,8)\to A(68,16,11)\geq 58$

Above the carried out computer calculations prove, that codes that we obtain using the theorem 2 it is impossible to present as juxtaposing two codes.

2 (2,2,1) +(18,16,9)	2 (2,2,1) +(18,16,10)	1	1	1	1	1	
2 (2,2,1) +(19,16,9)	2 (2,2,1) +(19,16,10)	2 (2,2,1) +(19,16,11)	1	1	1	1	
2 (2,2,1) +(20,16,9)	2 (2,2,1) +(20,16,10)	2 (2,2,1) +(20,16,11)	2 (2,2,1) +(20,16,12)	1	1	1	
2 (2,2,1) +(21,16,9)	2 (2,2,1) +(21,16,10)	2 (2,2,1) +(21,16,11)	2 (2,2,1) +(21,16,12)	2 (2,2,1) +(21,16,13)	1	1	
2 (2,2,1) +(22,16,9)	2 (2,2,1) +(22,16,10)	2 (2,2,1) +(22,16,11)	2 (2,2,1) +(22,16,12)	2 (2,2,1) +(22,16,13)	2 (2,2,1) +(22,16,14)	1	
2 (2,2,1) +(23,16,9)	2 (2,2,1) +(23,16,10)	2 (2,2,1) +(23,16,11)	2 (2,2,1) +(23,16,12)	2 (2,2,1) +(23,16,13)	2 (2,2,1) +(23,16,14)	2 (2,2,1) +(23,16,15)	
2 (2,2,1) +(24,16,9)	2 (2,2,1) +(24,16,10)	2 (2,2,1) +(24,16,11)	2 (2,2,1) +(24,16,12)	2 (2,2,1) +(24,16,13)	2 (2,2,1) +(24,16,14)	2 (2,2,1) +(24,16,15)	
3 (3,2,1) +(24,16,9)	3 (3,2,1) +(24,16,10)	3 (3,2,1) +(24,16,11)	3 (3,2,1) +(24,16,12)	3 (3,2,1) +(24,16,13)	3 (3,2,1) +(24,16,14)	3 (3,2,1) +(24,16,15)	
3 (3,2,1) +(25,16,9)	3 (3,2,1) +(25,16,10)	3 (3,2,1) +(25,16,11)	4 (4,2,1) +(24,16,12)	4 (4,2,2) +(24,16,12)	4 (4,2,3) +(24,16,12)	3 (3,2,1) +(25,16,15)	
3 (3,2,1) +(26,16,9)	3 (3,2,1) +(26,16,10)	3 (3,2,1) +(26,16,11)	4 (4,2,1) +(25,16,12)	4 (4,2,1) +(25,16,13)	4 (4,2,2) +(25,16,13)	4 (4,2,3) +(25,16,13)	
3 (3,2,1) +(27,16,9)	3 (3,2,1) +(27,16,10)	5 (10,6,4) +(20,12,8)	5 (10,6,4) +(20,12,9)	5 (10,6,4) +(20,12,10)	6 (10,6,5) +(20,12,10)	5 (10,6,4) +(20,12,12)	
3 (3,2,1) +(28,16,9)	3 (3,2,1) +(28,16,10)	5 (10,6,4) +(21,12,8)	5 (4,2,2) +(27,16,11)	6 (7,4,3) +(24,14,11)	6 (4,2,2) +(27,16,13)	6 (4,2,2) +(27,16,14)	
3 (3,2,1) +(29,16,9)	4 (4,2,1) +(28,16,10)	5 (5,2,1) +(27,16,11)	6 (7,4,3) +(25,14,10)	7 (7,4,3) +(25,14,11)	8 (14,8,7) +(18,10,8)	8 (14,8,7) +(18,10,9)	
3 (3,2,1) +(30,16,9)	4 (4,2,1) +(29,16,10)	6 (7,4,3) +(26,14,9)	7 (7,4,3) +(26,14,10)	9 (15,8,6) +(18,10,8)	11 (11,6,5) +(22,12,10)	11 (11,6,5) +(22,12,11)	
4 (4,2,1) +(30,16,9)	4 (4,2,1) +(30,16,10)	6 (4,2,2) +(30,16,10)	8 (8,4,3) +(26,14,10)	10 (11,6,5) +(23,12,9)	12 (15,8,7) +(19,10,8)	15 (15,8,7) +(19,10,9)	
4 (4,2,1) +(31,16,9)	5 (5,2,1) +(30,16,10)	7 (7,4,3) +(28,14,9)	10 (15,8,6) +(20,10,7)	12 (16,8,6) +(19,10,8)	16 (16,8,6) +(19,10,9)	17 New(16,8,9)|(19,10,9	
4 (4,2,1) +(32,16,9)	6 (6,2,1) +(30,16,10)	9 (12,6,4) +(24,12,8)	11 New(13,6,9)|(23,12,9)	16 (16,8,6) +(20,10,8)	17 (17,8,6) +(19,10,9)	22 (12,6,6) +(24,12,10)	
4 (4,2,1) +(33,16,9)	6 (6,2,1) +(31,16,10)	9 (12,6,4) +(25,12,8)	14 New(13,6,9)|(24,12,9)	17 (17,8,6) +(20,10,8)	22 (12,6,6) +(25,12,9)	24 (13,6,6) +(24,12,10)	
4 (4,2,1) +(34,16,9)	7 (17,8,5) +(21,10,6)	10 (13,6,4) +(25,12,8)	16 (16,8,6) +(22,10,7)	18 (13,6,5) +(25,12,9)	25 (13,6,6) +(25,12,9)	28 (10,4,4) +(28,14,12)	
4 (4,2,1) +(35,16,9)	7 (17,8,5) +(22,10,6)	13 (13,6,4) +(26,12,8)	16 (16,8,6) +(23,10,7)	26 New(14,6,9)|(25,12,9)	28 (14,6,5) +(25,12,10)	30 (16,8,8) +(23,10,8)	
5 (5,2,1) +(35,16,9)	8 (12,6,4) +(28,12,7)	13 (13,6,4) +(27,12,8)	17 (17,8,6) +(23,10,7)	27 New(14,6,9)|(26,12,9)	28 (14,6,5) +(26,12,10)	38 (20,8,6) +(20,10,10)	
5 (5,2,1) +(36,16,9)	8 (13,6,4) +(28,12,7)	14 (14,6,4) +(27,12,8)	20 (16,6,4) +(25,12,9)	29 New(14,6,9)|(27,12,9)	39 (14,6,6) +(27,12,9)	39 (14,6,6) +(27,12,10)	
5 (5,2,1) +(37,16,9)	9 (18,8,5) +(24,10,6)	16 (20,8,5) +(22,10,7)	21 (18,8,6) +(24,10,7)	39 (15,6,5) +(27,12,9)	39 (14,6,6) +(28,12,9)	45 (19,8,7) +(23,10,9)	
5 (5,2,1) +(38,16,9)	9 (12,6,4) +(31,12,7)	16 (20,8,5) +(23,10,7)	24 (19,8,6) +(24,10,7)	39 (15,6,5) +(28,12,9)	42 (15,6,5) +(28,12,10)	52 (19,8,7) +(24,10,9)	
5 (4,2,2) +(40,16,8)	10 (19,8,5) +(25,10,6)	20 (21,8,5) +(23,10,7)	28 (19,8,6) +(25,10,7)	39 (16,6,5) +(28,12,9)	48 (16,6,5) +(28,12,10)	56 (20,8,7) +(24,10,9)	
6 (6,2,1) +(39,16,9)	12 (19,8,5) +(26,10,6)	21 (21,8,5) +(24,10,7)	28 (19,8,6) +(26,10,7)	40 (20,8,6) +(25,10,8)	56 (21,8,6) +(24,10,9)	72 (20,8,7) +(25,10,9)	
6 (6,2,1) +(40,16,9)	13 (20,8,5) +(26,10,6)	21 (21,8,5) +(25,10,7)	28 (19,8,6) +(27,10,7)	48 (21,8,6) +(25,10,8)	56 (21,8,6) +(25,10,9)	80 (20,8,7) +(26,10,9)	
6 (6,2,1) +(41,16,9)	14 (20,8,5) +(27,10,6)	23 (23,8,5) +(24,10,7)	36 (20,8,6) +(27,10,7)	54 (21,8,6) +(26,10,8)	72 (22,8,6) +(25,10,9)	91 (21,8,7) +(26,10,9)	
6 (6,2,1) +(42,16,9)	16 (20,8,5) +(28,10,6)	24 (24,8,5) +(24,10,7)	37 (20,8,6) +(28,10,7)	56 (21,8,6) +(27,10,8)	78 (20,8,7) +(28,10,8)	118 (21,8,7) +(27,10,9)	
7 (7,2,1) +(42,16,9)	16 (20,8,5) +(29,10,6)	25 (19,8,6) +(30,10,6)	37 (21,8,6) +(28,10,7)	66 (22,8,6) +(27,10,8)	78 (21,8,7) +(28,10,8)	120 (21,8,7) +(28,10,9)	
7 (7,2,1) +(43,16,9)	20 (21,8,5) +(29,10,6)	28 (19,8,6) +(31,10,6)	39 (20,8,6) +(30,10,7)	77 (22,8,6) +(28,10,8)	78 (22,8,7) +(28,10,8)	132 (22,8,7) +(28,10,9)	
7 (7,2,1) +(44,16,9)	21 (21,8,5) +(30,10,6)	31 (20,8,6) +(31,10,6)	40 (20,8,6) +(31,10,7)	77 (23,8,6) +(28,10,8)	91 (25,8,6) +(26,10,9)	132 (23,8,7) +(28,10,9)	
8 (8,2,1) +(44,16,9)	21 (21,8,5) +(31,10,6)	31 (20,8,6) +(32,10,6)	42 (21,8,6) +(31,10,7)	78 (24,8,6) +(28,10,8)	100 (25,8,6) +(27,10,9)	132 (24,8,7) +(28,10,9)	
8 (8,2,1) +(45,16,9)	23 (23,8,5) +(30,10,6)	31 (21,8,6) +(32,10,6)	42 (22,8,6) +(31,10,7)	78 (25,8,6) +(28,10,8)	118 (26,8,6) +(27,10,9)	132 (25,8,7) +(28,10,9)	
9 (9,2,1) +(45,16,9)	24 (24,8,5) +(30,10,6)	31 (22,8,6) +(32,10,6)	42 (23,8,6) +(31,10,7)	78 (26,8,6) +(28,10,8)	130 (26,8,6) +(28,10,9)	167 (22,8,7) +(32,10,9)	
11 Optimal z17	25 (25,8,5) +(30,10,6)	35 (20,8,6) +(35,10,6)	42 (24,8,6) +(31,10,7)	78 (27,8,6) +(28,10,8)	130 (27,8,6) +(28,10,9)	176 (22,8,7) +(33,10,9)	
?	30 (25,8,5) +(31,10,6)	35 (21,8,6) +(35,10,6)	42 (25,8,6) +(31,10,7)	78 (28,8,6) +(28,10,8)	130 (28,8,6) +(28,10,9)	198 (23,8,7) +(33,10,9)	
?	30 (25,8,5) +(32,10,6)	35 (22,8,6) +(35,10,6)	42 (26,8,6) +(31,10,7)	?	?	232 (23,8,7) +(34,10,9)	
?	31 (27,8,5) +(31,10,6)	35 (23,8,6) +(35,10,6)	42 (27,8,6) +(31,10,7)	?	130 (26,8,6) +(32,10,9)	253 (23,8,7) +(35,10,9)	
?	31 (27,8,5) +(32,10,6)	38 (15,6,5) +(44,12,7)	42 (14,6,6) +(45,12,7)	?	130 (26,8,6) +(33,10,9)	253 (23,8,7) +(36,10,9)	
?	31 (28,8,5) +(32,10,6)	42 (15,6,5) +(45,12,7)	45 (15,6,6) +(45,12,7)	?	130 (26,8,6) +(34,10,9)	254 (25,8,7) +(35,10,9)	
?	31 (29,8,5) +(32,10,6)	45 (16,6,5) +(45,12,7)	45 (16,6,6) +(45,12,7)	78 (13,2,2) +(48,16,12)	130 (26,8,6) +(35,10,9)	257 (26,8,7) +(35,10,9)	
?	31 (27,8,5) +(35,10,6)	45 (17,6,5) +(45,12,7)	45 (17,6,6) +(45,12,7)	78 (13,2,2) +(49,16,12)	130 (26,8,6) +(36,10,9)	273 (27,8,7) +(35,10,9)	
?	33 (28,8,5) +(35,10,6)	45 (18,6,5) +(45,12,7)	48 (15,6,6) +(48,12,7)	78 (13,2,2) +(50,16,12)	130 (11,2,3) +(52,16,12)	278 (27,8,7) +(36,10,9)	
?	34 (29,8,5) +(35,10,6)	48 (16,6,5) +(48,12,7)	56 (15,6,6) +(49,12,7)	91 (14,2,2) +(50,16,12)	137 (32,8,6) +(32,10,9)	296 (28,8,7) +(36,10,9)	
?	35 (30,8,5) +(35,10,6)	48 (16,6,5) +(49,12,7)	56 (16,6,6) +(49,12,7)	100 (15,2,2) +(50,16,12)	137 (32,8,6) +(33,10,9)	300 (29,8,7) +(36,10,9)	
?	37 (22,6,4) +(44,12,7)	56 (17,6,5) +(49,12,7)	56 (17,6,6) +(49,12,7)	100 (15,2,2) +(51,16,12)	150 (33,8,6) +(33,10,9)	320 (30,8,7) +(36,10,9)	
11 (11,2,1) +(56,16,9)	38 (23,6,4) +(44,12,7)	56 (18,6,5) +(49,12,7)	56 (18,6,6) +(49,12,7)	105 (15,2,2) +(52,16,12)	150 (33,8,6) +(34,10,9)	327 (30,8,7) +(37,10,9)	
12 (12,2,1) +(56,16,9)	40 (23,6,4) +(45,12,7)	56 (19,6,5) +(49,12,7)	57 (11,4,5) +(57,14,8)	120 (16,2,2) +(52,16,12)	163 (34,8,6) +(34,10,9)	362 (31,8,7) +(37,10,9)	
18 Optimal z18	42 (24,6,4) +(45,12,7)	56 (20,6,5) +(49,12,7)	57 (12,4,5) +(57,14,8)	130 (17,2,2) +(52,16,12)	163 (34,8,6) +(35,10,9)	372 (32,8,7) +(37,10,9)	
21 Optimal z18	45 (25,6,4) +(45,12,7)	57 (13,4,4) +(57,14,8)	60 (22,8,6) +(48,10,7)	130 (17,2,2) +(53,16,12)	178 (35,8,6) +(35,10,9)	403 (32,8,7) +(38,10,9)	

In particular, codes that were obtain using the theorem 2 is close to optimal codes. For instance

$$28 \geq A(35,18,16) \geq 17\ ;\ 12 \geq A(36,18,13) \geq 11$$
$$15 \geq A(37,18,13) \geq 14\ ;\ 39 \geq A(39,18,14) \geq 26$$

References

1. I Gachkov “Error Correcting codes with Mathematica”, Lecture note in Computer science LNCS 2657 s.737-746 (2003)
2. E. Brouwer, J.B. Shearer, N.J.A .Sloane, “A new table of Constant weight codes”. IEEE Transactions of information theory, v 36, No 6 (1990)
3. E.Agrell,A. Vardy, and K. Zeger,“Upper bounds for constant-weight codes,” IEEE Transactions on Information Theory,vol. 46, no. 7, pp. 2373-2395, Nov. 2000.
4. F.J. MacWilliams and N.J.A. Sloane, “The Theory of Error-Correcting Codes”, Amsterdam: North-Holland, 1977

Graph Coloring with web*Mathematica*

Ünal Ufuktepe, Goksen Bacak, and Tina Beseri

Izmir Institute of Technology, Department of Mathematics
Urla, Izmir, TURKEY
{unalufuktepe,goksenbacak,tinabeseri}@iyte.edu.tr

Abstract. Coloring of a graph is an assignment of colors either to the edges of the graph G, or to vertices, or to maps in such a way that adjacent edges/vertices/maps are colored differently. We consider the problem of coloring graphs by using web*Mathematica* which is the new web-based technology. In this paper, we describe some web-based interactive examples on graph coloring with web*Mathematica*.

1 Introduction

A graph G=(V,E) is a mathematical structure consisting of two sets V and E. The elements of V are called vertices and the elements of E are called edges.

web*Mathematica* is based on a standard Java technology called servlets. It allows a site to deliver HTML pages that are enhanced by the addition of *Mathematica* commands. When a request is made for one of these pages the *Mathematica* commands are evaluated and the computed result is placed in the page. This is done with the standard Java templating mechanism, Java Server Pages (JSPs), making use of a library of tag extensions called the MSP Taglib; examples of these for graph coloring are given in a later section [8]. We developed some modules and used Combinatorica package [5] to color the graphs by webMathematica.

1.1 Edge Coloring

Edge coloring is an optimization problem: Given a graph, how many colors are required to color its edges in such a way that no two edges which share an endpoint receive the same color? A k-edge coloring with k colors G is k-edge colorable if a k-edge colors exists. The smallest k for which G is k-edge colorable is called the edge-coloring number of G.

Definition 1. *The edge-coloring number of G is called the chromatic index of G and is denoted by* $\chi'(G)$.

$\Delta(G)$ is the maximum vertex degree in G. An obvious lower bound for $\chi'(G)$ is $\Delta(G)$ since the edges incident with one vertex must be differently colored. It follows that $\chi'(G) \geq \Delta(G)$. On the other hand, Vizing has proved in 1964 that any simple graph G has an edge coloring with at most $\Delta(G) + 1$ colors:

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 376–381, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Proposition 1 (Vizing, 1964). *Every graph G satisfies,*

$$\Delta(G) \leq \chi'(G) \leq \Delta(G) + 1$$

1.2 Vertex Coloring

The most applications involving vertex-coloring are concerned with determining the minimum number of colors required under the condition that the end points of an edge can't have the same color. A proper vertex-coloring of a graph is an assignment from its vertex set to a color set such that the endpoints of each edge are assigned two different colors. The chromatic number of a graph G, denoted by $\chi(G)$, is the minimum number of different colors required for a proper vertex coloring of G. Thus, $\chi(G) = k$ if graph G is k-colorable but not $(k-1)$-colorable. A graph G has $\chi(G) = 1$ if and only if G has no edges. A graph G has $\chi(G) = 2$ if and only if it is bipartite. The study of vertex coloring of graphs is customarily restricted to simple graphs. A graph with a self-loop is regarded as uncolorable, since the endpoint of the self-loop is adjacent to itself. Moreover, a multiple adjacency has no more effect on the colors of its endpoints than a single adjacency. Although the chromatic number is one of the most studied parameters in graph theory, no formula exists for the chromatic number of an arbitrary graph. Thus, we must try to find bounds for the chromatic number.

1.3 Map Coloring

A map on a surface is an imbedding of a graph on that surface. An k-coloring of a map is a coloring with k-colors. A map coloring is proper if each $e \in E_G$, the regions that meet an edge e are colored differently. The chromatic number of a map is the minimum number of colors needed for a proper coloring.[4,7]

The chromatic number of a map equals the chromatic number of its dual graph.

2 Applications with web*Mathematica*

To understand what web*Mathematica* can do regarding coloring planner graph and maps we need to know Combinatorica, the standard package that has many functions for dealing graphs.

We mainly use this package by adding the following ColorVertices, ColorEdges, and DrawG modules:

```
ColorVertices[g_] := Module[{c, p, s},
c = VertexColoring[g];
p = Table[Flatten[Position[c, i]], {i, 1, Max[c]}];
s = ShowLabeledGraph[ Highlight[g, p]]]
```

The module ColorVertices colors vertices of the given graph g.

```
ColorEdges[g_] := Module[{k, e, m, kk, r, s},
k = EdgeColoring[g]; e = Edges[g];
m = Max[k]; kk = Table[i, {i, 1, 1000}];
r = For[i = 1, i <= m ,
For[j = 1, j <= M[g], If[k[[j]] == i,
kk[[j]] = Hue[i/m]]; j++]; i++] ;
s = ShowGraph[g, Table[{e[[i]], EdgeColor -> kk[[i]]}, {i, 1,
M[g]}],
VertexNumber -> On, EdgeStyle -> Thick]]
```

The module ColorEdges colors vertices of the given graph g.

```
DrawG[elist_]:=Module[{edgelist=elist,size,vertlist,vnum},
size=Length[edgelist];
vertlist=Union[Flatten[edgelist]];
vnum=Length[vertlist];
Do[edgelist[[i]]={edgelist[[i]]},{i,size}];
vertlist=CompleteGraph[vnum][[2]];
Graph[edgelist,vertlist]]
```

The module DrawG draw the simple graph without isolated points. DrawG takes as input the list of edges of a graph. The vertices of the graph of order n must be labeled consecutively $1, 2, 3, \ldots, n$. This module must be added to the package DiscreteMath`Combinatorica`.

The package Combinatorica must be loaded before running the program.

webMathematica allows the generation of dynamic web content with Mathematica. The following example draws the given graph and colors edges and vertices

```
<%@ page language="java" %>
<%@ taglib uri="/webMathematica-taglib" prefix="msp" %>
<html><head> <title>Graph Coloring </title>
</head> <body bgcolor="# ffffff">
<h1>Graph Coloring </h1>
<FORM ACTION="graphcolor.jsp" METHOD="POST">
Input the list of the edges in order as follows:
<msp:allocateKernel>
<INPUT type="text" name="v" ALIGN="LEFT" size="50"
Value ="<msp:evaluate> MSPValue[ $$v, "{{1,4},{2,3},
{2,4},{3,4}}"]</msp:evaluate>" />
<msp:evaluate> <<DiscreteMath`Combinatorica`
<<Graphics`Colors`
</msp:evaluate>
<h3>Vertex Coloring and Edge coloring are </h3><p>
```

```
<msp:evaluate> input=True;
If[MSPValueQ[$$v],
n=MSPToExpression[$$v]; g=DrawG[n];
MSPShow[ColorVertices[g]],input=False;]
</msp:evaluate>
<msp:evaluate>
If[MSPValueQ[$$v],
MSPShow[ColorEdges[g]],input=False;]
</msp:evaluate> </p> <p>


   The Chromatic Number of the given graph is
<msp:evaluate>
kn=ChromaticNumber[g]
</msp:evaluate></p>
<INPUT TYPE="Hidden" NAME="formNo" VALUE="1">
<INPUT TYPE="Submit"NAME="taskValue"
VALUE="Color the graph's edges and vertices" >
</msp:allocateKernel>
</form></body></html>
```

A form element is a block of HTML that may contain input elements. A form may be activated with an input of type submit. The action attribute refers to a URL that accessed when the form is activated. The method attribute tells the browser what HTTP method to use, in this case, a post method. This example has two input tags. The first allows the user of the page to enter the list of edges of the graph, and the second specifies a button that, when pressed, will submit the form. When the form is submitted, it will send information from input elements to the URL specified by the action attribute. This information is sent to a Mathematica kernel and assigned to a Mathematica symbol (see Fig.1). The name of the symbol is given by $$ to the value of the name attribute. When a value entered in the text field and the "Color the graph's edges and vertices" button pressed,the text is displayed. This example also shows the use of the MSP functions MSPShow, MSPValue, and MSPToExpression.

To color edges and vertices of the standard graphs (Complete graph, wheel, tree, cycle, and the others) we need more inputs. If the web user selects one of the standard graphs and its number of vertices then they can get easily colored graph and its chromatic numbers as follows:

```
   <FORM ACTION="familycolor.jsp" METHOD="POST">
<p>Please select one of the following graphs and
input into the box:(1,2,3,or 4) </p> <p>1-CompleteGraph,</p>
<p>2-RandomTree, </p> <p>3-Wheel,</p> <p>4-Cycle)</p>
<msp:allocateKernel>
```

```
<INPUT type="text" name="m" ALIGN="LEFT" size="6"
value="<msp:evaluate> MSPValue[$$m,"1"]</msp:evaluate>" />
Input the number of the vertices for the selected graph:
<INPUT type="text" name="n" ALIGN="LEFT" size="6"
value="<msp:evaluate> MSPValue[$$n,"5"]</msp:evaluate>" />
<msp:evaluate> <<DiscreteMath`Combinatorica`
<<Graphics`Colors`</msp:evaluate>
<h3>Vertex Coloring and Edge Coloring are </h3>
<msp:evaluate> MSPBlock[{$$m,$$n},
Which[$$m==1,MSPShow[ColorVertices[CompleteGraph[$$n]]],
$$m==2, MSPShow[ColorVertices[a=RandomTree[$$n]]],
$$m==3, MSPShow[ColorVertices[Wheel[$$n]]],
$$m==4, MSPShow[ColorVertices[Cycle[$$n]]]]]
</msp:evaluate>
<msp:evaluate> MSPBlock[{$$m,$$n},
Which[$$m==1, MSPShow[ColorEdges[CompleteGraph[$$n]]],
$$m==2, MSPShow[ColorEdges[a]],
$$m==3, MSPShow[ColorEdges[Wheel[$$n]]],
$$m==4, MSPShow[ColorEdges[Cycle[$$n]]]]]
</msp:evaluate>
The Chromatic Number of the selected graph is <msp:evaluate>
MSPBlock[{$$m,$$n},
Which[$$m==1, ChromaticNumber[CompleteGraph[$$n]],
$$m==2, ChromaticNumber[a],
$$m==3, ChromaticNumber[Wheel[$$n]],
$$m==4, ChromaticNumber[Cycle[$$n]]]]
</msp:evaluate>
<p> The Edge Chromatic Number of the selected graph is
<msp:evaluate>
MSPBlock[{$$m,$$n},
Which[$$m==1, EdgeChromaticNumber[CompleteGraph[$$n]],
$$m==2, EdgeChromaticNumber[a],
$$m==3, EdgeChromaticNumber[Wheel[$$n]],
$$m==4, EdgeChromaticNumber[Cycle[$$n]]]]
</msp:evaluate> </p> <input type="submit" name="button"
value="Graph Color"> </msp:allocateKernel>
```

We need one more input tag to color edges and vertices of the bipartite graph one in this script. Combinatorica package uses Brelaz's heuristic to find a good, but not necessarily minimal, edge coloring of graph G. Then when the web user enters the huge number of the vertices he/she might get interesting results or time out error for the edge coloring. We also tried to color maps by using GraphColoring package of Stan Wagon and J.P.Hutchinson with webMathematica but since there are some version problems we left it for the future works [4,7].

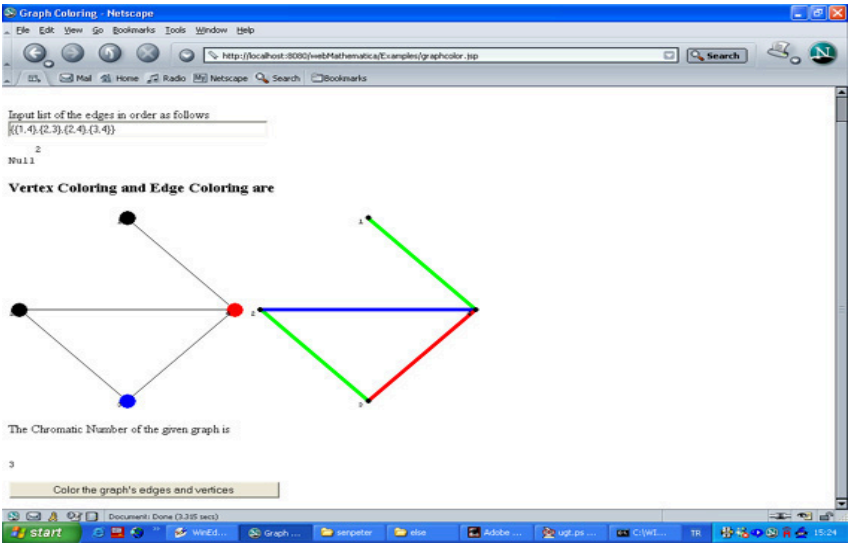


Fig. 1. A view of graphcolor.jsp after information entered by the user

Please check
http://gauss.iyte.edu.tr:8080/webMathematica/Examples/graphcolor.jsp
http://gauss.iyte.edu.tr:8080/webMathematica/Examples/familycolor.jsp
for the details of these JSP scripts.

References

1. Appel,K.I., Haken,W., and Koch,J.: Every planar map is four colorable I: Discharing, Illinois J.Math, (1977)
2. Birkhoff,G.D. and Lewis,D.C.: Chromatic polinomials, Trans.Amer.Math.Soc., 60:355-451, (1946)
3. Gross,J. and Yellen,J.: Graph Theory and Its Applications, CRC Press, (1999)
4. Hutchinson,J.P. and Wagon,S.: The four-color theorem, Mathematica in Education and Research, 42-51, 6:1 (1997)
5. Skiena,S.: Implementing Discrete Mathematics-Combinatorics and Graph Theory with Mathematica, Addison-Wesley Publishing Company, (1990)
6. Soaty,T.L. and Kainen,P.C.: The four color problem, Dover, New York, (1986)
7. Wagon,S.: An April Fool's hoax, Mathematica in Education and Research, 46-52, 7:1 (1998)
8. Wickham,T.: webMathematica A User Guide, Wolfram Research, Inc., (2002)
9. Wolfram,S.: The Mathematica Book, Cambrigde Univ. Press, (1996)

Construction of Solutions for Nonintegrable Systems with the Help of the Painlevé Test

S.Y. Vernov

Skobeltsyn Institute of Nuclear Physics, Moscow State University,
Vorob'evy Gory, Moscow, 119992, Russia.
svernov@theory.sinp.msu.ru

Abstract. The generalized Hénon–Heiles system with an additional nonpolynomial term has been considered. In two nonintegrable cases with the help of the Painlevé test new special solutions have been found as converging Laurent series, depending on three parameters. For some values of these parameters the obtained Laurent series coincide with the Laurent series of the known elliptic solutions. The calculations have been made with use of computer algebra system REDUCE. The obtained local solutions can assist to find the elliptic three parameters solutions. The corresponding algorithm has been realized in REDUCE and Maple.

1 The Painlevé Test

When we study some mechanical problem time is assumed to be real, whereas the integrability of motion equations is connected with the behavior of their solutions as functions of complex time. Solutions of a system of ODE's are regarded as analytic functions, maybe with isolated singular points. A singular point of a solution is said *critical* (as opposed to *noncritical*) if the solution is multivalued (single-valued) in its neighborhood and *movable* if its location depends on initial conditions. *The general solution* of an ODE of order N is the set of all solutions mentioned in the existence theorem of Cauchy, i.e. determined by the initial values. It depends on N arbitrary independent constants. *A special solution* is any solution obtained from the general solution by giving values to the arbitrary constants. *A singular solution* is any solution which is not special, i.e. which does not belong to the general solution. A system of ODE's has ***the Painlevé property*** if its general solution has no movable critical singularity point [1].

The Painlevé test is any algorithm, which checks some necessary conditions for a differential equation to have the Painlevé property. The original algorithm, developed by P. Painlevé and used by him to find all the second order ODE's with Painlevé property, is known as the α-method. The method of S.V. Kovalevskaya [2] is not as general as the α–method, but much more simple. The remarkable property of this test is that it can be checked in a finite number of steps. In 1980, motivated by the work of S.V. Kovalevskaya [2], M.J. Ablowitz, A. Ramani and H. Segur [3] developed a new algorithm of the Painlevé test for ODE's. This algorithm appears very useful to find solutions as a formal

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 382–387, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Laurent series. First of all, it allows to determine the dominant behavior of a solution in the neighborhood of the singularity point t_0. If the solution tends to infinity as $(t-t_0)^\beta$, where β is a negative integer number, then substituting the Laurent series expansions one can transform nonlinear differential equations into a system of linear algebraic equations on coefficients of the Laurent series. If a single-valued solution of autonomous system depends on not only the parameter t_0, which characterizes the syngularity point location, but also on other parameters, then some coefficients of its Laurent series have to be arbitrary and the corresponding systems have to have zero determinants. The numbers of such systems (named *resonances* or *Kovalevskaya exponents*) can be determined due to the Painlevé test. In such a way we obtain solutions only as formal series, but for some nonintegrable systems, for example, the generalized Hénon–Heiles system [4], the convergence of the Laurent- and psi-series solutions has been proved. Such solutions also assist to find the elliptic solutions [5].

2 The Hénon–Heiles Hamiltonian

Let us consider the generalized Hénon–Heiles system with an additional nonpolynomial term, which is described by the Hamiltonian:

$$H = \frac{1}{2}\Big(x_t^2 + y_t^2 + \lambda x^2 + y^2\Big) + x^2 y - \frac{C}{3}\,y^3 + \frac{\mu}{2x^2} \tag{1}$$

and the corresponding system of the motion equations:

$$\begin{cases} x_{tt} = -\lambda x - 2xy + \dfrac{\mu}{x^3}, \\ y_{tt} = -y - x^2 + Cy^2, \end{cases} \tag{2}$$

where $x_{tt} \equiv \frac{d^2x}{dt^2}$ and $y_{tt} \equiv \frac{d^2y}{dt^2}$, λ, μ and C are arbitrary numerical parameters. If $C=1$, $\lambda=1$ and $\mu=0$, then (1) is the initial Hénon–Heiles Hamiltonian [6].

The general solutions in the analytic form are known [7] only in the three integrable cases $\big(C=-1$ and $\lambda=1$; $C=-6$ and λ is an arbitrary number; $C=-16$ and $\lambda=1/16\big)$, in other cases not only four-, but even three-parameter exact solutions have yet to be found.

In all above-mentioned cases system (2) is integrable at any value of μ. Moreover the function y, solution of system (2), satisfies the following fourth-order equation, which does not include μ:

$$y_{tttt} = (2C-8)y_{tt}y - (4\lambda+1)y_{tt} + 2(C+1)y_t^2 + \frac{20C}{3}y^3 + (4C\lambda-6)y^2 - 4\lambda y - 4H, \tag{3}$$

where H is the energy of the system. If $x_0 = Cy_0^2 - y_0 - y_{0tt} \neq 0$, then H is not an arbitrary parameter, but a function of initial data: y_0, y_{0t}, y_{0tt} and y_{0ttt}. The form of this function depends on μ.

The Painlevé test of eq. (3) gives the following dominant behaviors and resonance structures:

1. The function y tends to infinity as $b_{-2}(t-t_0)^{-2}$, where $b_{-2}=-3$ or $b_{-2}=\frac{6}{C}$.

2. For $b_{-2}=-3$ (the *Case 1*) the values of resonances are $r=-1,\ 10,\ (5\pm\sqrt{1-24(1+C)})/2$. In the *Case 2* ($b_{-2}=\frac{6}{C}$) $r=-1,\ 5,\ 5\pm\sqrt{1-48/C}$.

The resonance $r=-1$ corresponds to arbitrary parameter t_0. Other values of r determine powers of t (their values are $r-2$), at which new arbitrary parameters can appear as solutions of the linear systems with zero determinant. For integrability of system (2) all values of r have to be integer and all systems with zero determinants have to have solutions at any values of included in them free parameters. It is possible only in the integrable cases.

For the search for special solutions, it is interesting to consider such values of C, for which r are integer numbers either only in *Case 1* or only in *Case 2*. If there exist a negative integer resonance, different from $r=-1$, then such Laurent series expansion corresponds rather to special than general solution. We demand that all values of r, but one, are nonnegative integer numbers and all these values are different. From these conditions we obtain the following values of C: $C=-1$ and $C=-4/3$ (*Case 1*), or $C=-16/5$, $C=-6$ and $C=-16$ (*Case 2*, $\alpha=\frac{1-\sqrt{1-48/C}}{2}$), and also $C=-2$, in which these two *Cases* coincide. It has been shown in [8] (for $\mu=0$) and [9] (for an arbitrary value of μ) that single-valued three-parameter special solutions can exist only in two nonintegrable cases: $C=-16/5$ and $C=-4/3$ (λ is arbitrary).

Using the method of construction of the Laurent series solutions for nonlinear differential equations describing in [8], we obtain single-valued local solutions of eq. (3). At $C=-4/3$ these solutions are:

$$y=-3\frac{1}{t^2}+b_{-1}\frac{1}{t}+\frac{29}{24}b_{-1}^2+\frac{1}{2}\lambda-\frac{3}{4}+\left(\frac{17}{6}b_{-1}^2+\frac{5}{3}\lambda-\frac{5}{4}\right)b_{-1}t+b_2t^2+\ldots$$

There exist four possible values of the parameter b_{-1}:

$$b_{-1}=\pm\sqrt{\frac{105-140\lambda\pm\sqrt{7(1216\lambda^2-1824\lambda+783)}}{385}}. \tag{4}$$

The parameters b_2 and b_8, coefficients at t^2 and t^8 correspondingly, are arbitrary. The energy H enters in coefficients beginning from b_4.

At $C=-16/5$ we obtain the following solutions:

$$y=-\frac{15}{8t^{-2}}+\tilde{b}_{-1}-\frac{5}{32}+\frac{62}{45}\tilde{b}_{-1}^2+\left(\frac{5}{12}\lambda+\frac{632}{225}\tilde{b}_{-1}^2-\frac{25}{192}\right)\tilde{b}_{-1}t+\ldots$$

with

$$\tilde{b}_{-1}=\pm\frac{3\sqrt{6872250-21991200\lambda\pm52360\sqrt{35(2048\lambda^2-1280\lambda+387)}}}{41888}. \tag{5}$$

The coefficients $\tilde{b}_3$ and $\tilde{b}_8$ are arbitrary parameters. Beginning from $\tilde{b}_4$ some coefficients include the energy H. So, the obtained local solutions depend on

four independent parameters: t_0, H and two coefficients (b_2 and b_8 or $\tilde{b}_3$ and $\tilde{b}_8$). With the help of some computer algebra system, for example, REDUCE [12], these solutions can be obtained with arbitrary accuracy. When a formal series is obtained the question about its convergence arises. The convergence of psi-series solutions of the generalized Hénon–Heiles system with $\mu = 0$ on some real time interval has been proved in [4].

3 Global Single-Valued Solutions

We have found local single-valued solutions. Of course, existence of local single-valued solutions is necessary, but not sufficient condition to exist global ones, because solutions, which are single-valued in the neighborhood of one singularity point, can be multivalued in the neighborhood of another singularity point. So, we can only assume that global three-parameter solutions are single-valued. If we assume this and moreover that these solutions are elliptic functions (or some degenerations of them), then we can seek them as solutions of some polynomial first order equations. There are a few methods to construct such solutions, representing them as the finite Taylor or Laurent series of elliptic functions or degenerate elliptic functions, for example, $\tanh(t)$. These methods use results of the Painlevé test, but don't use the obtained Laurent-series solutions. In 2003 R. Conte and M. Musette [5] have proposed the method, which uses such solutions.

The classical theorem, which was established by Briot and Bouquet [10], proves that if the general solution of a polynomial autonomous first order ODE is single-valued, then this solution is either an elliptic function, or a rational function of $e^{\gamma x}$, γ being some constant, or a rational function of x. Note that the third case is a degeneracy of the second one, which in its turn is a degeneracy of the first one. It has been proved by Painlevé [1] that the necessary form of the polynomial autonomous first order ODE with the single-valued general solution is

$$\sum_{k=0}^{m} \sum_{j=0}^{2m-2k} h_{jk}\, y^j y_t^k = 0, \qquad h_{0m} = 1, \tag{6}$$

in which m is a positive integer number and h_{jk} are constants.

Rather than to substitute eq. (6) in some nonintegrable system, one can substitute the Laurent series of unknown special solutions in eq. (6) and obtain a system, which is linear in h_{jk} and nonlinear in parameters which are included in the Laurent coefficients.

Let us compare two methods to find special analytic solutions for the generalized Hénon–Heiles system.

The first way is the following:

1) Transform system (2) into eq. (3)

2) Assume that y satisfies some more simple equation, substitute this equation in (3) and obtain a nonlinear algebraic system.

3) Solve the obtained system.

The second, proposed by R. Conte and M. Musette, way is the following:

1) Choose a positive integer m and define the first order ODE (6), which contains unknown constants h_{jk}.

2) Compute coefficients of the Laurent series solutions for (2) with some fixed C. The number of coefficients has to be greater than the number of unknowns.

3) Substituting the obtained coefficients, transform eq. (8) in linear and overdetermined system in h_{jk} with coefficients depending on arbitrary parameters.

4) Exclude h_{jk} and obtain the nonlinear system in five parameters.

5) Solve the obtained system.

To obtain the explicit form of the elliptic function, which satisfy the known first order ODE, one can use the classical method due to Poincaré, which has been implemented [11] in Maple [13].

The second way has a few preferences and can be automatized. The first preference is that one does not need to transform system (2) in one differential equation either in y or in x. Moreover at $C = -16/5$ not x, but x^2 may be an elliptic function. To construct the Laurent series for x^2 is easier than to find the fourth order equation in x^2. The main preference of the second method is that the number of unknowns in the resulting algebraic system does not depend on number of coefficients of the first order equation. For example, eq. (6) with $m = 8$ includes 60 unknowns h_{jk}, and it is not possible use the first way to find similar solutions. Using the second method we obtain nonlinear system in five variables: λ_1, λ_2, H and two arbitrary coefficients of the Laurent-series solutions independently of the value of m.

The first way also has one important preference. It allows to obtain solutions for an arbitrary C, whereas using the second method one has to fix value of C to construct the Laurent series solutions, because of the resonance structure depends on C. Using the first way one can also construct equations, which forms are different from (6). For example, the substitution of the eqution

$$y_t^2 = \tilde{\mathcal{A}}(y-P_0)^3 + \tilde{\mathcal{B}}(y-P_0)^{5/2} + \tilde{\mathcal{C}}(y-P_0)^2 + \tilde{\mathcal{D}}(y-P_0)^{3/2} + \tilde{\mathcal{E}}(y-P_0) \quad (7)$$

gives two-parameter solutions [9]:

$$y(t-t_0) = \left(\frac{a\wp(t-t_0)+b}{c\wp(t-t_0)+d}\right)^2 + P_0, \quad (8)$$

where $\wp(t-t_0)$ is the Weierstrass elliptic function, a, b, c and d are some constants. The parameter P_0 defines the energy of the system. Solutions (8) exist in both above-mentioned nonintegrable cases: $C = -16/5$ and $C = -4/3$. There exist two different solutions for each pair of values of C and λ. To find these solutions we a priori assume some class of possible solutions, then compute whether there exists some solution in the given class. We hope that the use of the second method and the corresponding computer algebra program allows to find the three-parameter elliptic solutions without any preliminary assumption. This method is implemented both in REDUCE and in Maple.

4 Conclusion

We have found the special solutions of the Hénon–Heiles system with $C = -16/5$ and $C = -4/3$ as the Laurent series. For some values of parameters the obtained solutions coincide with the known exact periodic solutions. The Painlevé test does not show any obstacle to the existence of three-parameter single-valued solutions, so, the probability to find exact, for example elliptic, three-parameter solutions, that generalize the solutions found in [9], is high.

The author is grateful to R. Conte, V. F. Edneral, A. K. Pogrebkov and E. I. Timoshkova for valuable discussions. This work has been supported by Russian Federation President's Grants NSh–1685.2003.2 and NSh–1450.2003.2 and by the grant of the scientific Program "Universities of Russia".

References

1. Painlevé, P.: Leçons sur la théorie analytique des équations différentielles, profeesées à Stockholm (1895) sur l'invitation de S. M. le roi de Suède et de Norwège, Hermann, Paris, 1897. Reprinted in: Oeuvres de P. Painlevé, vol. 1, ed. du CNRS, Paris, 1973. On-line version: The Cornell Library Historical Mathematics Monographs, http://historical.library.cornell.edu/
2. Kowalevski, S.: Sur le problème de la rotation d'un corps solide autour d'un point fixe. Acta Mathematica. **12** (1889) 177–232; Sur une properiété du sustème d'équations différentielles qui définit la rotation d'un corps solide autour d'un point fixe. Acta Mathematica. **14** (1890) 81–93, {in French}. Reprinted in: Kovalevskaya, S.V.: Scientific Works, AS USSR Publ. House, Moscow, 1948, {in Russian}.
3. Ablowitz M.J., Ramani A., Segur H., A Connection between Nonlinear Evolution Equations and Ordinary Differential Equations of P-type. I & II, J. Math. Phys. **21** (1980) 715–721 & 1006–1015
4. Melkonian, S.: Psi-series Solutions of the Cubic Hénon–Heiles System and Their Convergence. J. of Nonlin. Math. Phys. **6** (1999) 139–160, math.DS/9904186.
5. Conte, R., Musette, M.: Analytic solitary waves of nonintegrable equations. Physica **D181** (2003) 70, nlin.PS/0302051.
6. Hénon, M., Heiles, C.: The Applicability of the Third Integral of Motion: Some Numerical Experiments, Astronomical J. **69** (1964) 73–79.
7. Conte, R., Musette, M., Verhoeven, C.: Integration of a generalized Hénon–Heiles Hamiltonian, J. Math. Phys. **43** (2002) 1906–1915, nlin.SI/0112030; General solution for Hamiltonians with extended cubic and quartic potensials, TMF (Russ. J. Theor. Math. Phys.) **134** (2003) 148–159, nlin.SI/0301011.
8. Vernov, S.Yu.: Construction of solutions for the generalized Hénon–Heiles system with the help of the Painlevé test. TMF (Russ. J. Theor. Math. Phys.) 135 (2003) 409–419, {in Russian}, 792–801, {in English}; math-ph/0312048.
9. Vernov, S.Yu., Timoshkova, E.I.: On two nonintegrable cases of the generalized Hénon–Heiles system with an additional nonpolynomial term, math-ph/0402049.
10. Briot, C., Bouquet, T.: Théovie des fonctions doublement périodiques, 1859.
11. van Hoeij, M.: pack. 'algcurves', Maple V (1997), http://www.math.fsu.edu/hoeij/
12. Hearn, A.C.: REDUCE. User's and Contributed Packages Manual, Vers. 3.7, CA and Codemist Ltd, Santa Monica, California, 1999, http://www.zib.de/Symbolik/reduce/more/moredocs/reduce.pdf
13. Heck, A.: Introduction to Maple, 3rd Edition, Springer–Verlag, New York, 2003.

Computer Algebra Manipulations in Astronomy

Tamara Ivanova

Institute of Applied Astronomy
191187, St.Petersburg, Russia
itv@quasar.ipa.nw.ru

Abstract. The specialized Poisson Series Processors (PSPs) are proposed. They are a typical software for the implementation of analytical algorithms of celestial mechanics. The PSPs are designed for manipulating long Poisson (polynomial-trigonometric) and echeloned series with the rational, floating-point or arbitrary length (infinite precision) coefficients and with unlimited number of power and angular variables. The Keplerian processor and analytical generator of special celestial mechanics functions based on the PSPs are also developed.

1 Introduction

The main objects of celestial mechanics are so-called Poisson series, i.e., the polynomial-trigonometric series of the form:

$$S = \sum C_{i,j}\, x^i \begin{matrix}\sin\\ \cos\end{matrix} (jy)\,, \tag{1}$$

$$x = (x_1, \ldots, x_n),\quad y = (y_1, \ldots, y_m),\quad i = (i_1, \ldots, i_n),\quad j = (j_1, \ldots, j_m).$$

Here x and y are vectors of polynomial and trigonometric variables, respectively. The summation is performed over all integer values of indices i and j. Coefficients $C_{i,j}$ may be represented as rational, floating-point, arbitrary length or complex numbers. The last form of coefficients corresponds to the exponential Poisson series:

$$S = \sum C_{i,j}\, x^i\, exp\, \sqrt{-1}\ (jy). \tag{2}$$

The systems of operations on the series of this kind are known as Poisson processors. These specialized processors turn out to be very effective in performing the symbolic operations with series even of hundreds of thousand terms characteristic of high-accuracy analytical theories of celestial mechanics. Though in general they are constructed for the purpose of solving celestial mechanics problems they may be used equally well in other fields such as theoretical mechanics, physics, etc. Poisson processors have been developed in many centres of celestial mechanics. The brief review of them and their applications to solve the celestial mechanics problems are given by Brumberg (Brumberg, 1995). As a rule they are different in the representation of coefficients, in the number of power and trigonometric variables, in the range of associated indices of these variables, in storage techniques for the Poisson series, etc. Some of them allow

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 388–393, 2004.
© Springer-Verlag Berlin Heidelberg 2004

manipulation of a wider class of objects than the Poisson series, for instance, so-called echeloned series (very long polynomial-trigonometric series with literal frequencies):

$$S = \sum x^i \begin{matrix}\sin\\ \cos\end{matrix} (jy) \sum_{k=1}^{K_{i,j}} C_{i,j,k} \prod_{\nu=1}^{N_k} \left(\gamma^{(\nu,k)}\omega\right)^{q^{(\nu,k)}}, \tag{3}$$

$$x = (x_1, \ldots, x_n), \qquad y = (y_1, \ldots, y_m), \qquad i = (i_1, \ldots, i_n),$$

$$j = (j_1, \ldots, j_m), \qquad \omega = (\omega_1, \ldots, \omega_m), \qquad \gamma^{(\nu,k)} = (\gamma_1^{(\nu,k)}, \ldots, \gamma_m^{(\nu,k)}).$$

Each term of this series is characterized by the set of power x, angular y and frequency ω variables with associated multi-indices i, j and $\gamma^{(\nu,k)}$ together with the quantities $q^{(\nu,k)}$, respectively. The coefficients $C_{i,j,k}$ are rational, floating-point or arbitrary length numbers. The exponents $q^{(\nu,k)}$ are integers. The first summation is again performed here over all integer values of indices i and j. The values of $K_{i,j}$ and N_k are directly obtained as a result of symbolic operations on the series.

In this case the trigonometric variables y_r are represented by the linear functions of an argument t with the literal values of the frequencies w:

$$y_r = w_r\, t + y_r^{(0)}, \qquad r = (1, \ldots, m). \tag{4}$$

The Echeloned Poisson Series Processor EPSP (Ivanova, 2001) is designed for manipulating the echeloned series. This processor is developed on the basis of the Poisson series processor PSP (Ivanova, 1995). PSP is intended directly for operating on the Poisson series themselves, that is, the polynomial-trigonometric series where the frequencies are taken into account numerically in differentiating and integrating with respect to argument t. The echeloned series processor allows to solve a wider class of problems as compared with the Poisson series processor. This is a case, for instance, of analytical Lunar theory where one has to handle the divisors in symbolic form. From the user's point of view both processors are similar, except that the list of basic operations of EPSP includes the operations of the total differentiation and integration with respect to time in the analytical form relatively to the frequencies. It leads to more complex storage schemes in the echeloned series processor.

2 Basic Characteristics of PSPs

The most important characteristics of PSPs are as follows:

- PSPs are written in standard FORTRAN-77 language.
- PSPs have no restrictions on the number of power, angular and frequency variables and on the ranges of changing of associated indices to be prescribed by user for any specific problem. For comparison, processor ESP (Rom, 1971) deals with the echeloned series of 12 polynomial, six trigonometric and three frequency variables.

- Every term of the series is characterized by the analytical order of smallness N calculated by formula

$$N = \sum_{l=1}^{n} i_l W_l \,, \tag{5}$$

  where W_l is the weight function of the polynomial variable with number l, i_l is the corresponding index. The analytical order of smallness is a very important characteristic because the power variables are as a rule the small parameters resulting in the possibility of restricting the series by the terms with the total order of the power variables not more than some prescribed value.
- All mathematical operations with the series are formal. The criterion for rejecting a term is based on its smallness determined by the analytical order of the term, the numerical estimation of its coefficient and the admissible values of the indices.
- PSPs allow the user to write his/her own procedures without any difficulty. It is especially easy to perform the identical operations on every term of the series due to the availability of special scanning procedures executing a given action on each term of series.
- The hierarchical architecture of PSPs allow rather easily the adaptation of the system on different computers and the modification of it for the objects slightly different from the standard Poisson series (for instance, for manipulating the exponents instead of the trigonometric functions or for changing the type of coefficients) or in the case of using different storage techniques for series. The hierarchical principle involves the subdivision of all system procedures into three levels. The procedures of the lowest level realize the operations on coefficients of terms and their indices. The subroutines of the middle level realize the operations on individual terms of the series and depend on specific representation of Poisson series in computer memory. Finally, the analytical operations on the series are realized by the highest level procedures taking into account only general mathematical structure of objects.
- The list of basic operations of PSPs includes the standard arithmetic operations with the series, the partial differentiation and integration with respect to polynomial and trigonometric variables, the total differentiation and integration with respect to time, raising to any integer, rational or real exponent, binomial and Taylor expansion up to some prescribed order, substitution of the series in place of any set of the power and trigonometric variables, fast evaluation of the series for numerical values of any variables, conversion of the series, different sortings and selections, etc. PSPs enable the realization of the input and output of the series in any format or unformatted mode and typing them in natural mathematical form.

All algorithms of the operations with the series imply the standard lexicographic order of the input series and do not alter it in the resulting series.

3 Storage Schemes of PSPs

The storage area consists of two or three global arrays which keep the series terms, their coefficients and divisors or factors for the EPSP. The memory for coefficients is represented by the standard one-dimensional array where rational, floating-point and complex coefficients are located consecutively taking into account their length depending on their type. Two other arrays for keeping the series, divisors and factors and first array for arbitrary length coefficients use the one-way linked list structure for dynamical memory allocation (Figure 1) permitting the application of an algorithm for quick searching and inserting of terms into series.

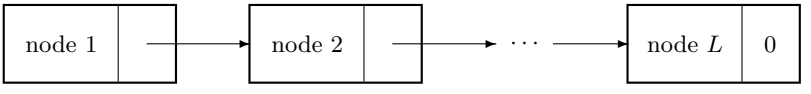


Fig. 1. Structure of the one-way linked list memory allocation for the series, the divisors and the arbitrary length coefficients of L nodes.

The series consists of a passport and a sequence of its terms connected in circular chain with one link (Figure 2). The series passport contains its general characteristics: the number of terms in the series (N) and the link on its last term ($LAST$).

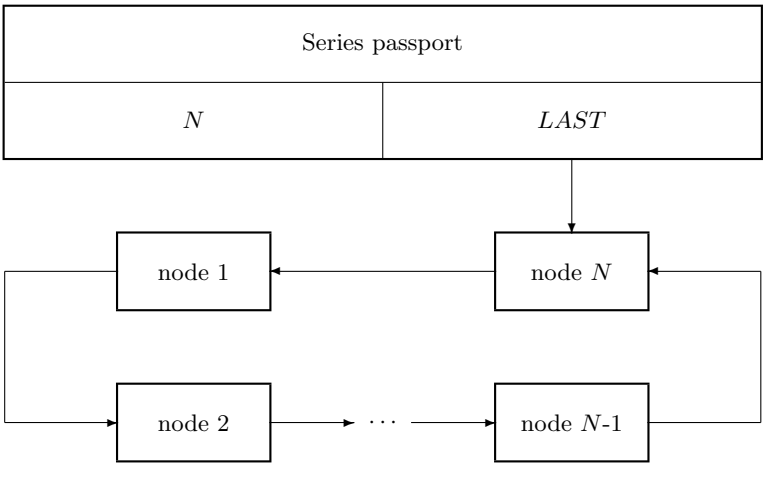


Fig. 2. Structure of a series

The node is an elementary structure unit. It includes the main characteristics of the separate term of a series (Figure 3):

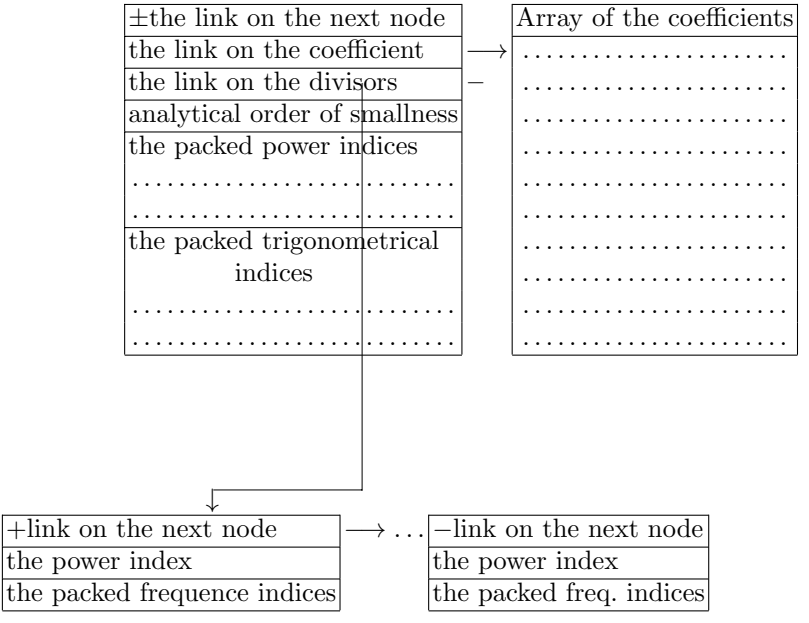


Fig. 3. Structure of a node.

∗ the link on the next node;
∗ the name of the trigonometric function defined by the sign '+' for a cosine term and the sign '−' for a sine term assigned to the above-mentioned link;
∗ the link on the beginning of the coefficient;
∗ the link on the beginning of the symbolic divisors or factors. If this link is equal to zero it means that this node does not contain any divisors;
∗ analytical order of smallness;
∗ the packed power and trigonometric indices.

Each term of the symbolic divisors or factors includes:
∗ the link on the next node;
∗ a flag assigned to the link and expressed by signs '+' or '−'. It indicates whether the divisors are finished or not for the given node;
∗ an exponent index for one divisor. It can be positive or negative;
∗ the packed frequency indices.

In the case of rational coefficients, the double precision floating-point numbers for correct operations on integer numbers are used. If a numerator or a denominator of rational coefficient falls outside the limits of maximum permissible number then they are transformed automatically to floating-point coefficients and the processor continues its work. If the coefficients are of arbitrary length then they are represented in the form like to divisors (Figure 4).

+link on the next node	$\Longrightarrow \ldots$	−link on the next node
numerator of rational number		numerator of rational number
denominator of rational number		denominator of rational number

Fig. 4. Structure of an arbitrary length coefficient.

Each term of the coefficient of arbitrary length includes:
* the link on the next node;
* a flag assigned to the link and expressed by signs '+' or '−'. It indicates whether the coefficient is finished or not for the given node;
* numerator of rational number;
* denominator of rational number.

The packed indices are able to occupy an arbitrary number of computer integer words dependent on the number of indices, the value of their range and the length of the computer integer word. All the operations performed on the nodes deal with the standard form of nodes. It means that the first trigonometric or frequency index not being zero, must be positive.

4 Conclusion

The Keplerian processor and analytical generator of special celestial mechanics functions based on PSPs are also proposed. They may be regarded as an illustration of the practical application of the processors. These systems are designed for implementing the expansions of the most important mathematical functions of celestial mechanics such as Legendre and Gegenbauer polynomials, Bessel and spherical functions, etc., for constructing the expansions of the elliptic motion functions of the unperturbed two-body problem and the expansions of the typical celestial mechanics functions such as Hansen coefficients, Newcomb operators, Kaula inclination functions and their generalizations, etc. The software of this sort is necessary to produce analytical solutions of various celestial mechanics problems.

PSPs are available on request in electronic form from itv@quasar.ipa.nw.ru.

References

1. Brumberg, V. A.: 1995, *Analytical Techniques of Celestial Mechanics*, Springer, Heidelberg.
2. Ivanova, T. V.: 1995, PSP: A New Poisson Series Processor, in: S. Ferraz-Mello, B. Morando and J.-E. Arlot (eds), *Proceedings of the 172nd Sympozium of the International Astronomical Union, Dynamics, Ephemerides and Astrometry of the Solar System*, Paris, France. Kluwer Academic Publishers.
3. Ivanova, T. V.: 2001, A New Echeloned Series Processor (EPSP), *Celest. Mech. and Dyn. Astron.*, **80**, Kluwer, 167–176.
4. Rom, A.: 1971, Echeloned Series Processor (ESP), *Celest. Mech.*, **3**, 331–345.

Higher Order Quadrature on Sparse Grids

Hans-Joachim Bungartz[1] and Stefan Dirnstorfer[2]

[1] IPVS, Universität Stuttgart, Universitätsstraße 38, D-70569 Stuttgart, Germany
bungartz@ipvs.uni-stuttgart.de
[2] Nagler & Company, D-92263 Schnaittenbach, Germany
dirnstor@web.de

Abstract. Sparse grids have turned out to be a very efficient discretization scheme that, to some extent, breaks the curse of dimensionality and, therefore, is especially well-suited for higher dimensional scenarios. Besides the classical sparse grid application, the numerical solution of partial differential equations, sparse grids have been used for various topics such as Fourier transform, image compression, numerical quadrature, or data mining, so far. In this paper, we summarize and assess recent results concerning the application of sparse grids to integrate functions of higher dimensionality, the focus being on the explicit and adaptive use of higher order basis polynomials.

1 Introduction

The representation of functions and the discretization of PDE by conventional methods is limited to time-depending 3 D problems due to storage requirements and computational cost. The reason is the so-called *curse of dimensionality*, saying that the cost to represent a function or to compute an approximation with some prescribed accuracy ε depends exponentially on the problem's dimensionality d. We, hence, encounter complexities of the order $O(\varepsilon^{-\alpha d})$ with $\alpha > 0$ depending on the respective approach (i.e. the polynomial degree of the ansatz functions, e.g.), the function's smoothness, and on the implementation. With sparse grids, we circumvent this problem by considering spaces of functions with bounded mixed derivatives [6], only, which is not a serious restriction due to the possibility of adaptive mesh refinement. Starting from some suitable 1 D multilevel basis on the unit interval $[0, 1]$, we obtain a multilevel basis for $\Omega := [0, 1]^d$ by a simple tensor product construction. If the resulting hierarchical basis functions are arranged by gathering those of a support of size $2 \cdot \mathbf{h_l}$ for level $\mathbf{l} = (l_1, \ldots l_d)$, $h_{l_j} = 2^{-l_j}$, $j = 1, ..., d$, in a hierarchical subspace $W_\mathbf{l}$, we get the multilevel subspace splitting illustrated for $d = 2$ in Fig. 1.

Now, for each of the $W_\mathbf{l}$, its cost (in terms of degrees of freedom) and its benefit (in terms of its contribution to the overall interpolant) can be estimated and used as input of a cost-benefit optimisation process selecting the most important subspaces with their respective grid points. This a priori optimisation leads us to grid patterns and corresponding approximation spaces that consist of $O(N \cdot \log(N)^{d-1})$ or even $O(N)$ grid points, only, instead of the $O(N^d)$ of

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 394–401, 2004.
© Springer-Verlag Berlin Heidelberg 2004

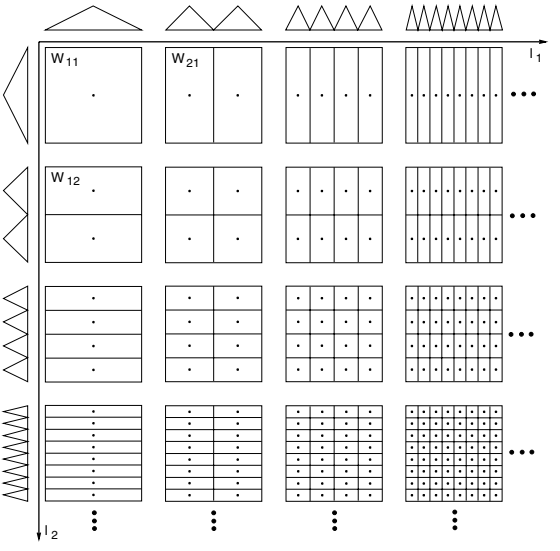


Fig. 1. Scheme of subspaces for $d = 2$: Each square represents one hierarchical subspace $W_{\mathbf{l}}$ with associated grid points and basis functions' support of size $2 \cdot \mathbf{h}_{\mathbf{l}}$.

a regular full grid (N denoting the maximum number of grid points in one direction). These grids correspond to simplizoidal selections of subspaces $W_{\mathbf{l}}$ in Fig. 1, and they are known as *sparse* grids (see Fig. 2). It turns out that the number of degrees of freedom needed for some prescribed accuracy does (up

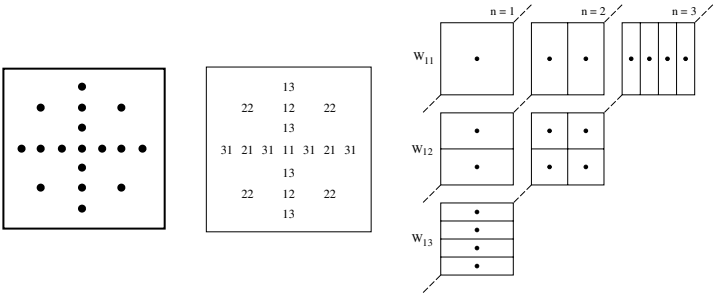


Fig. 2. 2 D regular sparse grid of level 3 and assignment of grid points to subspaces.

to logarithmic factors) no longer depend on d exponentially. This allows either to treat problems of low dimensionality substantially faster or to tackle higher dimensional problems. Extensions of the piecewise linear hierarchical basis to general polynomial degree p as well as interpolets or (pre-)wavelets have been successfully used as 1 D ingredient for the tensor product (see [4,7]).

Since its introduction [17], the sparse grid concept has been applied to most discretization schemes for PDE such as finite elements, finite differences, finite volumes, spectral methods, and splitting extrapolation [8]. In the FE context, special attention has been paid to adaptive mesh refinement [1,4] and to fast solvers [3]. Concerning fields of application, fluid flow was the first focus of sparse grids. Meanwhile, however, sparse grids are also used for problems from quantum mechanics, for problems in the context of stochastic differential equations, or for the discretization of differential forms in the context of the Maxwell equations. Apart from PDE, sparse grids have been and are applied to a variety of other problem classes. Among these problems are integral equations, general operator equations, eigenvalue problems, data mining, and numerical quadrature [10,11, 5], which is the topic of this contribution. For a more detailed introduction to sparse grids, we refer to the extensive survey in [7] and to the references therein.

2 Hierarchical Lagrangian Polynomials

In [4], the piecewise linear hierarchical basis was generalized to hierarchical bases of (piecewise) arbitrary polynomial degree. The main idea of the so-called hierarchical Lagrangian interpolation is to hold to (globally) $\mathcal{C}^0$ functions and elements with one degree of freedom, only, but to define local basis polynomials

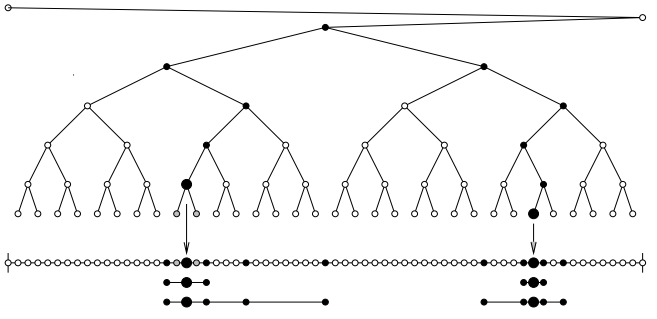


Fig. 3. Ancestors (solid) used for quartic hierarchical Lagrangian interpolation and descendants (dotted) of two grid points.

of higher degree as well as local interpolants by using interpolation conditions in a sufficient number of hierarchical ancestors of the respective grid point *outside* the local element, as illustrated in Fig. 3. That is, we use conditions outside the local support to define the basis polynomials, but we, afterwards, omit the polynomials' parts outside their respective support according to Fig. 1. Depending on the relative position of the ancestor points taken into account, different basis polynomials are obtained. Fig. 4 illustrates this construction principle as well as the resulting four different basis polynomials for the quartic case $p = 4$. Hence, hierarchical Lagrangian interpolation provides a straightforward access

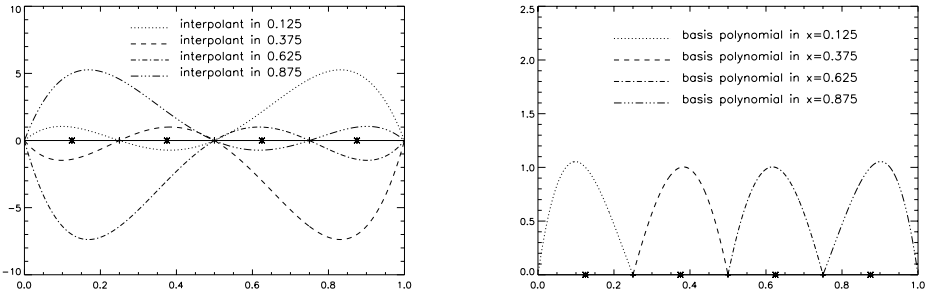


Fig. 4. Basis polynomials for $p = 4$: construction via hierarchical Lagrangian interpolation (left) and used restriction to the respective hierarchical support (right).

to higher order approximation without investing more than one degree of freedom per element. Actually, using such 1 D polynomial bases of degree p in the tensor product leads to accuracies of $O(p)$ [4].

3 Numerical Quadrature on Sparse Grids

Due to the roots of sparse grids going back to Archimedes' quadrature of a parabola and to work of Smolyak [16], quadrature has always been a hot topic in sparse grid research. Starting with the explicit use of the piecewise linear functions to calculate integrals [2], two principal strategies have been pursued. The first approach is based upon the direct application of Smolyak's formula

$$Q_n^{(d)} f \;:=\; \left(\sum_{i=0}^{n} \left(Q_i^{(1)} - Q_{i-1}^{(1)} \right) \otimes Q_{n-i}^{(d-1)} \right) f \,, \tag{1}$$

where $Q_n^{(d)}$ denotes a d-dimensional quadrature formula, to standard 1 D rules $Q_n^{(1)}$ such as the Newton-Cotes or Clenshaw-Curtis rules, the Gauss formulas [15], or the incremental Gauss-Patterson rules [10]. Recently, in [11] a generalization of the conventional sparse grid approach which is able to assess the dimensions adaptively according to their importance was developed.

The second approach does not start from 1 D rules as input for Smolyak's formula (1), but, basically, computes a sparse grid interpolant of the integrand and accumulates all occurring basis functions' contributions to the integral. To tackle large values of d and to mimic the increased degree of accuracy of Gaussian quadrature, the Lagrangian basis from the previous section has to be modified. Actually, the basis used in [5] does without boundary points, introducing the grid points' positions as degrees of freedom, themselves (see Fig. 5 for the 1 D case). In the following numerical examples of Sect. 4, the non-equidistant grids and bases (shown in the right part of Fig. 5 for 1 D and resulting from these for larger d) will be used, leading us to the so-called *piecewise Gauss quadrature*.

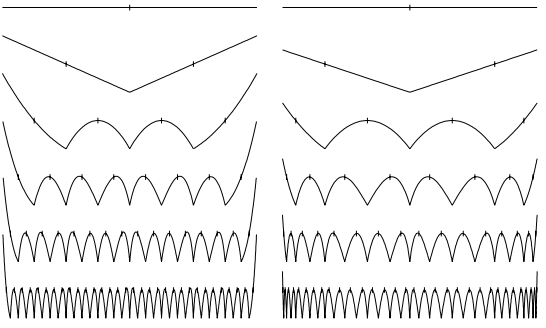


Fig. 5. Hierarchical polynomial bases according to [5]: equidistant (left) and piecewise Gauss version (right), both without boundary points.

A comparison of both approaches reveals the pros and cons of either method. The direct application of (1) allows to use 1 D quadrature rules of arbitrary *global* accuracy, such as the hierarchical Gauss-Kronrod or Gauss-Patterson rules (cf. [10]), whereas the piecewise Gauss quadrature only allows for *locally* higher order, due to the underlying hierarchical Lagrangian approach. On the other hand, in the Smolyak formula, there are no possibilities of a local grid adaptation, which is, in contrast to that, straightforward with piecewise Gauss quadrature. It is not clear how important local mesh refinement may be in the really high dimensional case of, say, $d \gg 20$, where, anyway, only a very small number of steps of refinement is feasible. However, since most of those scenarios are characterized by big differences in the dimensions' importance, some kind of *dimension adaptivity* as studied in [11] and in work in progress [12] in the sequel of [5] will be essential for efficiently tackling such problems.

4 Numerical Results

Next, we apply both sparse gird approaches of Sect. 3 to typical problems of moderate to high d, and we compare the results with standard techniques normally used in these scenarios, such as Monte Carlo quadrature or quasi Monte Carlo quadrature [14], also known as quadrature based on low-discrepancy sequences.

As a first example, we study a simple transport problem from computational physics (see [13] for details), described by the integral equation

$$y(x) \;=\; x + \int_x^1 \gamma y(z) dz \,, \tag{2}$$

for which the exact solution is known. Think of a particle travelling through a 1 D (z) slab of length one. In each step, the particle covers a distance which is uniformly distributed on $[0, 1]$. This may cause it to exit the slab; otherwise, it may be absorbed with probability $1-\gamma$. The function $y(x)$ denotes the probability for some particle at current position x to leave the slab. The solution of the integral equation (2) can also be represented as an infinite dimensional integral,

$$y(x) \;=\; \int_{[0,1]^\infty} \sum_{n=0}^{\infty} F_n(x,\mathbf{z}) d\mathbf{z}\,, \tag{3}$$

where the vector $\mathbf{z}$ contains the leap lengths and where F_n denotes the probability for the particle to leave the slab after exactly n steps. F_n, basically, is a product of two Heaviside functions, such that the overall integrand has a discontinuity along the diagonal (see Fig. 6), which is kind of a worst case for a regular sparse grid, of course. Concerning dimensionality, since higher dimensions do not contribute considerably to the integral, the sum can be truncated. Fig. 6 (right) shows numerical results for $y(0)$ with $\gamma = 0.5$ in twenty dimensions. The performance of the adaptive sparse grid lies within the range of the standard Monte Carlo method, but can not compete with quasi Monte Carlo. At least, our adaptivity rescues Monte Carlo performance in this worst case scenario.

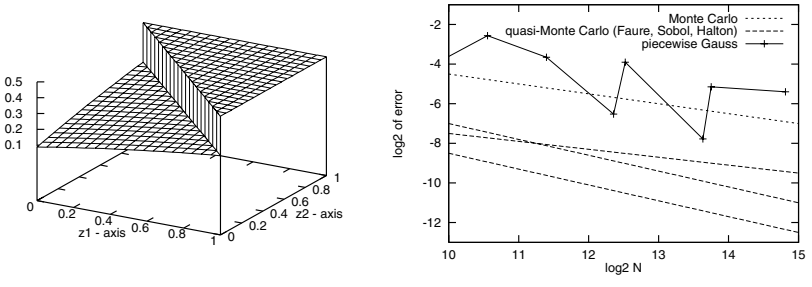


Fig. 6. The discontinuous integrand $\sum_{n=0}^{\infty} F_n(0,\mathbf{z})$ in the first two dimensions (left) and results for numerical quadrature taking into account twenty dimensions according to standard Monte Carlo, quasi-Monte Carlo, and piecewise Gauss quadrature (right).

Much better results can be obtained by changing the formulation once more. $F_n(x,\mathbf{z})$ in (3) can be replaced by a polynomial $F_n^*(x,\mathbf{z})$ describing the contribution of each jump, which, then, results in a smooth integrand. Now, the sparse grid properties can be exploited. Results for $d = 20$ are shown in Fig. 7. While quasi Monte Carlo gains about two digits of accuracy, the sparse grid gains eight. With 30000 evaluated points, the adaptive sparse grid outperforms quasi Monte Carlo by about four digits. Fig. 7 also gives the convergence behaviour if we further reduce dimensionality up to eight. The non-adaptive Gauss-Patterson on sparse grids according to [10] beats quasi Monte Carlo, but not the adaptive sparse grid. Hence, in spite of the smooth integrand, the different importance of the involved directions requires some adaptive strategy.

As a second example, we consider a problem from computational finance, a so-called collateralized mortgage obligation or CMO problem (see [9] for details). There, the present value (PV) of some security is defined as the expectation value over random variables involved in the interest rate fluctuations, each dimension

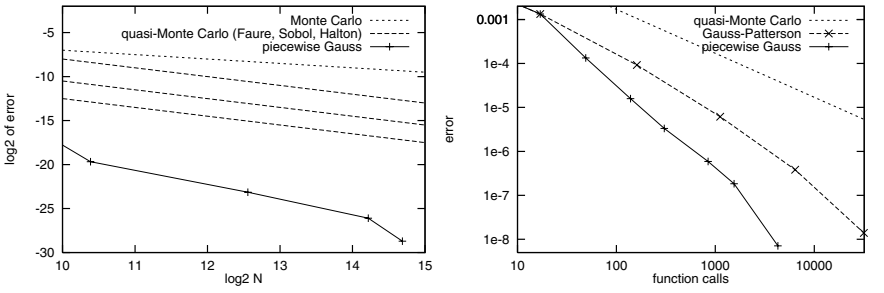


Fig. 7. Results for the smooth integrand in 20 (left) and 8 dimensions (right).

representing one month. After suitable transformations, PV can be written as

$$PV \;=\; \int_{[0,1]^d} v(G(x_1), \cdots, G(x_d))\, dx_1 \cdots dx_d\,, \tag{4}$$

where $d \geq 120$, typically. A more efficient way to compute (4) was suggested in [9], where the authors reduced the effective number of dimensions by representing the interest rate fluctuations as a Brownian motion $b(t)$. Here, a future value $b(t + \Delta t)$ can be generated from the current value $b(t)$ by a random jump:

$$b(t + \Delta t) = b(t) + \sqrt{\Delta t}\, \xi_{t+\Delta t}\,, \tag{5}$$

where ξ_t denotes some random noise. Using the so-called Brownian bridge (BB) discretization, where $b(t + \Delta t)$ is computed from both a past and a future value of b, most of the involved functions' variation can be put into the dimensions defining the first few values of b. In a first scenario with $d = 256$, we compare the adaptive piecewise Gauss and the Gauss-Patterson rule. Fig. 8 (left) shows how the BB formulation improves the performance of adaptive schemes. To make this accessible to Gauss-Patterson, too, a higher order formula is used in the 30 most important directions, found by some setup run up to level two (cf. [10]). The piecewise Gauss rule can find these important directions itself, but shows drawbacks where no use of adaptivity can be made, i.e. at the very beginning of the process and without the BB construction. Although the piecewise Gauss grid performs best for more than 10^6 points, there is no doubt that some dimension adaptivity as suggested in [11] will lead to further improved results. In our second CMO scenario ($d = 360$), we compare our approach with published results for Sobol's quasi Monte Carlo method. The right part of Fig. 8 shows that the adaptive sparse grid can not keep pace for the investigated sample sizes. An increasing downward slope can be expected with even more function evaluations. However, for such large d, the logarithmic factors in the standard (L_2-based) sparse grids' complexity cause problems. A remedy might be the energy-based sparse grids discussed in [6], doing without the logarithmic factors.

Finally, note that aspects of implementation are crucial for a memory- and runtime-efficient maintaining and processing of adaptive sparse grids, see [10,5].

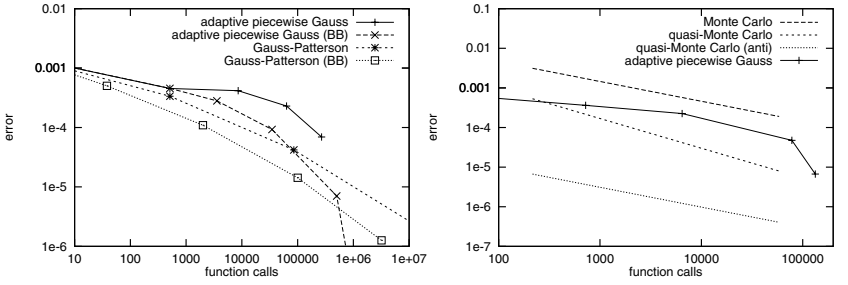


Fig. 8. The CMO problem for $d = 256$ (left), with and without the Brownian bridge, and for $d = 360$ with Brownian bridge (right, comparison with (quasi) Monte Carlo).

References

1. S. Achatz. Higher order sparse grids methods for elliptic partial differential equations with variable contraints. *Computing*, 71(1):1–15, 2003.
2. T. Bonk. A new algorithm for multi-dimensional adaptive numerical quadrature. In W. Hackbusch and G. Wittum, editors, *Adaptive Methods – Algorithms, Theory, and Applications*, volume 46 of *NNFM*, pages 54–68. Vieweg, 1994.
3. H.-J. Bungartz. A multigrid algorithm for higher order finite elements on sparse grids. *ETNA*, 6:63–77, 1997.
4. H.-J. Bungartz. *Finite Elements of Higher Order on Sparse Grids*. Shaker, 1998.
5. H.-J. Bungartz and S. Dirnstorfer. Multivariate quadrature on adaptive sparse grids. *Computing*, 71(1):89–114, 2003.
6. H.-J. Bungartz and M. Griebel. A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. *J. Complex.*, 15:167–199, 1999.
7. H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 2004.
8. H.-J. Bungartz, M. Griebel, and U. Rüde. Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems. *Comput. Meth. Appl. Mech. Eng.*, 116:243–252, 1994.
9. R. Caflisch, W. Morokoff, and A. Owen. Valuation of mortgage backed securities using brownian bridges to reduce effective dimension. *J. Comput. Finance*, 1, 1997.
10. T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numerical Algorithms*, 18:209–232, 1998.
11. T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71:65–87, 2003.
12. T. Kern. Dünngittertechniken zur hochdimensionalen numerischen Quadratur. Master's thesis, Universität Stuttgart, 2003.
13. W. Morokoff and R. Caflisch. Quasi-monte carlo integration. *J. Comp. Phys.*, 122:218–230, 1995.
14. H. Niederreiter. *Random Number Generation and Quasi-Monte-Carlo Methods*. SIAM, Philadelphia, 1992.
15. E. Novak and K. Ritter. High dimensional integration of smooth functions over cubes. *Numer. Math.*, 75(1):79–98, 1996.
16. S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Math. Dokl.*, 4:240–243, 1963.
17. C. Zenger. Sparse grids. In W. Hackbusch, editor, *Parallel Algorithms for Partial Differential Equations*, volume 31 of *NNFM*. Vieweg, 1991.

Application of Extrapolation Methods to Numerical Solution of Fredholm Integral Equations Related to Boundary Value Problems

Avram Sidi

Computer Science Department, Technion - Israel Institute of Technology,
Haifa 32000, Israel
asidi@cs.technion.ac.il

Abstract. Fredholm integral equations arise naturally in the context of ordinary and partial differential equations: Two-point boundary value problems can be reformulated as Fredholm integral equations, whose kernels are continuous but have finite jump discontinuities in their derivatives. Two-dimensional elliptic boundary problems can be reformulated as Fredholm integral equations with kernels that have singularities, some having logarithmic singularities. In this note, we describe quadrature methods whose accuracies can be improved at will. These are obtained by improving the underlying numerical quadrature formulas in a clever fashion. In the case of two-point boundary value problems, they are obtained by correcting the trapezoidal rule appropriately to the accuracy required. In the case of boundary integral equations, they are obtained by first correcting the basic trapezoidal rule and then extrapolating it to required accuracy.

1 Introduction

In this note, we describe two quadrature methods for the numerical solution of Fredholm integral equations that have been proposed recently. Despite the singular nature of the integral equations involved, these methods turn out to be very efficient in the sense that their accuracies can be improved at will, the increase in their computational costs being minor.

The relevant integral equations are those that arise in the context of ordinary and partial differential equations with boundary conditions. We give two examples of such problems now.

1. Let the two-point boundary value problem

$$y'' = f(x, y), \quad 0 < x < 1; \quad a_0 y(0) + b_0 y'(0) = c_0, \quad a_1 y(1) + b_1 y'(1) = c_1,$$

be given. Here $f(x, y)$ is a linear or nonlinear function in y. Subtracting $k^2 y$, with $k > 0$ chosen appropriately (even "optimally"), from both sides of the differential equations, and using Green's functions, this problem can be reformulated in the form

$$y(x) = r(x) + \int_0^1 g(x, t)\big[k^2 y(t) - f(t, y(t))\big]\, dt,$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 402–409, 2004.
© Springer-Verlag Berlin Heidelberg 2004

with appropriate $r(x)$ and $g(x,t)$. Here the Green's functions are continuous for $x = t$ but their partial derivatives have finite jump discontinuities across $x = t$; they are smooth everywhere else on $[0,1] \times [0,1]$. An important advantage of approximating $y(x)$ via the integral equation formulation is that, in case of singularly perturbed problems, boundary layers, if present, can be resolved with great precision. Recall that this is a difficult task when the differential equation is attacked with difference methods.
For example, with $y'(0) = 0$ and $y(1) = 1$ as boundary conditions, $r(x)$ and $g(x,t)$ are given by

$$r(x) = \frac{\cosh kx}{\sinh kx}, \quad g(x,t) = \frac{1}{k \cosh k} \begin{cases} \cosh kx \sinh k(1-t), & 0 \le x \le t \\ \sinh k(1-x) \cosh kt, & t \le x \le 1 \end{cases}.$$

This approach was initiated by Keller [1], who used it to prove constructive existence and uniqueness theorems for solutions of the nonlinear problems above and also for designing a quadrature method based on the trapezoidal rule for computing numerical solutions. It was refined and developed further by Pennline [2], [3], [4].

2. Let $u(x,y)$ be the solution of the two-dimensional Laplace's equation with Dirichlet boundary condition

$$\Delta u(P) = 0, \quad P \in \Omega; \quad u(P) = f(P) \quad P \in \partial\Omega,$$

where Δ is the two-dimensional Laplacian, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, Ω is a finite simply connected open domain in the x-y plane, and $\partial\Omega$ is its boundary. $u(x,y)$ can be obtained from the integral

$$u(P) = \int_{\partial\Omega} \rho(Q) \log |P - Q| \, dl_Q, \quad P \in \Omega,$$

where dl_Q is the line element on $\partial\Omega$, and $\rho(Q)$ is the solution of the Fredholm integral equation of the first kind

$$\int_{\partial\Omega} \rho(Q) \log |P - Q| \, dl_Q = f(P), \quad P \in \partial\Omega.$$

This is the so-called *boundary integral equation formulation* of the interior Dirichlet problem. A nice feature of this formulation is that its dimension is one less than that of the original differential equation formulation.
Now, the kernel $g(P,Q) = \log |P - Q|$ of the new formulation is smooth for all $Q \in \partial\Omega$, except when $Q = P$, where it has a logarithmic singularity. This is easily seen after parameterizing the closed curve $\partial\Omega$ appropriately. Also, in terms of the parameter used to represent $\partial\Omega$, $g(P,Q)$ is periodic in P and Q, and the solution $\rho(P)$ and the right-hand side $f(P)$ are periodic in P. In addition, if $\partial\Omega$ is smooth, then $\rho(P)$, $P \in \partial\Omega$, is also a smooth function. These facts can be used to advantage when treating the integral equation numerically.

2 Treatment of Kernels with Jump Discontinuities

We begin with integral equations of the form

$$y(x) = \int_0^1 g(x,t)F(t,y(t))\,dt, \quad 0 \le x \le 1, \tag{1}$$

where (i) $r(x)$ is continuous on $[0,1]$, (ii) $g(x,t)$ is continuous over the set $[0,1]\times[0,1]$, its partial derivatives $g_k(x,t) \equiv \frac{\partial^k}{\partial t^k} g(x,t)$, $k = 1,2,\dots$, are smooth for $0 \le t \le x$ and for $x \le t \le 1$, with finite jump discontinuities across $t = x$, namely, $g_k(x,x+) - g_k(x,x-) = \delta_k(x) \in C[0,1]$, $k = 1,2,\dots$, and (iii) $F(x,y)$ is linear or nonlinear in y and smooth on some set $\Delta = [0,1]\times[R_1,R_2]$, where $[R_1,R_2]$ is a finite, semi-infinite, or infinite interval.

The numerical treatment of this integral equation has been considered recently by Sidi and Pennline [8], and we follow this treatment here.

If a solution to (1) exists, then it can be shown that it is smooth over $[0,1]$. We assume that a unique solution $y(x)$ exists and that $R_1 \le y(x) \le R_2$ when $0 \le x \le 1$. A sufficient condition for this to be true is

$$\left[\max_{x\in[0,1]} \int_0^1 |g(x,t)|\,dt\right]\left[\max_{(x,w)\in\Delta} \left\|\frac{\partial}{\partial w}F(x,w)\right\|\right] < 1.$$

Choose a positive integer N and set $h = 1/N$ and $x_i = ih$, $i = 0,1,\dots,N$. Next, for $x = x_i$, $i \in \{0,1,\dots,N\}$, we approximate the integral

$$I(x) = \int_0^1 \phi(t)\,dt, \quad \phi(t) = g(x,t)F(t,y(t)),$$

via the standard trapezoidal rule

$$T(h) = h{\sum_{j=0}^{N}}''\phi(x_j), \quad {\sum_{j=0}^{N}}''\alpha_j = \frac{1}{2}\alpha_0 + \sum_{i=1}^{N-1}\alpha_j + \frac{1}{2}\alpha_N.$$

Now, by the fact that $F(t,y)$ is smooth in $[0,1]\times[R_1,R_2]$ and by the assumed properties of $g(x,t)$, $\phi(t)$ is smooth for $t \in [0,x]$ and $t \in [x,1]$, continuous for $t \in [0,1]$ with finite jump discontinuities across $t = x$ in all its derivatives. Consequently, for $x = x_i \in (0,1)$, i.e., for $i \in \{1,2,\dots,N-1\}$, $T(h)$ has the following Euler–Maclaurin expansion:

$$\begin{aligned} I(x) = T(h) &- \sum_{s=1}^{p-1} \frac{B_{2s}}{(2s)!}\left[\phi^{(2s-1)}(1) - \phi^{(2s-1)}(0)\right]h^{2s} \\ &+ \sum_{s=1}^{p-1} \frac{B_{2s}}{(2s)!}\left[\phi^{(2s-1)}(x+) - \phi^{(2s-1)}(x-)\right]h^{2s} + E(h), \end{aligned} \tag{2}$$

where $E(h) = O(h^{2p})$ as $h \to 0$, and uniformly in i. For $x = x_0 = 0$ and $x = x_N = 1$, the summation involving the terms containing $\left[\phi^{(2s-1)}(x+) - \phi^{(2s-1)}(x-)\right]$ is absent from (2).

Obviously, $I(x) - T(h) = O(h^2)$ as $h \to 0$, and thus the accuracy of $T(h)$ is simply too low. The formula

$$T_p(h) = T(h) - \sum_{s=1}^{p-1} \frac{B_{2s}}{(2s)!}\big[\phi^{(2s-1)}(1) - \phi^{(2s-1)}(0)\big]h^{2s} + \sum_{s=1}^{p-1} \frac{B_{2s}}{(2s)!}\big[\phi^{(2s-1)}(x+) - \phi^{(2s-1)}(x-)\big]h^{2s}, \tag{3}$$

on the other hand, satisfies $I(x) - T_p(h) = O(h^{2p})$ as $h \to 0$. Since $T_p(h)$ involves derivatives of the unknown function $y(x)$, we cannot make direct use of it, unfortunately. However, we can approximate $T_p(h)$ via another formula $\hat{T}_p(h)$, involving only the values of $y(x)$ and no derivatives of it, for which $I(x) - \hat{T}_p(h) = O(h^{2p})$ as $h \to 0$ as well.

We show how this is done only for $p = 2$, and refer the the reader to [8] for all other values of p. We have, for $x = x_i$, $i \in \{1, 2, \dots, N-1\}$,

$$T_2(h) = T(h) - \frac{B_2}{2}\big\{\big[\phi'(1) - \phi'(0)\big] - \big[\phi'(x+) - \phi'(x-)\big]\big\}h^2.$$

We first break up $\phi'(t)$ in the form

$$\phi'(t) = g_1(x,t)F(t,y(t)) + g(x,t)\frac{d}{dt}F(t,y(t)).$$

Then we approximate $\phi'(0)$ and $\phi'(1)$ by using one-sided numerical differentiation formulas for $\frac{d}{dt}F(t,y(t))$ at $t = 0$ and $t = 1$. These are

$$Q'(0) = \frac{1}{2h}\big[-3Q(0) + 4Q(h) - Q(2h)\big] + \frac{1}{3}Q'''(\xi)h^2, \quad 0 < \xi < 2h,$$

$$Q'(1) = \frac{1}{2h}\big[3Q(1) - 4Q(1-h) + Q(1-2h)\big] + \frac{1}{3}Q'''(\xi)h^2, \quad 1 - 2h < \xi < 1.$$

Next, by continuity of $g(x,t)$ across $t = x$, we have

$$\phi'(x+) - \phi'(x-) = \big[g_1(x,x+) - [g_1(x,x-)\big]F(x,y(x)) = \delta_1(x)F(x,y(x)).$$

$\hat{T}_2(h)$ is obtained by combining these in $T_2(h)$, and we obviously have $I(x) - \hat{T}_2(h) = O(h^4)$ as $h \to 0$. We finally use $\hat{T}_2(h)$ to define our quadrature method for the integral equation (1).

The resulting quadrature method is defined via the following $N+1$ equations:

$$\begin{aligned} y_i = r(x_i) &+ h{\sum_{j=0}^{N}}'' g(x_i,x_j)F_j \\ &- \frac{h}{24}\big[g(x_i,1)(3F_N - 4F_{N-1} + F_{N-2}) - g(x_i,0)(-3F_0 + 4F_1 - F_2)\big] \\ &- \frac{h^2}{12}\big[g_1(x_i,1)F_N - g_1(x_i,0)F_0\big] + \frac{h^2}{12}\delta_1(x_i)F_i, \quad i = 1,2,\dots,N-1, \end{aligned}$$

$$y_0 = r(x_0) + h\sum_{j=0}^{N}{}''g(x_0,x_j)F_j - \frac{h}{24}\big[g(0,1)(3F_N - 4F_{N-1} + F_{N-2}) - g(0,0)(-3F_0 + 4F_1 - F_2)\big] - \frac{h^2}{12}\big[g_1(0,1)F_N - g_1(0,0+)F_0\big]$$

$$y_N = r(x_N) + h\sum_{j=0}^{N}{}''g(x_N,x_j)F_j - \frac{h}{24}\big[g(1,1)(3F_N - 4F_{N-1} + F_{N-2}) - g(1,0)(-3F_0 + 4F_1 - F_2)\big] - \frac{h^2}{12}\big[g_1(1,1-)F_N - g_1(1,0)F_0\big].$$

Here $F_i \equiv F(x_i, y_i)$ and y_i is the approximation to $y(x_i)$.

Using precisely the same procedure, we can derive a quadrature method for (1) that is based on the quadrature formula $\hat{T}_p(h)$. Again, this method involves the numerical solution of $N+1$ equations in the unknowns y_j. As shown in [8], under appropriate conditions, this system has a unique solution for the y_j when N is sufficiently large, and that $y(x_j) - y_j = O(h^{2p})$ as $h \to 0$ (equivalently, $N \to \infty$), this error being uniform in j.

The quadrature method based on $\hat{T}_2(h)$ is illustrated with two examples involving nonlinear two-point boundary value problems in [8].

3 Treatment of Kernels with Logarithmic Singularity

We now turn to the numerical solution of Fredholm integral equations of the form

$$\omega f(t) + \int_a^b K(t,x)f(x)\,dx = g(t), \quad a \le t \le b; \quad \omega = 0, 1. \tag{4}$$

(Such equations are of the first or the second kind depending on whether $\omega = 0$ or $\omega = 1$, respectively.) The equations that are of interest to us here have the following important features: (i) The kernel $K(t,x)$ is periodic both in t and in x with period $T = b-a$ and is infinitely differentiable in $(-\infty,\infty)\setminus\{t+kT\}_{k=-\infty}^{\infty}$. It has a polar singularity and/or a branch singularity (of algebraic or logarithmic or algebraic-logarithmic type) at $x = t$ when $t \in [a,b]$. (ii) The input function $g(t)$ and the solution $f(t)$ are both periodic with period $T = b-a$ and infinitely smooth on $(-\infty,\infty)$.

In case $K(t,x)$ has an integrable singularity across $x = t$, (4) is said to be *weakly singular*. In case $K(t,x) \sim c/(x-t)$ as $x \to t$ for some constant $c \neq 0$, and the integral $\int_a^b K(t,x)f(x)\,dx$ is defined only in the Cauchy principal value sense, it is said to be *singular*.

The numerical treatment of these integral equations has been considered recently by Sidi and Israeli [7]. A comparison of the quadrature formulas proposed there with others has been given in [5]. For an extensive summary of the subject, see Sidi [6, Section 25.3 and Appendix D]. Here we restrict our attention to weakly singular integral equations with only logarithmic singularities across $x = t$. For other types of singularities, the reader is referred to [7] and [6].

A logarithmically singular kernel can be expressed in the form

$$K(t,x) = H_1(t,x)\log|t-x| + H_2(t,x), \tag{5}$$

where $H_1(t,x)$ and $H_2(t,x)$ are infinitely differentiable for all $t, x \in [a,b]$ (including $t = x$), but are not necessarily periodic. We assume that $H_1(t,t) \not\equiv 0$. Note that, in our method below, we do not require $H_1(t,x)$ and $H_2(t,x)$ for all x, but only for $x = t$; $H_1(t,t)$ and $H_2(t,t)$ can be obtained simply by expanding $K(t,x)$ about $x = t$.

Let n be a positive integer and set $h = T/n$ and $x_i = a + ih$, $i = 1, \dots, n$. In addition, let $t \in \{x_1, \dots, x_n\}$. Let us now define the "corrected" trapezoidal rule approximations $I[h;t]$ to the integral $I[t] = \int_a^b K(t,x)f(x)\,dx$ as in

$$I[h;t] = \sum_{i=1}^{n} w_n(t,x_i)f(x_i), \tag{6}$$

where

$$w_n(t,x) = hK(t,x) \quad \text{for } x \neq t, \quad w_n(t,t) = h\left[H_2(t,t) + H_1(t,t)\log\left(\frac{h}{2\pi}\right)\right]. \tag{7}$$

$I[h;t]$ has the asymptotic (Euler–Maclaurin) expansion

$$I[h;t] \sim I[t] + \sum_{k=1}^{\infty} \beta_k(t)h^{2k+1} \quad \text{as } h \to 0, \tag{8}$$

where

$$\beta_k(t) = -2\frac{\zeta'(-2k)}{(2k)!}\frac{\partial^{2k}}{\partial x^{2k}}[H_1(t,x)f(x)]\big|_{x=t}, \quad k = 1, 2, \dots\ . \tag{9}$$

Here $\zeta(s)$ is the Riemann Zeta function. From these expansions, it follows that $I[h;t] - I[t] = O(h^3)$ as $h \to 0$.

The quadrature method based on the rule $I[h;t]$ is now defined by the equations

$$\omega\tilde{f}_k + I[h;x_k] = g(x_k), \quad k = 1, 2, \dots, n. \tag{10}$$

More explicitly, these equations are

$$\omega\tilde{f}_k + \sum_{i=1}^{n} w_n(x_k,x_i)\tilde{f}_i = g(x_k), \quad k = 1, 2, \dots, n, \tag{11}$$

where $\tilde{f}_i$ is the approximation to $f(x_i)$. In general, the accuracy of the $\tilde{f}_i$ is the same as that of the underlying numerical quadrature formula, which is $I[h;t]$ in this case. We can increase the accuracy of the quadrature method by increasing that of $I[h;t]$, which we propose to achieve by using extrapolation. What makes this possible is the *periodicity* of the integrand $K(t,x)f(x)$ as a function of x. We turn to this subject next.

We start by using only one extrapolation to eliminate the term $\beta_1(t)h^3$ from the asymptotic expansion of $I[h;t]$. Let us choose $h = T/n$ for some even integer n and let $x_i = a + ih$, $i = 0, 1, \ldots, n$. Performing this single extrapolation, we obtain the Romberg-type quadrature rule

$$J_1[h;t] = \frac{8}{7}I[h;t] - \frac{1}{7}I[2h;t] \tag{12}$$

as the new approximation to $I[t]$. We also have

$$J_1[h;t] \sim I[t] + \sum_{k=2}^{\infty} \frac{2^3 - 2^{2k+1}}{7}\beta_k(t)h^{2k+1} \quad \text{as } h \to 0, \tag{13}$$

hence $J_1[h;t] - I[t] = O(h^5)$ as $h \to 0$. The quadrature method for (4) based on $J_1[h;t]$ is thus

$$\omega\tilde{f}_k + J_1[h;x_k] = g(x_k), \quad k = 1, 2, \ldots, n. \tag{14}$$

More explicitly,

$$\omega\tilde{f}_k + \sum_{i=1}^{n}\left[\frac{8}{7}w_n(x_k,x_i) - \frac{1}{7}\epsilon_{k,i}^{(1)}w_{n/2}(x_k,x_i)\right]\tilde{f}_i = g(x_k), \quad k = 1, 2, \ldots, n, \tag{15}$$

where

$$\epsilon_{k,i}^{(1)} = \begin{cases} 1 & \text{if } k-i \text{ even}, \\ 0 & \text{if } k-i \text{ odd}. \end{cases} \tag{16}$$

By applying two extrapolations, we can remove the terms $\beta_k(t)h^{2k+1}$, $k = 1, 2$, from the asymptotic expansion of $I[h;t]$. This time we choose $h = T/n$ for an integer n that is divisible by 4, and let $x_i = a+ih$, $i = 0, 1, \ldots, n$. Performing the two extrapolations, we obtain the Romberg-type quadrature rule

$$\begin{aligned} J_2[h;t] &= \frac{32}{31}J_1[h;t] - \frac{1}{31}J_1[2h;t] \\ &= \frac{256}{217}I[h;t] - \frac{40}{217}I[2h;t] + \frac{1}{217}I[4h;t] \end{aligned} \tag{17}$$

as the new approximation to $I[t]$. We also have

$$J_2[h;t] \sim I[t] + \sum_{k=3}^{\infty} \frac{2^3 - 2^{2k+1}}{7} \cdot \frac{2^5 - 2^{2k+1}}{31}\beta_k(t)h^{2k+1} \quad \text{as } h \to 0, \tag{18}$$

hence $J_2[h;t] - I[t] = O(h^7)$ as $h \to 0$. The quadrature method for (4) based on $J_2[h;t]$ is thus

$$\omega \tilde{f}_k + J_2[h; x_k] = g(x_k), \quad k = 1, 2, \ldots, n. \tag{19}$$

More explicitly,

$$\omega \tilde{f}_k + \sum_{i=1}^{n} \left[\frac{256}{217} w_n(x_k, x_i) - \frac{40}{217} \epsilon_{k,i}^{(1)} w_{n/2}(x_k, x_i) + \frac{1}{217} \epsilon_{k,i}^{(2)} w_{n/4}(x_k, x_i) \right] \tilde{f}_i = g(x_k), \quad k = 1, 2, \ldots, n, \tag{20}$$

where $\epsilon_{k,i}^{(1)}$ are as before and

$$\epsilon_{k,i}^{(2)} = \begin{cases} 1 & \text{if } k - i \text{ divisible by 4,} \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

For the development of Romberg-type formulas of all orders for all types of weak singularities, we refer the reader to Sidi and Israeli [7].

References

1. H.B. Keller. *Numerical Methods for Two-Point Boundary Value Problems*. Blaisdell, Waltahm, Mass., 1968.
2. J.A. Pennline. Improving convergence rate in the method of successive approximation. *Math. Comp.*, 37:127–134, 1981.
3. J.A. Pennline. Constructive existence and uniqueness for some nonlinear two-point boundary value problems. *J. Math. Anal. Appl.*, 96:584–598, 1983.
4. J.A. Pennline. Constructive existence and uniqueness for two-point boundary value problems with a linear gradient term. *Appl. Math. Comp.*, 15:233–260, 1984.
5. A. Sidi. Comparison of some numerical quadrature formulas for weakly singular periodic Fredholm integral equations. *Computing*, 43:159–170, 1989.
6. A. Sidi. *Practical Extrapolation Methods: Theory and Applications*. Number 10 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2003.
7. A. Sidi and M. Israeli. Quadrature methods for periodic singular and weakly singular Fredholm integral equations. *J. Sci. Comput.*, 3:201–231, 1988. Originally appeared as Technical Report No. 384, Computer Science Dept., Technion–Israel Institute of Technology, (1985), and also as ICASE Report No. 86-50 (1986).
8. A. Sidi and J.A. Pennline. Improving the accuracy of quadrature method solutions of Fredholm integral equations that arise from nonlinear two-point boundary value problems. *J. Integral Eqs. Appl.*, 11:103–139, 1999.

Extrapolation Techniques for Computing Accurate Solutions of Elliptic Problems with Singular Solutions

H. Koestler[1] and U. Ruede[2]

[1] University of Erlangen-Nuremberg, Germany,
Harald.Koestler@informatik.uni-erlangen.de
[2] University of Erlangen-Nuremberg, Germany,
Ulrich.Ruede@informatik.uni-erlangen.de

Abstract. Generalized functions occur in many practical applications as source terms in partial differential equations. Typical examples are point loads and dipoles as source terms for electrostatic potentials. For analyzing the accuracy of such computations, standard techniques cannot be used, since they rely on global smoothness. At the singularity, the solution tends to infinity and therefore standard error norms will not even converge.
In this article we will demonstrate that these difficulties can be overcome by using other metrics to measure accuracy and convergence of the numerical solution. Only minor modifications to the discretization and solver are necessary to obtain the same asymptotic accuracy and efficiency as for regular and smooth solutions. In particular, no adaptive refinement is necessary and it is also unnecessary to use techniques which make use of the analytic knowledge of the singularity. Our method relies simply on a mesh-size dependent representation of the singular sources constructed by appropriate smoothing. It can be proved that the pointwise accuracy is of the same order as in the regular case. The error coefficient depends on the location and will deteriorate when approaching the singularity where the error estimate breaks down. Our approach is therefore useful for accurately computing the global solution, except in a small neighborhood of the singular points. It is also possible to integrate these techniques into a multigrid solver exploiting additional techniques for improving the accuracy, such as Richardson and τ-Extrapolation.

1 Introduction

Typical error estimates for the numerical solution of boundary value problems depend on the smoothness of the true solution which is not given in many practical applications. Reasons for such singular solutions can for example be reentrant corners, discontinuous coefficients, singular functions in the boundary conditions or source terms with singularities. In this article we consider the last case. As application we choose for simplicity electrostatic potentials of point loads, physically modeled by the Maxwell equations in the vacuum (cf. [1]). The method

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 410–417, 2004.
© Springer-Verlag Berlin Heidelberg 2004

is extensible to more general situations, but the basic idea of the method is presented here in terms of this simple example. It leads to the **Poisson equation** with Dirichlet boundary conditions in the unit cube $\Omega = [0,1]^3 \subset \mathbb{R}^3$

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega \end{aligned} \quad , \tag{1}$$

where Δ is the Laplace operator, $\partial\Omega$ denotes the boundary of Ω and g is a smooth function. The source term f contains one or more singularities, because a point load is modeled by the Dirac δ-function. As an extension we will consider sources containing dipoles and quadrupoles. A dipole can formally introduced as a directional derivative of a point load, a quadrupole as a directional derivative of a dipole. So we have to study the Dirac δ-function and its derivatives in order to solve problem (1). Note that the Dirac δ-function is not a function in the common sense (cf. [2]). To be able to deal with it the concept of a function is extended to the concept of a distribution or generalized function (cf. [3]).

For the discretization of (1) we use finite differences on equidistant grids with mesh size $h = \frac{1}{N}$. The discrete Laplace operator is given by the usual 7-point stencil (cf. [4], p.71).

The question now is how to discretize the right hand side f containing the singularity. f is equal to zero on every point in Ω except at the singularity where it is ∞. If we assume that the singularity is not located at a grid point, simply set $f_h \equiv 0$ and then try to solve the above problem without any modifications and a standard solver, we would get poor results. On the one hand, standard error norms near the singularity will not converge, because the solution at the singularity is unbounded, and on the other hand the accuracy is destroyed in the whole domain by the singularity. This phenomenon is called **pollution effect**.

In section 2 we present the **Zenger Correction Method** to overcome the pollution effect. The idea is to represent the singular generalized components of f by grid-adapted B-Splines. Extending results from [5] we can show that the pollution effect can be eliminated leading to errors of the same quality and convergence rate as for smooth problems. Its great advantage is that we do not need to know the exact singular component of the solution of the problem to approximate the singularity. Therefore it can be used for a variety of problems.

In section 3 the core theorems for the error estimates of the discretization error for the Poisson equation in the unit cube are presented and the proof of the main result is outlined.

In section 4 we further improve the accuracy of the numerical solution by using extrapolation. Two extrapolation methods, namely Richardson extrapolation and τ-extrapolation are briefly described in combination with the Zenger Correction.

In the 5th section the experimental results for the numerical solution of the Poisson equation with Zenger Correction are summarized.

2 Singular Source Terms in Poisson's Equation

The Zenger Correction Method uses the following generalized functions (cf. [2]) in order to approximate the physical singularities contained in the right hand side f of equation (1).

Definition 1 (Generalized functions H_i). *Let $H_0 : \mathbb{R} \to \mathbb{R}$ be the **Heaviside-function***

$$H_0(x) := \begin{cases} 0 & : \quad x \leq 0 \\ 1 & : \quad x > 0 \end{cases} , \tag{2}$$

and the distributions H_i, $i \in \mathbb{Z}$, be recursively defined by

$$H_i(x) := \begin{cases} \dfrac{d}{dx} H_{i-1}(x) & : \quad i > 0 \\ \displaystyle\int_{-\infty}^{x} H_{i+1}(\xi) d\xi & : \quad i < 0 \end{cases} . \tag{3}$$

The family of functions H_i enables us to represent arbitrary physical multipoles. For example H_1 corresponds to the Dirac δ-function, resp. a point load. Remember that the directional derivative of a point load was defined as a special dipole. For higher dimensions we use tensor products of these functions. For the two multi indices $\boldsymbol{x}_0 = (x_{0,1}, x_{0,2}, \ldots, x_{0,n}) \in \mathbb{R}^n$ that indicates the location of the singularity and $\boldsymbol{a} = (a_1, a_2, \ldots, a_n) \in \mathbb{Z}^n$ that specifies the functions H_i we use the notation

$$\mathcal{H}^{\boldsymbol{a}}_{\boldsymbol{x}_0}(\boldsymbol{x}) = \prod_{i=1}^{n} H_{a_i}(x_i - x_{0,i}) \ . \tag{4}$$

The idea of the Zenger Correction Method is now to **integrate** the right hand side f **analytically** a number of times, until the result is a product of smooth functions H_i with $i < 0$. Then we **differentiate** this function **numerically** as often as we had integrated it. This results in a approximation to the singularity which becomes more accurate with smaller mesh size h. For the numerical differentiation we use finite differences with the notation

$$\delta^{\boldsymbol{i}} = \delta^{i_1}_{x_1} \circ \delta^{i_2}_{x_2} \circ \cdots \circ \delta^{i_n}_{x_n} \ , \tag{5}$$

for the multi index $\boldsymbol{i} \in \mathbb{Z}^n$ that indicates how often we differentiate in each direction. Dependent on the number of integrations resp. differentiation steps n we call that procedure the **Zenger Correction of n-th order**. In general for even n with $k - n < 0$ we have

$$\delta^n_x H_k(x) = \begin{cases} 0 & : \quad |x| \geq \dfrac{n}{2} h \\ \dfrac{1}{h^n} \displaystyle\sum_{i=0}^{n} (-1)^i \binom{n}{i} H_{k-n}(x + (i - \frac{n}{2})h) & : \quad |x| < \dfrac{n}{2} h \end{cases} . \tag{6}$$

One can show that (6) is identical to a B-Spline (cf. [6]).

3 Error Estimates

In this section we prove that if the singularity in the source term is replaced by the above approximation, we obtain an $O(h^2)$ discretization error as in the smooth case, except in a small area near the singularity.

Definition 2 (H-bounded). *A family of functions $u_h(\boldsymbol{x})$ is called **h-bounded** on the domain $\Omega \subset \mathbb{R}^n$, if there exists a real valued, continuous function $r(\boldsymbol{x})$ on Ω which is not necessarily bounded on Ω, so that for every $\boldsymbol{x} \in \Omega$ there exists a number $h_0 > 0$ with $|u_h(\boldsymbol{x})| \leq r(\boldsymbol{x})$ for all $h = \frac{1}{N} < h_0, N \in \mathbb{N}$, $\boldsymbol{x} \in \Omega_h$. If $r(\boldsymbol{x})$ is bounded on Ω, $u_h(\boldsymbol{x})$ is called **strictly h-bounded** (cf. [7], p.6).*

A h-bounded family of grid functions u_h may be unbounded on Ω for $h \to 0$, but because of the continuity of r be bounded for all $h > 0$ on every compact subset of Ω.

Theorem 1. *Let the solution of $\Delta_h u_h = f_h$ in Ω_h be bounded on $\overline{\Omega}_h$. If $\delta^{2\boldsymbol{i}} f_h$ is h-bounded on Ω and for all i_l, m_l with $0 \leq i_l \leq m_l$ for $0 < l \leq n$, then $\delta^{2\boldsymbol{m}} u_h$ is h-bounded on Ω.*

The proof of this theorem is found in [8] for the 2D case and will be generalized to 3D in a forthcoming paper.

Now we are prepared for the central theorem.

Theorem 2. *Let u^* be the (weak) solution of the boundary value problem*

$$\begin{array}{rl} \Delta u = \mathcal{H}^{\boldsymbol{a}}_{\boldsymbol{x}_0}(\boldsymbol{x}) & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{array} , \tag{7}$$

where $\mathcal{H}^{\boldsymbol{a}}_{\boldsymbol{x}_0}(\boldsymbol{x})$ is a singularity located in $\boldsymbol{x}_0 \in \Omega = [0,1]^n$. $n \in \{2,3\}$ is the dimension of the problem. Let u^*_h the solution of

$$\begin{array}{rl} \Delta_h u_h = f_h & \text{in } \Omega_h \\ u_h = 0 & \text{on } \partial\Omega_h \end{array} , \tag{8}$$

where $f_h = \delta^{2\boldsymbol{m}} \mathcal{H}^{\boldsymbol{a}-2\boldsymbol{m}}_{\boldsymbol{x}_0}(\boldsymbol{x})$, and where $\boldsymbol{m}$ is chosen componentwise such that $2m_l > a_l$ for $1 \leq l \leq n$. By definition it follows that $\mathcal{H}^{\boldsymbol{a}-2\boldsymbol{m}}_{\boldsymbol{x}}$ are continuous functions. Then:

$$u^*_h = u^* + h^2 r$$

where r is h-bounded on $\Omega \backslash \{\boldsymbol{x}_0\}$.

The proof can be found in [5], pp.15 and can be extended to the 3D case.

The advantages of the Zenger Correction Method are that no modification of the grid or the solver is necessary. Furthermore the number of points that have to be corrected is fixed and does not depend on the mesh size h. The analytic solution is not needed to construct the correction.

Note that the Zenger Correction Method eliminates the pollution effect. This results in a (pointwise) $\mathcal{O}(h^2)$ accuracy at any fixed distance from the singular point. However, the method cannot provide locally good accuracy. This is impossible since the true solution tends to infinity where the singularity is located.

4 Extrapolation

In this section we present two extrapolation methods in order to improve the discretization error from $\mathcal{O}(h^2)$ to $\mathcal{O}(h^4)$.

4.1 Richardson Extrapolation

Richardson Extrapolation can be used if there exist asymptotic expansions of the discretization error (cf. [9]). In this case the solutions of different mesh sizes can be combined to eliminate the lower order terms. For our problem we use the mesh sizes h und $H = 2h$. In order to get the higher accuracy on the coarse grid we change the values there by

$$\widehat{\boldsymbol{u}}_H^* = \frac{4}{3}\mathcal{I}_h^H \boldsymbol{u}_h^* - \frac{1}{3}\boldsymbol{u}_H^* \ , \tag{9}$$

where $\mathcal{I}_h^H$ is an injection operator. The existence of such asymptotic expansions can be proved even in the case of singularities by extension of Theorem 2.

4.2 τ-Extrapolation

τ-Extrapolation is a multigrid specific technique that in contrast to Richardson extrapolation works only on a single grid. It is based on the principle of **defect correction** and has been first mentioned by Brandt (cf. [10], see also Hackbusch [11], pp.278).

In the CS(correction scheme)-Multigrid algorithm two different iterations are used alternately, the smoother and the coarse grid correction (cf. [4]). These two iterations have a common fixed point described by $\boldsymbol{f}_h - \boldsymbol{A}_h \boldsymbol{u}_h = 0$ (cf. [5], p. 17f). The smoother converges fast for certain (usually the high frequency) solution components, but converges only slowly for the remaining (low frequency) modes. The coarse grid correction behaves vice versa. If these complementary properties are combined the typical multigrid efficiency is obtained.

Now we follow the idea of **double discretization**, i.e. in the coarse grid correction process higher order discretizations are used. Using a correction of the form

$$\boldsymbol{u}_h^{(k+1)} = \boldsymbol{u}_h^{(k)} + \boldsymbol{e}_h^{(k)} \ , \tag{10}$$

where $\boldsymbol{e}_h^{(k)}$ is computed as a coarse grid correction

$$\boldsymbol{e}_h^{(k)} = \mathcal{I}_H^h \boldsymbol{A}_H^{-1} \widehat{\mathcal{I}}_h^H (\boldsymbol{f}_h - \boldsymbol{A}_h \boldsymbol{u}_h^{(k)}) \ , \tag{11}$$

would lead to a standard multigrid method. τ-extrapolation consists in using a linear combination of fine and coarse grid residual to construct an extrapolated correction

$$\widehat{\boldsymbol{u}}_h^{(k+1)} = \boldsymbol{u}_h^{(k)} + \mathcal{I}_H^h \boldsymbol{A}_H^{-1} (\frac{4}{3} \widehat{\mathcal{I}}_h^H (\boldsymbol{f}_h - \boldsymbol{A}_h \boldsymbol{u}_h^{(k)}) - \frac{1}{3} (\widehat{\mathcal{I}}_h^H \boldsymbol{f}_h - \boldsymbol{A}_H \mathcal{I}_h^H \boldsymbol{u}_h^{(k)})) \ . \quad (12)$$

It can be shown that this modification of the coarse grid correction leads to a numerical error of order $\mathcal{O}(h^4)$ (cf. [12]). The modified coarse grid correction is only applied on the finest grid once per V-cycle. Additionally we have to take care when choosing the restriction and the interpolation operators. Normally trilinear interpolation for $\mathcal{I}_H^h$, full weighting for $\widehat{\mathcal{I}}_h^H$ and injection for $\mathcal{I}_h^H$ is used, but this can vary from problem to problem. One has also to pay attention not do do too many post smoothing steps, because this can destroy the higher accuracy. For the Poisson equation with singular source term we have to discretize the right hand side on each grid due to the fact that the restriction of the B-spline cannot approximate the right hand side well enough on the coarse grid. A concise analysis of the τ-extrapolation is e.g. found in [12].

5 Experimental Results for a Point Load in 3D

For the experiments we use CS-Multigrid as solver, e.g $CS(2, 2, 15)$ means that we do 2 presmoothing and 2 postsmoothing steps and a maximum of 15 V-cycles (cf. [13]). The singularity is located at $\boldsymbol{x}_0 = (0.26, 0.26, 0.26)^T$ in the domain $\Omega = [0, 1]^3$. To evaluate the accuracy away from the singularity we will consider $\Omega \setminus R$, where $R = [0.125, 0.375]^3$ is a fixed neighbourhood of $\boldsymbol{x}_0$.

The analytical solution of the Poisson equation in 3D with a point load as source term is given by

$$u_p^*(\boldsymbol{x}) = -\frac{1}{4\pi|\boldsymbol{x}|} \ . \quad (13)$$

The boundary value problem with Zenger Correction of 4th order is described by

$$\begin{aligned} -\Delta u(\boldsymbol{x}) &= \mathcal{H}_{\boldsymbol{x}_0}^{(1,1,1)}(\boldsymbol{x}) \quad \text{in } \Omega \\ u(\boldsymbol{x}) &= u_p^*(\boldsymbol{x}) + g(\boldsymbol{x}) \text{ on } \partial\Omega \end{aligned} \quad (14)$$

with its discretization

$$\begin{aligned} -\Delta_h u_h(\boldsymbol{x}) &= \delta^{(4,4,4)} \mathcal{H}_{\boldsymbol{x}_0}^{(-3,-3,-3)}(\boldsymbol{x}) \text{ in } \Omega_h \\ u_h(\boldsymbol{x}) &= u_{p,h}^*(\boldsymbol{x}) + g_h(\boldsymbol{x}) \qquad \text{on } \partial\Omega_h \end{aligned} \ , \quad (15)$$

where $g(\boldsymbol{x}) = \sin(x\pi)\sin(y\pi)\sinh(\sqrt{2}z\pi)$. Table 1 lists the numerical results. The first column shows the mesh size h, the second the maximum norm of the discretization error, then follow the L_1 resp. L_2 norms in the whole domain Ω and in the domain $\Omega \setminus R$. The small numbers between the rows of the table show the numerical convergence rates α which are for a point $\boldsymbol{p} \in \Omega$ computed by

$$\alpha = (\ln|u^*(\boldsymbol{p}) - u_h(\boldsymbol{p})| - \ln|u^*(\boldsymbol{p}) - u_{h/2}(\boldsymbol{p})|)/(\ln 2) \quad (16)$$

and analogous for the norms in the other columns.

Table 1. Convergence rates of the discretization error.

h	$(\frac{1}{2},\frac{1}{2},\frac{1}{2})$	L_∞	L_1 Ω	L_2 Ω	L_1 $\Omega\backslash R$	L_2 $\Omega\backslash R$
$\frac{1}{16}$	4.81e-02	2.67e+00	2.04e-02	5.28e-02	1.98e-02	2.61e-02
	2.0		2.1	2.4	2.1	2.1
$\frac{1}{32}$	1.21e-02	1.22e+00	4.61e-03	9.89e-03	4.52e-03	6.22e-03
	2.0		2.0	1.2	2.1	2.0
$\frac{1}{64}$	3.02e-03	1.69e+00	1.12e-03	4.34e-03	1.08e-03	1.52e-03
	2.0		2.0	-0.2	2.0	2.0
$\frac{1}{128}$	7.55e-04	6.92e+00	2.78e-04	5.07e-03	2.63e-04	3.74e-04
	2.0		2.0	1.9	2.0	2.0
$\frac{1}{256}$	1.89e-04	3.21e+00	6.91e-05	1.39e-03	6.50e-05	9.30e-05

Using an additional Richardson extrapolation or additional τ-extrapolation for solving problem (14) we obtain the improved accuracy rates, as shown in Table 2 and 3, respectively.

Table 2. Convergence rates of the discretization error with Richardson extrapolation.

h	$(\frac{1}{2},\frac{1}{2},\frac{1}{2})$	L_∞	L_1 Ω	L_2 Ω	L_1 $\Omega\backslash R$	L_2 $\Omega\backslash R$
$\frac{1}{16}$	5.77e-05	7.30e-01	3.89e-04	1.28e-02	3.44e-05	5.32e-05
	4.0		3.6	2.9	3.9	3.5
$\frac{1}{32}$	3.70e-06	2.55e-01	3.19e-05	1.69e-03	2.28e-06	4.66e-06
	4.0		2.4	1.0	3.9	3.7
$\frac{1}{64}$	2.33e-07	3.79e-01	6.14e-06	8.71e-04	1.48e-07	3.56e-07
	4.0		1.6	-0.0	4.0	3.9
$\frac{1}{128}$	1.44e-08	1.03e+00	1.97e-06	8.77e-04	9.21e-09	2.36e-08

Table 3. Convergence rates of the discretization error with τ-extrapolation.

h	$(\frac{1}{2},\frac{1}{2},\frac{1}{2})$	L_∞	L_1 Ω	L_2 Ω	L_1 $\Omega\backslash R$	L_2 $\Omega\backslash R$
$\frac{1}{16}$	1.67e-03	2.64e+00	2.54e-03	4.57e-02	1.54e-03	2.61e-03
	3.9		3.4	2.6	3.9	3.8
$\frac{1}{32}$	1.12e-04	1.18e+00	2.46e-04	7.43e-03	1.01e-04	1.85e-04
	4.0		2.5	0.9	4.0	3.9
$\frac{1}{64}$	7.24e-06	1.66e+00	4.30e-05	3.94e-03	6.46e-06	1.23e-05
	4.0		2.0	-0.3	4.0	3.9
$\frac{1}{128}$	4.58e-07	6.85e+00	1.10e-05	4.97e-03	4.09e-07	8.00e-07
	4.0		2.3	1.9	4.0	4.0
$\frac{1}{256}$	2.86e-08	3.09e+00	2.17e-06	1.30e-03	2.55e-08	5.02e-08

6 Conclusion

In this paper we have presented the basic idea of the Zenger Correction Method including some simple examples. More examples, i.e. problems with dipoles and quadrupoles can be found in [6]. Furthermore a concise mathematical analysis of the Zenger Correction will be presented in a forthcoming paper.

References

1. Fließbach, T.: Elektrodynamik. 3rd edn. Spektrum Verlag (2000)
2. Stackgold, I.: Green's Functions and Boundary Value Problems. 2nd edn. John Wiley & Sons (1998)
3. Walter, W.: Einführung in die Theorie der Distributionen. 3rd edn. BI-Wiss.-Verlag (1994)
4. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Academic Press (2001)
5. Rüde, U.: On the accurate computation of singular solutions of Laplace's and Poisson's equation. Multigrid Methods: Theory, Applications, Supercomputing: Proceedings of the Third Copper Mountain Conference on Multigrid Methods, April 5-10 (1987)
6. Köstler, H.: Akkurate Behandlung von Singularitäten bei partiellen Differentialgleichungen. Diplomarbeit, Friedrich-Alexander Universität Erlangen-Nürnberg (2003)
7. Fößmeier, R.: Differenzenverfahren hoher Ordnung für elliptische Randwertprobleme mit gekrümmten Rändern. PhD thesis, Technische Universität München (1984)
8. Zenger, C., Gietl, H.: Improved schemes for the Dirichtlet problem of Poisson's equation in the neighbourhood of corners. Numerische Mathematik **30** (1978) 315–332
9. Richardson, L.: The deferred approach to the limit. I. Single lattice. Phil. Trans. Roy. Soc. London A **226** (1927) 229–349
10. Brandt, A.: On the accurate computation of singular solutions of Laplace's and Poisson's equation. Multigrid Methods: 1984 guide with applications to fluid dynamics, GMD Studie Nr. 85, St. Augustin (1984)
11. Hackbusch, W.: Multi-Grid Methods and Applications. Springer Verlag (1985)
12. Rüde, U.: Multiple τ-extrapolation for multigrid methods. Technical Report I-8701, Technische Universität München (1987)
13. Briggs, W., Henson, V., McCormick, S.: A Multigrid Tutorial. Society for Industrial and Applied Mathematics, Philadelphia (1987)

Vandermonde–Type Matrices in Two Step Collocation Methods for Special Second Order Ordinary Differential Equations

Silvana Martucci[1] and Beatrice Paternoster[2]

[1] Dipartimento di Matematica e Applicazioni, Universitá di Napoli, Italy
[2] Dipartimento di Matematica e Informatica, Universitá di Salerno, Italy

Abstract. We consider a general family of two step nonlinear methods for the numerical integration of Ordinary Differential Equations of type $y'' = f(x, y)$. By applying a collocation technique, linear systems with a Vandermonde–type matrix arise during the construction of the methods. The computation of its determinant reduces to the computation of a recurrence formula depending on the collocation abscissas.

1 Introduction

We are concerned with the derivation of a general family of two–step collocation methods for the numerical integration of second order Ordinary Differential Equations (ODEs), in which the first derivative does not appear explicitly,

$$y''(t) = f(t, y(t)), \qquad y(t_0) = y_0, \quad y'(t_0) = y'_0, \quad y(t), f(t, y) \in R^s. \tag{1}$$

The idea behind polynomial collocation is well known and old [1,3]. After fixing a set of collocation abscissas, the numerical solutions is given by a polynomial which satisfies the differential equation at the given set of collocation points, and which reproduces the values already obtained in the previous step point. One–step collocation methods for ODEs (1) form a subset of implicit Runge–Kutta (RK) methods [3] and have been exstensively studied. Multistep collocation methods were first investigated by Guillon and Soulé [2]. Then Lie and Norsett [5] considered multistep RK of Gauss type, and Hairer and Wanner [4] those of Radau type. The parameters of the k–step s–stage multistep RK method of Radau type are listed in [6].

We extend the procedure indicated in [4], to obtain a general family of two step collocation methods for (1) within the family of two step Runge–Kutta–Nyström (TSRKN) methods, introduced in [8,9], providing numerical approximetions not only for the solution, but also to its first derivative at the step point.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 418–425, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Construction of the Method

The following definition generalizes the Definition 3.2 of [4], to obtain a general family of two step collocation methods for the ODEs (1):

Definition 1. *Let m real numbers $c_1, \ldots, c_m \in [0,1]$, the solution values y_n, y_{n-1} and the derivative values y'_n, y'_{n-1}. The collocation polynomial $P(x)$ of degree $2m+3$ is then defined by:*

$$P(x_{n-1}) = y_{n-1}, \quad P(x_n) = y_n, \quad P'(x_{n-1}) = y'_{n-1}, \quad P'(x_n) = y'_n, \tag{2}$$

$$P''(x_{n-1} + c_i h) = f(x_{n-1} + c_i h, P(x_{n-1} + c_i h)), \tag{3}$$

$$P''(x_n + c_i h) = f(x_n + c_i h, P(x_n + c_i h)). \tag{4}$$

Then the numerical solution of (1) is given by

$$y_{n+1} = P(x_{n+1}), \quad y'_{n+1} = P'(x_{n+1}) \tag{5}$$

(2)–(4) constitute a Hermite interpolation problem with incomplete data, because the function values at $x_n + c_i h$ are missing. Following [4,6], to compute the *collocation polynomial* $P(x)$ for $k = 2$, we introduce the dimensionless coordinate $t = (x - x_n)/h$, $x = x_n + th$, with nodes $t_1 = -1, t_2 = 0$, and define the following polynomials, which constitute a generalized Lagrange basis:

- $\phi_i(t)$, $i = 1, 2,$ of degree $2m+3$, defined by

$$\phi_i(t_j) = \delta_{ij}, \quad \phi'_i(t_j) = 0, \quad i, j = 1, 2, \tag{6}$$

$$\phi''_i(c_j - 1) = 0, \quad \phi''_i(c_j) = 0, \quad i = 1, 2, \quad j = 1, \ldots, m. \tag{7}$$

- $\psi_i(t)$, $i = 1, 2,$ of degree $2m+3$, defined by

$$\psi_i(t_j) = 0, \quad \psi'(t_j) = \delta_{ij}, \qquad i, j = 1, 2, \tag{8}$$

$$\psi''_i(c_j - 1) = 0, \quad \psi''_i(c_j) = 0, \qquad i = 1, 2, \quad j = 1, \ldots, m. \tag{9}$$

- $\chi_{i,n-1}(t)$ and $\chi_{i,n}(t)$, $i = 1, \ldots, m$, of degree $2m+3$, defined by

$$\chi_{i,n-1}(t_j) = 0, \quad \chi_{i,n}(t_j) = 0, \quad i = 1, \ldots, m, \quad j = 1, 2, \tag{10}$$

$$\chi'_{i,n-1}(t_j) = 0, \quad \chi'_{i,n}(t_j) = 0, \quad i = 1, \ldots, m, \quad j = 1, 2, \tag{11}$$

$$\chi''_{i,n-1}(c_j - 1) = \delta_{ij}, \quad \chi''_{i,n-1}(c_j) = 0, \quad i, j = 1, \ldots, m, \tag{12}$$

$$\chi''_{i,n}(c_j - 1) = 0, \quad \chi''_{i,n}(c_j) = \delta_{ij}, \quad i, j = 1, \ldots, m. \tag{13}$$

δ_{ij} denotes the Kronecker tensor. Then the expression of the collocation polynomial $P(x)$ in terms of these polynomials is given by:

$$P(x_n + th) = \phi_1(t)\, y_{n-1} \; + \; \phi_2(t)\, y_n + h(\, \psi_1(t)\, y'_{n-1} \; + \; \psi_2(t)\, y'_n) + h^2 \sum_{j=1}^{m} (\chi_{j,n-1}(t)\, P''(x_{n-1} + c_j h) \; + \chi_{j,n}(t)\, P''(x_n + c_j h)).$$

After constructing $\phi_i(t)$, $\psi_i(t)$, $\chi_{i,n-1}(t)$ and $\chi_{i,n}(t)$, by putting $t = c_i$, writing $P(x_n + c_i h) = Y_n^i$ and inserting the collocation conditions (2)–(4) for $k = 2$, we obtain the expression of the TSRKN collocation method, as the following theorem states:

Theorem 1. *The collocation method defined by (2)–(3)–(4) is equivalent to the TSRKN method:*

$$Y_n^j = u_{j,1} y_{n-1} + u_{j,2} y_n + h(\overline{u}_{j,1} y'_{n-1} + \overline{u}_{j,2} y'_n) + h^2 \sum_{s=1}^{m} (a_{js} f(x_{n-1} + c_s h, Y_{n-1}^s) + b_{js} f(x_n + c_s h, Y_n^j)),$$

$$y_{n+1} = \theta_1 y_{n-1} + \theta_2 y_n + h(\eta_1 y'_{n-1} + \eta_2 y'_n) + h^2 \sum_{j=1}^{m} (v_j f(x_{n-1} + c_j h, Y_{n-1}^j) + w_j f(x_n + c_j h, Y_n^j)),$$

$$h y'_{n+1} = \theta'_1 y_{n-1} + \theta'_2 y_n + h(\eta'_1 y'_{n-1} + \eta'_2 y'_n) + h^2 \sum_{j=1}^{m} (v'_j f(x_{n-1} + c_j h, Y_{n-1}^j) + w'_j f(x_n + c_j h, Y_n^j)),$$

where

$$\theta_i = \phi_i(1), \quad u_{j,i} = \phi_i(c_j), \quad i = 1, 2, \;\; j = 1, \ldots, m \tag{14}$$

$$\eta_i = \psi_i(1), \quad \overline{u}_{j,i} = \psi_i(c_j), \quad i = 1, 2, \;\; j = 1, \ldots, m \tag{15}$$

$$v_j = \chi_{j,n-1}(1), \quad a_{js} = \chi_{j,n-1}(c_s), \quad j, s, = 1, \ldots, m, \tag{16}$$

$$w_j = \chi_{j,n}(1), \quad b_{js} = \chi_{j,n}(c_s), \quad j, s, = 1, \ldots, m \tag{17}$$

$$\theta'_i = \phi'_i(1), \quad \eta'_i = \psi'_i(1), \quad i = 1, 2, \tag{18}$$

$$v'_j = \chi'_{j,n-1}(1), \quad w'_j = \chi'_{j,n}(1), \quad j, = 1, \ldots, m \tag{19}$$

and $\phi_i(t)$, $\psi_i(t)$, $\chi_{i,n-1}(t)$ *and* $\chi_{i,n}(t)$ *are the polynomials defined by the conditions (6)–(13).*

Let us note that the order of approximation of $P(x_n + c_i h)$, that is the so called *stage order* of the TSRKN method, is $2m + 3$.

Let us now show how to construct the polynomials $\phi_i(t)$, $\psi_i(t)$, $\chi_{i,n-1}(t)$ and $\chi_{i,n}(t)$. We will follow the procedure indicate in [6]. We expand the polynomials:

$$\phi_i(t) = \sum_{k=0}^{2m+3} d_k^{(i)} t^k, \quad \psi_i(t) = \sum_{k=0}^{2m+3} e_k^{(i)} t^k, \ \ i = 1, 2,$$

$$\chi_{i,n-1}(t) = \sum_{k=0}^{2m+3} p_k^{(i)} t^k, \quad \chi_{i,n}(t) = \sum_{k=0}^{2m+3} q_k^{(i)} t^k, \ \ i = 1, \ldots m.$$

The following linear systems arise: $Hd^{(i)} = N_0^{(i)}$, $He^{(i)} = N_1^{(i)}$, $Hp^{(i)} = N_{2,n-1}^{(i)}$ and $Hq^{(i)} = N_{2,n}^{(i)}$, of order $2m + 4$ with

$$H = \begin{pmatrix} 1 & t_1 & t_1^2 & t_1^3 & \ldots & t_1^{2m+3} \\ 1 & t_2 & t_2^2 & t_2^3 & \ldots & t_2^{2m+3} \\ 0 & 1 & 2t_1 & 3t_1^2 & \ldots & (2m+3)t_1^{2m+2} \\ 0 & 1 & 2t_2 & 3t_2^2 & \ldots & (2m+3)t_2^{2m+2} \\ 0 & 0 & 2 & 2 \cdot 3(c_1 - 1) & \ldots & (2m+2)(2m+3)(c_1-1)^{2m+1} \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 2 & 2 \cdot 3(c_m - 1) & \ldots & (2m+2)(2m+3)(c_m-1)^{2m+1} \\ 0 & 0 & 2 & 2 \cdot 3c_1 & \ldots & (2m+2)(2m+3)c_1^{2m+1} \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 2 & 2 \cdot 3c_m & \ldots & (2m+2)(2m+3)c_m^{2m+1} \end{pmatrix} \tag{20}$$

$$d^{(i)} = \begin{pmatrix} d_0^{(i)} \\ d_1^{(i)} \\ d_2^{(i)} \\ d_3^{(i)} \\ d_4^{(i)} \\ \vdots \\ \vdots \\ d_{2m+3}^{(i)} \end{pmatrix}, e^{(i)} = \begin{pmatrix} e_0^{(i)} \\ e_1^{(i)} \\ e_2^{(i)} \\ e_3^{(i)} \\ e_4^{(i)} \\ \vdots \\ \vdots \\ e_{2m+3}^{(i)} \end{pmatrix}, p^{(i)} = \begin{pmatrix} p_0^{(i)} \\ p_1^{(i)} \\ p_2^{(i)} \\ p_3^{(i)} \\ p_4^{(i)} \\ \vdots \\ \vdots \\ p_{2m+3}^{(i)} \end{pmatrix}, q^{(i)} = \begin{pmatrix} q_0^{(i)} \\ q_1^{(i)} \\ q_2^{(i)} \\ q_3^{(i)} \\ q_4^{(i)} \\ \vdots \\ \vdots \\ q_{2m+3}^{(i)} \end{pmatrix},$$

$$N_0^{(i)} = \begin{pmatrix} \delta_{i1} \\ \delta_{i2} \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad N_1^{(i)} = \begin{pmatrix} 0 \\ 0 \\ \delta_{i1} \\ \delta_{i2} \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad N_{2,n-1}^{(i)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \delta_{i1} \\ \vdots \\ \delta_{im} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad N_{2,n}^{(i)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ \delta_{i1} \\ \vdots \\ \delta_{im} \end{pmatrix}.$$

To compute the coefficients of the collocation method defined in Theorem 1, we must now evaluate the polynomials $\phi_i(t)$, $\psi_i(t)$, $\chi_{i,n-1}(t)$ and $\chi_{i,n}(t)$ according to (14)–(19). When H is not singular, the computation yields

$$\begin{aligned}
\phi_i(t) &= [1\ t\ \ldots\ t^{2m+3}]\ H^{-1}N_0^{(i)}, \quad i = 1,2 \\
\theta_i = \phi_i(1) &= [1\ \ldots\ 1]\ H^{-1}N_0^{(i)}, \\
u_{ji} = \phi_i(c_j) &= [1\ c_j \ldots c_j^{2m+3}]\ H^{-1}N_0^{(i)}, \ \ i = 1,2, \ \ j = 1,\ldots,m \\
\psi_i(t) &= [1\ t\ \ldots\ t^{2m+3}]\ H^{-1}N_1^{(i)}, \quad i = 1,2 \\
\eta_i = \psi_i(1) &= [1\ \ldots\ 1]\ H^{-1}N_1^{(i)}, \\
\overline{u}_{ji} = \psi_i(c_j) &= [1\ c_j \ldots c_j^{2m+3}]\ H^{-1}N_1^{(i)}, \ \ i = 1,2, \ \ j = 1,\ldots,m \\
\chi_{i,n-1}(t) &= [1\ t\ \ldots\ t^{2m+3}]\ H^{-1}N_{2,n-1}^{(i)}, \quad i = 1,\ldots,m \\
v_i = \chi_{i,n-1}(1) &= [1\ \ldots\ 1]\ H^{-1}N_{2,n-1}^{(i)}, \quad i = 1,\ldots,m \\
a_{is} = \chi_{i,n-1}(c_s) &= [1\ c_s \ldots c_s^{2m+3}]\ H^{-1}N_{2,n-1}^{(i)}, \quad i,s = 1,\ldots,m \\
\chi_{i,n}(t) &= [1\ t\ \ldots\ t^{2m+3}]\ H^{-1}N_{2,n}^{(i)}, \quad i = 1,\ldots,m \\
v_i = \chi_{i,n}(1) &= [1\ \ldots\ 1]\ H^{-1}N_{2,n}^{(i)}, \quad i = 1,\ldots,m \\
b_{is} = \chi_{i,n}(c_s) &= [1\ c_s \ldots c_s^{2m+3}]\ H^{-1}N_{2,n}^{(i)}, \quad i,s = 1,\ldots,m.
\end{aligned}$$

For $m \geq 2$, from the initial and collocation conditions (2)–(4), a linear system of $2m+4$ equations in $2m+4$ unknowns arises, having H in (20) as matrix of coefficients, where $t_1 = -1$ and $t_2 = 0$.

The computation of the determinant of H allows us to determine the exceptional values of the collocation abscissas c_i for which the two step collocation method cannot be constructed.

3 Computation of the Determinant of H

The determinant of H is computed through the tecnique which is usually applied to the Vandermonde matrices; in this way the computation of the determinant of H reduces to the computation of a 2×2 matrix, whose elements are defined through a recurrence formula, as described in the following.

Let us reduce the determinant of H to the determinant of the $(2m+2) \times (2m+2)$ matrix H', that is $det(H) = -det(H')$, where

$$H' = \begin{pmatrix} 1 & -1 & \dots & (-1)^{2m+3} \\ -2 & 3 & \dots & (2m+3)(-1)^{2m+2} \\ 2 & 2\cdot 3(c_1-1) & \dots & (2m+2)(2m+3)(c_1-1)^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 2\cdot 3(c_m-1) & \dots & (2m+2)(2m+3)(c_m-1)^{2m+1} \\ 2 & 2\cdot 3c_1 & \dots & (2m+2)(2m+3)c_1^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 2\cdot 3c_m & \dots & (2m+2)(2m+3)c_m^{2m+1} \end{pmatrix}$$

Then $det(H) = -(2m+2)!\,(2m+3)!\,detH''$,

$$H'' = \begin{pmatrix} \frac{1}{2} & (-1)\frac{1}{2\cdot 3} & \dots & \frac{1}{(2m+2)(2m+3)}(-1)^{2m+1} \\ -1 & \frac{1}{2} & \dots & \frac{-1}{2m+3}(-1)^{2m+1} \\ 1 & (c_1-1) & \dots & (c_1-1)^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (c_m-1) & \dots & (c_m-1)^{2m+1} \\ 1 & c_1 & \dots & c_1^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & c_m & \dots & c_m^{2m+1} \end{pmatrix}.$$

By setting

$$x_1 = c_1 - 1, \dots, x_m = c_m - 1, x_{m+1} = c_1, \dots, x_{2m} = c_m,$$

$$r^{(0)}(n) = \frac{1}{n(n+1)}(-1)^{n+1}, \quad q^{(0)}(n) = \frac{-1}{n}(-1)^{n+1},$$

the matrix H'' can be written as

$$H'' = \begin{pmatrix} r^{(0)}(1) & r^{(0)}(2) & \dots & r^{(0)}(n) \\ q^{(0)}(1) & q^{(0)}(2) & \dots & q^{(0)}(n) \\ 1 & x_1 & \dots & x_1^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & \dots & x_m^{2m+1} \\ 1 & x_{m+1} & \dots & x_{m+1}^{2m+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2m} & \dots & x_{2m}^{2m+1} \end{pmatrix}.$$

Then H'' is a Vandermonde matrix except the firt two rows. Its determinant is now computed by using the tecnique which is usually applied to the Vandermonde determinant. In details, we multiply each column by x_1 and subtract it from the following column, starting from the penultimate column. At the end a block matrix is obtained:

$$H'' = \prod_{i<j,i=1}^{2m} (x_i - x_j) \begin{pmatrix} A & B \\ V & 0 \end{pmatrix}$$

where A is a $2 \times m$ matrix, 0 is the zero matrix of dimension $2m \times 2$, V is a lower triangular matrix with 1 on the diagonal, and B is a 2×2 matrix containing the elements of H'' in the upper right corner, which result after the transformations performed on H'' to compute its determinant. In particular B takes the following form:

$$B = \begin{pmatrix} r^{(2m)}(2m+1) & r^{(2m)}(2m+2) \\ q^{(2m)}(2m+1) & q^{(2m)}(2m+2) \end{pmatrix}, \tag{21}$$

where the r and q elements in B are derived through the following recurrence formulas, which keep track of the trasformation on H'':

$$r^{(0)}(n) = \frac{1}{n(n+1)}(-1)^{n+1}, \quad r^{(i)}(n) = r^{(i-1)}(n) - r^{(i-1)}(n-1)x_i,$$

$$q^{(0)}(n) = \frac{-1}{n}(-1)^{n+1}, \quad q^{(i)}(n) = q^{(i-1)}(n) - q^{(i-1)}(n-1)x_i,$$

$$i = 1, \ldots, 2m, \;\; n = 1, \ldots, 2m.$$

By considering the elementary symmetric polynomials in $x_1, \ldots, x_{2m}$,

$$p_1 = \sum x_i, \quad p_2 = \sum_{i<j} x_i x_j, \quad p_s = \sum_{i_1<i_2<i_s} x_{i_1} x_{i_2} \ldots x_{i_s},$$

it is possible to prove by induction that

$$r^{(j)}(n+1) = \sum_{k=0}^{j} (-1)^k r^{(0)}(n-k)\, p_k, \quad q^{(j)}(n+1) = \sum_{k=0}^{j} (-1)^k q^{(0)}(n-k)\, p_k.$$

In conclusion

$$det(H) = -\,(2m+2)!\,(2m+3)! \prod_{i<j,i=1}^{2m} (x_j - x_i)\, det(B)$$

where

$$det(B) = r^{(2m)}(2m+1)\, q^{(2m)}(2m+2) - q^{(2m)}(2m+1)\, r^{(2m)}(2m+2)$$

In this way also the computation of the exceptional values of the collocation abscissas for the method defined in theorem 1 reduces to the computation of the c–values which annihilate the 2×2 determinant of B in (21).

For instance, the computation of det(H) for $m = 1$ allows us to state that the collocation–based TSRKN is defined only if $c \neq \pm\sqrt{\frac{3}{10} \pm \frac{\sqrt{\frac{11}{6}}}{5}}$.

The resulting expressions for a higher number of stages can be easily derived through the illustrated technique and the usage of symbolic computation.

4 Concluding Remark

In this paper the coefficients of the new two step collocation methods for ODEs (1) are defined through the computation of the inverse of matrix H in (20), which is of Vandermonde type. Different approaches for the derivation of collocation–based TSRKN methods are under consideration, following [4,5,7], that avoid the numerical treatment of Vandermonde–type matrices, involving on the contrary the integrals of polynomials in the generalized Lagrange basis. The analysis of the properties of the TSRKN methods, such as the obtainable order of convergence, stability properties, efficiency and so on, will be subject of a forthcoming paper.

References

1. J. C. Butcher, The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods, Wiley, Chichester, 1987.
2. A. Guillon and J. L. Soulé, La résolution numérique des problémes différentiels aux conditions initiales par des méthodes de collocation, *RAIRO Anal. Numér. Ser. Rouge* R–3, 17–44, 1969.
3. E. Hairer, S. P. Norsett and G. Wanner, Solving Ordinary Differential Equations I – Nonstiff Problems, Springer Series in Computational Mathematics **8**, Springer–Verlag, Berlin, 1987.
4. E. Hairer and G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems, Springer, Berlin, 1991.
5. I. Lie and S. P. Norsett, Superconvergence for multistep collocation, *Math. Comput.* **52**, 65–79, 1989.
6. E. Messina, J. J. B. de Swart and W. van der Veen, Parallel iterative linear solvers for multistep Runge–Kutta methods, *J. Comput. Appl. Math.* **85**, 145–167, 1997.
7. B. Paternoster, General two–step Runge–Kutta methods based on algebraic and trigonometric polynomials, *Int. J. Appl. Math.* **6**(4), 347–362, 2001.
8. B. Paternoster, Two step Runge-Kutta-Nyström methods for $y'' = f(x, y)$ and P-stability, *Computational Science - ICCS 2002, Lecture Notes in Computer Science 2331, Part III*, P.M.A.Sloot, C.J.K.Tan, J.J.Dongarra, A.G.Hoekstra Eds., 459–466, Springer Verlag, Amsterdam (2002).
9. B.Paternoster, Two step Runge-Kutta-Nystrom methods for oscillatory problems based on mixed polynomials, *Computational Science - ICCS 2003, Lecture Notes in Computer Science 2658, Part II*, P.M.A.Sloot, D.Abramson, A.V.Bogdanov, J.J.Dongarra, A.Y.Zomaya, Y.E.Gorbachev Eds., pp. 131-138, Springer, Berlin Heidelberg (2003).

Direct Optimization Using Gaussian Quadrature and Continuous Runge-Kutta Methods: Application to an Innovation Diffusion Model

Fasma Diele[1], Carmela Marangi[1], and Stefania Ragni[2]

[1] Istituto per le Applicazioni del Calcolo M. Picone, CNR, Via Amendola 122, 70126 Bari, Italy

[2] Facoltà di Economia, Università di Bari, Via Camillo Rosalba 56, 70100 Bari, Italy

Abstract. In the present paper the discretization of a particular model arising in the economic field of innovation diffusion is developed. It consists of an optimal control problem governed by an ordinary differential equation. We propose a direct optimization approach characterized by an explicit, fixed step-size continuous Runge-Kutta integration for the state variable approximation. Moreover, high-order Gaussian quadrature rules are used to discretize the objective function. In this way, the optimal control problem is converted into a nonlinear programming one which is solved by means of classical algorithms.

1 Introduction

Dynamic optimization represents a challenging problem in several fields of applied science. From biology to engineering or economics, a wide variety of phenomena can be described in terms of optimal control problems, where a cost functional has to be optimized with respect to variables which control the system dynamics.

Numerical methods for dynamical optimization fall essentially into two classes: classical, indirect methods, relying on the maximum Pontryagin principle, with the main drawback of a lack of robustness, and so-called direct methods which attempt to find a solution through a direct optimization of the cost functional. In this paper we focus on the last type of methods and introduce a numerical scheme based on the continuous Runge-Kutta integration ([9], [11]) to solve a specific model in economics. The rationale for this choice is that an efficient and accurate approximation of the system dynamics can be obtained with a reduced number of variable evaluations. A further improvement in accuracy is achieved by adopting high-order Gauss-Legendre quadrature rules ([1]) for the functional discretization.

The approach has been tested on a specific innovation diffusion model recently proposed in [5]. The early mathematical models in this field date back to 1950 (see [7]) and represent an interesting research subject in economics, sociology, marketing as well as in applied mathematics. The problem is easily stated as the one of maximizing the profit of a monopolistic firm, due to the sale of a

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 426–433, 2004.
© Springer-Verlag Berlin Heidelberg 2004

new product, by planning an advertising and price strategy. The specific model we are concerned with, is briefly described in the next section. In Section 3 we introduce the numerical approximation which combines the continuous Runge-Kutta integration with high-order Gauss-Legendre quadrature rules in order to discretize the constrained problem at hand. Numerical results for some values of the model parameters, shown in Section 4, validate the effectiveness of the proposed approach. Finally we draw our conclusions in Section 5.

2 The Continuous Model

We assume that a monopolistic firm, manufacturing a new durable good (i.e. an *innovation*), has to plan advertising and price strategies in order to maximize the profit due to the product sale over a finite time horizon $[0, T]$.

We suppose that the rate of adopters increase is affected by the selling price $p(t)$ according to a given price response function $q(p)$, which is usually positive and decreasing. Different price response functions have been proposed in the literature (see [6]); most of them belong to a specific class obeying the following law

$$q(p) = \begin{cases} [c - (1-\lambda)\alpha p]_+^{\frac{1}{1-\lambda}} & \text{if } \lambda \neq 1 \\ c \exp(-\alpha p) & \text{if } \lambda = 1 \end{cases} \tag{1}$$

with $c, \lambda, \alpha \in \mathbb{R}$, $\alpha > 0$. Moreover, a further demand effect is represented by the amount of advertising by media $\gamma(t)$, ranging from 0 to a maximum value $\gamma_{\max}$.

Following the same approach of [5], we consider the spatial-temporal innovation diffusion accounting for an interaction among the individuals and assuming that the demand effect is related to the so-called *word-of-mouth* phenomenon. More precisely, the adopters are assumed to have a local influence according to a given kernel $\mathcal{K}(z, z')$, with $z, z' \in \Omega$, where $\Omega \subset \mathbb{R}^2$ represents the geographical domain the population lives in. Under the further assumption that $\mathcal{K}(z, z')$ is independent of variable z', we denote

$$\mathcal{K} = \int_\Omega \mathcal{K}(z, z')\, dz.$$

As a consequence, the individuals adopting the new product are uniformly distributed with respect to space variables. Then, if $\mathcal{A}(t)$ represents the adopters cumulative number at time t, the related evolution dynamics can be modelled as follows

$$\frac{d\mathcal{A}}{dt}(t) = q(p(t))\,(\mathcal{K}\mathcal{A}(t) + \gamma(t))\,(1 - \mathcal{A}(t)), \qquad t \in [0, T] \tag{2}$$

provided with condition $\mathcal{A}(0) = 0$, that is no individual has adopted the innovation at initial time (for more details, see [3] and [5]).

Furthermore, we define the firm's payoff function as

$$J(p, \gamma) = \int_0^T \exp(-rt) \left((p(t) - c_p) \frac{d\mathcal{A}}{dt}(t) - c_a \gamma(t) \right) dt \tag{3}$$

where c_a, c_p and r denote the per unit advertising cost, the unitary production cost and the discount factor, respectively, all assumed to be constant during the whole time horizon.

The resulting optimal control model consists of maximizing (3) with respect to control parameters represented by price and advertising, governed by state equation (2). More precisely, defining $\mathcal{F}$ as the space of all the functions mapping $[0,T]$ into $\mathbb{R}$ and setting

$$\mathcal{U} = \{(u,v) \in \mathcal{F}^2 \mid u, v \text{ are piecewise continuous}, 0 \le u(t),\ 0 \le v(t) \le \gamma_{\max}\},$$

the problem reduces to

$$\max_{p,\gamma \in \mathcal{U}} J(p,\gamma) \tag{4}$$

constrained by equation (2) with condition $\mathcal{A}(0) = 0$. For a theoretical analysis of the model see [5]. Moreover, in [4] a preliminary numerical solution of the problem is given accounting for spatial effects of the new product diffusion in the market; in that case the optimal control model is governed by a partial differential equation.

3 Numerical Approximation

Our main interest consists of performing the numerical discretization of the previous continuous model based on the use of continuous extensions of Runge-Kutta methods combined with Gaussian quadrature rules.

Let us introduce a mesh $0 = t_0 < t_1 < \ldots < t_N = T$ of the given time interval $[0,T]$ with constant step length h, then $t_i = ih$ for each $i = 0, \ldots N$. Moreover, setting

$$\psi(\mathcal{A}(t), p(t), \gamma(t), t) = q(p(t))\,(\mathcal{K}\mathcal{A}(t) + \gamma(t))\,(1 - \mathcal{A}(t)),$$

we define

$$g(p(t), \gamma(t), \mathcal{A}(t), t) = \exp(-rt)((p(t) - c_p)\psi(\mathcal{A}(t), p(t), \gamma(t), t) - c_a\gamma(t)).$$

Hence, the integral in (3) can be written as

$$\begin{aligned}\int_0^T g(p(t),\gamma(t),\mathcal{A}(t),t)\ dt &= \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} g(p(t),\gamma(t),\mathcal{A}(t),t)\ dt \\ = \sum_{i=0}^{N-1} \frac{h}{2} \int_{-1}^{1} g\left(p\left(t_i + h\tfrac{s+1}{2}\right), \gamma\left(t_i + h\tfrac{s+1}{2}\right), \mathcal{A}\left(t_i + h\tfrac{s+1}{2}\right), t_i + h\tfrac{s+1}{2}\right) ds.\end{aligned} \tag{5}$$

We approximate every integral by means of the well-known Gauss-Legendre rules; thus, the right-hand term in (5) is discretized as

$$\frac{h}{2} \sum_{i=0}^{N-1} \sum_{l=0}^{L} w_l\ g\left(p(\tau_{i,l}), \gamma(\tau_{i,l}), \mathcal{A}(\tau_{i,l}), \tau_{i,l}\right) \tag{6}$$

where

$$\tau_{i,l} = t_i + h\frac{s_l + 1}{2}, \qquad i = 0, \ldots, N-1,\ l = 0, \ldots, L$$

and s_l, w_l represent the knots (i.e. the zeros of Legendre polynomial of order $L+1$) and the coefficients, respectively, of the chosen quadrature scheme (see [1]). Notice that the quantities involved in the previous approximation are the values of both the controls p, γ and the state variables $\mathcal{A}$ as sampled at each instant $\tau_{i,l}$.

If we assume to approximate control parameters in the class of piecewise linear functions

$$\mathcal{V} = \{v \in \mathcal{F} \mid v|_{[t_i, t_{i+1}]} \text{ is linear}\},$$

we have that

$$\begin{aligned} p(\tau_{i,l}) &= \rho_l\, p(t_i) + \xi_l\, p(t_{i+1}), \\ \gamma(\tau_{i,l}) &= \rho_l\, \gamma(t_i) + \xi_l\, \gamma(t_{i+1}) \end{aligned}$$

with $\rho_l = \dfrac{1 - s_l}{2}$ and $\xi_l = 1 - \rho_l$ for every $l = 0, \ldots, L$.

Moreover, in order to evaluate $\mathcal{A}(\tau_{i,l})$'s, the state equation is discretized by applying continuous extensions of Runge-Kutta methods. We consider an explicit m-stage scheme specified by the Butcher array

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}$$

and its continuous extension characterized by coefficients $\widetilde{b}_j(\theta)$ $(j = 1, \ldots, m)$ which are suitable polynomials in $\theta \in [0, 1]$ obtained as in [11]. The resulting scheme for the state system is the following

$$\begin{aligned} \mathcal{A}(\tau_{i,l}) &= \mathcal{A}(t_i) + h \sum_{j=1}^{m} \widetilde{b}_j(\xi_l)\psi(\widetilde{\mathcal{A}}_j, p(t_i + hc_j), \gamma(t_i + hc_j), t_i + hc_j), \\ \mathcal{A}(t_{i+1}) &= \mathcal{A}(t_i) + h \sum_{j=1}^{m} \widetilde{b}_j(1)\psi(\widetilde{\mathcal{A}}_j, p(t_i + hc_j), \gamma(t_i + hc_j), t_i + hc_j), \\ \widetilde{\mathcal{A}}_j &= \mathcal{A}(t_i) + h \sum_{v=1}^{j-1} a_{j,v}\psi(\widetilde{\mathcal{A}}_v, p(t_i + hc_v), \gamma(t_i + hc_v), t_i + hc_v),\ j = 1, \ldots, m \end{aligned} \tag{7}$$

starting from $\mathcal{A}(t_0) = 0$.

Note that condition $0 \le \xi_l \le 1$ holds for each $l = 0, \ldots L$. Furthermore, in every interval (t_i, t_{i+1}) we should add extra control variables at each inner time step $\sigma_{i,j} := t_i + hc_j$ since, in general, $\sigma_{i,j} \neq \tau_{i,l}$. However, additional controls might be avoided simply by choosing Runge-Kutta scheme in such a way that $c_1 = 0$, $c_m = 1$ and every inner c_j is equal to ξ_l for some $l = 0, \ldots, L$.

Finally, let us consider the following vector

$$\mathbf{u} = (p(t_0), p(t_1), \ldots, p(t_N), \gamma(t_0), \gamma(t_1), \ldots, \gamma(t_N))^T \in \mathbb{R}^{2(N+1)}.$$

In this notation, the sum in (6) can be defined as

$$\widetilde{J}(\mathbf{u}) = \frac{h}{2} \sum_{i=0}^{N-1} \sum_{l=0}^{L} w_l \; g(\rho_l \, u_{i+1} + \xi_l \, u_{i+2}, \; \rho_l \, u_{N+2+i} + \xi_l \, u_{N+3+i}, \; \mathcal{A}(\tau_{i,l}), \; \tau_{i,l}) \quad (8)$$

where each u_j denotes the j-th entry of vector $\mathbf{u}$ and the values $\mathcal{A}(\tau_{i,l})$, depending on $\mathbf{u}$ itself, are computed by recursive formula (7). In this way, the numerical approximation of the optimal control model described in Section 2 yields a solution to the nonlinear programming problem

PROBLEM 1 *Maximize* $\widetilde{J}(\mathbf{u})$ *defined in (8) and constrained by*

$$\begin{aligned} &0 \le u_i, && i = 1, \ldots, 2(N+1), \\ &u_{N+1+i} \le \gamma_{\max}, && i = 1, \ldots, N+1. \end{aligned}$$

We note the proposed numerical discretization follows an idea similar to the one provided by authors in [2].

4 Numerical Tests

In order to test the effectiveness of the numerical procedure described in the previous section, we use the Gauss-Legendre rule with 2 knots (i.e. $L = 1$); furthermore, the chosen Runge-Kutta scheme is the continuous extension of the 2-stage *improved Euler* method (see [8]) where

$$b_1(\theta) = -\frac{1}{2}\theta^2 + \theta \quad \text{and} \quad b_2(\theta) = \frac{1}{2}\theta^2.$$

Discrete solutions have been achieved solving nonlinear programming Problem 1 in Matlab environment by means of `fmincon` routine.

In our numerical examples, the price response function is defined in correspondence with $\lambda = c = 1$ in (1). We set the discount factor $r = 0$, since in this case the theoretical solution of the continuous model is known; in particular, in [5] it has been proven that optimal price is characterized by a concave behaviour, while advertising strategy is a bang-bang function.

In Figures 1 and 2, we plot the approximate time evolution of price and advertising strategies as well as adopters cumulative number obtained for different values of parameters specified in each caption. As it is shown in figures, numerical results are in agreement with theoretical ones. We remark that the case in Figure 2 shows an evident bang-bang behaviour which cannot be catched by classical gradient algorithms. A further improvement could be obtained by exploiting the advantages of global stochastic optimization approaches.

5 Conclusions

A numerical approximation has been provided in order to solve a specific economic model which describes the spread of a new product in a market and deals

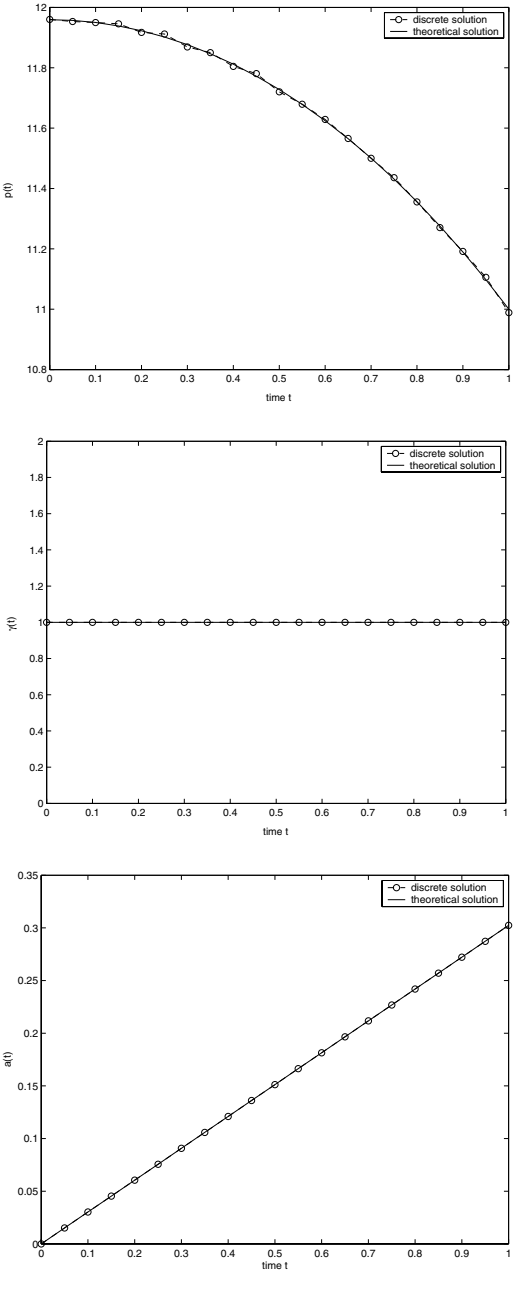


Fig. 1. Optimal price (on the top), advertising (in the middle) and adopters cumulative number (on the bottom) in correspondence with $T = 1$, $\gamma_{max} = 1$, $\mathcal{K} = 1$, $\alpha = 0.1$, $c_a = 1$ and $c_p = 1$. Time step length is $h = 0.05$.

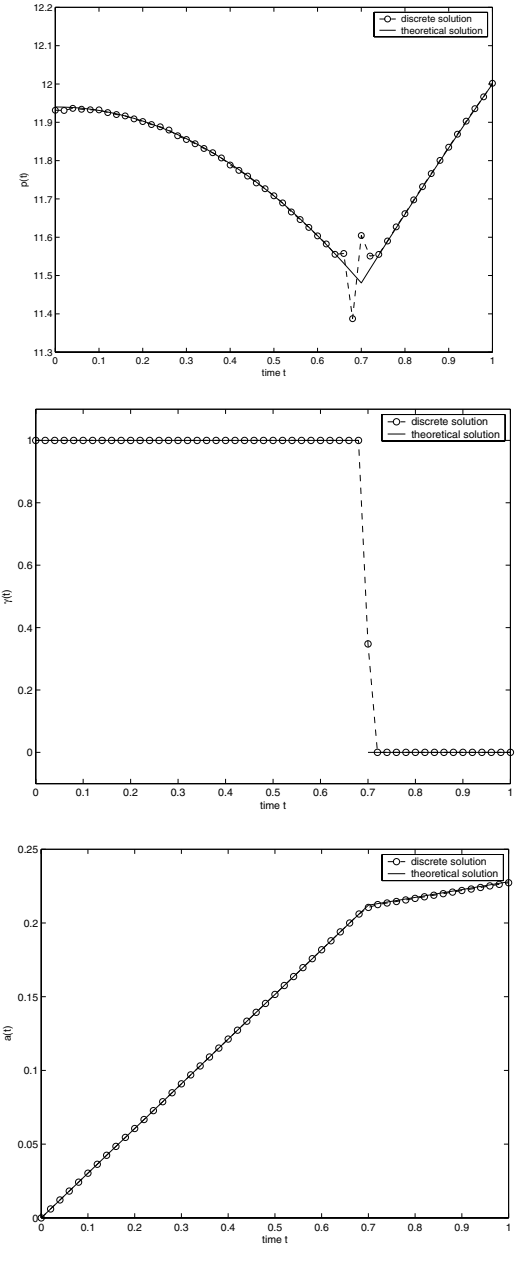


Fig. 2. Optimal price (on the top), advertising (in the middle) and adopters cumulative number (on the bottom) in correspondence with $T = 1$, $\gamma_{\max} = 1$, $\mathcal{K} = 1$, $\alpha = 0.1$, $c_a = 2.5$ and $c_p = 2$. Time step length is $h = 0.02$.

with the choice of price and advertising strategies maximizing the profit coming from its sale. The discrete procedure reduces a given optimal control problem governed by ordinary differential equations to a nonlinear programming one. More precisely, the discrete scheme is characterized by a direct optimization procedure obtained by using Gauss-Legendre quadrature rules in order to approximate the integral optimizing functional and continuous Runge-Kutta methods for the state variables differential constraints. The main advantage of this approach is that it allows the reduction of nodal variable values necessary for the chosen discretization while retaining a great accuracy in the solution. Finally, the numerical results confirm the effectiveness of the proposed algorithm.

References

1. Davis, P.J., Rabinowitz, P.: Methods of numerical integration, 2nd Ed., Academic, New York (1984).
2. Diele, F., Marangi, C., Ragni, S.: Numerical methods based on Gaussian quadrature and continuous Runge-Kutta integration for optimal control problems. IAC Thecnical Report N.33 (2/2004).
3. De Cesare, L., Di Liddo, A.: A Bolza optimal control problem for innovation diffusion. Dynamic Systems and Appl., **9** (2000) 269–280.
4. De Cesare, L., Di Liddo, A., Ragni, S.: Numerical solutions to some optimal control problems arising from innovation diffusion. Computational Economics, **22** (2003) 173–186.
5. De Cesare, L., Di Liddo, A., Ragni, S.: On the shape of optimal price and advertising strategies arising in innovation diffusion. Accepted for publication in Dynamic Systems and Appl.
6. Kalish, S., Lilien, G. L.: Optimal price subsidy for accelerating the diffusion of innovation. Marketing Science, **2** (1983) 407–420.
7. Kendall, D. G.: Discussion on Bartlett, M. S., Measles, periodicuty and community size'. J. Roy. Stat. Soc. Ser. A, **120** (1957) 48–70.
8. Lambert, J. D.: Numerical methods for ordinary differential systems. Wiley & Sons, Chichester (1997).
9. Owren, B., Zennaro, M.: Derivation of efficient, continuous, explicit Runge-Kutta methods. SIAM J. Sci. Stat. Comput., **Vol. 13**, No. 6 (1992) 1488–1501.
10. Schwartz, A.: Theory and implementation of numerical methods based on Runge-Kutta integration for solving optimal control problems. PhD Thesis, U.C. Berkeley (1996).
11. Zennaro, M.: Natural continuous extensions of Runge-Kutta methods. Mathematics of Computation, **Vol. 46**, No. 173 (1986) 119–133.

The ReLPM Exponential Integrator for FE Discretizations of Advection-Diffusion Equations*

Luca Bergamaschi[1], Marco Caliari[2], and Marco Vianello[3]

[1] Dip.to di Met. e Mod. Mat. per le Scienze Appl., Università di Padova
[2] Dip.to di Informatica, Università di Verona
[3] Dip.to di Mat. Pura ed Appl., Università di Padova, marcov@math.unipd.it

Abstract. We implement an exponential integrator for large and sparse systems of ODEs, generated by FE (Finite Element) discretization with mass-lumping of advection-diffusion equations. The relevant exponential-like matrix function is approximated by polynomial interpolation, at a sequence of real Leja points related to the spectrum of the FE matrix (ReLPM, Real Leja Points Method). Application to 2D and 3D advection-dispersion models shows speed-ups of one order of magnitude with respect to a classical variable step-size Crank-Nicolson solver.

1 The Advection-Diffusion Model

We consider the classical evolutionary advection-diffusion problem

$$\begin{cases} \dfrac{\partial c}{\partial t} = \mathrm{div}(D\nabla c) - \mathrm{div}(cv) + \phi & x \in \Omega,\ t > 0 \\ c(x,0) = c_0(x) & x \in \Omega \\ c(x,t) = g_{\mathrm{D}}(x,t) & x \in \Gamma_{\mathrm{D}},\ t > 0 \\ \langle D\nabla c(x,t), \nu\rangle = g_{\mathrm{N}}(x,t) & x \in \Gamma_{\mathrm{N}},\ t > 0 \end{cases} \tag{1}$$

with mixed Dirichlet and Neumann boundary conditions on $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} = \partial\Omega$, $\Omega \subset \mathbb{R}^d$, $d = 2, 3$; cf. [1]. Equation (1) represents, e.g., a simplified model for solute transport in groundwater flow (advection-dispersion), where c is the solute concentration, D the hydrodynamic dispersion tensor, $D_{ij} = \alpha_{\mathrm{T}}|v|\delta_{ij} + (\alpha_{\mathrm{L}} - \alpha_{\mathrm{T}})v_i v_j/|v|$, $1 \le i, j \le d$, v the average linear velocity of groundwater flow and

* Work supported by the research project CPDA028291 "Efficient approximation methods for nonlocal discrete transforms" of the University of Padova, and by the subproject "Approximation of matrix functions in the numerical solution of differential equations" (co-ordinator M. Vianello, University of Padova) of the MIUR PRIN 2003 project "Dynamical systems on matrix manifolds: numerical methods and applications" (co-ordinator L. Lopez, University of Bari). Thanks also to the numerical analysis group at the Dept. of Math. Methods and Models for Appl. Sciences of the University of Padova, for having provided FE matrices for our numerical tests, and for the use of their computing resources.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 434–442, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ϕ the source (cf. [2,3]). The standard Galerkin FE discretization of (1) nodes $\{x_i\}_{i=1}^N$ and linear basis functions $\{\varphi_i\}_{i=1}^N$ [4], gives the linear system of ODEs

$$\begin{cases} P\dot{\boldsymbol{z}} = H\boldsymbol{z} + \boldsymbol{b} + \boldsymbol{q}, & t > 0 \\ \boldsymbol{z}(0) = \boldsymbol{c}_0 \end{cases} \tag{2}$$

where $\boldsymbol{z} = [z_1(t), \ldots, z_N(t)]^{\mathrm{T}}$, $\boldsymbol{c}_0 = [c_0(x_1), \ldots, c_0(x_N)]^{\mathrm{T}}$, P is the symmetric positive-definite mass matrix, H the (nonsymmetric) stiffness matrix and

$$b_i \approx \int_{\mathrm{supp}(\varphi_i)} \phi\ \varphi_i\ \mathrm{d}\Omega, \quad q_i \approx \int_{\Gamma_{\mathrm{N}} \cap \mathrm{supp}(\varphi_i)} g_{\mathrm{N}}\ \varphi_i\ \mathrm{d}\Gamma_{\mathrm{N}}, \quad i = 1, \ldots, N\ . \tag{3}$$

As it is known, such a basic FE discretization makes sense numerically only on sufficiently fine grids (small grid Péclet numbers). Otherwise, special stabilization techniques should be adopted, like Petrov-Galerkin upwinding; see [1].

Actually, system (2) does not take into account yet of possible Dirichlet conditions: we discuss in detail their treatment within the next sections.

2 An Exponential Integrator via Mass-Lumping

In the sequel we consider stationary velocity, source and boundary conditions in (1), which give a system of ODEs like (2) with constant H, $\boldsymbol{b}$ and $\boldsymbol{q}$. Such a system can be rewritten in the form

$$\begin{cases} \dot{\boldsymbol{z}} = P^{-1}H\boldsymbol{z} + \boldsymbol{\phi} + P^{-1}\boldsymbol{q}, & t > 0 \\ \boldsymbol{z}(0) = \boldsymbol{c}_0 \end{cases}$$

which is suitable for the application of exponential integrators (cf. [5,6,7]). Observe that $\boldsymbol{\phi} = P^{-1}\boldsymbol{b}$ since we chose $\boldsymbol{b} = P[\phi(x_1), \ldots, \phi(x_N)]^{\mathrm{T}}$ in (3). In order to impose the lacking Dirichlet conditions we make vanishing the i-th row of P^{-1} and of $\boldsymbol{\phi}$ (recall that g_{D} is stationary) and substitute the i-th component of the initial vector $\boldsymbol{c}_0$ with $g_{\mathrm{D}}(x_i)$, if $x_i \in \Gamma_{\mathrm{D}}$, obtaining $\widehat{P^{-1}}$, $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{c}_0}$. In practice, the system above is modified into

$$\begin{cases} \dot{\boldsymbol{c}} = \widehat{P^{-1}}H\boldsymbol{c} + \widehat{\boldsymbol{\phi}} + \widehat{P^{-1}}\boldsymbol{q}, & t > 0 \\ \boldsymbol{c}(0) = \widehat{\boldsymbol{c}_0} \end{cases} \tag{4}$$

System (4) is now the discrete approximation of the PDE (1), where the Dirichlet conditions have been artificially imposed also to the initial data $\boldsymbol{c}_0$.

As known, the solution can be written explicitly in the exponential form

$$\boldsymbol{c}(t) = \widehat{\boldsymbol{c}_0} + t\varphi(t\widehat{P^{-1}}H)[\widehat{P^{-1}}H\widehat{\boldsymbol{c}_0} + \widehat{\boldsymbol{\phi}} + \widehat{P^{-1}}\boldsymbol{q}]\ ,$$

where $\varphi(z)$ is the *entire* function $\varphi(z) = (e^z - 1)/z$ if $z \neq 0$, $\varphi(0) = 1$. Clearly, availability of matrix P^{-1} (and thus $\widehat{P^{-1}}$) is a computationally expensive task:

so, we can apply the well known *mass-lumping* technique (sum on the diagonal of all the row elements) to P in order to obtain a diagonal matrix P_{L} (and, immediately, P_{L}^{-1} and $\widehat{P_{\mathrm{L}}^{-1}}$); cf. [1,8]. Now we can consider the transformed system (4) with $\widehat{P^{-1}}$ replaced by $\widehat{P_{\mathrm{L}}^{-1}}$, and apply the *exact* and *explicit* time-marching scheme (see the work by Schaefer [5] for FD spatial discretization)

$$\boldsymbol{c}_{k+1} = \boldsymbol{c}_k + \Delta t_k \varphi(\Delta t_k H_{\mathrm{L}}) \boldsymbol{v}_k, \quad k = 0, 1, \ldots, \quad \boldsymbol{c}_0 = \widehat{\boldsymbol{c}_0} \ , \tag{5}$$

where we have defined $H_{\mathrm{L}} = \widehat{P_{\mathrm{L}}^{-1}} H$ and $\boldsymbol{v}_k = H_{\mathrm{L}} \boldsymbol{c}_k + \widehat{\boldsymbol{\phi}} + \widehat{P_{\mathrm{L}}^{-1}} \boldsymbol{q}$. Exactness of the exponential integrator (5) entails that the time-steps Δt_k can be chosen, at least in principle, arbitrarily large with no loss of accuracy, making it an appealing alternative to classical finite-difference integrators (cf. [6,7,9]).

However, the practical application of (5) rests on the possibility of approximating efficiently the *exponential operator* $\varphi(\Delta t H_{\mathrm{L}})\boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^N$. To this aim, two classes of *polynomial* methods are currently used. We have Krylov-like methods, which are based on the idea of projecting the operator on a "small" Krylov subspace of the matrix via the Arnoldi process, and typically involve long-term recurrences in the nonsymmetric case; see, e.g., [10,11,12], and [13, 14] for other (nonstandard) Krylov-like approaches. The second class consists of methods based on polynomial interpolation or series expansion of the entire function φ on a suitable compact subset containing the spectrum (or in general the field of values) of the matrix (e.g. Faber and Chebyshev series, interpolation at special points like Faber, Fejér and Leja points). They typically require some preliminary estimate of the underlying spectral structure, but, despite of this, this second class of methods turned out to be competitive with Krylov-based approaches, especially on very large nonsymmetric matrices, cf. [15,16,17, 18,19]. In this work we adopt the Real Leja Points Method (shortly ReLPM), recently proposed in the frame of FD spatial discretization of advection-diffusion equations [9].

3 Computing the Exponential Operator by the ReLPM (Real Leja Points Method)

Sequences of Leja points $\{z_j\}_{j=0}^{\infty}$ for the compact $K \subset \mathbb{C}$ are defined recursively as follows: if z_0 is an arbitrary fixed point in K (usually such as $|z_0| = \max_{z \in K} |z|$, cf. [20]), the z_j are chosen in such a way that

$$\prod_{k=0}^{j-1} |z_j - z_k| = \max_{z \in K} \prod_{k=0}^{j-1} |z - z_k|, \quad j = 1, 2, \ldots$$

By the maximum principle, the Leja points for K lie on ∂K. We recall that an efficient algorithm for the computation of a sequence Leja points, the so-called Fast Leja Points, has been recently proposed in [21].

Now, for any fixed compact set K with more than one point, there is a function Φ which maps the exterior of K conformally onto the exterior of the unit

disk, and satisfies in particular the condition $\lim_{z\to\infty} \Phi(z)/z = \gamma$, where $\gamma > 0$ is called *capacity* of K (cf. [22]). For any $R > 0$, define $\Gamma_R = \{z : |\Phi(z)| = R/\gamma\}$ and K_R the bounded domain with boundary Γ_R; observe that $K_\gamma = K$, K_R has capacity R, and $K_{R_1} \subseteq K_{R_2}$ if $R_1 \leq R_2$. Let f be an entire function: it is well known [22,20] that the sequence of polynomials p_m of degree m that interpolate f on the Leja points $\{z_j\}_{j=0}^m$ for K converges *maximally* to f on every K_R, i.e. asymptotically like the best uniform approximation polynomials, and thus *superlinearly*, that is

$$\limsup_{m\to\infty} \|f - p_m\|_{K_R}^{1/m} = 0 \ .$$

Moreover, Leja sequences are attractive for interpolation at high-degree, in view of the stability of the corresponding algorithm in the Newton form (cf. [20]).

From these properties, we can derive a stable and efficient polynomial approximation method for the matrix operator φ in (5). In fact, cf. [17], if $\{p_m\}$ converges maximally to the entire function f on a compact K, then

$$\limsup_{m\to\infty} \|f(A)\boldsymbol{v} - p_m(A)\boldsymbol{v}\|_2^{1/m} = 0 \ . \tag{6}$$

In particular, if R is such that the spectrum $\sigma(A)$ of the matrix $A \in \mathbb{R}^{N\times N}$ is contained in K_R, and if $A = X^{-1}\Lambda X$ is diagonalizable, we have the estimate

$$\|f(A)\boldsymbol{v} - p_m(A)\boldsymbol{v}\|_2 \leq \mathrm{cond}_2(X)\cdot\|f - p_m\|_{K_R}\cdot\|\boldsymbol{v}\|_2 = \mathcal{O}\left(\left(\frac{e\cdot R}{m}\right)^{m+1}\right) \ , \tag{7}$$

for $m \geq m_0 > R$. More refined convergence estimates (even in the nondiagonalizable case) can be obtained by resorting to the notions of pseudospectra and field of values of A; cf. [17,9].

Following [9], an algorithm for the approximation of the advection-diffusion FE propagator $\varphi(\Delta t H_{\mathrm{L}})\boldsymbol{v}$ can be now easily developed, by means of Newton interpolation at "spectral" Leja points. In the sequel, the compact subset used for estimating the spectrum of H_{L} in (5) will be an ellipse in a suitable family of *confocal ellipses* $\{K_R\}$, where $K = K_{c/2} = [d-c, d+c] \subset \mathbb{R}$ is the common focal interval. Note that we restrict our attention to ellipses symmetric with respect to the real axis, since in our application we deal with real matrices. Moreover, it makes sense to consider a real focal interval, since the numerically evaluated spectra of H_{L} with small grid Péclet number have an horizontal projection larger than the vertical projection. The advantage of working with such confocal ellipses stems from complex approximation theory: if the spectrum $\sigma(\Delta t H_{\mathrm{L}})$ is contained in K_R for some R, then a sequence of polynomials converging maximally to φ on $K_{c/2} = [d-c, d+c]$ converges maximally also on K_R (cf. [22]), and thus the corresponding sequence of matrix polynomial operators converges maximally to $\varphi(\Delta t H_{\mathrm{L}})\boldsymbol{v}$; cf. (6)–(7). Thus we are entitled to interpolate on Leja points of the focal interval $[d-c, d+c] \subset \mathbb{R}$, working with *real* instead of complex arithmetic (as it would be required interpolating directly on the complex Leja points of some ellipse of the family). Clearly, a key step in this procedure is given by

estimating at low cost the reference focal interval for the spectrum of $\Delta t H_{\mathrm{L}}$. Following [5] and [9], which deal with FD discretizations, we adopt the simplest estimate given directly by Gershgorin's theorem.

We can now sketch the algorithm for Leja points interpolation (ReLPM) of the advection-diffusion propagator $\varphi(\Delta t H_{\mathrm{L}})\boldsymbol{v}$ in Table 1. The ReLPM algorithm

Table 1. Algorithm ReLPM (Real Leja Points Method)

- INPUT: H_{L}, $\boldsymbol{v}$, Δt, tol
- Estimate the spectral focal interval $[d-c, d+c]$ for $\Delta t H_{\mathrm{L}}$, by Gershgorin's theorem
- Compute a sequence of Fast Leja Points $\{\xi_j\}$ in $[d-c, d+c]$ as in [21]
- $d_0 := \varphi(\xi_0)$, $\boldsymbol{w}_0 := \boldsymbol{v}$, $\boldsymbol{p}_0 := d_0 \boldsymbol{w}_0$, $m := 0$
- WHILE $e_m^{\mathrm{Leja}} := |d_m| \cdot \|\boldsymbol{w}_m\|_2 >$ tol
  - $\boldsymbol{w}_{m+1} := (\Delta t H_{\mathrm{L}} - \xi_m I)\boldsymbol{w}_m$
  - $m := m + 1$
  - compute the next divided difference d_m
  - $\boldsymbol{p}_m := \boldsymbol{p}_{m-1} + d_m \boldsymbol{w}_m$
- OUTPUT: the vector $\boldsymbol{p}_m$: $\|\boldsymbol{p}_m - \varphi(\Delta t H_{\mathrm{L}})\boldsymbol{v}\|_2 \approx e_m^{\mathrm{Leja}} \leq$ tol

---

turns out to be quite simple and efficient. Indeed, being based on *two-term* vector recurrences in *real arithmetic*, its storage occupancy and computational cost are very small. For implementation details not reported in Table 1, we refer to [9]. We only stress that ReLPM is very well structured for a possible *parallel implementation*, since it uses only sparse matrix-vector multiplications and basic vector operations, but no inner product.

4 Application: 2D and 3D Advection-Dispersion Models

We present now three examples (cf. [13]), concerning application of the ReLPM exponential integrator (5) to advection-dispersion models like (1), together with the comparison with the classical variable step-size Crank-Nicolson solver.

Example 1. As first numerical test, we consider problem (1) on the 2D rectangular domain $\Omega = [0,1] \times [0,0.5]$, with a regular grid of $N = 161 \times 81 = 13041$ nodes and $M = 25600$ triangular elements. Here, $\phi \equiv 0$ and $c_0 \equiv 1$. Dirichlet boundary conditions $c = 1$ for $0.2 \leq y \leq 0.3$ and $c = 0$ elsewhere are imposed on $\Gamma_{\mathrm{D}} = \{0\} \times [0, 0.5]$; the Neumann condition $\partial c/\partial \nu = 0$ is prescribed on $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$. The velocity is $v = (v_1, v_2) = (1, 0)$, and $\alpha_{\mathrm{L}} = \alpha_{\mathrm{T}} = 0.00625$.

Example 2. The second numerical test is the extension of the first on a 3D domain $\Omega = [0,1] \times [0,0.5] \times [0,1]$, with a regular grid of $N = 81 \times 41 \times 9 = 29889$

nodes and $M = 153600$ tetrahedral elements. The boundary conditions of the previous test are extended along the z axis, while we have taken the peaked initial data $\boldsymbol{c}_0 = [1, \ldots, 1, 100, 1, \ldots, 1]^{\mathrm{T}}$. The velocity is $v = (v_1, v_2, v_3) = (1, 0, 0)$, and $\alpha_{\mathrm{L}} = \alpha_{\mathrm{T}} = 0.0125$.

Example 3. In the last numerical test, the domain is the unit circle with a mesh consisting of $N = 35313$ nodes and $M = 245997$ triangular elements, generated by the `triangle` software of Shewchuk (available at `www.netlib.org/voronoi`). Dirichlet boundary condition $c = 0$ are imposed on the whole boundary. The velocity is $v = (v_1, v_2) = (1, 1)$, and $\alpha_{\mathrm{L}} = \alpha_{\mathrm{T}} = 0.00625$; moreover $c_0 \equiv 1$, and the source is constant and negative, $\phi \equiv -1$.

4.1 Crank-Nicolson (CN) Method

Although Crank-Nicolson (CN) method might not be considered the best choice for time integration of advection-diffusion problems, it is a robust method still widely used in engineering applications, and a sound baseline benchmark for any advection-diffusion solver (cf. [13]). In the case of the relevant ODEs system (2) (with stationary $\boldsymbol{b}$ and $\boldsymbol{q}$), its variable step-size version writes as

$$\left(P - \frac{h_k}{2}H\right)\boldsymbol{u}_{k+1} = \left(P + \frac{h_k}{2}H\right)\boldsymbol{u}_k + h_k\,(\boldsymbol{b} + \boldsymbol{q}), \quad k = 0, 1, \ldots, \quad \boldsymbol{u}_0 = \boldsymbol{c}_0 \ .$$

In order to impose Dirichlet conditions, we change the i-th row of the system matrix above with the basis vector $\boldsymbol{e}_i$ and the i-th component of the right hand side with the value of $g_{\mathrm{D}}(x_i)$, if $x_i \in \Gamma_{\mathrm{D}}$. The linear system is solved by the biconjugate gradient stabilized method, preconditioned at each step (since the system matrix depends on h_k) with the incomplete triangular factor and no fill-in. As for estimation of the local truncation error $\mathcal{O}(\|\dddot{\boldsymbol{c}}(t_k)\|_2 h_k^3)$ and step-size control, we have used standard finite-difference approximation of the third derivatives.

4.2 Numerical Tests and Comparisons

In Table 2 we have compared the absolute and relative errors with respect to the "exact" solution for Example 1 at the "steady" state $t = 1.3$ (where $\|\dot{\boldsymbol{c}}(1.3)\|_2 \leq 10^{-2} \cdot \|\boldsymbol{c}_0\|_2$). The reference solution has been computed by CN with a local tolerance equal to 10^{-6}, whereas the comparison of the errors is made using a local tolerance of 10^{-4} for both methods (namely "tol" for the ReLPM algorithm in Table 1), which guarantees an error of the order of the spatial discretization error. Note that ReLPM is more accurate than CN at the final time, which shows that the mass-lumping technique does not significantly degrade the accuracy of the exponential integrator (5).

While for CN the local time-step is selected adaptively, in order to guarantee a local error below the given tolerance, for scheme (5) there is no restriction on the choice of Δt_k, since it is *exact* for autonomous linear systems of ODEs. To

Table 2. Comparison of absolute and relative errors for Example 1

	CN	ReLPM $\eta = 0.1$	ReLPM $\eta = 0.5$
abs. err.	$3.5 \cdot 10^{-2}$	$4.7 \cdot 10^{-3}$	$4.8 \cdot 10^{-3}$
rel. err.	$1.0 \cdot 10^{-3}$	$1.4 \cdot 10^{-4}$	$1.4 \cdot 10^{-4}$

follow with some accuracy the evolution of the solution, we propose as in [9] to select the local time-step in (5) in such a way that the relative variation of the solution be smaller than a given percentage η, that is

$$\|\boldsymbol{c}_{k+1} - \boldsymbol{c}_k\|_2 \leq \eta \cdot \|\boldsymbol{c}_k\|_2, \quad 0 < \eta < 1 \ . \tag{8}$$

If condition (8) is not satisfied, the time step Δt_k is halved and $\boldsymbol{c}_{k+1}$ recomputed; if it is satisfied with $\eta/2$ instead of η, the next time-step Δt_{k+1} is doubled.

Tables 3–5 show that the ReLPM exponential integrator is always faster than CN method, with speed-ups (ratio between CPU times) ranging from 5.1 to 16.2.

Table 3. Comparison of CN and ReLPM for Example 1 (constant initial data)

CN		ReLPM			
# time-steps	CPU s.	η	# time-steps	CPU s.	Speed-up
451	37.47	0.1	41	7.33	5.1
		0.25	17	4.62	8.1
		0.5	11	4.02	9.3
		0.75	8	3.68	10.2

Table 4. Comparison of CN and ReLPM for Example 2 (peaked initial data)

CN		ReLPM			
# time-steps	CPU s.	η	# time-steps	CPU s.	Speed-up
385	198.24	0.1	44	27.53	7.2
		0.25	22	19.94	9.9
		0.5	11	15.37	12.9
		0.75	9	12.26	16.2

Table 5. Comparison of CN and ReLPM for Example 3 (constant initial data)

CN		ReLPM			
# time-steps	CPU s.	η	# time-steps	CPU s.	Speed-up
1079	507.36	0.1	90	66.24	7.7
		0.25	38	41.53	12.2
		0.5	21	36.09	14.0
		0.75	15	32.09	15.8

Note that the local tolerance for Examples 1 and 2 has been chosen equal to 10^{-4}, that is the order of spatial discretization error, whereas in Example 3

equal to 10^{-5}, to match the finer accuracy of the FE discretization. Finally, Fig. 1 shows that even the choice of the variation percentage $\eta = 0.5$ allows to track with sufficient accuracy the evolution of the solution, with much less steps than CN, exhibiting speed-ups of one order of magnitude.

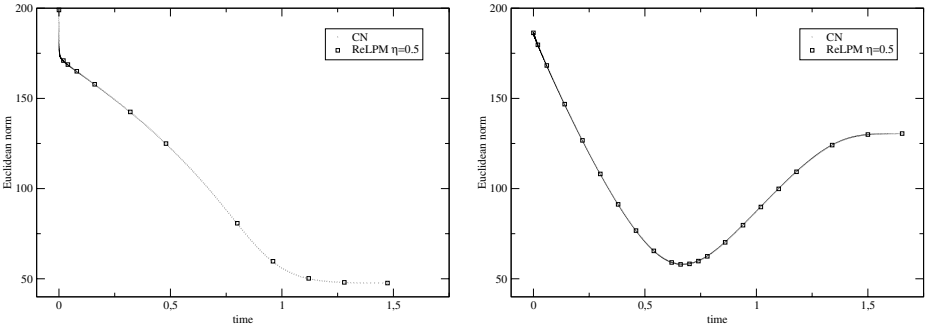


Fig. 1. Evolution of the norm of the solution computed by the CN method and the ReLPM ($\eta = 0.5$) for Example 2 (left) and Example 3 (right)

References

1. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Volume 33 of Springer series in Computational Mathematics. Springer-Verlag, Berlin (2003)
2. Wood, W.L.: Introduction to numerical methods for water resources. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York (1993)
3. Fetter, C.W.: Applied Hydrogeology. Prentice Hall (2000)
4. Thomée, V.: Galerkin finite element methods for parabolic problems. Springer-Verlag, Berlin (1997)
5. Schaefer, M.J.: A polynomial based iterative method for linear parabolic equations. J. Comput. Appl. Math. **29** (1990) 35–50
6. Gallopoulos, E., Saad, Y.: Efficient solution of parabolic equations by Krylov subspace methods. SIAM J. Sci. Statist. Comput. **13** (1992) 1236–1264
7. Hochbruck, M., Lubich, C., Selhofer, H.: Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput. **19** (1998) 1552–1574
8. Vanselow, R.: Error estimates of a FEM with lumping for parabolic PDEs. Computing **68** (2002) 131–141
9. Caliari, M., Vianello, M., Bergamaschi, L.: Interpolating discrete advection-diffusion propagators at Leja sequences. Preprint, to appear in J. Comput. Appl. Math. (2003)
10. Hochbruck, M., Lubich, C.: On Krylov subspace approximations to the matrix exponential. SIAM J. Numer. Anal. **34** (1997) 1911–1925
11. Saad, Y.: Analysis of some Krylov subspace approximations to the matrix exponential operator. SIAM J. Numer. Anal. **29** (1992) 209–228

12. Sidje, R.B.: Expokit. A Software Package for Computing Matrix Exponentials. ACM Trans. Math. Software **24** (1998) 130–156
13. Pini, G., Gambolati, G.: Arnoldi and Crank-Nicolson methods for integration in time of the transport equation. Int. J. Numer. Meth. Fluids **35** (2001) 25–38
14. Moret, I., Novati, P.: RD rational approximation of the matrix exponential. Preprint, to appear in BIT (2003)
15. Bergamaschi, L., Vianello, M.: Efficient computation of the exponential operator for large, sparse, symmetric matrices. Numer. Linear Algebra Appl. **7** (2000) 27–45
16. Bergamaschi, L., Caliari, M., Vianello, M.: Efficient approximation of the exponential operator for discrete 2D advection-diffusion problems. Numer. Linear Algebra Appl. **10** (2003) 271–289
17. Moret, I., Novati, P.: The computation of functions of matrices by truncated Faber series. Numer. Funct. Anal. Optim. **22** (2001) 697–719
18. Moret, I., Novati, P.: An interpolatory approximation of the matrix exponential based on Faber polynomials. J. Comput. Appl. Math. **131** (2001) 361–380
19. Novati, P.: A polynomial method based on Fejér points for the computation of functions of unsymmetric matrices. Appl. Numer. Math. **44** (2003) 201–224
20. Reichel, L.: Newton interpolation at Leja points. BIT **30** (1990) 332–346
21. Baglama, J., Calvetti, D., Reichel, L.: Fast Leja points. Electron. Trans. Numer. Anal. **7** (1998) 124–140
22. Walsh, J.L.: Interpolation and approximation by rational functions in the complex domain. Volume XX of Amer. Math. Soc. Colloq. Publ. AMS (1935)

Function Fitting Two–Step BDF Algorithms for ODEs

Liviu G. Ixaru[1] and Beatrice Paternoster[2]

[1] Institute of Physics and Nuclear Engineering, Bucharest, Romania
[2] Dipartimento di Matematica e Informatica, Universitá di Salerno, Italy

Abstract. We investigate the problem of how big would be the additional accuracy gain from a two-step bdf algorithm for ordinary differential equations if its weights are constructed via function fitting. We find that (*i*) the order of the algorithm is increased by three units (from two to five), (*ii*) this enhancement can be achieved not only in the frame of the traditional exponential fitting but also in the frame of a new, more general approach, which results more flexible and safer than the other one.

1 Introduction

It is well known that the quality of a multistep algorithm to solve first order ODEs depends on the set of reference functions chosen for the determination of the algorithm weights. The classical form of these algorithms is constructed in terms of power functions but other functions may be used as well, such as exponential functions or mixtures of power and exponential functions, see e.g. [3], [1], [8], [11], and the references therein.

The old problem of how the frequencies of the exponential functions should be tuned in order to obtain a maximal gain in accuracy received a pertinent answer only recently, [6]. (According to the usual terminology in the field, the frequency is the parameter λ in the function $\exp(\lambda x)$.) In this paper our main interest consists in searching for alternatives to the exponential fitting. Multistep algorithms based on other functions than the power and/or exponential functions were published before, mainly in connection with the Schrödinger equation (see [2] for functions of the form $x^{l+k},\ l > 0,\ k = 1,\ 2,\ \ldots$ (i.e. an ad hoc subset of power functions) or [7,9,10] for the Bessel and Neumann functions). Now we choose these functions in terms of the behaviour of the derivatives of the solution, implementing the function fitting technique to derive a fifth order version of the method.

2 Preliminaries

Let X be some point on the real axis and $[X-h, X+h]$ an interval around it, $[X-h, X+h]$. We consider the third order linear and homogeneous differential equation, where $\alpha_m(x),\ m = 0,\ 1,\ 2$, are low degree polynomials:

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 443–450, 2004.
© Springer-Verlag Berlin Heidelberg 2004

$$y^{(3)} + \alpha_2(x)y'' + \alpha_1(x)y' + \alpha_0(x)y = 0,\ X - h \le x \le X + h, \tag{1}$$

We denote the three linear independent solutions of equation (1) as $\phi_1(x)$,$\phi_2(x)$ and $\phi_3(x)$. Each of them is constructed by solving the initial value problem for a specified set of initial conditions at $x = X$. We take:

$$\begin{array}{llll} y(X) = 1,\ y'(X) = 0,\ y''(X) = 0 & \text{for} & \phi_1(x), \\ y(X) = 0,\ y'(X) = 1,\ y''(X) = 0 & \text{for} & \phi_2(x), \\ y(X) = 0,\ y'(X) = 0,\ y''(X) = 2 & \text{for} & \phi_3(x). \end{array}$$

The solution of eq.(1) can be constructed through a power series expansion around X:

$$y(x) = \sum_{n=0} y_n \Delta^n, \quad \text{with} \quad \Delta = x - X,\ \Delta \in [-h,\ h]. \tag{2}$$

We are interested only in five particular cases; each of these cases leads to a specific set of weights in the two–step bdf algorithm. The case abbreviated below as C0 leads to the classical algorithm, C1, C2 and C3 provide three versions which form together the new, flexible fitting algorithm while C4 leads to a typical exponential fitting algorithm.

C0 . This is the particular case of eq.(1) when $\alpha_0(x) = \alpha_1(x) = \alpha_2(x) = 0$. In this case the three solutions are simply $\phi_1(x) = 1$, $\phi_2(x) = x - X$ and $\phi_3(x) = (x - X)^2$.

C1 . This is $\alpha_2(x) = p_0 + p_1\Delta + p_2\Delta^2,\ \alpha_1(x) = 0,\ \alpha_0(x) = 0$, where $p_0,\ p_1$ and p_2 are real constants. Eq.(1) reads

$$y^{(3)} + (p_0 + p_1\Delta + p_2\Delta^2)y'' = 0. \tag{3}$$

Upon inserting (2) in (3) and conveniently organizing the terms, the following recurrence relation is obtained:

$$\begin{aligned} y_{n+3} = &-\tfrac{1}{(n+3)(n+2)(n+1)}[(n+2)(n+1)p_0 y_{n+2} \\ &+ (n+1)np_1 y_{n+1} + n(n-1)p_2 y_n],\ n = 0,\ 1,\ 2, \ldots \end{aligned}$$

Since in this case we have $\phi_1(x) = 1$ and $\phi_2(x) = x - X$ directly, the recurrence relation has to be used only for the construction of $\phi_3(x)$. The starting values consistent with the mentioned initial value set for $\phi_3(x)$ are:

$$y_0 = 0,\ y_1 = 0,\ y_2 = 1.$$

C2 . This corresponds to $\alpha_2(x) = 0,\ \alpha_1(x) = p_0 + p_1\Delta + p_2\Delta^2,\ \alpha_0(x) = 0$, where again $p_0,\ p_1$ and p_2 are real constants. Eq.(1) now becomes

$$y^{(3)} + (p_0 + p_1\Delta + p_2\Delta^2)y' = 0.$$

We have $\phi_1(x) = 1$ directly, while for $\phi_2(x)$ and $\phi_3(x)$ we use the recurrence formula

$$y_{n+3} = -\frac{1}{(n+3)(n+2)(n+1)}[(n+1)p_0 y_{n+1} + np_1 y_n + (n-1)p_2 y_{n-1}],\ n = 0,\ 1, \ldots$$

with the starting values

$$y_0 = 0,\ y_1 = 1,\ y_2 = 0 \quad \text{for} \quad \phi_2(x),$$
$$y_0 = 0,\ y_1 = 0,\ y_2 = 1 \quad \text{for} \quad \phi_3(x),$$

and $y_{-1} = 0$ by default.

C3 . This corresponds to $\alpha_2(x) = 0,\ \alpha_1(x) = 0,\ \alpha_0(x) = p_0 + p_1\Delta + p_2\Delta^2$, where, as before, $p_0,\ p_1$ and p_2 are real constants. Eq.(1) reads

$$y^{(3)} + (p_0 + p_1\Delta + p_2\Delta^2)y = 0.$$

with the folllowing recurrence relation for $\phi_1(x)$, $\phi_2(x)$ and $\phi_3(x)$:

$$y_{n+3} = -\frac{1}{(n+3)(n+2)(n+1)}[p_0 y_n + np_1 y_{n-1} + n(n-1)p_2 y_{n-2}],\ n = 0,\ 1,\ 2, \ldots$$

with the starting values

$$\begin{aligned} &y_0 = 1,\ y_1 = 0,\ y_2 = 0, \quad \text{for} \quad \phi_1(x), \\ &y_0 = 0,\ y_1 = 1,\ y_2 = 0, \quad \text{for} \quad \phi_2(x), \\ &y_0 = 0,\ y_1 = 0,\ y_2 = 1, \quad \text{for} \quad \phi_3(x), \end{aligned} \tag{4}$$

and with the default values $y_{-2} = y_{-1} = 0$ in all these.

C4 . Here $\alpha_0(x) = q_0,\ \alpha_1(x) = q_1,\ \alpha_2(x) = q_2$, where $q_0,\ q_1$ and q_2 are real constants. The solution of

$$y^{(3)} + q_2 y'' + q_1 y' + q_0 y = 0 \tag{5}$$

is given by the recurrence relation

$$y_{n+3} = -\frac{1}{(n+3)(n+2)(n+1)}[(n+2)(n+1)q_2 y_{n+2} + (n+1)q_1 y_{n+1} + q_0 y_n],$$

$n = 0,\ 1,\ 2, \ldots$ with the starting values (4).

This case admits also an analytic solution. Three linear independent solutions are $\exp(\lambda_1 x),\ \exp(\lambda_2 x)$ and $\exp(\lambda_3 x)$, where $\lambda_1,\ \lambda_2$ and λ_3 are the roots of the polynomial $P(\lambda) = \lambda^3 + q_2\lambda^2 + q_1\lambda + q_0$. Our $\phi_1(x),\ \phi_2(x)$ and $\phi_3(x)$ are some linear combinations of these exponential functions. For all cases C1, C2, C3 and C4 the number of terms to be retained in the series (2) in order to reach some predetermined accuracy in the results depends on h and on the numerical values of the parameters $p_0,\ p_1,\ p_2$ or $q_0,\ q_1,\ q_2$. For given p–s or q–s this number decreases with h.

Weights of the two–step bdf algorithm

We consider the initial value problem

$$y' = f(x, y),\ x \in [a,\ b],\ y(a) = y_0\ , \tag{6}$$

and its solution by a two-step bdf algorithm,

$$a_0\bar{y}_k + a_1\bar{y}_{k+1} + \bar{y}_{k+2} = hb_2 f(x_{k+2}, \bar{y}_{k+2}),\ k = 0,\ 1,\ 2, \ldots\ , \tag{7}$$

on an equidistant partition with the stepsize h. $\bar{y}_j$ is an approximation to $y(a + jh)$. The weights $a_0,\ a_1$ and b_2 will differ from one interval to another. Their

construction on the current interval centered at $X = x_{k+1}$ (and then $x_k = X - h,\ x_{k+2} = X + h$) is done as it follows.

The linear functional

$$\mathcal{L}[y(x); X, h, a_0, a_1, b_2] =: a_0 y(X-h) + a_1 y(X) + y(X+h) - h b_2 y'(X+h)$$

is associated to algorithm (7) and we require that this functional is identically vanishing when $y(x) = \phi_1(x), \phi_2(x),\ \phi_3(x)$. We have

$$a_0\phi_i(X-h) + a_1\delta_{i1} + \phi_i(X+h) - b_2 h\phi_i'(X+h) = 0,\ i = 1,\ 2,\ 3. \qquad (8)$$

Here δ_{ij} is the usual Kronecker symbol and it appears because, by the very construction, the three functions satisfy $\phi_1(X) = 1$ and $\phi_2(X) = \phi_3(X) = 0$. The linear system (8) has the solution

$$\begin{aligned} D &= h[-\phi_2(X-h)\phi_3'(X+h) + \phi_3(X-h)\phi_2'(X+h)], \\ a_0 &= h[\phi_2(X+h)\phi_3'(X+h) - \phi_3(X+h)\phi_2'(X+h)]/D, \\ b_2 &= [-\phi_2(X-h)\phi_3(X+h) + \phi_3(X-h)\phi_2(X+h)]/D, \\ a_1 &= b_2 h\phi_1'(X+h) - a_0\phi_1(X-h) - \phi_1(X+h). \end{aligned}$$

When $\phi_1(x) = 1,\ \phi_2(x) = x - X$ and $\phi_3(x) = (x-X)^2$, as for C0, we get

$$a_0 = \frac{1}{3},\ a_1 = -\frac{4}{3} \ \text{ and } \ b_2 = \frac{2}{3}, \qquad (9)$$

that is the classical weights.

3 Error Analysis

For the error analysis it is convenient to express $\mathcal{L}$ in terms of the moments. The moment L_m is the expression of $\mathcal{L}$ when $y(x)$ is the power function $(x-X)^m$, i.e.

$$L_m =: \mathcal{L}[(x-X)^m; X, h, a_0, a_1, b_2].$$

In fact, the knowledge of the moments allow writing

$$\mathcal{L}[y(x); X, h, a_0, a_1, b_2] = \sum_{m=0}^{\infty} \frac{1}{m!} L_m y^{(m)}(X) \qquad (10)$$

for any function $y(x)$ which admits a series expansion in power functions.

If $y(x)$ is the solution of (6), the series (10) furnishes the local error of the method. Omitting the details, it is possible to draw two conclusions:

1. If $p_0 = p_1 = p_2$, as for the classical algorithm, the first nonvanishing moment is L_3. The leading term of the error then reads

$$lte = \frac{1}{3!} L_3 y^{(3)}(X) = -\frac{2}{9} h^3 y^{(3)}(X),$$

a well known result. The order of this version is therefore two.

2. The order is still two also when $p_0,\ p_1$ and/or p_2 are different from zero. This is because the free term in L_2 is absent i.e. $L_2 \sim h^3$.

$p_0,\ p_1$ and p_2 should be chosen in order to obtain an increased accuracy. To this purpose we introduce the function

$$s(x) =: \frac{y^3(x)}{y''(x)}, \tag{11}$$

where $y(x)$ is the exact solution of eq.(6). We have

$$y^{(3)}(x) = -s(x)y''(x),\ y^{(4)}(x) = (s^2(x) - s'(x))y''(x),$$
$$y^{(5)}(x) = (s^3(x) - 3s(x)s'(x) + s''(x))y''(x),\ \ldots$$

An analysis of $\mathcal{L}$ suggests searching for a determination of the p-s in terms of $s(x)$. We just require that the parabola $\alpha_2(x) = p_0 + p_1\Delta + p_2\Delta^2$ is interpolating $s(x)$ at the three knots X and $X \pm h$ to obtain

$$p_0 = s_0,\ p_1 = \frac{1}{2h}(s_1 - s_{-1}),\ p_2 = \frac{1}{2h^2}(s_1 - 2s_0 + s_{-1})$$

where $s_j = s(X + jh)$. With this determination, finally, $\mathcal{L} \sim h^6$.

To summarize, we have obtained the following result: *The C1 version of the two step bdf algorithm (7) is in general of the second order. However, if the coefficients $p_0,\ p_1$ and p_2 are fixed by the parabolic interpolation of function $s(x)$ defined by eq.(11), the order becomes five.*

The above error analysis can be repeated for versions C2 and C3, as well. The result remains the same, i.e. each of these is in general of the second order, but if the p-s are determined by interpolating the functions

$$s(x) =: \frac{y^3(x)}{y'(x)},$$

for C2, and

$$s(x) =: \frac{y^3(x)}{y(x)},$$

for C3, then these versions become of the fifth order.

Seen from practical point of view each of these three versions exhibits its own limitation. For instance, C1 version has to be avoided when the second derivative of the solution of (6) has a zero inside the quoted interval. Likewise, the use of optimal C2 (or C3) versions should be avoided on the intervals where the first derivative of the solution (or the solution itself) has a zero. For the (exponential fitting) C4–based algorithm, the interpolation procedure to obtain a fifth order algorithmis is necessarily replaced by that of solving the linear algebraic system

$$q_2y''(X+jh) + q_1y'(X+jh) + q_0y(X+jh) = -y^{(3)}(X+jh),\ j = -1,\ 0,\ 1 \tag{12}$$

for the unknowns $q_0,\ q_1$ and q_2, which becomes increasingly ill-conditioned and therefore the determination of the parameters is less and less accurate.

4 The Flexible Function Fitting Algorithm

In spite of the fact that each of these version has its own limitation, the very existence of three function fitting versions of the same fifth order makes a choice possible in terms of a safety criterion. The rule of choosing between various versions in the flexible fitting algorithm is in the following.

Let us assume that we know the exact values at the points $X-h,\ X$ and $X+h$ for the solution of the differential equation (6) and for its derivatives.

We introduce

$$s^1 = -\frac{\sum_{j=-1}^{1} y^{(3)}(X+jh)}{\sum_{j=-1}^{1} y''(X+jh)}, s^2 = -\frac{\sum_{j=-1}^{1} y^{(3)}(X+jh)}{\sum_{j=-1}^{1} y'(X+jh)},$$

$$s^3 = -\frac{\sum_{j=-1}^{1} y^{(3)}(X+jh)}{\sum_{j=-1}^{1} y(X+jh)},$$

and evaluate the maximal deviation of the input data for each version,

$$\Delta s^m = \max[|s^m - s^m_{-1}|, |s^m - s^m_0|, |s^m - s^m_1|],\ m = 1,\ 2,\ 3.$$

We compare $\Delta s^1,\ \Delta s^2$ and Δs^3 and let $\bar{m}$ be that m such that $\Delta s^{\bar{m}}$ is the smallest of the three. The version C$\bar{m}$ is selected for application in that step. In the case of systems of differential equations the selection of the optimal version is operated on each component separately because only in this way the method remains of the fifth order.

In a real run the exact values of the solution and of its derivatives are obviously not known but we just rely on the numerical values calculated up to that interval, that is on $\bar{y}_0,\ \bar{y}_1,\ \bar{y}_2, \ldots,\ \bar{y}_k,\ \bar{y}_{k+1}$. On effectively using the last three of them, an extrapolated value at the new point $x_{k+2} = X + h$, denoted y^*_{k+2}, is generated by the four point Milne–Simpson formula

$$-\bar{y}_{k-1} + \frac{27}{11}(-\bar{y}_k + \bar{y}_{k+1}) + y^*_{k+2} = h[\frac{3}{11}(f_{k-1} + f(x_{k+2}, y^*_{k+2})) + \frac{27}{11}(f_k + f_{k+1})],$$

which is sufficiently accurate for this purpose; its error behaves as h^7. $\bar{y}_k,\ \bar{y}_{k+1}$ and y^*_{k+2} are then accepted as sufficiently reliable representations of the exact values of the solution at the three mesh points of the current interval. The values of the derivatives at the same points (these are needed for the interpolation) are generated via the analytic expressions of f and of its total first and second derivative with respect to x.

5 A Numerical Illustration

We present one test case, whose exact solution is known. We compare three algorithms: *(i)* The classical two-step bdf algorithm (C0); it is based on the three solutions of eq.(1) with $\alpha_0(x) = \ \alpha_1(x) = \ \alpha_2(x) = 0$ and it has the weights in (9). *(ii)* The optimal exponential fitting algorithm (CEF); this is based on the

solutions of eq.(5) where the three parameters $q_0,\ q_1$ and q_2 are calculated by solving the system (12) in each step. *(iii)* The optimal flexible fitting algorithm (CFF); in each step this chooses between versions C1, C2 and C3 in terms of the safety criterion explained before. We assume that $y(a)$ and $y(a+h)$ are given for C0 while three starting values, $y(a),\ y(a+h)$ and $y(a+2h)$ are available for CEF and CFF. The third data is needed for the activation of the Milne–Simpson extrapolation.

Let us consider the system of two equations

$$\begin{aligned} y_1' &= -2y_1 + y_2 + 2\sin(x), \\ y_2' &= y_1 - 2[y_2 + \sin(x) - \cos(x)],\ x \in [0,\ 2],\ y_1(0) = 2,\ y_2(0) = 1. \end{aligned} \tag{13}$$

Its solution is

$$\begin{aligned} y_1(x) &= \exp(-x) + \exp(-3x) + \sin(x), \\ y_2(x) &= \exp(-x) - \exp(-3x) + \cos(x). \end{aligned}$$

Table 1. Absolute errors $\Delta y_i(x) = y_i(x) - y_i^{comput}(x)$ at $x =$ 1.5 and 2.0 from the classical two–point bdf algorithm C0, its optimal exponential fitting extension and its optimal flexible fitting extension for the two components of system (13).

	$x = 1.5$		x=2.0	
h	Δy_1	Δy_2	Δy_1	Δy_2
C0				
0.0500	0.822(−03)	−0.353(−03)	0.260(−03)	−0.230(−03)
0.0250	0.198(−03)	−0.853(−04)	0.604(−04)	−0.579(−04)
0.0125	0.485(−04)	−0.210(−04)	0.145(−04)	−0.145(−04)
CEF				
0.0500	0.131(−07)	0.151(−07)	0.115(−07)	0.143(−07)
0.0250	0.409(−09)	0.482(−09)	0.329(−09)	0.326(−09)
0.0125	0.128(−10)	0.152(−10)	0.163(−06)	0.522(−06)
CFF				
0.0500	−0.121(−07)	−0.154(−07)	−0.400(−08)	−0.113(−07)
0.0250	−0.832(−09)	−0.884(−09)	−0.393(−09)	−0.597(−09)
0.0125	−0.286(−10)	−0.301(−10)	−0.139(−10)	−0.201(−10)

Each of the two components is a linear combination of four exponential functions, with the frequencies $-1,\ -3$ and $\pm\, i$, respectively. For this reason, one may be tempted to admit that the exponential fitting version is the method of choice. The reality is however different. At each step, and for each of the two components of the equation, the version CEF determines its optimal parameters by first solving the linear system (12) but the accuracy of this evaluation depends on whether the system is well or badly conditioned which, at its turn, depends of the magnitude of h. When h is still big the effect is negligible but when h is

further decreased it becomes more and more important. In the case of system (13) this effect is negligible for both components of the solution when $h = 0.05$ or $h = 0.025$ but when $h = 0.0125$ it becomes important around $x = 1.7$ for the second component. Note also that $y_2(x)$ changes the sign in that region.

The consequencies are clearly seen in table 2. The errors from CEF are conforming the theoretical fifth order when $x = 1.5$ but when $x = 2$ the errors at $h = 0.0125$ are abnormally big and they will remain so at any bigger x. This does not happen with CFF because this algorithm is flexible, and also because the parameters are calculated by interpolation. In the interval around $x = 1.7$ the version C3 is excluded for the second component but the two others are still available. As a matter of fact, in that region, it activates the version C3 for the first component but C2 for the second.

The result is very encouraging but more investigations are needed for a complete understanding of the properties of the new versions and for improving the quality of the present code, together with analysis of the stability properties of the CFF algorithm.

References

1. L. Gr. Ixaru, Numerical Methods for Differential Equations and Applications, Reidel, Dordrecht-Boston-Lancaster, 1984.
2. L. Gr. Ixaru, The Numerov method and singular potentials, *J. Comput. Phys.* **72**, 270–274, 1987.
3. L. Gr. Ixaru, Operations on oscillatory functions, *Comput. Phys. Commun.* **105**, 1–19, 1997.
4. L. Gr. Ixaru and B. Paternoster, A Gauss quadrature rule for oscillatory integrands, *Comput. Phys. Commun.* **133**, 177–188, 2001.
5. L. Gr. Ixaru, M. Rizea, H. De Meyer and G. Vanden Berghe, Weights of the exponential fitting multistep algorithms for first order ODEs, *J. Comput. Appl. Math.* **132**, 83–93, 2001.
6. L. Gr. Ixaru, G. Vanden Berghe and H. De Meyer, Frequency evaluation in exponential fitting multistep algorithms for ODEs, *J. Comput. Appl. Math.* **140**, 423–434, 2002.
7. T. E. Simos, A sixth order Bessel and Neumann fitted method for the numerical solution of the Schrödinger equation, *Molecular Simulation* **21**, 191–204, 1999.
8. T. E. Simos, An exponentially fitted eight-order method for the numerical solution of the Schrödinger equation, *J. Comput. Appl. Math.* **108**, 177–194, 1999.
9. T. E. Simos and A. D. Raptis, A fourth order Bessel fitting method for the numerical solution of the Schrödinger equation, *J. Comput. Appl. Math.* **43**, 313–322, 1992.
10. T. E. Simos and P. S. Williams, Bessel and Neumann fitted methods for the numerical solution of the radial Schrödinger equation, *Computers and Chemistry* **21**, 175–179, 1997.
11. G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke, Exponentially-fitted explicit Runge-Kutta methods, *Comput. Phys. Commun.* **123**, 7–15, 1997.

Pseudospectral Iterated Method for Differential Equations with Delay Terms

J. Mead[1] and B. Zubik–Kowal[2]

[1] Department of Mathematics, Boise State University, Boise, Idaho 83725
mead@math.boisestate.edu
[2] Department of Mathematics, Boise State University, Boise, Idaho 83725
zubik@math.boisestate.edu

Abstract. New efficient numerical methods for hyperbolic and parabolic partial differential equations with delay terms are investigated. These equations model a development of cancer cells in human bodies. Our goal is to study numerical methods which can be applied in a parallel computing environment. We apply our new numerical method to the delay partial differential equations and analyse the error of the method. Numerical experiments confirm our theoretical results.

Keywords: Linear delay equations, numerical approximations, parallel computing, error estimations, numerical experiments.

1 Introduction

1.1 Delay Partial Differential Problems

In this paper investigate numerical solutions to the linear initial boundary value problem with a delay term

$$\begin{aligned} &\frac{\partial}{\partial t}u(x,t)=\epsilon\frac{\partial^2}{\partial x^2}u(x,t)+c\frac{\partial}{\partial x}u(x,t)-\nu u(x,t-\tau_0)+g(x,t),\ 0<t\leq T,\\ &\quad u(x,t)=f_0(x,t),\quad -\tau_0\leq t\leq 0,\ -L\leq x\leq L. \end{aligned} \tag{1.1}$$

Here, $\epsilon>0$, $c\in\mathbb{R}$, $\tau_0\geq 0$, $L>0$ and $T>0$ are given constants. The choice of $\epsilon=0$ gives the hyperbolic equation, $c=0$ gives the parabolic equation and the choice of $\epsilon\neq 0$ and$c\neq 0$ gives the parabolic advection-diffusion equation.

Different types of boundary conditions are required for the two cases $\epsilon\neq 0$ and $\epsilon=0$. For the parabolic case ($\epsilon\neq 0$) there are two boundary conditions

$$u(\pm L,T)=f_{\pm}(t),$$

while for the hyperbolic case ($\epsilon=0$, $c\neq 0$) there is one boundary condition, either

$$u(L,t)=f_+(t)\ (\text{if } c>0)\quad \text{or}\quad u(-L,t)=f_-(t)\ (\text{if } c<0).$$

Here, f_0, $f_{\pm}$ and g are given continuous functions.

Delay problems like (1.1) are used to model cancer cells in human tumors, see [1]. For other applications in population dynamics see [4].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 451–458, 2004.
© Springer-Verlag Berlin Heidelberg 2004

1.2 Purpose of the Paper

We study the Chebyshev pseudospectral (ChPS) spatial discretization of (1.1) (see e.g. [3], [5]) with the Kosloff and Tal-Ezer (KTE) transformation [6], together with Jacobi waveform relaxation methods for time integration (see e.g. [2], [8]). The ChPS method is known from its high accuracy. Another advantage of ChPS is that convergence of waveform relaxation (WR) is faster when it is applied to ChPS semi-discrete systems than it is applied to finite difference semi-discrete systems, see [8]. The advantage of WR methods is that they are efficient in parallel computing environments and for linear equations like (1.1) they allow direct application of implicit methods for integration in time (they allow the use of much larger stepsizes for time integration which fulfill stability restrictions, as compared with the explicit methods).

The goal of this paper is to show that since the KTE transformation makes most of the entries of the differentiation matrices significantly smaller, the convergence of WR is faster with the KTE transformation than without. Morever, for every type of equation the rate of WR convergence increases with increasing parameter $\alpha \in [0,1]$. We show this by using error bounds and actual errors which occur in the process of computations. Using extensive numerical data we also compare WR errors with their error bounds. Our results confirm the conclusions derived in [7].

2 Numerical Solution to Delay Problem (1.1)

We apply the process of pseudospectral spatial discretization ([3], [5]) with KTE transformation ([6]) and replace (1.1) by the following numerical scheme

$$\begin{aligned} \frac{d}{dt}U(t) &= Q_\alpha U(t) - \nu U(t-\tau_0) + f_\alpha(t), \quad 0 < t \le T, \\ U(t) &= \tilde{f}_0^\alpha(t), \quad -\tau_0 \le t \le 0. \end{aligned} \tag{2.2}$$

Here, Q_α is a matrix which depends on the parameter $\alpha \in [0,1]$ of the KTE transformation and the constants ϵ and c (see [3], [5] and [6]). The components of the vector function $U(t)$ provide us with approximations to the values $u(x,t)$ of the exact solution to problem (1.1), that is,

$$U_i(t) \approx u(x_i^\alpha, t),$$

where x_i^α, $i = 0, \ldots, N$, are the transformed Chebyshev grids (see [6]). Further, the vector function $f_\alpha(t)$ is related to $g(x,t)$ and $f_\pm(t)$; the vector function $\tilde{f}_0^\alpha(t)$ is related to the initial function $f_0(x,t)$.

We simplify the system (2.2) by splitting the matrix Q_α into two matrices:

$$\tilde{A}_\alpha = diag(Q_\alpha), \qquad \tilde{B}_\alpha = Q_\alpha - diag(Q_\alpha).$$

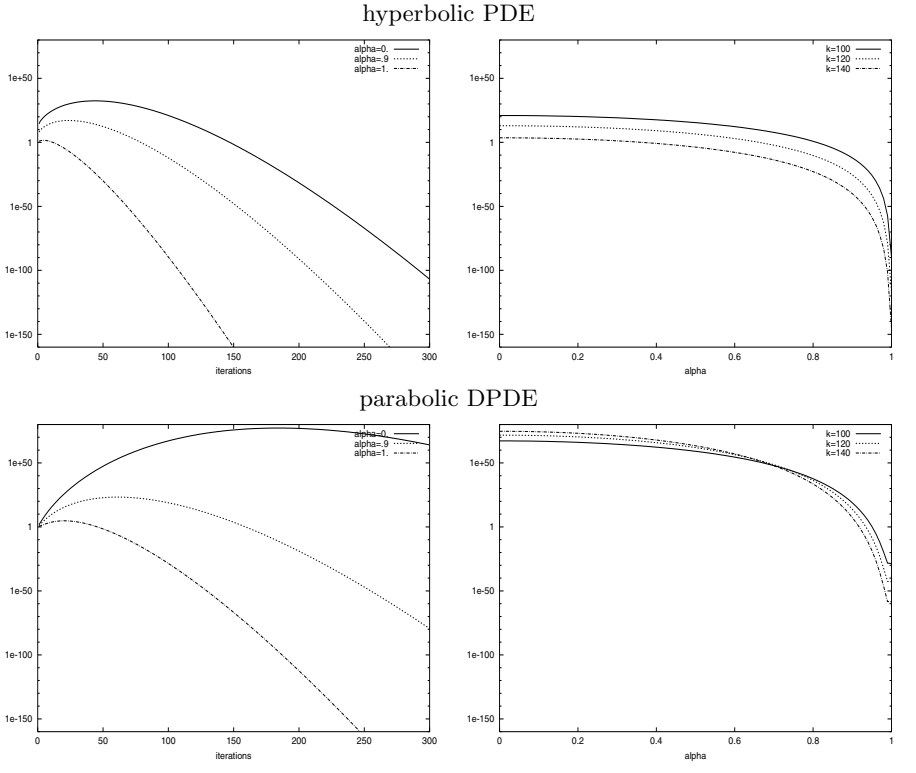


Fig. 1. Error bounds (2.5) as functions of the iteration index k and as functions of the parameter α.

Then the solution $U(t)$ to (2.2) can be approximated by successive iterates $U^k(t)$ computed according to the following Jacobi WR scheme

$$\begin{aligned} &\frac{d}{dt}U^{k+1}(t)=\tilde{A}_\alpha U^{k+1}(t)+\tilde{B}_\alpha U^k(t)-\nu U^k(t-\tau_0)+f(t), \quad 0<t\le T,\\ &U^{k+1}(t)=\tilde{f}_0^\alpha(t), \quad -\tau_0\le t\le 0, \end{aligned} \tag{2.3}$$

(see [2], [8] and [7]). Here, $k=0,1,\ldots$ is an iteration index and U^0 is an arbitrary starting function. Since the matrix $A_\alpha^{(l)}$ is diagonal, each equation of the system (2.3) can be solved independently by a different processor.

To study convergence of the waveform relaxation process (2.3) we consider the error defined by

$$e_\alpha^k(t)=U^k(t)-U(t). \tag{2.4}$$

To investigate an error estimation for (2.4) we denote by $\|\cdot\|$ an arbitrary vector norm or the induced matrix norm. It is shown in [8] that error estimations for (2.4) are more delicate if the following logarithmic norm

$$\mu(M) = \lim_{\delta \to 0^+} \frac{\|I + \delta M\| - 1}{\delta},$$

defined for an arbitrary matrix M, is used. Here, I is an identity matrix.

An error estimation for nonlinear delay differential equations (more general than the linear equation in (1.1)) is derived in [7, Theorem 4.3]. The estimation presented in [7, Theorem 4.3] is derived under the assumption that the right-hand sides of the nonlinear delay differential equations are Lipschitz continuous with respect to the delay terms. Since the delay equation (1.1) is linear, the coefficient ν is the Lipschtiz constant, which we use to apply [7, Theorem 4.3]. This leads to the following error estimation

$$\|e_\alpha^{k+1}(t)\| \leq \frac{(\|\tilde{B}_\alpha\| + \nu)^{k+1}}{k!} \int_0^t s^k \exp\Big(s\mu(\tilde{A}_\alpha)\Big)\, ds \max_{0 \leq \tau \leq t} \|e_\alpha^0(\tau)\|, \tag{2.5}$$

for $k = 0, 1, \ldots$, $t \in [0, T]$.

The estimation (2.5) has an advantage over the traditional estimation which, when applied to (1.1), after some computations (different than these included in the proof of [7, Theorem 4.3]) gives

$$\|e_\alpha^{k+1}(t)\| \leq \frac{\Big(t(\|\tilde{B}_\alpha\| + \nu)\Big)^{k+1}}{(k+1)!} \exp\Big(t\|\tilde{A}_\alpha\|\Big) \max_{0 \leq \tau \leq t} \|e_\alpha^0(\tau)\|. \tag{2.6}$$

It is easily seen that estimation (2.5) is sharper than estimation (2.6). We use the sharper estimation (2.5) and confirm the conclusion derived in [7] that the error bound (2.5) decreases for increasing α.

Figure 1 presents the error bounds (2.5) with $N = 32$ as functions of k for $\alpha = 0, 0.9, 1$ and as functions of $\alpha \in [0, 1]$ for $k = 100, 120, 140$. The error bounds are plotted for problems posed for $x \in [-10, 10]$ and $t \in [0, 1]$.

The error bounds are shown for delay partial differential equations (DPDEs) and for partial differential equations (PDEs) without delay terms. They are plotted for the hyperbolic PDE with $\epsilon = 0$, $c = 1$, $\nu = 0$ and the parabolic DPDE with $\epsilon = 1$, $c = 0$, $\nu = 5$. The pictures for the mixed problems with $\epsilon = 1$, $c = 1$ are similar to the pictures for the parabolic DPDE. We refer the reader to [7] for the pictures with parabolic PDEs, hyperbolic DPDEs and mixed PDEs and DPDEs.

We see significant improvement in convergence of WR when $\alpha = 0.9$ and $\alpha = 1$. In the next section we present the errors $\|e_\alpha^k(t)\|$ by means of extensive numerical data. It is shown in Section 3 that the errors $\|e_\alpha^k(t)\|$ behave like their error bounds (i.e. they decrease as α increases).

3 Numerical Experiments

In this section we present results of numerical experiments for the test problem (1.1). We choose $L = 10$ and $T = 1$ and consider six problems: the advective-diffusive problem with $\epsilon = c = 1$, the diffusive problem with $\epsilon = 1$, $c = 0$ and

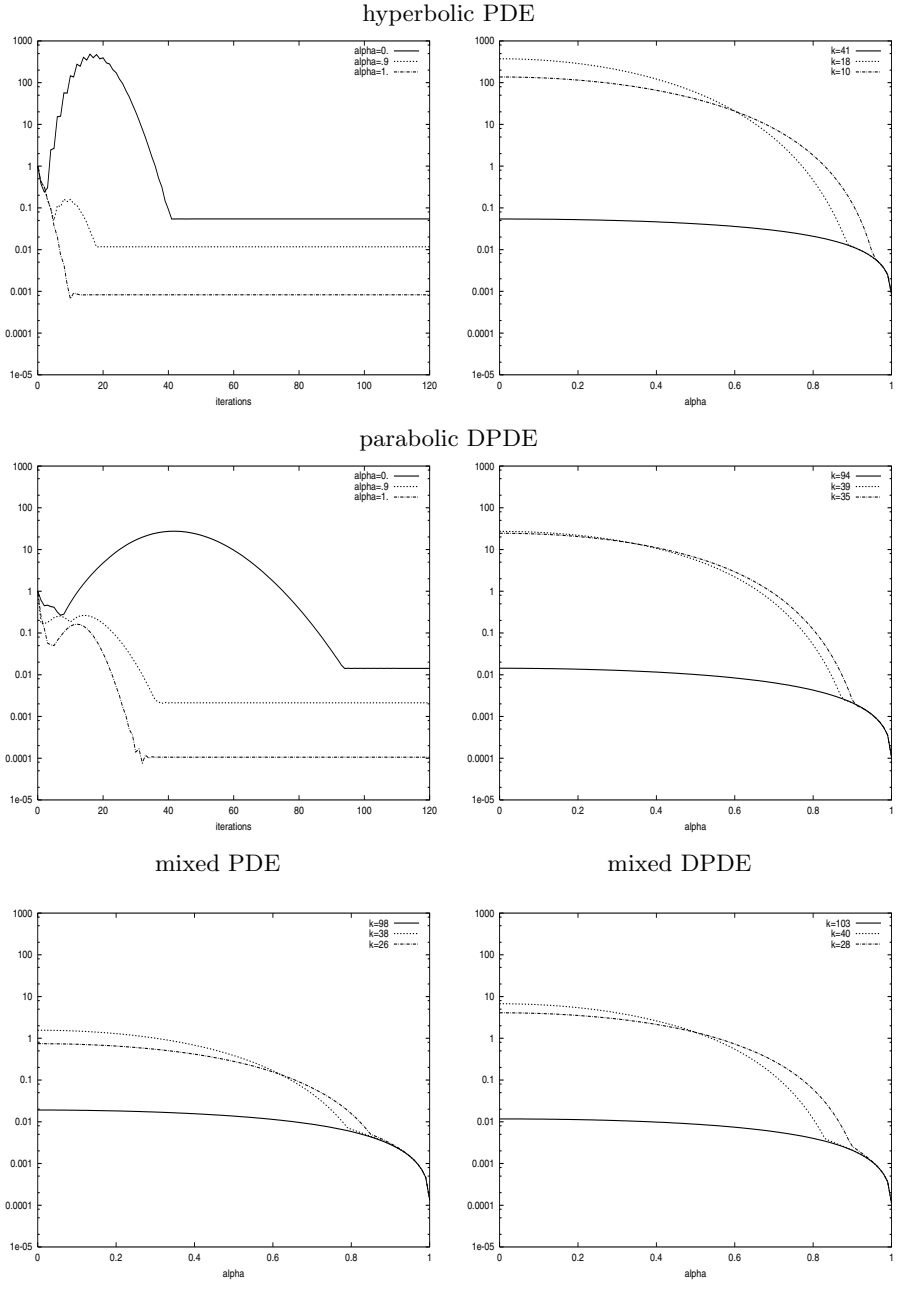


Fig. 2. Errors (3.7) with $N = 32$ as functions of k and as functions of α.

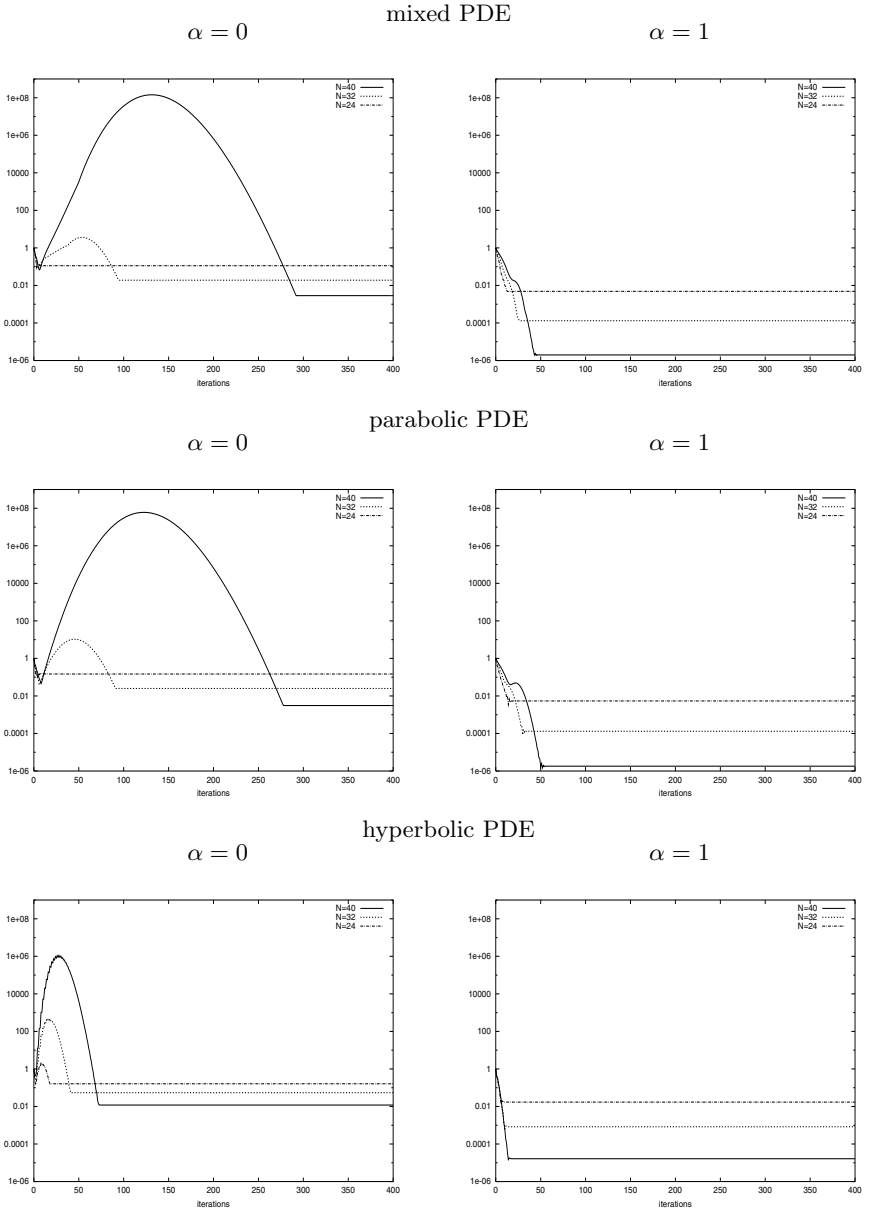


Fig. 3. Errors (3.7) for $N = 40, 32, 24$ as functions of k with $\alpha = 0$ (left pictures) and $\alpha = 1$ (right pictures).

the hyperbolic problem with $\epsilon = 0$, $c = 1$; each problem with $\nu = 0$ and $\tau_0 = 0$ for non-delay case and with $\nu = 5$ and $\tau_0 = 0.1$ for delay case.

To integrate the systems (2.3) in time we use the BDF3 method with the time step $\Delta t = 10^{-2}$. Integration of the systems (2.3) results in the approximations $U_{i,n}^k \approx U_i^k(n\Delta t)$, $n = 1, 2, \ldots$. We study the errors

$$\max_{i=0,\ldots,N} |U_{i,n}^k - u(y_i^\alpha, n\Delta t)| \tag{3.7}$$

measured at $n\Delta t = T = 1$ which can be compared with the upper bounds (2.5) plotted in Figures 1.

To compare the errors (3.7) with the error bounds from Figure 1 we present the errors (3.7) with $N = 32$ in Figure 2. They are presented as functions of the iteration index k for $\alpha = 0$, $\alpha = 0.9$, $\alpha = 1$ and as functions of the parameter α for fixed values of k. The fixed values of k are chosen to be the values where the error curves as functions of k become horizontal.

Figure 2 shows that for a fixed k the error (3.7) decreases as α increases with the smallest value at $\alpha = 1$. Figure 3 presents the errors (3.7) as functions of k for fixed values of α and for different values of N. The errors (3.7) are presented for the extreme values of parameter choices $\alpha = 0$ and $\alpha = 1$. Pictures for delay case are presented in [7].

4 Concluding Remarks

We applied the ChPS spatial discretization with the KTE transformation to delay and non-delay partial differential equations. Jacobi WR was then applied to the resulting semi-discrete systems. Our method is new for both kinds of equations. We conclude that the method works equally well for delay and non-delay equations.

Since Jacobi WR was used, our method can be efficiently used in parallel computing environments. We studied the relation between the WR convergence and the parameter α used for the spatial discretization. Using error bounds we conclude that WR converges more quickly as α increases from 0 to 1. This conclusion is confirmed by numerical experiments with delay and nondelay equations.

Since our method is successful for the test problem (1.1), our future work will address the numerical solution of the delay differential problem from [1].

References

1. B. Basse, B. C. Baguley, E. S. Marshall, W. R. Joseph, B. van Brunt, G. Wake, D. J. N. Wall, A mathematical model for analysis of the cell cycle in human tumors, to appear in *J. Mathematical Biology*.
2. K. Burrage, Z. Jackiewicz, R. A. Renaut, The performance of preconditioned waveform relaxation techniques for pseudospectral methods, *Numer. Methods Partial Differential Equations* **12** (1996) 245-263.
3. C. Canuto, M. Y. Hussaini, A. Quarteroni, T. A. Zang, Spectral Methods in Fluid Dynamics, Springer-Verlag, 1988.
4. C. J. Chyan, G. F. Webb, A model of proliferating cell populations with correlation of mother-daughter mitotic times, Ann. Mat. Pura Appl. **157**, (1991) 1-11.

5. B. Fornberg, A Practical Guide to Pseudospectral Methods, Cambridge University Press, 1996.
6. D. Kosloff and J. Tal-Ezer, A modified Chebyshev pseudospectral method with an $O(N^{-1})$ time step restriction, *J. Comput. Phys.* **104**, (1993) 457-469.
7. J. Mead and B. Zubik-Kowal, An iterated pseudospectral method for delay partial differential equations, submitted.
8. B. Zubik-Kowal, Chebyshev pseudospectral method and waveform relaxation for differential and differential-functional parabolic equations, *Appl. Numer. Math.* **34**, (2000) 309-328.

A Hybrid Numerical Technique for the Solution of a Class of Implicit Matrix Differential Equation

Nicoletta Del Buono and Luciano Lopez

Dipartimento di Matematica, Università degli Studi di Bari,
Via E. Orabona, 4, I-70125 Bari, ITALY [delbuono,lopezl]@dm.uniba.it

Abstract. This paper is concerned with the numerical solution of an implicit matrix differential system of the form $Y^T\dot{Y} - F(t, Y) = 0$, where $Y(t)$ is a $n \times n$ real matrix which may converge to a singular matrix. We propose a hybrid numerical technique based on an implicit second order Runge Kutta scheme which derives a particular algebraic Riccati equation and via its solution approximates the solutions of the differential problem at hand. Numerical examples demonstrating the behavior of the proposed approach are also reported.

1 Introduction

Many physical and industrial processes are most naturally and easily modeled as sets of implicit differential equations (IDEs) or as mixed systems of differential and algebraic equations (DAEs) and in the last decades there has been an increasing interest in exploiting the advantages of working directly with these implicit models ([2], [5], [11]). In this paper we concentrate our attention on problems whose general form is that of an implicit differential equation

$$G(t, Y(t), \dot{Y}(t)) = Y^T(t)\dot{Y}(t) - F(t, Y(t)) = 0, \quad t \geq t_0, \tag{1}$$

where G and Y are square real matrices of the same dimension $(n \times n)$ and G is supposed to have sufficiently many bounded derivatives (i.e. F is assumed to be a sufficiently smooth matrix function mapping $\mathbb{R} \times \mathbb{R}^{n\times n}$ into $\mathbb{R}^{n\times n}$). The initial value $Y(t_0)$ is supposed to be specified and the solution $Y(t)$ is sought on a bounded interval $[t_0, T]$.

If the partial derivative of G with respect to $\dot{Y}$ is invertible for all $t \in [t_0, T]$ then we can formally solve $\dot{Y}$ in (1) to obtain a set of ordinary differential equations:

$$\dot{Y}(t) = Y^{-T}(t)F(t, Y(t), \qquad Y(t_0) = Y_0. \tag{2}$$

Examples of the above ODEs appear in some matrix inverse eigenvalue problems, [6], feedback control problems [12], in the context of geometric integration on matrix manifolds [8], in multivariate data analysis [17].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 459–466, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Here we are interested in the case when $\frac{\partial G}{\partial \dot{Y}} = Y^T$ crosses a singular matrix at some instant $\tau \leq T$. In fact it has been proved in [7] that the property of the solution $Y(t)$ to be non singular for all $t \in [t_0, T]$ is not guaranteed a priori since the presence of a finite escape point $\tau \leq T$, where $Y(\tau)$ may become singular, is not precluded.

Numerical approaches for the solution of the problem at hand, must therefore take into account this behavior and may be divided roughly into two classes: direct discretization of the given system and methods which involve a reformulation combined with a discretization.

The desire for as direct a discretization as possible arises because a reformulation may be costly, may require additional input (enlarging the dimension of the problem) and may be more cumbersome than the original problem. In fact, generally, any implicit differential equations of the form (1) can be written in a semi-explicit form upon defining $\dot{Y} = Z$, which leads to

$$\begin{aligned} \dot{Y} &= Z \\ 0 &= G(t, Y, Z) = Y^T Z - F(t, Y). \end{aligned}$$

The latter differential algebraic equation with constraint may be solved carrying out a regularization technique, that is replacing the algebraic constraint by an ordinary differential equation depending on a small parameter $0 \leq \varepsilon \ll 1$ to get then

$$\begin{aligned} \dot{Y} &= Z \\ \varepsilon \dot{Z} &= G(t, Y, Z). \end{aligned}$$

The ordinary differential system derived from this regularization techniques is very stiff (see for instance [2] and [5]), and its numerical solution is typically more arduous. Moreover, as observed in [7], the solution $Y(t)$ can present an additional structure to be preserved during the integration, which might be destroyed by regularization. Hence, the above discussion justifies the research of a direct discretization of (1).

In the remainder of the paper we detail how a direct discretization formula is carried out to derive from (1) a general Riccati equation which can be easily solved. The numerical treatment of this Riccati equation is also discussed. Finally, some numerical tests illustrate the behavior of the proposed hybrid approach and some concluding remarks are sketched.

2 Hybrid Numerical Technique

As observed before, when the solution of (1) approaches a singular matrix at some isolated point τ, the Jacobian matrix of (1) becomes singular too. In this case, since the explicit ordinary differential equation (2) cannot be derived, a direct discretization of $G(t, Y(t), \dot{Y}(t)) = 0$ have to be considered approximating $Y(t)$ and $\dot{Y}(t)$, for instance, by a Runge Kutta scheme.

Runge Kutta methods have been originally conceived for numerical solution of ordinary differential equations. From an approximation Y_n of the solution at

the instant t_n, they construct an approximation of Y_{n+1} at $t_{n+1} = t_n + h$, where h is a constant step-size, via the formula:

$$Y_{n+1} = Y_n + h \sum_{i=1}^{s} b_i \dot{Y}_{ni} \tag{3}$$

where $\dot{Y}_{ni}$ is explicitely defined as a function of the internal stages Y_{ni} given by:

$$Y_{ni} = Y_n + h \sum_{i=1}^{s} a_{ij} \dot{Y}_{ni}, \qquad i = 1, \ldots, s, \tag{4}$$

where a_{ij}, b_i, c_i are the coefficients defining the specific method and s is the number of the stages. To extend Runge Kutta method to implicit differential equation $G(t, Y, \dot{Y}) = 0$ we define Y_{n+1} as the solution of (3), (4) and

$$G(t_n + c_i h, Y_{ni}, \dot{Y}_{ni}) = 0. \tag{5}$$

Explicit Runge Kutta methods are not directly suitable for the implicit equation in interest because, at each step, Y_{n+1} have to be determined as the solution of a linear system which can become very ill conditioned.
This can be explicitely observed applying, for instance, the forward Euler method to (1), that is

$$G(t_n, Y_n, \frac{Y_{n+1} - Y_n}{h}) = 0.$$

This provides the following system of linear equation for Y_{n+1}

$$Y_n^T Y_{n+1} - Y_n^T Y_n - hF(t_n, Y_n) = 0.$$

Unfortunately this simple method does not work when the numerical solution is approximating a singular matrix (i.e. $\det(Y_n) \simeq 0$), therefore stability considerations indicate an implicit discretization rule as the most suitable to be used.

An interesting result can be obtained applying to (1) the second order Gauss Legendre Runge Kutta scheme defined by the Butcher array

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

which applied to the implicit differential equation (1) provides

$$Y_{n+1}^T Y_{n+1} + Y_n^T Y_{n+1} - Y_{n+1}^T Y_n - Y_n^T Y_n - 2hF(t_n + \frac{h}{2}, \frac{Y_n + Y_{n+1}}{2}) = 0. \tag{6}$$

In the following we will show how to solve the nonlinear equation (6) in order to get at each step a numerical approximation of $Y(t_{n+1})$.

2.1 On the Solution of the Algebraic Riccati Equation

To avoid the nonlinearity embedded in F, we solve recursively equation (6) starting from a guess approximation $Y_{n+1}^{(0)}$, which can be chosen equals to Y_n, that is we resolve:

$$Y_{n+1}^T Y_{n+1} + Y_n^T Y_{n+1} - Y_{n+1}^T Y_n - Y_n^T Y_n - 2hF(t_n + \frac{h}{2}, \frac{Y_n + Y_{n+1}^{(0)}}{2}) = 0. \quad (7)$$

Setting $A = Y_n$ and $C = (Y_n^T Y_n + 2hF(t_n + \frac{h}{2}, \frac{Y_n + Y_{n+1}^{(0)}}{2}))$ the latter equation reads as a general algebraic Riccati equation

$$\mathcal{R}(X) = X^T X + A^T X - X^T A + C = 0. \quad (8)$$

Algebraic Riccati equation have been largely treated ([1], [10], [15]).We will assume that (7) possesses a unique solution $X \in \mathbb{R}^{n\times n}$.

To approximate the solution of the Riccati equation we apply the Newton iteration. The Riccati function $\mathcal{R}$ is clearly a mapping from $\mathbb{R}^{n\times n}$ into itself. Moreover, the first Frechet derivative of $\mathcal{R}$ at a matrix X is the linear map $\mathcal{R}'_X : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, which applied to $H \in \mathbb{R}^{n\times n}$ is given by

$$\mathcal{R}'_X(H) = H^T X + X^T H + A^T H - H^T A = H^T(X - A) + (X + A)^T H.$$

Then the Newton iteration for (8) is as follows:

- starting from X_0
- the solution of $\mathcal{R}(X) = 0$ can be obtained recursively as

$$X_{j+1} = X_j + D_j$$

with D_j solution of the Sylvester equation

$$\mathcal{R}'_X(D_j) = -\mathcal{R}(X_j) \Leftrightarrow (X_j + A)^T D_j + D_j^T (X_j - A) = -\mathcal{R}(X_j). \quad (9)$$

It appears clear that at each step we have to focus on the numerical treatment of the linear matrix equation

$$\mathcal{A}X + X^T \mathcal{B} = \mathcal{C} \quad (10)$$

with $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathbb{R}^{n\times n}$ given matrices. The first question arising in examinating the above linear equation is about its solvability. An answer to this question can be found in [16] (see also [3], [14]); we report it in the following.

Theorem 1. *Let $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathbb{R}^{n\times n}$; there exists some matrix $X \in \mathbb{R}^{n\times n}$ such that $\mathcal{A}X + X^T\mathcal{B} = \mathcal{C}$ if and only if*

$$\operatorname{rank} \begin{bmatrix} \mathcal{C} & \mathcal{A} \\ \mathcal{B} & O \end{bmatrix} = \operatorname{rank} \begin{bmatrix} O & \mathcal{A} \\ \mathcal{B} & O \end{bmatrix}.$$

(Note that this condition is equivalent to ask that there exist nonsingular matrices $P \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ such that

$$P \begin{bmatrix} \mathcal{C} & \mathcal{A} \\ \mathcal{B} & O \end{bmatrix} Q = \begin{bmatrix} O & \mathcal{A} \\ \mathcal{B} & O \end{bmatrix}.$$

Conditions on the uniqueness of the solution of (10) and on the construction of such a solution X can be obtained rewriting (10) as an ordinary linear system of n^2 equations in n^2 unknowns, using the Kronecker product $\otimes$ and the function **vec** from the space of $n \times n$ matrices to the space of n^2 vectors.

In fact, (10) is equivalent to:

$$(I \otimes \mathcal{A})\mathbf{vec}(X) + (\mathcal{B}^T \otimes I)\mathbf{vec}(X^T) = \mathbf{vec}(\mathcal{C}). \tag{11}$$

Note that there exists a permutation matrix (see [14]) which maps the vector $\mathbf{vec}(X^T)$ into $\mathbf{vec}$(X),that is:

$$\mathbf{vec}(X^T) = P(n,n)\mathbf{vec}(X)$$

being $P(n,n)$ the permutation matrix such that

$$P(n,n) = \sum_{i=1}^{n}\sum_{j=1}^{n} E_{ij} \otimes E_{ij}^T$$

where $E_{ij} \in \mathbb{R}^{n\times n}$ with elements 1 in the ij position and zeros elsewhere. Hence, rearranging (11) we obtain

$$[(I \otimes \mathcal{A}) + (\mathcal{B}^T \otimes I)P(n,n)]\mathbf{vec}(X) = \mathbf{vec}(\mathcal{C}). \tag{12}$$

Hence it follows immediately that the solution X of (10) exists and is unique if the coefficient matrix $[(I \otimes \mathcal{A}) + (\mathcal{B}^T \otimes I)P(n,n)]$ has full rank n^2.

It should be pointed out that using the Newton's method (9), the coefficient matrix is of full rank and therefore a unique solution is ensured. About the convergence of the Newton's procedure, since the starting matrix for the iteration is an approximation of the solution of the differential system at a previous step, with an appropriate choice of the step size h, the algorithm converges in a reasonable number of iteration.

The linear system involved in the solution of the algebraic Riccati equation can be solved using direct methods such as Gaussian elimination or QR-decomposition or, when the dimension increases, any iterative solver which preserves also the sparsity of the coefficient matrix. Equation (10) can be also solve applying a subspace method as suggested in [4].

The following algorithm summerizes the complete hybrid scheme described before for the solution of the implicit differential equation (1):

GL2Riccati method

Given a step size $h > 0$, an approximation Y_n of $Y(t_n)$, and a user defined tolerance *tol*, compute an approximation Y_{n+1} of the solution $Y(t_n + h)$ as follows:

(1) Set $k = 0$ and $Y_{n+1}^{(k)} = Y_n$;

(2) set $A = Y_n$ and compute $C = Y_n^T Y_n + 2hF(t_n + \frac{h}{2}, \frac{Y_n + Y_{n+1}^{(k)}}{2})$

(3) apply the Newton method to the Riccati equation

$$\mathcal{R}(X) = X^T X + A^T X - X^T A + C = 0$$

(3a) Set $j = 0$ and $X_j = Y_n$
(3b) Compute the solution D_j of the equation

$$(X_j + A)^T D_j + D_j^T (X_j + A) = -\mathcal{R}(X_j)$$

Use a direct or an iterative algorithm for linear system (12).
(3c) Compute $X_{j+1} = X_j + D_j$
(4) Set $j = j + 1$ and iterate until $\|\mathcal{R}(X_{j+1})\| \leq tol$
(5) Set $k = k + 1$ and $Y_{n+1} = Y_{n+1}^{(k)} = X_j$
(6) Repeat from (2) until:

$$\|Y_{n+1}^T Y_{n+1} - AY_{n+1}^T - A^T Y_{n+1} - Y_n^T Y_n - 2hF(t_n + \frac{h}{2}, \frac{Y_n + Y_{n+1}^{(k)}}{2})\| \leq tol$$

3 Numerical Tests

In this section we report some numerical tests in order to illustrate the behavior of the proposed approach. All the numerical results have been obtained by Matlab 6.1 codes implemented on Pentium IV 1.6GHz with 1GbRAM. We compare the GL2Riccati scheme with SVDRK2, the second order Runge Kutta scheme applied to the singular value decomposition approach illustrated in [7] and [9].

3.1 Example 1

As first example we consider the differential system

$$Y^T Y' - \begin{pmatrix} 0 & -\frac{\delta}{2} \\ -\frac{\delta}{2} & 0 \end{pmatrix} = 0, \qquad Y(0) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \tag{13}$$

with $\delta \neq 0$, which has solution given by

$$Y(t) = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{1+\delta t} & -\sqrt{1+\delta t} \\ \sqrt{1-\delta t} & \sqrt{1-\delta t} \end{pmatrix},$$

existing in the neighborhood $(-1/\delta, 1/\delta)$ of $t_0 = 0$. In this case the matrix function $G(Y)$ exists and is a continuous and Lipschitz matrix function only in a neighborhood of $Y(0)$. We solve a problem with $\delta = 1/2$. Figure 1 depicts the behavior of the global error on the interval $[1, 2]$ of GL2Riccati scheme (solid line) and SVDRK2 (dotted line). The two methods have been applied with the same step size and show the same performance inside the chosen interval.

3.2 Example 2

As second example we consider the differential system

$$Y^T \dot{Y} - \begin{pmatrix} -\sin(t)\cos(t) & \cos(t) \\ -t\sin(t) & t \end{pmatrix} = 0$$

whose matrix solution

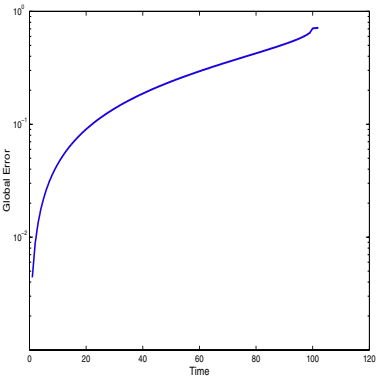


Fig. 1. Behavior of the global error

$$Y(t) = \begin{pmatrix} \cos(t) & t \\ 0 & 1 \end{pmatrix}$$

defined for all t, but periodically, at each multiple of the point $\tau_k = \pi/2$ becomes a singular matrix. We integrate this system in the interval $[\pi/4, \pi/2]$, with a step size $h = 0.01$. Figure 2 provides the semilog plot of the global error on the integration interval of GL2Riccati scheme (solid line) and SVDRK2 (dotted line). In this case the GL2Riccati approach shows a reduced global error with respect that of SVDRK2, moreover while this latter method blows up at the singularity, the proposed procedure is able to integrate the system at the singularity even if an order reduction can occur.

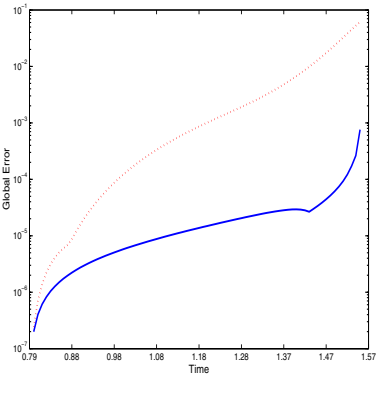


Fig. 2. Behavior of the global error

4 Conclusion

Implicit differential equations play a key role in many applications. In this paper we have considered a particular class of IDE whose solution matrix not guaranteed to be nonsingular for all t. We have demonstrated the relationship between the numerical solution of this class of IDE and a general Riccati equation, suggesting a hybrid techniques which provides a numerical approximation of the solution also at escape points, moreover it should be highlighted that since the proposed procedure is based on a second order Gauss Legendre Runge Kutta scheme it is able to preserve any eventual quadratic structure of the theoretical solution.

References

1. Ammar G., Benner P., and Mehrmann V., A multishift algorithm for the numerical solution of algebraic Riccati equations, ETNA **1** (1993) 33-48.
2. Ascher U.M., Petzold L.R., Computer methods for Ordinary Differential Equations and Differential-Algebraic Equations SIAM 1998.
3. Baksalary J.K., Kala R., The matrix equation $AX - YB = C$. Linear Algebra Appl. **25** (1979) 41-43.
4. Brands J., Computing tall skinny solutions of $AX - XB = C$. Mathematics and Computers in Simulation **61** (2003) 385-397.
5. Brenan W.E., Campbell S.L., Petzold L.R., Numerical Solution of Inital-value Problems in Differential Algebraic Equations. North-Holland: New York, 1989.
6. Chu, M.T., Inverse Eigenvalue Problems, SIAM Review **40** (1998) 1-39.
7. Del Buono N., Lopez L., Numerical integration of a class of ordinary differential equations on the general linear group of matrices. Numerical Algorithms **34**, (2003) 271-282.
8. Del Buono N., Lopez L., Geometric integration on manifold of square oblique rotation matrices. SIAM J. Matrix Anal. Appl. **23**(4) (2002) 974-989.
9. Del Buono N., Lopez L., Numerical solution of matrix ordinary differential equations with singular solutions. Tech. Report, Department of Mathematics, University of Bari, Italy, (2004).
10. Guo C.H., Laub A.J., On a Newton-like method for solving algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **21**(2) (2000) 694-698.
11. Hairer E., Lubich C., Roche M., The numerical solution of differential algebraic systems by Runge Kutta methods, Lecture Notes in Mathematics, **1409**, Springer-Verlag 1989
12. Helmke U., Moore J.B., Optimization and dynamical systems. Springer-Verlag: London, 1994.
13. Horn R.A., Johnson, C.A., Matrix Analysis. Cambridge University Press: New York, 1985.
14. Horn R.A., Johnson, C.A., Topics in Matrix Analysis. Cambridge University Press: New York, 1991.
15. Lancaster P., Rodman L., Algebraic Riccati Equations. Clarenton Press: Oxford, 1995.
16. Roth W.E., The equation $AX - YB = C$ and $AX - XB = C$ in matrices. Proc. Am. Soc. **3**, (1952) 392-396.
17. Trendafilov N.T., Lippert R.A.,The Multi-Mode Procrustes problem. Linear Algebra and Applications **349**(1-3), (2002) 245-264.

A Continuous Approach for the Computation of the Hyperbolic Singular Value Decomposition

T. Politi

Dipartimento di Matematica, Politecnico di Bari,
Via Amendola 126/B, I-70126 Bari (Italy).
politi@poliba.it

Abstract. In this paper a continuous approach based on the Projected Gradient Flow technique is presented in order to find a generalization of the Singular Value Decomposition (SVD) of a rectangular matrix called Hyperbolic SVD. If A is a $m \times n$ real matrix with full column rank and if G is a $n \times n$ diagonal sign matrix, i.e. $g_{ii} = \pm 1$, the Hyperbolic Singular Value Decomposition of the pair (A, G) is defined as $A = U\Sigma V$, where U is orthogonal, Σ is diagonal with positive entries and V is hypernormal (or G-orthogonal), i.e. $V^T GV = G$. In this work we use a continuous approach based on the projected gradient technique obtaining two differential systems, the first one evolving on group of orthogonal matrices and the second on the quadratic group related to G. A numerical test is reported in order to show the effectiveness of the approach.

1 Introduction

Recently there has been a growing interest in numerical analysis community in the exploitation of the concept of structure associated to matrices. Examples of these structures are orthogonality, simplecticity, skew-symmetry and so on. Some of these properties are related to group structures (and sometimes to Lie-group structures) or to algebra structure (for example the skew-symmetry) or, more simplicity to algebraic properties (it is the case of the obliqueness, i.e. matrices Y such that $\mathrm{diag}(Y^T Y) = I$). In particular a great attention has been devoted to the numerical solution of differential equations evolving on matrix groups (see [7] for quadratic groups, [6] for Lie groups, [3,4] for orthogonal matrices). The experience in the effective solution of differential equations on matrix groups can be used also to solve some linear algebra problems, such as the computation of factorizations of time depending matrices (see [13]) or the computation of singular value decomposition factors as limit point of continuous flows (see [2]). In this paper we consider this last problem in the case of the Hyperbolic Singular Value Decomposition. The work is organized as follows: in Section 2 we define the Hyperbolic Singular Value Decomposition and recall some important features and applications, in Section 3 we use the projected gradient technique in order to obtain two differential flows having, respectively, the unitary and the hypernormal factors as limit point. Finally in the Section 4 a numerical test is described in order to show the effectiveness of the differential approach.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 467–474, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 The Hyperbolic Singular Value Decomposition

In this section we define the Hyperbolic SVD of a $m \times n$ real matrix A with a couple of its applications, but first we give some important definitions.

Definition 1. *Let G be a $m \times m$ diagonal matrix with entries ± 1, then a matrix $U \in \mathbb{R}^{m \times m}$ is said* hypernormal *if $U^T G U = G$.*

In [5] the hypernormal matrices are called $G-$orthogonal.

Definition 2. *If G is a $m \times m$ diagonal matrix with entries ± 1, then a matrix $U \in \mathbb{R}^{m \times m}$ is said* hyperexchange *if*

$$U^T G U = \widehat{G} \tag{1}$$

where $\widehat{G}$ is another real diagonal matrix with entries ± 1.

It is easy to observe that there is a strict relation between hypernormal and hyperexchange matrices. In fact if V is an hyperexchange matrix there exists a permutation matrix P such that $W = VP$ is hypernormal. In fact, from (1) since G and $\widehat{G}$ have the same inertia (i.e. the same number of $+1$ and -1) and there exists a permutation matrix P such that $PGP^T = \widehat{G}$, hence

$$V^T G V = \widehat{G} = PGP^T \qquad \Rightarrow \qquad (VP)^T G (VP) = G$$

and $W = VP$ is hypernormal. Considering the quadratic group related to matrix G, i.e. the set

$$\mathcal{H}_G = \{Y \in \mathbb{R}^{n \times n} \mid \det(Y) \neq 0,\ YGY^T = G\}$$

we observe that it coincides with the set of hypernormal matrices. Moreover we shall denote by $\mathcal{O}(n)$ as the set of orthogonal matrices of order n. The following result states the existence of the Hyperbolic SVD (see [10]).

Theorem 1. *Let $A \in \mathbb{R}^{m \times n}$ and G be a square diagonal real matrix of order n with entries equal to ± 1, if the rank of AGA^T is equal to $\min\{m, n\}$ then there exist a $m \times m$ unitary matrix U, an hypernormal $n \times n$ matrix V and a $m \times n$ diagonal matrix Σ with positive entries such that*

$$A = U \Sigma V. \tag{2}$$

We observe that in [1,8] the matrix V is requested to be an hyperexchange matrix but we have already shown that the two sets of matrices are strictly related. The Hyperbolic SVD has some interesting applications: for example in [12] it has been used to solve the symmetric indefinite eigenvalues problem $A\mathbf{x} = \lambda \mathbf{x}$, where A is a square matrix. The algorithm proposed consists of two steps:

- A is decomposed by symmetric indefinite factorization $A = LGL^T$ (see [9]), being G a diagonal matrix with entries equal to ± 1;

• the second step is the computation of the Hyperbolic SVD of L, i.e. $L = U\Sigma V$. Since

$$A = LGL^T = U\Sigma VGV^T\Sigma U^T = UG\Sigma^2 U^T,$$

the eigenvalues of A are $\lambda_i = g_{ii}\sigma_i^2$, while the columns of U are the corresponding eigenvectors. If $G = \mathrm{diag}(I_k, -I_{m-k})$ and we divide the matrix A in block form $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, with $A_1 \in \mathbb{R}^{m\times k}$ and $A_2 \in \mathbb{R}^{m\times(m-k)}$, then the Hyperbolic SVD could be used to find the eigenvalues of the matrix

$$H = AGA^T = A_1A_1^T - A_2A_2^T = UG\Sigma^2U^T$$

without forming explicitly the matrix.

3 A Gradient Flow Approach

The aim of this section is to use the projected gradient in order to construct a continuous steepest descent flow on the manifold $\mathcal{O}(m) \times \mathcal{H}_G$. To apply this approach it is necessary to transform the problem into a minimization one with manifold constraints. From (2) it is

$$\Sigma = U^{-1}AV^{-1} = U^TAGV^TG.$$

Since Σ must be diagonal it is natural to minimize the distance between the matrix U^TAGV^TG and the manifold of real diagonal matrices and to choose as projection the operator giving the main diagonal of the matrix. Defining the function

$$F(U,V) = \left\|U^TAGV^TG - \mathrm{diag}(U^TAGV^TG)\right\|_F^2 \tag{3}$$

where $U \in \mathcal{O}(m)$ and $V \in \mathcal{H}_G$ and, following the same approach as in [2] for the usual SVD, we have to solve the constrained problem:

$$\begin{aligned} &\text{Minimize} \quad F(U,V) = \langle U^TAGV^TG - P(A), U^TAGV^TG - P(A)\rangle \\ &\text{subject to} \quad U^TU = I_n, \qquad V^TGV = G \end{aligned}$$

where $P(A) = \mathrm{diag}(U^TAGV^TG)$, and $\langle A, B\rangle$ denotes the Frobenius inner product of two matrices $A, B \in \mathbb{R}^{m\times n}$ defined as

$$\langle A, B\rangle = \mathrm{trace}\left(AB^T\right) = \sum_{i,j} a_{ij}b_{ij}. \tag{4}$$

As seen in the previous section the set of hypernormal matrices is the quadratic group related to matrix G, that is a particular Lie group and having as Lie algebra the set

$$\mathfrak{h}_G = \{A \in \mathbb{R}^{n\times n} \mid AG + GA^T = 0\}.$$

We observe that if G is the identity matrix of order n then $\mathfrak{h}_G$ is the set of real skew-symmetric matrices. A property of the Lie algebra is that it is the tangent

space of $\mathcal{H}_G$ at any hypernormal matrix Q is defined as $\mathcal{T}_Q\mathcal{H}_G = Q\mathfrak{h}_G$. If $K \in \mathfrak{h}_G$ and S is a $G-$self adjoint matrix (i.e. $S = GS^TG$) it is

$$\text{trace}(SK) = 0$$

then S is orthogonal to any K with respect to metric (4). Then the normal space $\mathcal{N}_Q\mathcal{H}_G$ of $\mathcal{H}_G$ at any Q, hypernormal matrix, can be expressed as $\mathcal{N}_Q\mathcal{H}_G = Q\mathcal{S}_G$, where $\mathcal{S}_G$ is the set of $G-$self adjoint matrices. Function $F(U,V)$ is defined on the Cartesian product $\mathcal{O}(m) \times \mathcal{H}_G$. Taking $H \in \mathbb{R}^{m\times m}$ and $K \in \mathbb{R}^{n\times n}$, the Fréchet derivative of F acting on $(H,K) \in \mathbb{R}^{m\times m} \times \mathbb{R}^{n\times n}$ can be considered as

$$F'(U,V).(H,K) = \frac{\partial F}{\partial U}.H + \frac{\partial F}{\partial V}.K, \tag{5}$$

where $\Lambda.\eta$ denotes the result of the action by the linear operator Λ on η. We now calculate each action in (5) as follows. First

$$\begin{aligned}
\frac{\partial F}{\partial V}.K &= \langle U^TAGK^TG - \text{diag}(U^TAGK^TG), U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle = \\
&= \langle U^TAGK^TG, U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle + \\
&\quad -\langle \text{diag}(U^TAGK^TG), U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle = \\
&= \langle U^TAGK^TG, U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle = \\
&= \langle U^TAGK^T, U^TAGV^T - \text{diag}(U^TAGV^T)\rangle = \\
&= \langle VGA^TU - \text{diag}(VGA^TU), KGA^TU\rangle = \\
&= \langle (VGA^TU - \text{diag}(VGA^TU))U^TAG, K\rangle.
\end{aligned}$$

It follows, from the Riesz representation theorem, that the partial gradient with respect to the Frobenius inner product can be represented as

$$\frac{\partial F}{\partial V} = (VGA^TU - \text{diag}(VGA^TU))U^TAG.$$

Then

$$\begin{aligned}
\frac{\partial F}{\partial U}.H &= \langle H^TAGV^TG - \text{diag}(H^TAGV^TG), U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle = \\
&= \langle H^TAGV^TG, U^TAGV^TG - \text{diag}(U^TAGV^TG)\rangle = \\
&= \langle H^TAGV^T, U^TAGV^T - \text{diag}(U^TAGV^T)\rangle = \\
&= \langle H^T, (U^TAGV^T - \text{diag}(U^TAGV^T))VGA^T\rangle = \\
&= \langle AGV^T(VGA^TU - \text{diag}(VGA^TU)), H\rangle,
\end{aligned}$$

and

$$\frac{\partial F}{\partial U} = AGV^T(VGA^TU - \text{diag}(VGA^TU)).$$

The gradient $\nabla F(U,V)$ now must be projected in the Cartesian product of the tangent spaces of the two manifolds, i.e. $\mathcal{T}_Q\mathcal{O}(m) \times \mathcal{T}_Q\mathcal{H}_G$. The tangent space $\mathcal{T}_Q\mathcal{O}(m)$ is $Q\mathfrak{h}$, where $\mathfrak{h}$ is the set of real skew-symmetric matrices of order m. It is well known that any general matrix $X \in \mathbb{R}^{n\times n}$ can be uniquely splitted as

$$X = Q\frac{Q^TX + X^TQ}{2} + Q\frac{Q^TX - X^TQ}{2},$$

then the projection $\mathcal{P}_{\mathcal{O}(m)}(X)$ onto the tangent space $\mathcal{T}_Q\mathcal{O}(m)$ is given by

$$\mathcal{P}_{\mathcal{O}(m)}(X) = Q\frac{Q^TX - X^TQ}{2}.$$

Similarly it is easy to verify that any matrix $X \in \mathbb{R}^{n\times n}$ has a unique splitting

$$X = Q\left[\frac{1}{2}(GQ^TGX - GX^TGQ) + \frac{1}{2}(GQ^TGX + GX^TGQ)\right]$$

where $Q \in \mathcal{H}_G$, $GQ^TGX - GX^TGQ \in \mathfrak{h}_G$ and $GQ^TGX + GX^TGQ \in \mathcal{S}_G$. The projection of the gradient of $F(U,V)$ into the tangent space $\mathcal{T}_Q\mathcal{H}_G$ is

$$\mathcal{P}_{\mathcal{H}_G}(X) = Q\frac{GQ^TGX - GX^TGQ}{2}.$$

Hence the differential systems that must be solved are

$$\frac{dU}{dt} = -\mathcal{P}_{\mathcal{O}(m)}\left(\frac{\partial F}{\partial U}\right) \tag{6}$$

$$\frac{dV}{dt} = -\mathcal{P}_{\mathcal{H}_G}\left(\frac{\partial F}{\partial V}\right). \tag{7}$$

4 Numerical Tests

In this section we show a numerical example obtained applying the projected gradient flow approach described previously. We consider a 5×3 real random matrix A having as main diagonal of Σ the vector $(0.5, 1, 1.5)$, and taking $G = \text{diag}(I_2, -1)$. The differential systems (6)-(7) have been solved numerically in the interval $[0, 30]$. In Figure 1 we show the behaviour of the objective function (3), while in Figure 2 we show the behaviour of the diagonal entries of the matrix $X_n = U_n^TAGV_n^TG$. Matrices U_n and V_n are the numerical approximations of the solutions $U(t)$ and $V(t)$, computed at $t = t_n$, obtained first integrating the differential systems (6)-(7) with the MatLab ode routine `ode113` and then projecting the numerical solutions on the manifolds. For the orthogonal flow (6) the projection has been computed taking the orthogonal factor of the QR decomposition (see [3] for more details), while for the flow (7) the hypernormal factor of hyperbolic QR decomposition has been taken (see [11]). In [3] has been

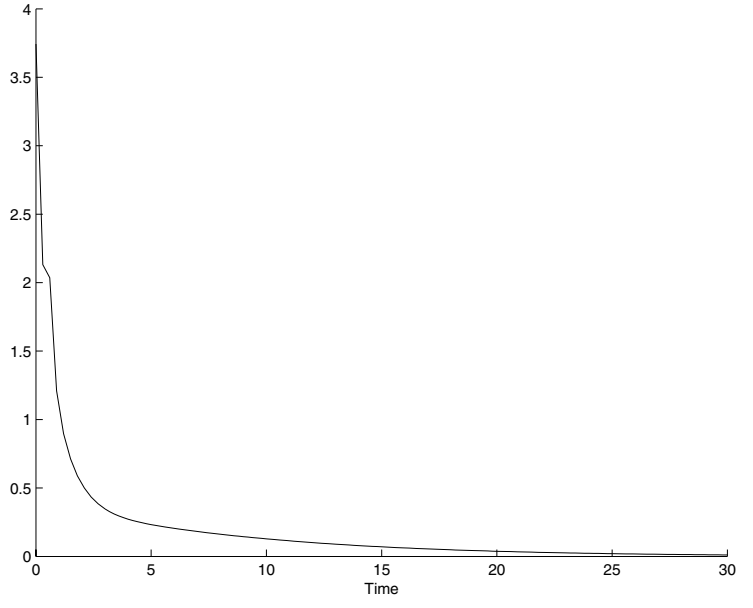


Fig. 1. Evolution of the objective function.

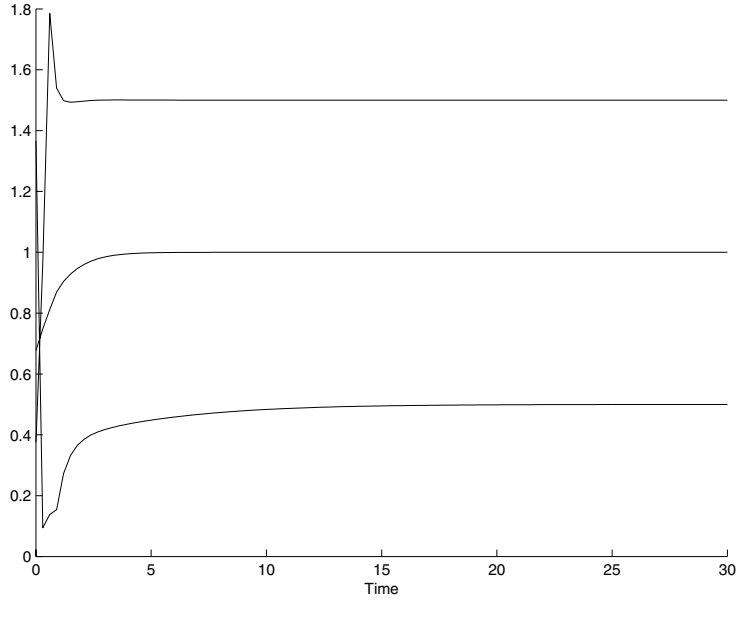


Fig. 2. Diagonal elements of matrices X_n.

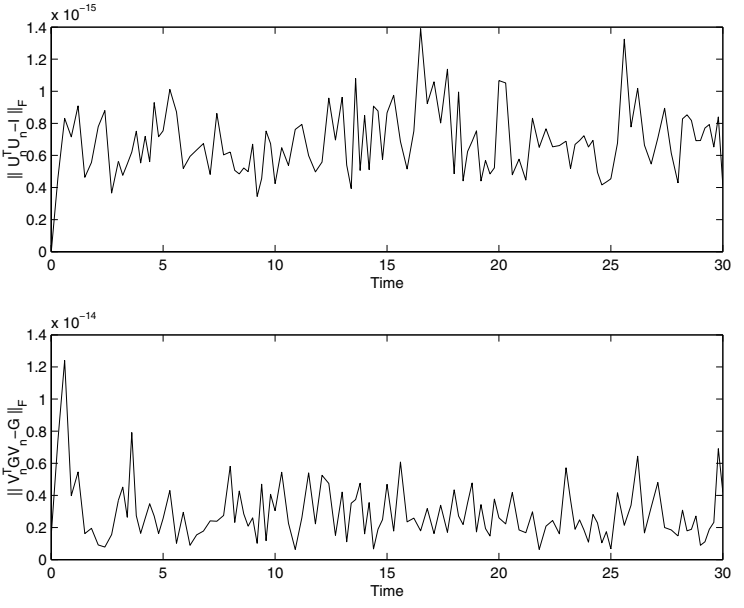


Fig. 3. Errors in the orthogonal and in the quadratic group related to G.

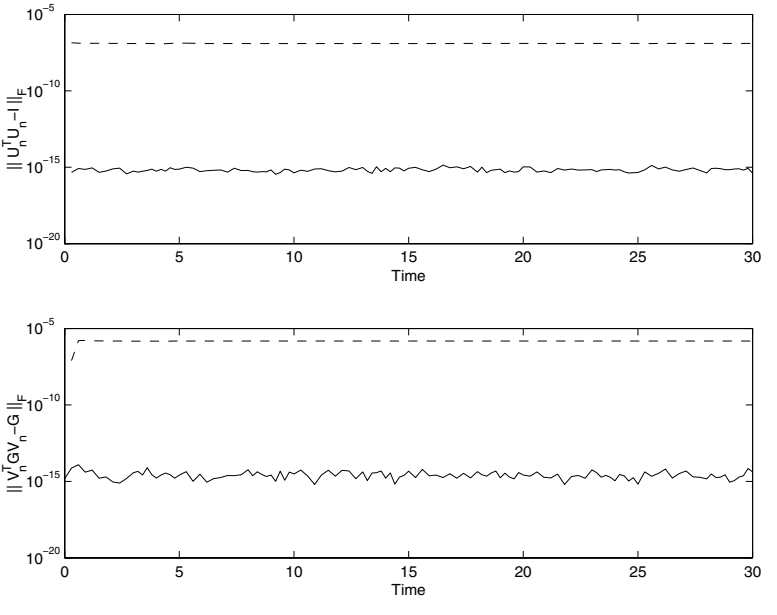


Fig. 4.

proved that the order of the ODE integrator is preserved. The initial conditions for (6) and (7) are random matrices in the manifolds. We observe that the behaviour of the diagonal elements of X_n approaches the theoretical values already when $t \simeq 8$. In Figure 3 we show the departure from the orthogonal manifold and the quadratic group related to G for U_n and V_n respectively, computed as $\|U_n^T U_n - I_n\|_F$ and $\|V_n^T G V_n - G\|_F$. Finally in Figure 4 we show that the use of the projection of the numerical solution on the manifolds is necessary. In the picture the solid lines denote the manifold errors of the numerical solutions for $U(t)$ and $V(t)$ using the projection while the dashed lines denote the errors given by the MatLab integrator. The solution is computed by routine `ode113` with a relative tolerance set to 10^{-6} but it departes from the manifold very soon.

References

1. Bojanczyk A.W., Onn R., Steinhardt A.O.: Existence of the Hyperbolic Singular Value Decomposition. Lin. Alg. Appl. **185** (1993) 21–30
2. Chu M.T., Driessel K.R.: The projected gradient method for least square matrix approximations with spectral constraints. SIAM J. Numer. Anal. **27** (1990) 1050–1060
3. Dieci L., Russell R.D., Van Vleck E.S.: Unitary integrators and applications to continuous orthonormalization techniques. SIAM J. Numer. Anal. **31** (1994) 261–281
4. Diele F., Lopez L., Peluso R.: The Cayley transform in the numerical solution of unitary differential systems. Adv. Comput. Math. **8** (1998) 317–334
5. Higham N.: $J-$Orthogonal matrices: properties and generation. SIAM Rev. **45** (3) (2003) 504–519
6. Iserles A., Munthe-Kaas H., Nørsett S.P., Zanna A.: Lie-group methods. Acta Numerica **9** (2000) 215–365
7. Lopez L., Politi T.: Applications of the Cayley approach in the numerical solution of matrix differential systems on quadratic groups. Appl. Num. Math. **36** (2001) 35–55
8. Onn R., Steinhardt A.O., Bojanczyk A.W.: The Hyperbolic Singular Value Decomposition and Applications. IEEE Trans. Sign. Proc. **39** (7) (1991) 1575–1588
9. Slapničar I.: Componentwise analysis of direct factorization of real symmetric and Hermitian matrices. Lin. Alg. Appl. **272** (1998) 227–275
10. Slapničar I.: Highly accurate symmetric eigenvalue decomposition and Hyperbolic SVD. Lin. Alg. Appl. **358** (2003) 387–424
11. Stewart M., Stewart G.W.: On hyperbolic triangularization: stability and pivoting. SIAM J. Matrix Anal. Appl. **19** (4) (1998) 847–860.
12. Veselić K.: A Jacobi eigenreduction algorithm for definite matrix pairs. Numer. Math. **64** (1993) 241–269
13. Wright K.: Differential equations for the analytic singular value decomposition of a matrix. Numer. Math. **63** (1992) 283–295

Using P-GRADE for Monte Carlo Computations in a Distributed Environment

Vassil N. Alexandrov[1], Ashish Thandavan[1], and Péter Kacsuk[2]

[1] Department of Computer Science, University of Reading, Reading, UK
[2] MTA SZTAKI Research Institute, Budapest, Hungary

Abstract. Computations involving Monte Carlo methods are, very often, easily and efficiently parallelized. P-GRADE is a parallel application development environment which provides an integrated set of programming tools for development of general message-passing applications to run in heterogeneous computing environments or supercomputers. In this paper, we show how Monte Carlo algorithms for solving Systems of Linear Equations and Matrix Inversion can easily be parallelized using P-GRADE.

1 Introduction

The problem of inverting a real $n \times n$ matrix (MI) and solving system of linear algebraic equations (SLAE) is of an unquestionable importance in many scientific and engineering applications: e.g. communication, stochastic modelling, and many physical problems involving partial differential equations. For example, the direct parallel methods of solution for systems with dense matrices require $O(n^3/p)$ steps when the usual elimination schemes (e.g. non-pivoting Gaussian elimination, Gauss-Jordan methods) are employed [4].

We concentrate on Monte Carlo methods for MI and solving SLAEs, since, firstly, only $O(NL)$ steps are required to find an element of the inverse matrix, where N is the number of chains and L is an estimate of the chain length in the stochastic process, which are independent of matrix size n and secondly, these stochastic methods are inherently parallel.

Several authors have proposed different coarse grained Monte Carlo parallel algorithms for MI and SLAE [6,7,8,9,10]. In this paper, we investigate how Monte Carlo can be used for diagonally dominant and some general matrices via a general splitting and how efficient mixed (stochastic/deterministic) parallel algorithms can be derived for obtaining an accurate inversion of a given non-singular matrix A. We employ either uniform Monte Carlo (UM) or almost optimal Monte Carlo (MAO) methods [6,7,8,9,10].

Note that the algorithms are built under the requirement $\|T\| < 1$. Therefore to develop efficient methods we need to be able to solve problems with matrix norms greater than one. Thus we developed a spectrum of algorithms for MI and solving SLAEs ranging from special cases to the general case. Parallel MC methods for SLAEs based on Monte Carlo Jacobi iteration have been presented

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 475–482, 2004.
© Springer-Verlag Berlin Heidelberg 2004

by Dimov [10]. Parallel Monte Carlo methods using minimum Makrov Chains and minimum communications are presented in [1]. Most of the above approaches are based on the idea of balancing the stochastic and systematic errors [10]. In this paper we have presented a hybrid algorithms for MI and solving SLAEs by combining two ideas: iterative Monte Carlo methods based on the Jacobi iteration and deterministic procedures for improving the accuracy of the MI or the solution vector of SLAEs in Sections 2 and 3. Further the parallel approach using P-GRADE and some numerical experiments are presented in Section 4 and 5 respectively.

2 Monte Carlo and Matrix Computation

Assume that the system of linear algebraic equations (SLAE) is presented in the form:

$$Ax = b \tag{1}$$

where A is a real square $n \times n$ matrix, $x = (x_1, x_2, ..., x_n)^t$ is a $1 \times n$ solution vector and $b = (b_1, b_2, ..., b_n)^t$.

Assume the general case $\|A\| > 1$. We consider the splitting $A = D - C$, where off-diagonal elements of D are the same as those of A, and the diagonal elements of D are defined as $d_{ii} = a_{ii} + \gamma_i ||A||$, choosing in most cases $\gamma_i > 1, i = 1, 2, ..., n$. We further consider $D = B - B_1$ where B is the diagonal matrix of D, e.g. $b_{ii} = d_{ii} i = 1, 2, ..., n$. As shown in [1] we could transform the system (1) to

$$x = Tx + f \tag{2}$$

where $T = D^{-1}C$ and $f = D^{-1}b$. The multipliers γ_i are chosen so that, if it is possible, they reduce the norm of T to be less than 1. In the general case we consider finding D^{-1} using MC and after that finding A^{-1}. Then, if required, the solution vector is found by $x = A^{-1}b$.

Consider first the stochastic approach. Assume that $\|T\| < 1$ and that the system is transformed to its iterative form (2). Consider the Markov chain given by:

$$s_0 \rightarrow s_1 \rightarrow \cdots \rightarrow s_k, \tag{3}$$

where the $s_i, i = 1, 2, \cdots, k$, belongs to the state space $S = \{1, 2, \cdots, n\}$. Then for $\alpha, \beta \in S, p_0(\alpha) = p(s_0 = \alpha)$ is the probability that the Markov chain starts at state α and $p(s_{j+1} = \beta | s_j = \alpha) = p_{\alpha\beta}$ is the transition probability from state α to state β. The set of all probabilities $p_{\alpha\beta}$ defines a transition probability matrix $P = \{p_{\alpha\beta}\}_{\alpha,\beta=1}^{n}$ [3,8,9]. We say that the distribution $(p_1, \cdots, p_n)^t$ is acceptable for a given vector g, and that the distribution $p_{\alpha\beta}$ is acceptable for matrix T, if $p_\alpha > 0$ when $g_\alpha \neq 0$, and $p_\alpha \geq 0$, when $g_\alpha = 0$, and $p_{\alpha\beta} > 0$ when $T_{\alpha\beta} \neq 0$, and $p_{\alpha\beta} \geq 0$ when $T_{\alpha\beta} = 0$ respectively. We assume $\sum_{\beta=1}^{n} p_{\alpha\beta} = 1$, for all $\alpha = 1, 2, \cdots, n$. Generally, we define

$$W_0 = 1, W_j = W_{j-1} \frac{T_{s_{j-1}s_j}}{p_{s_{j-1}s_{j-1}}} \tag{4}$$

for $j = 1, 2, \cdots, n$.

Consider now the random variable $\theta[g] = \frac{g_{s_0}}{p_{s_0}} \sum_{i=1}^{\infty} W_i f_{s_i}$. We use the following notation for the partial sum:

$$\theta_i[g] = \frac{g_{s_0}}{p_{s_0}} \sum_{j=0}^{i} W_j f_{s_j}. \tag{5}$$

Under condition $\|T\| < 1$, the corresponding Neumann series converges for any given f, and $E\theta_i[g]$ tends to (g, x) as $i \to \infty$. Thus, $\theta_i[g]$ can be considered as an estimate of (g, x) for i sufficiently large. To find an arbitrary component of the solution, for example, the r^{th} component of x, we should choose, $g = e(r) = (\underbrace{0, ..., 1}_{r}, 0, ..., 0)$ such that $e(r)_\alpha = \delta_{r\alpha} = \begin{cases} 1 & \text{if} \quad r = \alpha \\ 0 & otherwise \end{cases}$

It follows that $(g, x) = \sum_{\alpha=1}^{n} e(r)_\alpha x_\alpha = x_r$.

The corresponding Monte Carlo method is given by:

$$x_r = \hat{\Theta} = \frac{1}{N} \sum_{s=1}^{N} \theta_i[e(r)]_s,$$

where N is the number of chains and $\theta_i[e(r)]_s$ is the approximate value of x_r in the s^{th} chain. It means that using Monte Carlo method, we can estimate only one, few or all elements of the solution vector. We consider Monte Carlo with uniform transition probability (UM) $p_{\alpha\beta} = \frac{1}{n}$ and Almost optimal Monte Carlo method (MAO) with $p_{\alpha\beta} = \frac{|T_{\alpha\beta}|}{\sum_{\beta=1}^{n} |T_{\alpha\beta}|}$, where $\alpha, \beta = 1, 2, \ldots, n$. Monte Carlo MI is obtained in a similar way [3].

To find the inverse $A^{-1} = C = \{c_{rr'}\}_{r,r'=1}^{n}$ of some matrix A, we must first compute the elements of matrix $M = I - A$, where I is the identity matrix. Clearly, the inverse matrix is given by $C = \sum_{i=0}^{\infty} M^I$ which converges if $\|M\| < 1$. To estimate the element $c_{rr'}$ of the inverse matrix C, we let the vector f be the following unit vector $f_{r'} = e(r')$.

We then can use the following Monte Carlo method for calculating elements of the inverse matrix C:

$$c_{rr'} \approx \frac{1}{N} \sum_{s=1}^{N} \left[\sum_{(j|s_j = r')} W_j \right], \tag{6}$$

where $(j|s_j = r')$ means that only W_j for which $s_j = r'$ are included in the sum.

The *probable error* of the method, is defined as $r_N = 0.6745\sqrt{D\theta/N}$, where $P\{|\bar{\theta} - E(\theta)| < r_N\} \approx 1/2 \approx P\{|\bar{\theta} - E(\theta)| > r_N\}$, if we have N independent realizations of random variable (r.v.) θ with mathematical expectation $E\theta$ and average $\bar{\theta}$ [5].

3 The Hybrid MC Algorithm

Consider now the algorithm which can be used for the inversion of a general non-singular matrix A. Note that in some cases to obtain a very accurate inversion of matrix D some filter procedures can be applied.

Algorithm: Finding A^{-1}.

1. **Initial data:** Input matrix A, parameters γ and ϵ.
2. **Preprocessing:**
   - 2.1 **Split** $A = D - (D - A)$, where D is a diagonally dominant matrix.
   - 2.2 **Set** $D = B - B_1$ where B is a diagonal matrix $b_{ii} = d_{ii}$ $i = 1, 2, ..., n$.
   - 2.3 **Compute** the matrix $T = B^{-1}B_1$.
   - 2.4 **Compute** $||T||$, the Number of Markov Chains $N = (\frac{0.6745}{\epsilon} \cdot \frac{1}{(1-||T||}))^2$.
3. **For** i=1 to n;
   - 3.1 **For** j=1 to j=N;

     Markov Chain Monte Carlo Computation:
     - 3.1.1 **Set** $t_k = 0$(stopping rule), $W_0 = 1$, $SUM[i] = 0$ and $Point = i$.
     - 3.1.2 **Generate** an uniformly distributed random number $nextpoint$.
     - 3.1.3 **If** $T[point][netxpoint]! = 0$.

     LOOP
     - 3.1.3.1 **Compute** $W_j = W_{j-1} \frac{T[point][netxpoint]}{P[point][netxpoint]}$.
     - 3.1.3.2 **Set** $Point = nextpoint$ and $SUM[i] = SUM[i] + W_j$.
     - 3.1.3.3 **If** $|W_j| < \gamma$, $t_k = t_k + 1$
     - 3.1.3.4 **If** $t_k \geq n$, end LOOP.

     - 3.1.4 **End If**
     - 3.1.5 **Else** go to step 3.1.2.
   - 3.2 **End of loop j**.
   - 3.3 **Compute** the average of results.
4. **End of loop i**.
5. **Obtain** The matrix $V = (I - T)^{-1}$.
6. **Therefore** $D^{-1} = VB^{-1}$.
7. **Compute** the MC inversion $D^{-1} = B(I - T)^{-1}$.
8. **Set** $D_0 = D^{-1}$ (approximate inversion) and $R_0 = I - DD_0$.
9. **use filter procedure** $R_i = I - DD_i$, $D_i = D_{i-1}(I + R_{i-1})$, $i = 1, 2, ..., m$, where $m \leq k$.
10. **Consider the accurate inversion of D** by step 9 given by $D_0 = D_k$.
11. **Compute** $S = D - A$ where S can be any matrix with all non-zero elements in diagonal and all of its off-diagonal elements are zero.
12. **Main function** for obtaining the inversion of A based on D^{-1} step 9:
    - 12.1 **Compute** the matrices $S_i, i = 1, 2, ..., k$, where each S_i has just one element of matrix S.
    - 12.2 **Set** $A_0 = D_0$ and $A_k = A + S$
    - 12.3 **Apply** $A_k^{-1} = A_{k+1}^{-1} + \frac{A_{k+1}^{-1} S_{i+1} A_{k+1}^{-1}}{1 - trace(A_{k+1}^{-1} S_{i+1})}$, $i = k-1, k-2, ..., 1, 0$.
13. **Print**the inversion of matrix A.
14. **End** of algorithm.

The basic idea is to use MC to find the approximate inverse of matrix D, refine the inverse (filter) and find A^{-1}. According to the general definition of a regular splitting [2], if A, M and N are three given matrices and $A = M - N$, then the pair of matrices M, N are called regular splitting of A, if M is nonsingular and M^{-1} and N are non-negative.

Therefore, let A be a nonsingular diagonal dominant matrix. If we find a regular splitting of A such that $A = D - C$, the SLAE $x^{(k+1)} = Tx^{(k)} + f$, where $T = D^{-1}C$, and $f = D^{-1}b$ converges to the unique solution x^* if and only if $\|T\| < 1$ [2].

4 Parallelisation Using P-GRADE

The Parallel GRaphical Application Development Environment is, as the name suggests, a parallel programming environment which supports the whole life-cycle of parallel program development. All the stages from initial design to execution and debugging to performance visualisation and tuning of the parallel application are supported by P-GRADE.

It uses a combination of graphics and program statements to describe what the application does. The execution environment can be a varied one, ranging from clusters of workstations to supercomputers. A parallel application typically consists of two or more processes which communicate via messages. Two popular message passing libraries used for parallel programming are Parallel Virtual Machine (PVM) and Message Passing Interface (MPI). P-GRADE allows the developer to choose which library he / she wishes to use without needing to know the syntax of the underlying message passing system. All the messages are generated automatically from the graphics. Compilation and distribution of the executables are performed automatically in the heterogeneous environment. An integrated debugger allows the program to be methodically debugged during runtime and monitoring and visualisation tools provide performance information.

4.1 Tools in P-GRADE

P-GRADE consists of a few main components. The application developer uses the GRED [12] editor to design and construct the parallel program. The program flow is described by a special graphical programming language called GRAPNEL. The GRP2C precompiler compiles the graphical information into C code with PVM or MPI. It also creates additional makefiles which are used by the UNIX make utility to build the application executables.

Once the executables have been created, the parallel program can be executed either in debugging mode or in trace mode. In the debugging mode, the execution of the program is under the control of the DIWIDE [11] distributed debugger which provides options to create breakpoints, perform step-by-step execution, animation of the flow of control, etc. Once the program has been successfully debugged, it can be executed in trace mode. GRM [13], a distributed

monitoring tool, is responsible for generating a file containing trace events defined by the developer. The collected data can then be visualised by the PROVE [13] visualization tool, which is invaluable in assisting the developer to locate performance bottlenecks in the running application.

P-GRADE currently supports job execution in interactive as well as batch mode. An application could be executed interactively on a cluster of workstations the workstations involved are declared beforehand and processes are assigned to run on them by PVM or MPI. On the other hand, the application could be submitted as a job to a batch scheduling system like Condor, which would be responsible for assigning processes to resources. In future versions of P-GRADE, the target execution environment could be a computational grid managed by Globus.

4.2 Parallel Approach

Inherently, Monte Carlo methods for solving SLAE allow us to have minimal communication, i.e. to partition the matrix A, pass the non-zero elements of the dense (sparse) matrix (or its partitions) to every processor, to run the algorithm in parallel on each processor computing $\lceil n/p \rceil$ rows (components) of MI or the solution vector and to collect the results from slaves at the end without any communication between sending non-zero elements of A and receiving partitions of A^{-1} or x. The splitting procedure and refinement are also parallelised and integrated in the parallel implementation. Even in the case, when we compute only k components ($1 \leq k \leq n$) of the MI (solution vector) we can divide evenly the number of chains among the processors, e.g. distributing $\lceil kN/p \rceil$ chains on each processor. The only communication is at the beginning and at the end of the algorithm execution which allows us to obtain very high efficiency of parallel implementation.

In addition, an iterative filter process is used to improve the accuracy of the Markov Chain Monte Carlo calculated inverse.

In P-GRADE we employed a master/slave approach, where the main process had to read the data from a file, partition it, send it out to the slaves and collect the results from them at the end of the computation. The slaves were defined using the Process Farm template in P-GRADE which allows scaling to larger number of processes when more compute resources are available. The GRM and PROVE tools were extremely useful in fine-tuning the performance of the application.

5 Numerical Experiments

The algorithms ran on partition of a 32 processor IBM SP3 machine as well as a workstation cluster over a 100 Mbps Ethernet network. Each workstation had an Intel Pentium IV processor with 256 MB RAM and a 30 GB harddisk and was running SUSE Linux 8.1. The MPI environment used was LAM MPI 7.0.

We have carried out tests with low precision $10^{-1} - 10^{-2}$ and higher precision $10^{-5} - 10^{-6}$ in order to investigate the balance between stochastic and deterministic components of the algorithms based on the principle of balancing of errors (e.g. keeping the stochastic and systematic error of the same order) [6]. Consider now, finding the solution to SLAE using Monte Carlo and applying the filter procedures with precision $10^{-5} - 10^{-6}$:

Table 1. MC with filter procedures on the cluster

Matrix Size	Time (Dense Case) in seconds			
	4 proc.	8 proc.	12 proc.	16 proc.
250	59.269	24.795	16.750	14.179
500	329.072	177.016	146.795	122.622
1000	1840.751	989.423	724.819	623.087

Table 2. MC with filter procedures on the miniGrid

Matrix Size	Time (MC, Dense Case) in seconds	
	16 proc. (4 SP and 12 cluster)	16 proc. (8 SP and 8 cluster)
250	729.208	333.418
500	4189.225	1945.454

The above results show that all the algorithms scale very well. The second table shows that it is important to balance computations in a Grid environment and communicate with larger chunks of data. For example, in this case this can lead to a substantial reduction of computational time.

6 Conclusion

In this paper we have considered how we can efficiently use P-GRADE for programming a hybrid Monte Carlo/deterministic algorithms for Matrix Computation for any non-singular matrix. We have compared the efficiency of the algorithm on a cluster of workstations and in a Grid environment. The results show that the algorithms scale very well in such setting, but a careful balance of computation should be maintained.

References

1. B. Fathi, B.Liu and V. Alexandrov, *Mixed Monte Carlo Parallel Algorithms for Matrix Computation* , Lecture Notes in Computer Science, No 2330, Springer-Verlag, 2002, pp 609-618
2. Ortega, J., *Numerical Analysis*, SIAM edition, USA, 1990.
3. Alexandrov V.N., *Efficient parallel Monte Carlo Methods for Matrix Computation,* Mathematics and computers in Simulation, Elsevier **47** pp. 113-122, Netherlands, (1998).
4. Golub, G.H., Ch., F., Van Loan, *Matrix Computations,* The Johns Hopkins Univ. Press, Baltimore and London, (1996)
5. Sobol I.M. *Monte Carlo Numerical Methods.* Moscow, Nauka, 1973 (in Russian).
6. Dimov I., Alexandrov V.N. and Karaivanova A., *Resolvent Monte Carlo Methods for Linear Algebra Problems,* Mathematics and Computers in Simulation, Vo155, pp. 25-36, 2001.
7. Fathi Vajargah B. and Alexandrov V.N., *Coarse Grained Parallel Monte Carlo Algorithms for Solving Systems of Linear Equations with Minimum Communication,* in Proc. of PDPTA, June 2001, Las Vegas, 2001, pp. 2240-2245.
8. Alexandrov V.N. and Karaivanova A., *Parallel Monte Carlo Algorithms for Sparse SLAE using MPI,* LNCS 1697, Springer 1999, pp. 283-290.
9. Alexandrov V.N., Rau-Chaplin A., Dehne F. and Taft K., *Efficient Coarse Grain Monte Carlo Algorithms for matrix computation using PVM,* LNCS 1497, pp. 323-330, Springer, August 1998.
10. Dimov I.T., Dimov T.T., et all, *A new iterative Monte Carlo Approach for Inverse Matrix Problem,* J. of Computational and Applied Mathematics **92** pp 15-35 (1998).
11. Kacsuk P., Lovas R. and Kovács J., *Systematic Debugging of Parallel Programs in DIWIDE Based on Collective Breakpoints and Macrosteps*, Proc. of the 5th International Euro-Par Conference, Toulouse, France, 1999, pp. 90-97.
12. Kacsuk P., Dózsa G., Fadgyas T. and Lovas R. *The GRED Graphical Editor for the GRADE Parallel Program Development Environment*, Journal of Future Generation Computer Systems, Vol. 15(1999), No. 3, pp. 443-452.
13. Balaton Z., Kacsuk P. and Podhorszki N., *Application Monitoring in the Grid with GRM and PROVE* , Proc. of the International Conference on Computational Science, ICCS 2001, San Francisco, CA., USA. pp. 253-262.

Calculating Activation Energies in Diffusion Processes Using a Monte Carlo Approach in a Grid Environment

Mark Calleja and Martin T. Dove

Mineral Physics Group, Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ, United Kingdom

Abstract. A Monte Carlo based method due to Mishin [1] for obtaining activation energies for ionic transport processes has been implemented for empirical-potential models and used in conjunction with Condor and Globus tools. Results are shown for Na^+ migrating in quartz, and O^{2-} vacancy hopping in $CaTiO_3$. We also describe the extensions at the Condor/Globus interface we found necessary to implement in order to facilitate transparent file transfer with Condor-G job submission.

1 Introduction

An understanding of the activation energies present in processes involving atomistic motions in crystalline materials is necessary if an accurate quantitative study of the kinetic properties in such systems is to be made. For example, such information is of use in determining ion-vacancy diffusion behaviour [2]. One approach to obtain an activation energy for diffusion in a crystal is via lattice energy minimisation, where an atom is pushed through the structure in steps, and the energy of the relaxed structure calculated at each step. The main disadvantage of this approach is that it is strictly only applicable for zero temperature, and therefore cannot be used when there is significant thermal disorder of where the structure changes with temperature. Moreover, for very large systemsy, the energy minimisation process may take a long time due to the existance of a complex and shallow energy surface. Alternatively it may be possible to study diffusion in crystalline materials using traditional molecular dynamics (MD) schemes [3], which have been succesfully used for the study of atomic diffusion in fluid phases. However, these methods generally sample many nano-seconds of simulation time, which is not always sufficient to observe such processes. Some effort has been directed to modifying MD schemes to measure these quantities [4]. Temperature accelerated dynamics methods can be considered to fall in this category [5, 6]. In some cases high-temperature behaviour can be inferred albeit in an approximate way [7], such as by driving diffusion with an applied field.

Mishin suggested a scheme based on the Monte Carlo (MC) method which has been applied to the study of ionic diffusion at grain boundaries [1]. The attraction of this approach is that since it is MC based, the method parallelises trivially

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 483–490, 2004.
© Springer-Verlag Berlin Heidelberg 2004

and can be deployed on the resources in our mini-grid. The rest of this article is hence divided as follows: we first describe how we have implemented this method to determine the activation energies asssociated with moving a tracer atom from a given crystallographic site towards a target site. Next, the simulation lifecycle is decribed, from job submission using Condor-G to the problems we faced with automatic file retrieval from the globus gatekeeper, and how we set about solving this problem. We illustrate the work with sample results from studies of the diffusion of alkali cations in a perfect crystal of quartz and within domain walls in quarz, and of studies of oxygen vacancies in the perovskite $CaTiO_3$.

2 The Physical Model

The method studied here is a modification on the standard Metropolis Monte Carlo method [8]. In this scheme, an ion is selected for migration towards a chosen location, *e.g.* a Na^+ ion migrating along a channel in quartz, or an O^{2-} ion hopping into a vacant site. The diffusing ion is marched towards its target in small steps. At every step the entire crystal is allowed to relax through the MC algorithm, except that the migrating ion is constrained to relax perpendicular to its last jump direction. This is important since the locus of equilibrium position of the jumping atom is not necessarily a straight line. In practise, the migrating ion is only forcibly marched to within $\sim 90\%$ of its destination before being allowed to relax without constraint. This is to allow the true energy minimum to be reached, which will generally not be at the specified target location. Sample temperature is incorporated through the MC algorithm, and the structure will relax in a way that incorporates thermal fluctuations corresponding to the preset temperature.

Although this method should work independently of the potential models employed, we have used simple empirical potentials. These are much quicker, though not necessarily as accurate, as *ab initio* methods (*e.g.* see [9], although only the $T = 0$ K case is studied by these authors). The interatomic potentials used here include the simple two-body Buckingham plus Coulombic potentials,with the following functional form for ions i and j separated by distance r:

$$V(r) = A_{ij} \exp(-b_{ij} r) - \frac{C_{ij}}{r^6} + \frac{q_i q_j}{4\pi\epsilon_o r} \tag{1}$$

Long range interactions are handled using the well-known Ewald method [10]. Note that the method described here is not limited to two-body potentials, and our implementation supports three-body potentials (such as bond-angle dependent potentials).

After every hop of the migrating ion, the whole system is allowed to equilibrate before statistics are gathered. The number of moves per ion required for equilibration are highly system dependent, but are generally ~ 1000 per ion for every migration hop. The hop distances employed for the migrating ion are typically 0.1-0.2 Å, depending on the amount of detail required from the profile.

At the end of all the parallel runs it is then possible to obtain an average energy value at each hop point. Hence note that each point is the result of two averaging processes, the first due to the statistics gathering from an individual simulation at a given hop, and the second when the values due to different simulation runs are compared for the same point. Due to the increased thermal fluctuations at high temperatures, the variance on the averages will increase for higher temperatures, meaning that if a uniform level of accuracy is required across a range of operating temperatures it will be necessary to run more configurations at the higher temperatures.

It is important to note that what we measure here is the interaction energy associated with this diffusive process, and not the free energy. To measure the latter one needs to take into account a small entropic contribution, which can be done using the method of thermodynamic integration [11]. This requires many runs at different temperatures, starting with the model T = 0 K system up to the required temperature.

3 Job Submission on a Minigrid

Since the MC method is inherently statistical in nature, sampling the whole of phase space and not only the minimum-energy pathway, it is essential to run many calculations and obtain averages across many configurations for each step along the diffusion pathway. This approach lends itself easily to parallelisation, running different calculations on different processors.

In our studies we generally set off an identical number of initial configurations, but with different seeds for the random number generators, on a number of processors. Individual jobs were submitted to a local machine which had the required Condor [12] submission daemons running as well as a Globus installation [13]. Various versions of these tools were utilized, though as of writing they currently are 2.4.3 (Globus) and 6.6.0 (Condor). The jobs themselves were submitted using Condor-G rather than Globus's own interface since this allowed for simple workflows to be implemented using DAGMan (Directed Acyclic Graph Manager), a meta-scheduler for Condor. Figure 1 shows the scheme employed in this work. This approach was chosen since it provides a single point of access for remote facilities via the relevant globus server which helps keep firewall issues to a minimum (as opposed to, say, using Condor glideins).

The process begins with the submission of a condor job script (step 1), which specifies that the job should be run in the globus universe. The only non-standard feature of this script is the extra RSL value, condor_home_dir, with which we specify the directory we would like the output files to be returned to on this submitting machine. This tag is our own addition, and we'll say more about it below. The job is now submitted to the local condor pool. In step 2, the job is forwarded to the remote gatekeeper, invariably across a firewall. It is the gatekeeper's task to fork a relevant jobmanager (e.g. a condor job for an Intel/Linux machine) which submits the job to the condor pool local to the gatekeeper. For our set-up we found that the default condor.pm file used by the jobamanger (usually

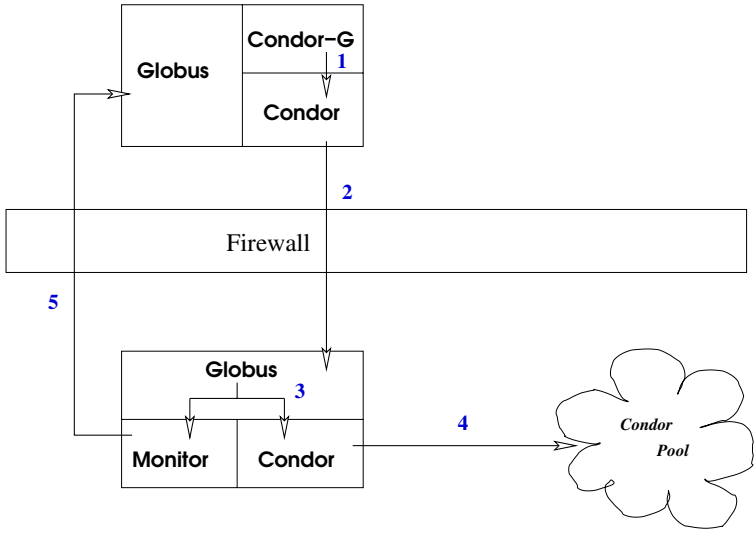


Fig. 1. Job lifecycle employed in this work. See text for description of workflow

located in $GLOBUS_LOCATION/lib/perl/Globus/GRAM/JobManager) does not provide sufficient functionality for file transfer other than stdin, stdout or stderr. Hence, to circumvent this we modified condor.pm so that it forks off a new process (step 3) which is then exec'd with a new monitoring job whose duty is to periodically check for the completion of the job in condor pool. Also, the condor_home_dir attribute is added to the condor.rvf file in $GLOBUS_LOCATION/share/globus_gram_job_manager. All files generated by such a job are placed in a temporary directory so that they cannot be automatically removed by condor when the job completes, or it may do so before we've had a chance to send the output files back to the submitter. In the meantime, the original job is commited to a suitable resource within the pool (step 4). After periodically sleeping, the monitoring process detects that the condor job has finished and scans the temporary directory designated for all job files, discriminating between input from output files. It then tries to send the output files back to the submitting machine in the directory given in the condor_home_dir value using gsiftp (step 5). The latter is used since the https socket used for initially sending input data over may now be closed and no longer available. This means that we require access through any firewall protecting the submitting machine. On the successful completion of all file transfer the monitor cleans up by removing the temporary directory and all its contents before exiting.

4 Results

We present some typical examples that we have studied (all at 0 GPa), with each plot being the result of eight parallel runs. Figure 2 shows the energy profile for

a Na^+ ion migrating along the *c*-axis in alpha bulk and twinned quartz at 10 K (potential details are given in [14]). The abscissa denotes the hop count, and cannot readily be associated with a jump distance since the hopping atom does not necessarily follow a straight path, although the paths followed in the quartz systems studied here are relatively straight (unlike the perovskite model studied below). The ordinate gives the net change in crystal energy for the processes. Note how the three small energy barriers due to the three sets of oxygen pairs in the helical structure of the bulk can be easily distinguished. By comparison, the energy barrier seen in a twin wall is an order of magnitude greater, explaining why cationic conductivity is poisoned along this axis by twin walls [7].

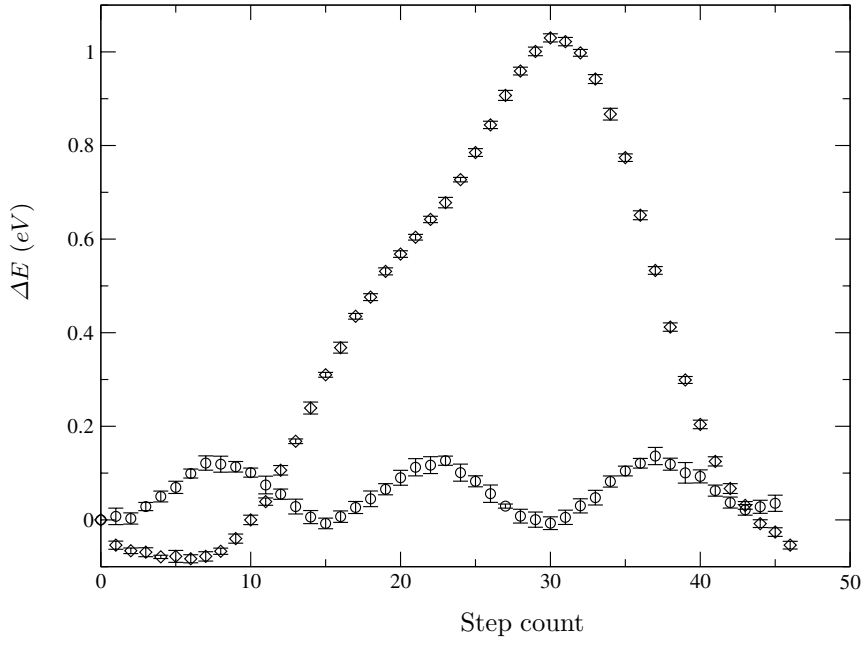


Fig. 2. Activation energy for Na^+ ion diffusing for one unit cell along [001] in quartz at 10 K. Circles are for motion in the bulk, diamonds in a wall

In Figure 3 we show a similar process, but this time the sodium ion is migrating along the [100] direction at 10, 150 and 300 K in bulk quartz. An interesting, temperature-dependent, feature is observed with the shoulder at $\sim 0.75eV$, which appears to develop on heating to 150 K before subsiding again on further heating to 300 K. Note the small bump at around the 23rd step in the 10 K profile; this is an example of the diffusing ion being pushed too far, so that by

this step it's being displaced away from its equilibrium position. This situation is rectified by the next step, when it is allowed to relax without any constraint.

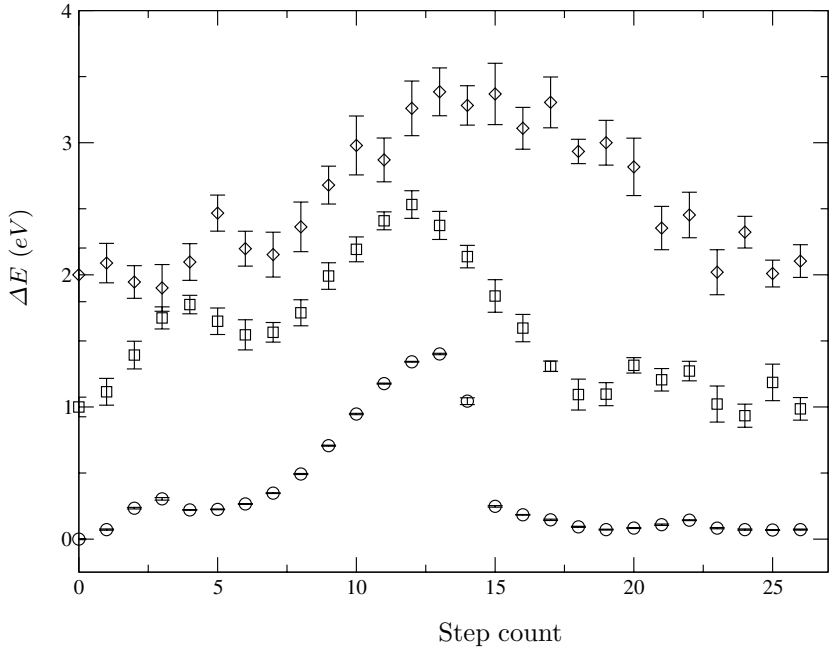


Fig. 3. Energy profile for Na^+ ion migrating along [100] in quartz at 10 (bottom), 150 (middle) and 300 (top) K

For the next system studied we show the energetics associated with an O^{2-} ion migrating towards a nearby vacancy in bulk orthorhombic $CaTiO_3$ for various temperatures (potential details are given in [15]). Figure 4 shows the corresponding profiles for runs at 100, 300, 500 and 1000 K. Even at the highest temperature, the saddle point energy can be identified reasonably clearly. Note the apparent weak temperature dependence, with the peak values rising with increasing T.

5 Discussion

A simple Monte Carlo based method due to Mishin has been shown to provide accurate activation and saddle point energies even at relatively high temperatures. Obviously the temperature range is highly system dependent, with larger

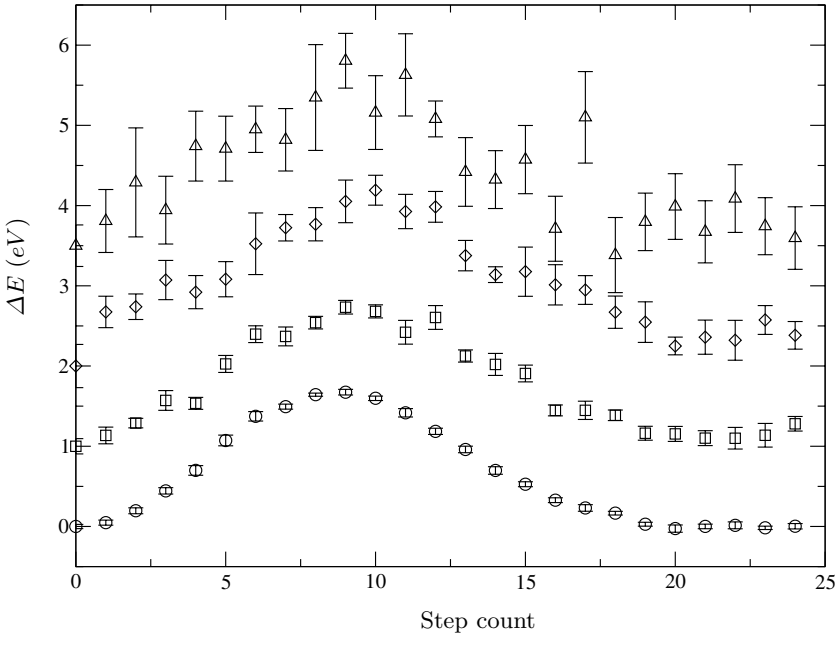


Fig. 4. Energy profile for O^{2-} ion migrating $CaTiO_3$ at (from bottom to top) 100, 300, 500 and 1000 K

activation energies being easier to characterise at high T than small barriers. Furthermore, the method readily lends itself to be deployed within a grid infrastructure, though not without extending current technologies. However, we generally find that the combination of Condor, Globus and DAGMan tools provide a ready fabric for making the most of Monte Carlo simulations on a grid, with only minor modifications and exstensions.

Acknowledgements. We would like to thank Mark Hayes (CeSC) and Rik Tyer (CCLRC) for useful discussions and the NERC (UK) (grant: Environment from the molecular level) for funding this work.

References

1. Y. Mishin, *Defect and Diffusion Forum*, **143 - 147** (1997) 1357
2. R. J. Borg and G. J. Dienes, *An Introduction to Solid State Diffusion,* Academic Press (1988)
3. W. Smith, C. W. Yong, P. M. Rodger, *Mol. Simulat.*, **28** (2002) 385

4. S. C. Parker, N. H. de Leeuw, E. Bourova and D. J. Cooke, *Rev. Mineral Geochem.,* **42** (2001) 63
5. A. F.Voter *J. Chem. Phys*, **106** (1997) 11
6. R. A. Miron and K. A. Fichthorn *J. Chem. Phys*, **119** (2003) 6210
7. M. Calleja, M. T. Dove and E. K. H. Salje, *J. Phys.: Condens. Matter*, **13** (2001) 9445.
8. J. M. Thijssen, *Computational Physics*, Cambridge (1999).
9. A. I. Lichtenstein, R. O. Jones, H. Xu and P. J. Heaney, *Phys. Rev. B* **58** (1998) 6219
10. P. Ewald, *Ann. Phys.* **64** (1921) 253
11. M. C. Warren, M. T. Dove, E. R. Myers, A. Bosenick, E. L. Palin, C. I. Sainz-Diaz, B. S.Guitton and S. A. T. Redfern *, Mineral. Mag.*, **65** (2001) 221–248
12. T. Tannenbaum, D. Wright, K. Miller and M. Livny, "Condor - A Distributed Job Scheduler", in *Beowulf Cluster Computing with Linux*, MIT, (2002)
13. I. Foster and C. Kesselman, *Intl J. Supercomputer Applications*, **11(2)**, (1997) 115-128
14. G. J. Krammer, N. P. Farragher, B. W. H. van Beest and R. A. van Santen, *Phys. Rev. B* **43** (1991) 5068
15. M. Calleja, M. T. Dove and E. K. H. Salje, *J. Phys.: Cond. Matt.* **15** (2003) 2301

Using Parallel Monte Carlo Methods in Large-Scale Air Pollution Modelling

V.N. Alexandrov[1] and Z. Zlatev[2]

[1] Advanced Computing and Emergent Technologies Centre, University of Reading, Reading, UK (v.n.alexandrov@reading.ac.uk)
[2] National Environmental Research Institute, Frederiksborgvej 399, P. O. Box 358, DK-4000 Roskilde, Denmark (zz@dmu.dk)

Abstract. Large-scale air pollution models can successfully be used in different environmental studies. These models are described mathematically by systems of partial differential equations. Splitting procedures followed by discretization of the spatial derivatives lead to several large systems of ordinary differential equations of order up to 80 millions. These systems have to be handled numerically at up to 250 000 time-steps. Furthermore, many scenarios are often to be run in order to study the dependence of the model results on the variation of some key parameters (as, for example, the emissions). Such huge computational tasks can successfully be treated only if (i) fast and sufficiently accurate numerical methods are used and (ii) the models can efficiently be run on parallel computers. Efficient Monte Carlo methods for some subproblems will be presented and applications of the model in the solution of some environmental tasks will also be made.

1 Introduction

The control of the pollution levels in different highly polluted regions of Europe and North America (as well as in other highly industrialized parts of the world) is an important task for the modern society. Its relevance has been steadily increasing during the last two-three decades. The need to establish reliable control strategies for the air pollution levels will become even more important in the future. Large-scale air pollution models can successfully be used to design reliable control strategies. Many different tasks have to be solved before starting to run operationally an air pollution model. The following tasks are most important:

- describe in an adequate way all important physical and chemical processes,
- apply fast and sufficiently accurate numerical methods in the different parts of the model,
- ensure that the model runs efficiently on modern high-speed computers (and, first and foremost, on different types of parallel computers),
- use high quality input data (both meteorological data and emission data) in the runs,
- verify the model results by comparing them with reliable measurements taken in different parts of the space domain of the model,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 491–498, 2004.
© Springer-Verlag Berlin Heidelberg 2004

– carry out some sensitivity experiments to check the response of the model to changes of different key parameters

and

– visualize and animate the output results to make them easily understandable also for non-specialists.

The performance of the model on high-speed computers will be discussed in this paper.

1.1 Main Physical and Chemical Processes

Five physical and chemical processes have to be described by mathematical terms in the beginning of the development of an air pollution model. These processes are: (i) horizontal transport (advection), (ii) horizontal diffusion, (iii)chemical transformations in the atmosphere combined with emissions from different sources, (iv) deposition of pollutants to the surface and (v) vertical exchange (containing both vertical transport and vertical diffusion).

It is important to describe in an adequate way all these processes. However, this is an extremely difficult task; both because of the lack of knowledge for some of the processes (this is mainly true for some chemical reactions and for some of the mechanisms describing the vertical diffusion) and because a very rigorous description of some of the processes will lead to huge computational tasks which may make the treatment of the model practically impossible. The main principles used in the mathematical description of the main physical and chemical processes as well as the need to keep the balance between the rigorous description of the processes and the necessity to be able to run the model on the available computers are discussed in [6].

1.2 Mathematical Formulation of a Large Air Pollution Model

The description of the physical and chemical processes by mathematical terms leads to a system of partial differential equations (PDEs) of the following type:

$$\begin{aligned}\frac{\partial c_s}{\partial t} = &-\frac{\partial (uc_s)}{\partial x} - \frac{\partial (vc_s)}{\partial y} - \frac{\partial (wc_s)}{\partial z} \\ &+\frac{\partial}{\partial x}\left(K_x \frac{\partial c_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s}{\partial z}\right) \\ &+E_s - (\kappa_{1s} + \kappa_{2s})c_s + Q_s(c_1, c_2, \ldots, c_q), \quad s = 1, 2, \ldots, q,\end{aligned} \tag{1}$$

where (i) the concentrations of the chemical species are denoted by c_s, (ii) u, v and w are wind velocities, (iii) K_x, K_y and K_z are diffusion coefficients, (iv) the emission sources are described by E_s, (v) κ_{1s} and κ_{2s} are deposition coefficients

and (vi) the chemical reactions are denoted by $Q_s(c_1, c_2, \ldots, c_q)$. The CBM IV chemical scheme, which has been proposed in [4], is actually used in the version of DEM (the Danish Eulerian Model; [6], [7]) that will be considered in this paper.

2 Achieving Parallelism

Delivering efficient parallel algorithms for treating large scale air pollution is very important. Note, for example, that the size of the computational tasks can be formidable if we need finer resolution, e.g. 480×480 mesh, leading to solving 8064000 equations per time step and depending on the number of species and time steps potentially to a systems of ordinary differential equations of order up to 80 millions.

Therefore the preparation of a parallel code is by no means an easy task. Moreover, it may happen that when the code is ready the computing centre exchanges the computer which has been used in the preparation of the code with another (hopefully, more powerful) computer. This is why it is desirable to use only standard tools in the preparation of the code. This will facilitate the transition

of the code from one computer to another when this becomes necessary. Only standard MPI ([3]) tools are used in the parallel versions of DEM.

2.1 Development of MPI Versions of DEM

The approach used when MPI tools are to be implemented is based in dividing the space domain of the model into p sub-domains, where p is the number of processors which are to be used in the run. Two specific modules are needed in the MPI versions: (i) a pre-processing module and (ii) a post-processing module.

- **The pre-processing module.** corresponding to the p sub-domains obtained in the division of the space domain. In this way, each processor will work during the whole computational process with its own set of input data.
- **The post-processing module.** Each processor prepares its own set of output data. During the post-processing the p sets of
  output data corresponding to the p sub-domains are collected and common output files are prepared for future use.
- **Benefits of using the two modules.** Excessive communications during the computational process are avoided when the two modules are used. It should be stressed, however, that not all communications during the computational process are avoided. Some communications along the inner boundaries of the sub-domains are still needed. However, these communications are to be carried only once per step and only a few data are to be communicated. Thus, the actual communications that are to be carried out during the computations are rather cheap when the pre-processing and the post-processing modules are proper implemented.

It is important to emphasize here that the introduction of p sub-domains leads to a reduction of the main arrays by a factor of p. Consider as an illustrations the major arrays used in the chemical sub-model. The dimensions of these arrays are reduced from $(N_x \times N_y, N_s)$ to $(N_x \times N_y/p, N_s)$. It is clear that this is equivalent to the use of p chunks. Chunks of length $N_x \times N_y/p$ are still very large. Therefore, the following algorithm has also to be used (in each sub-domain) when the MPI versions are used:

```
DO ICHUNK=1,NCHUNKS
    Copy chunk ICHUNK from some of the eight
    large arrays into small two-dimensional
    arrays with leading dimension NSIZE
    DO J=1,NSPECIES
        DO I=1,NSIZE
            Perform the chemical reactions involving
            species J for grid-point I
        END DO
    END DO
    Copy some of the small two-dimensional
    arrays with leading dimension NSIZE
    into chunk ICHUNK of the corresponding
    large arrays
END DO
```

However, the reduction of the arrays leads to a reductions of the copies that are to be made in the beginning and in the end of the algorithm. Thus, the reduction of the arrays leads to a better utilization of the cache memory.

The automatic reduction of the sizes of the involved arrays, and the resulting from this reduction better utilization of the cache memory, make the MPI versions attractive also when shared memory machines are available.

3 Description of the Grid of Sun Computers

Sun computers located at the Danish Centre for Scientific Computing (the Danish Technical University in Lyngby) were used in the runs. The computers and the their characteristics are shown in Table 1. All these computers were connected with a 1Gbit/s Switch.

The computers are united in a grid (consisting of 216 processors) so that a job sent without a special demand will be assigned on the computer on which there are sufficiently many free processors. The different computers have processors of different power (therefore, it is in principle possible to use the grid as a heterogeneous architecture, but this option is not available yet).

We are in general allowed to use no more than 16 processors, but several runs on more that 16 processors were performed with a special permission from the Danish Centre for Scientific Computing. In the runs in this section we used only "newton" (i.e. we had always a requirement specifying the particular computer on which the job must be run)

More details about the high speed computers that are available at the Technical University of Denmark can be found in [5].

Table 1. The computers available at the Sun grid

Computer	Type	Power	RAM	Processors
Bohr	Sun Fire 6800	UltraSparc-III 750 MHrz	48 GB	24
Erlang	Sun Fire 6800	UltraSparc-III 750 MHrz	48 GB	24
Hald	Sun Fire 12k	UltraSparc-III 750 MHrz	144 GB	48
Euler	Sun Fire 6800	UltraSparc-III 750 MHrz	24 GB	24
Hilbert	Sun Fire 6800	UltraSparc-III 750 MHrz	36 GB	24
Newton	Sun Fire 15k	UltraSparc-IIIcu 900 MHrz	404 GB	72

4 Running the MPI Versions of DEM

Four MPI versions of DEM have been tested: (i) the 2-D model on a coarse grid, (ii) the 3-D version on a coarse grid, (iii) the 2-D version on a fine grid and (iv) the 3-D version on a fine grid.

The problems were run with three different sizes $NSIZE$ of chunks: (a) the minimal size of the chunks, $NSIZE = 1$ for all cases, (b) a medium size of the chunks, $NSIZE = 24$ for all cases and (c) the maximal size of the chunks, which is $NSIZE = 1152$ for the coarse grid when 8 processors are used and $NSIZE = 28800$ for the fine grid (again when 8 processors are used).

Finally, in most of the cases both 1 processor and 8 processors were used. Some of the jobs were also run on more than 8 processors.

All runs of the versions discretized on the coarse grid were run for the typical period of one year (in which case it is possible to study seasonal variations). The 2-D version of DEM discretized on the fine grid was run over a period of one month. Finally, the 3-D version of DEM discretized on the fine grid was run over a time period of 42 hours. This is a rather short period, but it is still meaningful to a certain degree because several changes from day to night and from night to day occur in this period, which is important for the test of the photo-chemical reactions.

The computing times in all tables are given in seconds. The abbreviations used in the tables can be explained as follows:

- ADV stands for the horizontal transport + diffusion process,
- CHEM stands for the process uniting the chemical reactions, the treatment of the emissions and the deposition part,
- COMM stands for the part needed to perform communications along the inner boundaries,
- VERT stands for the vertical exchange processes
- TOTAL stands for the total computing time (including the sum of the times given in the same column above the last item + the computing times needed for performing input-output operations, pre-processing, post-processing, etc.)

Table 2. Running DEM discretized on a $96 \times 96 \times 10$ grid on one processor

	$NSIZE = 1$		$NSIZE = 24$		$NSIZE = 1152$	
Process	Time	Part	Time	Part	Time	Part
ADV	169776	31.5%	159450	37.8%	169865	30.9%
CHEM	337791	62.7%	233471	55.3%	348769	63.4%
VERT	23221	4.3%	21473	5.1%	23014	4.2%
COMM	2	0.0%	2	0.0%	2	0.0%
TOTAL	538953	100.0%	421763	100.0%	549835	100.0%

Table 3. Running DEM discretized on a $96 \times 96 \times 10$ grid on eight processors

	$NSIZE = 1$			$NSIZE = 24$			$NSIZE = 1152$		
Process	Time	Part	Speed-up	Time	Part	Speed-up	Time	Part	Speed-up
ADV	18968	27.4%	9.0	18498	33.3%	8.6	18641	26.3%	9.1
CHEM	41334	59.6%	8.2	29189	52.3%	8.0	43291	61.3%	8.1
VERT	1213	1.7%	19.1	1200	2.2%	17.9	1240	1.8%	18.6
COMM	911	1.3%	-	878	1.6%	-	973	1.4%	-
TOTAL	69325	100.0%	7.8	55723	100.0%	7.6	70653	100.0%	7.8

The percentages of the computing times for the different processes related to the total computing times are given in the columns under "Part". The "Speed-up" is the ratio of the computing time on one processor and the computing time on p processors (where p is the number of processors that are used in the run under considerations; as mentioned above, eight processors were as a rule used in our experiments).

Running the 3-D MPI version discretized on the coarse grid. Results from the six runs with this code are shown in Table 2 (runs on one processor performed by using three values of $NSIZE$) and Table 3 (runs on 8 processors performed again with three values of $NSIZE$).

Running the 3-D MPI version discretized on the fine grid. Results from the six runs with this code are shown in Table 4 (runs on one processor performed by using three values of $NSIZE$) and Table 5 (runs on 8 processors performed again with three values of $NSIZE$).

Further, in the next table we present a comparison of the efficiency of using Monte Carlo algorithms [1,2] for solving subproblems (e.g. linear systems arising after discretization) in the model. Comparisons of the efficiency of these methods with some of the traditional ones such as LU is made. It is seen in Table 6 that with the growth of the problem size Monte Carlo calculations become more efficient than LU for the sparse structured matrices treated in the advection submodel.

Several runs were performed by using up to 60 processors. The 3-D refined version, where high efficiency is most desirable, was used in this runs. The results

Table 4. Running DEM discretized on a $480 \times 480 \times 10$ grid on one processor

	$NSIZE = 1$		$NSIZE = 24$		$NSIZE = 28800$	
Process	Time	Part	Time	Part	Time	Part
ADV	261631	67.0%	271419	72.9%	268337	49.8%
CHEM	86317	22.1%	56797	15.3%	228216	42.3%
VERT	40721	10.4%	42320	11.4%	41223	7.6%
COMM	1	0.0%	1	0.0%	1	0.0%
TOTAL	390209	100.0%	372173	100.0%	539319	100.0%

Table 5. Running DEM discretized on a $480 \times 480 \times 10$ grid on eight processors

	$NSIZE = 1$			$NSIZE = 24$			$NSIZE = 28800$		
Process	Time	Part	Speed-up	Time	Part	Speed-up	Time	Part	Speed-up
ADV	13606	46.2%	19.2	13515	52.7%	20.1	13374	28.9%	20.1
CHEM	10398	35.3%	8.3	6681	26.0%	8.5	25888	56.0%	8.8
VERT	2830	9.6%	14.4	2802	10.9%	15.1	2709	5.9%	15.2
COMM	2316	7.9%	-	2340	9.1%	-	3925	8.5%	-
TOTAL	29449	100.0%	13.3	25654	100.0%	14.5	46210	100.0%	11.7

are given in Table 7 and indicate that the parallel algorithms applied in DEM scale very well.

Major conclusions from the runs. It is seen that the exploitation of the cache memory is always giving good results (compare the results for $NSIZE = 24$ with the results for $NSIZE = 1$ and $NSIZE = 1152(28800)$. The speed-ups for the physical processes are super-linear (greater for ADV and VERT than for CHEM, which should be expected, because chunks are used in the chemical parts). The speed-ups for the total computing time are lower, but anyway at least close to linear.

5 Conclusion

It has been shown in this paper, that based on our parallelization approach, different submodels of DEM scale very well when the number of the processors used is increased. For some parts of the model we continue to observe super-linear speedup. In addition, the experiments with Monte Carlo show that for larger problems these algorithms are more efficient that the traditional ones, scale well and can lead to a reduction of computational time.

Further work is required, for example, the improvement of the fine resolution versions of DEM, especially the 3-D fine resolution version, is an important task which must be resolved in the near future. It is necessary both to improve the performance of the different versions of the model and to have access to more processors (and/or to more powerful computers) in order to be able to run operationally fine resolution versions of DEM.

Table 6. Running DEM with Monte Carlo in the Advection part for various grid refinements with time step 200, 800 and 2400 respectively

Pr. Size	Method	error	Startup Time	ADVEC	Total Time
8×8	LU	0.001	0.0	0.090	0.090
8×8	MC	0.001	2.1	0.090	2.190
32×32	LU	0.001	0.01	31.06	31.07
32×32	MC	0.001	3.52	29.230	32.72
96×96	LU	0.001	0.01	227.57	227.58
96×96	MC	0.001	63.8	88.8	152.6

Table 7. Running DEM discretized on a $480 \times 480 \times 10$ on different numbers of processors

Processors	Time	Speed-up
1	372173	-
15	12928	28.79
30	7165	51.94
60	4081	91.20

Acknowledgements. A grant (CPU-1101-17) from the Danish Centre for Scientific Computing (DCSC) gave us access to the Sun computers at the Technical University of Denmark. The members of the staff of DCSC helped us to resolve some difficult problems related to the efficient exploitation of the grid of Sun computers.

References

1. Alexandrov V.N., *Efficient parallel Monte Carlo Methods for Matrix Computation,* Mathematics and computers in Simulation, Elsevier **47** pp. 113-122, Netherlands, (1998).
2. Dimov I., Alexandrov V.N. and Karaivanova A., *Resolvent Monte Carlo Methods for Linear Algebra Problems,* Mathematics and Computers in Simulation, Vo155, pp. 25-36, 2001.
3. W. Gropp, E. Lusk and A. Skjellum, *Using MPI: Portable programming with the message passing interface*, MIT Press, Cambridge, Massachusetts (1994).
4. M. W. Gery, G. Z. Whitten, J. P. Killus and M. C. Dodge, *A photochemical kinetics mechanism for urban and regional computer modeling,* Journal of Geophysical Research, Vol. 94 (1989), 12925–12956.
5. WEB-site of the Danish Centre for Scientific Computing at the Technical University of Denmark, *Sun High Performance Computing Systems*, http://www.hpc.dtu.dk, 2002.
6. Z. Zlatev, *Computer treatment of large air pollution models*, Kluwer Academic Publishers, Dordrecht-Boston-London (1995).
7. Z. Zlatev, *Massive data set issues in air pollution modelling*, In: Handbook on Massive Data Sets (J. Abello, P. M. Pardalos and M. G. C. Resende, eds.), pp. 1169-1220, Kluwer Academic Publishers, Dordrecht-Boston-London (2002).

Parallel Importance Separation for Multiple Integrals and Integral Equations

Sofiya Ivanovska and Aneta Karaivanova

IPP – Bulgarian Academy of Sciences
Acad. G. Bonchev St., Bl.25A, 1113 Sofia, Bulgaria
{sofia,anet}@parallel.bas.bg

Abstract. In this paper we present error and performance analysis of a Monte Carlo variance reduction method for solving multidimensional integrals and integral equations. This method, called importance separation, combines the idea of separation of the domain into uniformly small subdomains with the approach of importance sampling. The importance separation method is originally described in our previous works, here we generalize our results and discuss the performance in comparison with crude Monte Carlo and importance sampling. Based on our previous investigation we propose efficient parallelizations of the importance separation method. Numerical tests implemented on PowerPC cluster using MPI are provided. The considered algorithms are carried out using pseudorandom numbers.

1 Introduction

Multidimensional numerical quadratures are of great importance in many practical areas, ranging from atomic physics to finance. The crude Monte Carlo method has rate of convergence $O(N^{-1/2})$ which is independent of the dimension of the integral, and that is why Monte Carlo integration is the only practical method for many high-dimensional problems.
Much of the efforts to improve Monte Carlo method (MCM)are in construction of variance reduction methods which speed up the computation. Importance sampling is probably the most widely used Monte Carlo variance reduction method, [11,6,13,14]. One use of importance sampling is to emphasize rare but important events, i.e., small regions of space in which the integrand is large. One of the difficulties in this method is that sampling from the importance density is required, but this can be performed using acceptance-rejection.
In [8] a method called importance separation (IS) was introduced. This method combines the ideas from importance sampling and stratification. The IS method has the best possible rate of convergence for certain class of functions but its disadvantage is that it gives better accuracy only for low dimensions and its increased computational complexity. This method was applied for evaluation of multidimensional integrals [3] and for solving integral equations [5].
In this paper we consider both problems, solving multiple integrals and integral equations through unified point of view converting the problem of solving of

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 499–506, 2004.
© Springer-Verlag Berlin Heidelberg 2004

integral equations into approximate calculation of a finite number of integrals (linear functionals of iterative functions), then importance separation is applied to the each of the integrals. We also describe the parallel implementation of the two algorithms based on IS; it has some difficulties due to hierarchical structure of the method.

2 Formulation of the Problem

2.1 Calculation of Multidimensional Integrals

Consider the problem of approximate calculation of the multiple integral

$$I = \int_G f(x)p(x)\,\mathrm{d}x, \;\; G \equiv [0;1]^d \tag{1}$$

where $f(x)$ is an integrable function for any $x \in G \subset \mathbb{R}^d$ and $p(x) \geq 0$ is a probability density function, such that $\int_G p(x)\,\mathrm{d}x = 1$.
The Monte Carlo quadrature formula is based on the probabilistic interpretation of an integral. If $\{x_n\}$ is a sequence in G sampled with density $p(x)$, then the Monte Carlo approximation to the integral is, [12],

$$I \approx I_N[f] = \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

with the integration error $\varepsilon_N = |I - I_N| \approx \sqrt{\frac{Var(f)}{N}}$.

2.2 Solving Integral Equations

Consider the Fredholm integral equation of the second kind:

$$u(x) = \int_\Omega k(x,x')u(x')\,\mathrm{d}x' + f(x)$$

or

$$u = \mathcal{K}u + f \;\; (\mathcal{K} \text{ is an integral operator}), \quad \text{where}$$

$k(x,x') \in L_2(\Omega \times \Omega), f(x) \in L_2(\Omega)$ are given functions and $u(x) \in L_2(\Omega)$ is an unknown function, $x, x' \in \Omega \subset \mathbb{R}^d$ (Ω is a bounded domain).
We are interested in Monte Carlo method for evaluation of linear functionals of the solution of the following type:

$$J(u) = \int \varphi(x)u(x)\,\mathrm{d}x = (\varphi, u). \tag{2}$$

It is assumed that $\varphi(x) \in L_2(\Omega)$. We can apply successive approximation method for solving integral equations:

$$u^{(i)} = \sum_{j=0}^{i} \mathcal{K}^{(j)} f = f + \mathcal{K}f + \ldots + \mathcal{K}^{(i-1)}f + \mathcal{K}^{(i)}f, \quad i = 1, 2, \ldots \tag{3}$$

where $u^{(0)}(x) \equiv f(x)$. It is known that the condition $\|\mathcal{K}\|_{L_2} < 1$ is a sufficient condition for convergence of the Neumann series. Thus, when this condition is satisfied, the following statement holds:

$$u^{(i)} \longrightarrow u \quad \text{as} \quad i \to \infty.$$

Therefore,

$$J(u) = (\varphi, u) = \lim_{i\to\infty}(\varphi, u^{(i)}) = \lim_{i\to\infty}\left(\varphi, \sum_{j=0}^{i}\mathcal{K}^{(j)}f\right) = \lim_{i\to\infty}\sum_{j=0}^{i}\left(\varphi, \mathcal{K}^{(j)}f\right).$$

An approximation of the unknown value (φ, u) can be obtained using a truncated Neumann series (3) for sufficiently large i:

$$(\varphi, u^{(i)}) = (\varphi, f) + (\varphi, \mathcal{K}f) + \ldots + (\varphi, \mathcal{K}^{(i-1)}f) + (\varphi, \mathcal{K}^{(i)}f).$$

So, we transform the problem for solving integral equations into a problem for approximate evaluation of a finite number of multidimensional integrals. We will use the following denotation $(\varphi, \mathcal{K}^{(j)}f) = I(j)$, where $I(j)$ is a value, obtained after integration over $\Omega^{j+1} = \Omega \times \ldots \times \Omega$, $j = 0, \ldots, i$. It is obvious that the calculation of the estimate $(\varphi, u^{(i)})$ can be replaced by evaluation of a sum of linear functionals of iterative functions of the following type $(\varphi, \mathcal{K}^{(j)}f), j = 0, \ldots, i$, which can be presented as:

$$\begin{aligned}(\varphi, \mathcal{K}^{(j)}f) &= \int_{\Omega}\varphi(t_0)\mathcal{K}^{(j)}f(t_0)\,\mathrm{d}t_0 = \\ &= \int_{G}\varphi(t_0)k(t_0,t_1)\ldots k(t_{j-1},t_j)f(t_j)\,\mathrm{d}t_0\ldots\,\mathrm{d}t_j,\end{aligned} \tag{4}$$

where $t = (t_0, \ldots, t_j) \in G \equiv \Omega^{j+1} \subset \mathbb{R}^{d(j+1)}$. If we denote by $F(t)$ the integrand function

$$F(t) = \varphi(t_0)k(t_0,t_1)\ldots k(t_{j-1},t_j)f(t_j), \quad t \in \Omega^{j+1},$$

then we will obtain the following expression for (4):

$$I(j) = (\varphi, \mathcal{K}^{(j)}f) = \int_{G}F(t)\,\mathrm{d}t, \quad t \in G \subset \mathbb{R}^{d(j+1)}. \tag{5}$$

So, from now on we will consider the problem for approximate calculation of multiple integrals of the type (5). We will first review briefly the most widely used Monte Carlo methods for integrals and integral equations. It is well-known that Monte Carlo methods reduce the problem to the approximate calculation of mathematical expectation which coincides with the unknown functional defined by (2).

3 Importance Separation for Integrals

The importance separation is a Monte Carlo method which combines the idea of separation of the domain of integration into uniformly small subdomains (stratification, [4]) and the Kahn approach to implement more samples in those subdomains where the integrand is large (importance sampling for integrals, [7], and

for integrals equations, [2,9]). This method has the best rate of convergence for the class of functions with bounded derivatives (see [1]).
One approach how to make a partition of the given domain into subdomains was studied in [8] where the problem for evaluation of the integral $I(j) = \int_G F(t)\,\mathrm{d}t$ is considered. The suggested there partition scheme of the domain $G = [a; b]$ into M subintervals (one-dimensional case) is the following one:

$$\begin{aligned} G &= \bigcup_{l=1}^{M} G_l, \quad G_l \equiv [x_{l-1}, x_l], \quad l = 1, \ldots, M-1, \\ C_i &= \frac{1}{2}[F(x_{i-1}) + F(x_M)](x_M - x_{i-1}), \quad i = 1, \ldots, M-1, \\ x_i &= x_{i-1} + \frac{C_i}{F(x_{i-1})(M-i+1)}, \qquad x_0 = a,\ x_M = b. \end{aligned} \tag{6}$$

It is known (see [12]) that

$$\mathbf{E}\theta_N^*(j) = I(j),$$

where

$$\theta_N^*(j) = \sum_{i=1}^{M} \frac{\mathrm{V}(G_i)}{N_i} \sum_{l=1}^{N_i} F(\xi_l^{(i)}), \quad \sum_{i=1}^{M} N_i = N,$$

and $\xi_l^{(i)}$ is a random point in the i-th subdomain of G.
In the general case of multidimensional integrals ($G \subset \mathbb{R}^n$) the following integration error (the probable error) holds [8]:

$$r_N \le \sqrt{2}n \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{L_i}c_{1_i}c_{2_i}^{\hat{}})^2\right]^{\frac{1}{2}} N^{-\frac{1}{2}-\frac{1}{n}}, \quad M = N, \tag{7}$$

where n is the dimension of the domain of integration, M is the number of subdomains, the integrand is a positive function $F(t)$, which belongs to $W^{(1)}(L, G)$. This means that $F(t)$ is continuous on G with partially continuous first derivatives and

$$\left|\frac{\partial F}{\partial t_l}\right| \le L_{i_l},\ l = 1, \ldots, d,\ t \in G_i,\ L_i = (L_{i_1}, \ldots, L_{i_d}),\ \hat{L_i} = \max_l L_{i_l}.$$

The constants $c_{1_i} (i = 1, \ldots, M)$ and the vectors of constants $c_{2_i} \in \mathbb{R}^d$ are determined from the requirement the subdomains $G_i,\ i = 1, \ldots, M$ have to be uniformly small in probability and in geometrical size, and it is also assumed that $c_{2_i}^{\hat{}} = \max_l c_{2_{i_l}}$.
From (7) it is clear that the error of the importance separation method which has the order $\mathcal{O}(N^{-1/2-1/n})$ asymptotically goes to $\mathcal{O}(N^{-1/2})$ for large dimensions n. This estimation of integration error shows that importance separation can be considered as a good method for approximate calculation of integrals only if n is not very large. Therefore when we translate this conclusion in the terms of integral equation, it means that the von Neumann series has to converge quickly.

4 Parallel Implementation

In this section we present the parallel importance separation for evaluation of multiple integrals and solving integral equations. The crude Monte Carlo possesses inherent parallelism which is based on the possibility to calculate simultaneously realizations of the random variable on different processors. For our algorithm (importance separation) we have some additional work: partitioning of the domain. We consider a multiprocessor configuration with p nodes.
N uniformly distributed random points $x_i \in [0;1]^d,\ \ i = 1, \ldots, N$ are used to obtain an approximation with given accuracy of the integral (1). For generation of $d-$dimensional random point we need d random numbers. To estimate the performance of the parallel algorithms we use:

$\mathbf{ET_p(A)}$ mathematical expectation of time, required for a set of p processing elements to solve the problem using algorithm A

$\mathbf{S_p(A)} = \dfrac{\mathbf{ET_1(A)}}{\mathbf{ET_p(A)}}$ speed-up

$\mathbf{E_p(A)} = \dfrac{\mathbf{S_p(A)}}{\mathbf{p}}$ parallel efficiency.

5 Numerical Experiments

We present the numerical results (accuracy, CPU-time in seconds, parallel efficiency, speed-up) for the considered algorithm, importance separation, applied to solving of multidimensional integrals and integral equations using pseudorandom number sequences. The numerical tests are implemented on a cluster of 4 two-processor computers Power Macintosh using MPI.

5.1 Calculation of Integrals

Here we present the results of solving of a multidimensional integral, which is used as a test example in [10].
Example 1. This example is Monte Carlo integration over $I^5 = [0,1]^5$ of the function

$$f_1(x) = exp\left(\sum_{i=1}^{5} a_i\, x_i^2\, \frac{2 + sin(\sum_{j=1, j\neq i}^{5} x_j)}{2}\right),$$

where $\mathbf{a} = (1, \frac{1}{2}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$. The numerical results for the accuracy of the described methods for computing the multidimensional quadratures are presented in Table 1. The results are presented as a function of N, number of samples, and as a function of the error, which is computed with respect to the exact solution. The importance separation method leads to smaller errors. The most important fact is that using importance separation we have very good accuracy even using small sample. The superior behavior of importance separation is illustrated also by Figure 1. Table 2 presents the achieved efficiency of the parallel implementation (using MPI) for the considered method. The speed-up is almost linear and the efficiency grows with the increase number of samples.

Table 1. Comparison between Crude MCM, Importance sampling, Importance separation for Example 1 (calculations are implemented on one processor)

N	Crude MCM		Imp. sampling		Imp. separation	
	$\|I - I_N\|$	T_1	$\|I - I_N\|$	T_1	$\|I - I_N\|$	T_1
100	0.009532	0.001	0.081854	0.008	0.000316	6
500	0.092960	0.004	0.007102	0.036	0.000003	31
2500	0.009027	0.020	0.006381	0.175	0.000068	152
10000	0.006611	0.076	0.004673	0.697	0.000061	610
50000	0.008443	0.386	0.003212	3.489	0.000021	3047

Table 2. Implementation of IS for Example 1 using MPI ($I = 2.923651$)

Importance separation					
$N = 1000$			$N = 10000$		
p	I_N	E_p	p	I_N	E_p
1	2.923604	1	1	2.923590	1
2	2.923603	0.979	2	2.923573	0.985
3	2.920636	0.967	3	2.923336	0.983
4	2.923804	0.941	4	2.923638	0.980
5	2.923463	0.934	5	2.923602	0.979
6	2.911825	0.925	6	2.922537	0.977

5.2 Solving Integral Equations

We use the following integral equation as a test **Example 2**:

$$u(x) = \int_{\Omega} k(x, x')u(x')\,\mathrm{d}x' + f(x), \qquad \text{where}$$

$$k(x, x') = \frac{0.055}{1 + \mathrm{e}^{-3x}} + 0.07, \qquad (\|\mathcal{K}\|_{L_2} \approx 0.2) \tag{8}$$

$$f(x) = 0.02(3x^2 + \mathrm{e}^{-0.35x}), \qquad \Omega \equiv [-2; 2].$$

This kind of equation describes some neuron networks procedures. We are interested in an approximate calculation of (φ, u), where $\varphi(x) = 0.7((x + 1)^2 \cos(5x) + 20)$. The results for parallel efficiency are presented as a function of the number of nodes p. The importance separation algorithm is constructed so that only one sample of the random variable is chosen in every subdomain. The number of iterations d is fixed, but it has been chosen in advance according to the L_2-norm of the kernel (8). For the approximate computation of any integral $I(j)$, $j = 0, \ldots, i$ different number of samples are used in order to have error balancing. The proposed parallel algorithm for importance separation shares the obtained subdomains among the processors. Thus, every node processes the corresponding set of subdomains independently. This fact and insignificant quantity of data that is transferred determine comparatively high parallel efficiency of the algorithm. The results for the achieved efficiency are given in Table 3, which illustrates the inherent parallelism of Monte Carlo methods.

Table 3. Relative error, CPU-time (in seconds) and parallel efficiency in the approximate calculation of (φ, u) for Example 2. The number of samples used for calculation of each of the integrals $I(j)$, $j = 0, \ldots, 5$ is denoted by N_j

$N_0 = 240$, $N_1 = 18^2$ $N_2 = 12^3$, $N_3 = 6^4$ $N_4 = 6^5$, $N_5 = 6^6$				$N_0 = 480$, $N_1 = 24^2$ $N_2 = 18^3$, $N_3 = 12^4$ $N_4 = 6^5$, $N_5 = 6^6$				$N_0 = 480$, $N_1 = 30^2$ $N_2 = 24^3$, $N_3 = 12^4$ $N_4 = 12^5$, $N_5 = 6^6$			
p	Rel. error	Time	E_p	p	Rel. error	Time	E_p	p	Rel. error	Time	E_p
1	0.1221	0.63	1	1	0.0014	0.81	1	1	0.0009	3.18	1
2	0.1079	0.31	0.997	2	0.0016	0.40	0.997	2	0.0005	1.59	0.999
3	0.0994	0.21	0.988	3	0.0036	0.27	0.989	3	0.0005	1.06	0.996
4	0.0272	0.16	0.976	4	0.0122	0.21	0.979	4	0.0010	0.80	0.994
6	0.1986	0.11	0.962	6	0.0046	0.14	0.967	6	0.0036	0.53	0.990

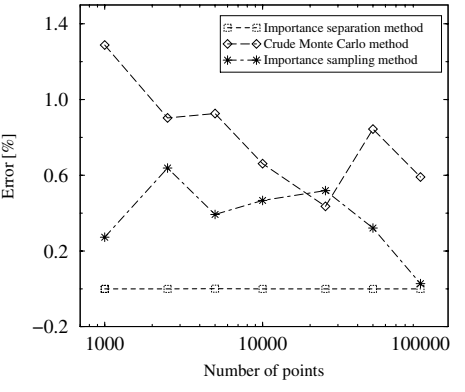


Fig. 1. Comparison of the accuracy of Crude MCM, Importance sampling, and Importance separation for Example 1

Acknowledgments. This work is supported by Center of Excellence BIS-21 Grant ICA1-2000-70016 and by the Ministry of Education and Science of Bulgaria under Grants # I 1201/02 and # MM 902/99.

References

1. N. S. Bahvalov. On the optimal estimations of convergence of the quadrature processes and integration methods, *Numerical Methods for Solving Differential and Integral Equations*, Nauka, Moscow, 5–63, 1964, (in Russian).
2. I. Dimov. Minimization of the probable error for some Monte Carlo methods, Mathematical Modelling and Scientific Computations, Andreev, Dimov, Markov, Ulrich (Eds.), Bulgarian Academy of Sciences, Sofia, 159-170, 1991.
3. I. Dimov, A. Karaivanova, R. Georgieva, and S. Ivanovska, Parallel Importance Separation and Adaptive Monte Carlo Algorithms for Multiple Integrals, Numerical Methods and Applications (I. Dimov, I.Lirkov, S. Margenov, and Z. Zlatev - Eds.), LNCS 2542, 99-107, Springer, 2003.

4. V. Dupach. Stochasticke pocetni metody, Cas. pro pest. mat. 81(1), 55-68, 1956.
5. Rayna Georgieva and Sofiya Ivanovska. Importance Separation for Solving Integral Equations, Large-Scale Scientific Computing (I. Lirkov, S. Margenov, J. Wasniewski, and P. Yalamov - Eds.), LNCS 2907, 144-152, Springer, 2004.
6. T. Hesterberg. Weighted average importance sampling and defensive mixture distributions, *Technometrics*, 37(2), 185–194, 1995.
7. H. Kahn. Random sampling (Monte Carlo) techniques in neutron attenuation problems, Nucleonics 6(5), 27-33, 1950; 6(6), 60-65, 1950.
8. A. Karaivanova. Adaptive Monte Carlo methods for numerical integration, *Mathematica Balkanica*, 11, 391–406, 1997.
9. G. A. Mikhailov. Optimization of the "weight" Monte Carlo methods, Moskow, 1987.
10. B. Moskowitz and R. E. Caflisch. Smoothness and dimension reduction in quasi-Monte Carlo methods, *J. Math. Comput. Modeling*, **23**: 37–54, 1996.
11. A. Owen and Y. Zhou. Safe and effective importance sampling, Technical report, Stanford University, Statistics Department, 1999.
12. I. M. Sobol. *Monte Carlo Numerical Methods*, Nauka, Moscow, 1973, (in Russian).
13. E. Veach and L. J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering, *Computer Graphics Proceedings*, Annual Conference Series, *ACM SIGGRAPH '95*, 419–428, 1995.
14. E. Veach. *Robust Monte Carlo Methods for Light Transport Simulation*, Ph.D. dissertation, Stanford University, 1997.

Investigation of the Sensitivity of the Monte Carlo Solution for the Barker-Ferry Equation with Sequential and Parallel Pseudo-Random Number Generators*

T.V. Gurov[1] and P.A. Whitlock[2]

[1] IPP - BAS, Acad. G. Bonchev st, bl. 25 A,
1113 Sofia, Bulgaria, gurov@parallel.bas.bg
[2] Dep. of CIS, BC-CUNY, 2900 Bedford Ave,
Brooklyn, NY 11210, USA, whitlock@sci.brooklyn.cuny.edu

Abstract. In this work a set of several sequential and parallel pseudo-random number generators (prng's) are tested when we apply a MC approach for solving a quantum-kinetic equation derived from ultra-fast transport in semiconductos. The aim is an optimization of the MC solver for the equation which accounts for quantum effects in the electron-phonon interaction. We consider so-called the Barker-Ferry (B-F) equation written for the case of zero electric field. The original formulation of this equation accounts for the action of the electric field during the process of collision.
The sensitivity of the MC solution for the electron energy distribution is investigated empirically, using prng's under consideration. The results obtained for the computational cost of the MC algorithm, the accuracy and the bias in the MC solution can be used to guide the choice of rng in the general case.

1 The Quantum-Kinetic Equation

The Barker-Ferry equation [1] can be utilized as a relevant physical model of the femtosecond relaxation of initially excited by a laser pulse electrons. For zero electrical field, the equation can be written in the following integral form [2]:

$$f(\mathbf{k},t) = \int_0^t dt' \int_0^{t'} dt'' \int d^3\mathbf{k}' \{S(\mathbf{k}',\mathbf{k},t'-t'')f(\mathbf{k}',t'') - S(\mathbf{k},\mathbf{k}',t'-t'')f(\mathbf{k},t'')\} + \phi(\mathbf{k}), \tag{1}$$

$$S(\mathbf{k}',\mathbf{k},t'-t'') = \frac{2V}{(2\pi)^3\hbar^2}|g_{\mathbf{k}'-\mathbf{k}}|^2 \exp(-\Gamma(\mathbf{k}',\mathbf{k})(t'-t'')) \times\{(n_{\mathbf{q}}+1)\cos(\Omega(\mathbf{k}',\mathbf{k})(t'-t'')) + n_{\mathbf{q}}\cos(\Omega(\mathbf{k},\mathbf{k}')(t'-t''))\}, \tag{2}$$

* Supported by ONR Grant N00014-96-1-1-1057 and by Center of Excellence BIS-21 grant ICA1-2000-70016, as well as by the NSF of Bulgaria through grant number I-1201/02.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 507–514, 2004.
© Springer-Verlag Berlin Heidelberg 2004

where $\mathbf{k}$ is the momentum, $f(\mathbf{k},t)$ is the distribution function and $\phi(\mathbf{k})$ is the positive initial condition. In the kernel (2), $n_{\mathbf{q}}$ is the Bose function [2], V is the volume and $\Omega(\mathbf{k}',\mathbf{k}) = (\varepsilon(\mathbf{k}') - \varepsilon(\mathbf{k}) - \hbar\omega_{\mathbf{q}})/\hbar$. The phonon energy is $\hbar\omega_{\mathbf{q}}$, which generally depends on $\mathbf{q} = \mathbf{k}' - \mathbf{k}$, and $\varepsilon(\mathbf{k}) = (\hbar\mathbf{k})^2/2m$ is the electron energy. The coupling $g_{\mathbf{k}'-\mathbf{k}} = -i\left[\frac{2\pi e^2\hbar\omega_{\mathbf{q}}}{V}\left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s}\right)\frac{1}{(\mathbf{k}'-\mathbf{k})^2}\right]^{\frac{1}{2}}$ applies to the Fröhlich interaction, and (ϵ_∞) and (ϵ_s) are the optical and static dielectric constants. The damping factor $\Gamma(\mathbf{k}',\mathbf{k}) = \Gamma(\mathbf{k}') + \Gamma(\mathbf{k})$ is related to the finite carrier lifetime for the scattering process: $\Gamma(\mathbf{k}) = \int d^3\mathbf{k}' \frac{V}{2^3\pi^2\hbar} \sum_{\pm} \|g_{\mathbf{k}'-\mathbf{k}}\|^2 \delta(\varepsilon(\mathbf{k}') - \varepsilon(\mathbf{k}) \pm \hbar\omega_{\mathbf{q}})(n_{\mathbf{q}} + \frac{1}{2} \pm \frac{1}{2})$. In spherical coordinates (k,θ,φ), with the k'_z axis oriented along $\mathbf{k}$ and zero lattice temperature ($n_{\mathbf{q}} = 0$), the equation (1) becomes one-dimensional with respect to the momentum integration [3]:

$$f(k,t) = \int_0^t dt'' \int_0^Q dk' K(k,k') \times \tag{3}$$

$$\times \left[K_1(k,k',t,t'')f(k',t'') + K_2(k,k',t,t'')f(k,t'')\right] + \phi(k),$$

$$K(k,k') = c_1 \frac{k'}{k} \ln\left(\frac{k+k'}{|k-k'|}\right),$$

$$K_1(k,k',t,t'') = -K_2(k',k,t,t'') = \frac{1}{\Omega_{k',k}^2 + \Gamma_{k',k}^2}\left\{\Gamma_{k',k} + \right.$$

$$\left. + \exp(-\Gamma_{k',k}(t-t''))\left[\Omega_{k',k} sin(\Omega_{k',k}(t-t'')) - \Gamma_{k',k} cos(\Omega_{k',k}(t-t''))\right]\right\}$$

and $c_1 = e^2\omega \left|\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s}\right| / (\pi\hbar)$. The functions $\Gamma(\mathbf{k}',\mathbf{k})$ and $\Omega(\mathbf{k}',\mathbf{k})$ depend only on the radial variables k and k' and are denoted by $\Gamma_{k',k}$ and $\Omega_{k',k}$, respectively where

$$\Gamma_k = \begin{cases} c_2 \ln\left((k + \sqrt{k^2 - \omega_1})/\sqrt{\omega_1}\right)/k, & \text{if } k^2 \geq \omega_1 \\ 0, & \text{if } k^2 < \omega_1, \end{cases}$$

with $\omega_1 = 2m\omega_q/\hbar$, $\quad c_2 = (me^2\omega_q/\hbar^2)|1/\epsilon_\infty - 1/\epsilon_s|$. The Neumann series corresponding to equation (3) converges [3] and a MC approach can be applied to evaluate the electron energy distribution. We note that this approach can be generalized for finite temperatures in a straightforward way.

2 Monte Carlo Approach

Define a terminated Markov chain $(\kappa_0,\tau_0) \to \ldots \to (\kappa_j,\tau_j) \to \ldots \to (\kappa_{l_\varepsilon},\tau_{l_\varepsilon})$, such that every point $(\kappa_j,\tau_j) \in (0,Q) \times (0,\tau_{j-1}), \quad j = 1,2,\ldots,l_\varepsilon$ (ε is the truncation parameter) is sampled using an arbitrary transition density function $r(k,k',t,t'')$ which is tolerant[1] to both kernels in equation (3).

The biased Monte Carlo estimator for the solution of equation (3) at the fixed point $k = \kappa_0$ at the time $t = \tau_0$ using backward time evolution of the numerical trajectories has the following form:

[1] $r(x)$ is tolerant of $g(x)$ if $r(x) > 0$ when $g(x) \neq 0$ and $r(x) \geq 0$ when $g(x) = 0$.

$$\xi_{l_\varepsilon}[\kappa_0,\tau_0] = \phi(\kappa_0) + \sum_{j=1}^{l_\varepsilon} W_j^\alpha \phi_\alpha(\kappa_j), \tag{4}$$

$$W_j^\alpha = W_{j-1}^\alpha \frac{K(\kappa_{j-1},\kappa_j)K_\alpha(\kappa_{j-1},\kappa_j,\tau_{j-1},\tau_j)}{p_\alpha r(\kappa_{j-1},\kappa_j,\tau_{j-1},\tau_j)},\ W_1^\alpha = 1,\ \alpha = 1,2,\ j = 0,\ldots,l_\varepsilon.$$

The probabilities p_α $(\alpha = 1, 2)$ are related to the choice of one of the kernels. Now we can define a Monte Carlo method

$$\frac{1}{N}\sum_{i=1}^{N}(\xi_{l_\varepsilon}[\kappa_0,\tau_0])_i \xrightarrow{P} f(\kappa_0,\tau_0), \tag{5}$$

where $\xi_{l_\varepsilon}[\kappa_0,\tau_0])_1, \xi_{l_\varepsilon}[\kappa_0,\tau_0])_2,\ldots,\xi_{l_\varepsilon}[\kappa_0,\tau_0])_N$ are independent values of the estimator (4) and $\xrightarrow{P}$ means stochastic convergence as $N \to \infty$. The relation (5) still does not determine the computation algorithm: we must specify the modeling function (sampling rule) $\xi_{l_\varepsilon}[\kappa_0,\tau_0] = g(\beta_1,\ldots,\beta_n)$, where $\beta_1,\ldots,\beta_n$ are uniformly distributed random numbers in the interval (0, 1). Now both relations (5) and the sampling rule define a Monte Carlo algorithm for (4).

Thus we can say [4] the constructive dimension (c.d.) of the algorithm is n, i.e. $c.d. = n$. Clearly, the variance of the MC estimator (4) does not depend on the c.d. Nevertheless, the c.d. has suggested a classification of sampling rules and an ordering of tests for pseudo-random numbers.

The transition density function in the Markov chain can be chosen in the following way $r_\alpha(k,t,k',t'') = r(k,k')r(t,t''/k,k')$, $\alpha = 1, 2$, where

$$r(k,k') = \overline{C}\frac{k'}{k}\ln\left(\frac{k+k'}{|k-k'|}\right) \quad \text{and} \quad r(t,t''/k,k') = \frac{\Gamma_{k,k'}\exp(-\Gamma_{k,k'}(t-t''))}{1-\exp(-\Gamma_{k,k'}t)}.$$

The normalized density function $r(k,k')$ can be expressed as an infinite weighted sum of other density functions by expanding $(k'/k)\ln((k+k')/(|k-k'|))$, i.e.

$$r(k,k') = \sum_{i=0}^{\infty}\overline{C}_i r_i(k,k'),\quad \overline{C}_i \geq 0,\quad \sum_{i=0}^{\infty}\overline{C}_i = 1,$$

$$r_i(k,k') = \begin{cases} (2i+3)\frac{(k')^{2i+2}}{k^{2i+3}}, & \text{when } 0 \leq k' < k \\ (2i-1)\left[\frac{(Qk)^{2i-1}}{Q^{2i-1}-k^{2i-1}}\frac{1}{(k')^{2i}}\right], & \text{when } k < k' \leq Q, \end{cases}$$

$$\overline{C}_i = \begin{cases} \frac{2}{(2i+1)(2i+3)}, & \text{when } 0 \leq k' < k \\ \frac{\frac{4k^2}{(4i^2-1)}\left(1-(\frac{k}{Q})^{2i-1}\right)}{(Q-k)\left[2k+(Q+k)\ln\left(\frac{Q+k}{Q-k}\right)\right]}, & \text{when } k < k' \leq Q. \end{cases}$$

The decomposition MC approach can be applied to sample k': 1. Generate β_1, β_2, β_3 uniform on [0, 1]; 2. Define $\overline{C}_i$ by β_1 using decomposition MC techniques. 3. Sample k' with the i-th density function $r_i(k,k')$, namely, $k' = k(\beta_3)^{\frac{1}{2i+1}}$, if $\beta_2 Q < k$. Otherwise, $k' = k/[1-\beta_3(1-(k/Q)^{2i-1})]^{\frac{1}{2i-1}}$.

Using the normalized conditional probability density function $r(t,t''/k,k')$ we can sample $t'' = \log(\beta_4(\exp(\Gamma_{k,k'}t)-1)+1)/\Gamma_{k,k'}$, where $\beta_4 \in (0, 1)$. Finally, we

generate $\beta_5 \in (0, 1)$ and choose one of the kernels $K_\alpha(k, k', t, t''),\ \alpha = 1, 2$ using probabilities $p_\alpha = |K_\alpha(k, k', t, t'')|/(|K_1(k, k', t, t'')| + |K_2(k, k', t, t'')|)$. Summarizing, we have used 5 uniform random numbers $\beta_1, \ldots, \beta_5$ in order to construct the MC estimator (4) for one transition $(k, t) \to (k', t'')$ in the Markov chain.

The computational complexity of the obtained iterative MC algorithm can be measured by the quantity $F = N \times t_{n_0} \times E(l_\varepsilon)$. We note that the number of the random walks, N, and the average number of transitions in the Markov chain, $E(l_\varepsilon)$, are connected with the stochastic and systematic errors [3]. However the mean time for modeling one transition, t_{n_0}, $(n_0 = 5)$ depends on the complexity of the transition density functions and the choice of the random number generator. It is strongly recommended that all simulations be done with two or more different generators, and the results compared to check whether the prng is introducing a bias. The c.d. of this algorithm can be defined as the average number of uniformly distributed random numbers necessary for carrying out one trial, i.e. $c.d. = n_0 E(l_\varepsilon)$. Thus we can use parallel prng's that produce $n_0 = 5$ independent and non-overlapping random sequences in order to compute every transition in the Markov chain as well as sampling 5 consecutive pseudo-random numbers from a sequential generator.

3 Numerical Results and Discussions

The simulation results are obtained for $GaAs$ with material parameters taken from [2]. The initial condition is a Gaussian function of the energy. The solution $f(k, t)$ is estimated by the MC estimator in 60 points of the simulation domain between 0 and $Q = 66 \times 10^7/m$.

The iterative MC algorithm is realized using the following sequential prng's: 1. CLCG-PL, Combined linear congruential generator with parameters recommended by P. L'Ecuyer [5]; 2. EICG, Explicit inversive congruential generator [6]; 3. ICG, Inversive congruential generator [7]; 4. LCG-F, Linear congruential generator with parameters recommended by Fishman [8]; 5. LCG-PM , Linear congruential generator with parameters recommended by Park and Miller [9]; 6. MT-MN, Mersenne Twister generator by Matsumoto and Nishimura [10]; as well as the following parallel prng's: 1. SNWS, Shuffled nested Weyl sequences [11] with a multiplier $M = 1234567$. To produce 5 random sequences we use the following seeds: $\gamma = \{2^{1/2}\}, \{3^{1/2}\}, \{5^{1/2}\}, \{7^{1/2}\}$ and $\{11^{1/2}\}$. 2. SPRNG, the Scalable Parallel Random Number Generator Library [12].

The MC algorithm were implemented in the C language. Numerical tests were performed on a Sun Ultra Enterprise 450 with 4 Ultra-SPARC, 400 MHz CPUs running Solaris. In all our tests $\varepsilon = 0.0001$. Such a choice of the truncation parameter allow us to ignore the systematic error [3] and to investigate whether any generator under consideration is introducing a bias when different stochastic errors are fixed. The quantity presented on the y-axes in all figures below, $kf(k, t)$, is proportional to the electron energy distribution function multiplied by the density of states. The quantity k^2 given on the x-axes in units of $10^{14}/m^2$ is proportional to the electron energy. Figure 1 compares the solutions for evolu-

Table 1. Comparison of the accuracy of the solution obtained with the SNWS and SPRNG generators for the 3 points with the biggest variance. The evolution time is $100fs$ in the case $N = 1$ million (mln), $150fs$ in the case $N = 5$ million (mln) and $200fs$ in the case $N = 10$ million (mln), respectively.

		SNWS		SPRNG		
N	k	kf_1	μ	kf_2	μ	$\lvert f_1 - f_2\rvert$
	48.00	9.8192	± 0.0038	9.7896	± 0.0038	6.2×10^{-4}
1	48.75	10.4557	± 0.0039	10.4595	± 0.0039	7.8×10^{-5}
mln	49.50	10.7340	± 0.0039	10.7024	± 0.0039	6.4×10^{-4}
	48.00	14.6451	± 0.0041	14.7029	± 0.0042	1.2×10^{-3}
5	48.75	15.6877	± 0.0043	15.7370	± 0.0043	1.0×10^{-3}
mln	49.50	15.4964	± 0.0042	15.5394	± 0.0042	8.7×10^{-4}
	48.00	17.1504	± 0.0066	16.9636	± 0.0066	3.9×10^{-3}
10	48.75	18.2430	± 0.0066	18.0536	± 0.0067	3.9×10^{-3}
mln	49.50	17.7436	± 0.0064	17.6170	± 0.0064	2.6×10^{-3}

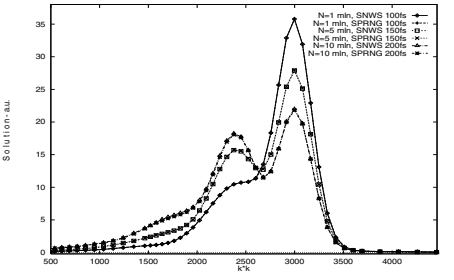


Fig. 1. Comparison of the electron energy distribution $kf(k,t)$ versus k^2 obtained by using of SNWS and SPRNG.

Table 2. Comparison of the computational complexity of the algorithm using the sequential and parallel prng's.

	$150fs,\ N = 150000$		$200fs,\ N = 750000$	
generator	CPU time	$E(l_\varepsilon)$	CPU time	$E(l_\varepsilon)$
MT-MN	15m41.80s	15.5084	1h20m51.78s	15.8995
LCG-PM	20m18.46s	15.5205	1h44m49.75s	15.9055
CLCG-PL	23m08.77s	15.5162	2h0m45.19s	15.9052
SNWS	23m17.10s	15.5300	1h59m54.47s	15.9155
SPRNG	24m50.11s	15.5085	2h6m20.08s	15.8982
EICG	30m51.42s	15.5265	2h38m47.89s	15.9095
LCG-F	30m56.13s	15.4899	2h38m58.54s	15.8891
ICG	47m31.30s	15.5153	3h2m44.73s	15.9023

tion times 100 femtoseconds (fs), $150fs$ and $200fs$ obtained by using the SNWS and, SPRNG parallel prng's. The number of realizations of the MC estimator (4) are 1 million (mln), 5 mln and 10 mln, respectively. We see that the solutions coincide. Table 1 shows the mean square error, μ, and the absolute error for the

3 values of the momentum k with the biggest variance using the SNWS and SPRNG generators. In this "the worst" case of the variance compared with the

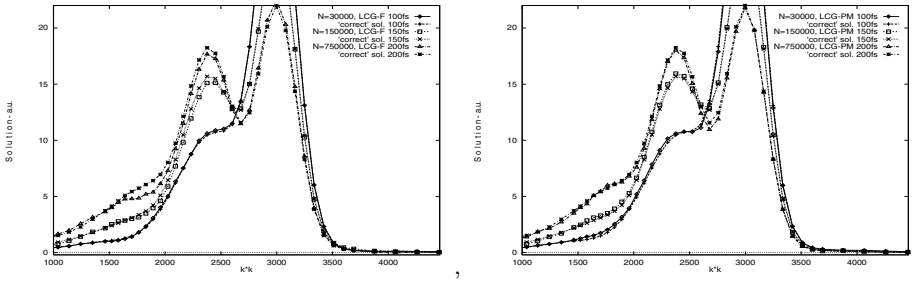


Fig. 2. Comparison of the electron energy distribution obtained by using of LCG-F and LCG-PM generators with the "correct" solution on the left and the right pictures, respectively.

variance at the other points, we have $\mu = O(10^{-3})$ and absolute errors are in agreement with the mean square error. Let us note that the exact solution of the B-F equation is unknown. Given the excellent agreement and similar variances, we can take any MC solution from Fig. 1 as a "correct" solution.

Figures 2 – 5 compare "correct" solutions (using the results with the SNWS generator) for the evolution times $100fs, 150fs$ and $200fs$ with the quantum solutions obtained using all the sequential and parallel prng's when the mean square error is $O(10^{-2})$. The number of realizations of the MC estimator are $N = 30000$, 150000 and 750000. Results obtained when $k^2 < 2000$ for $kf(k,t)$ using the sequential generators when compared with the "correct" solution show systematic differences. The best case occurs when using the CLG-PM, minimal standard, generator. However, for all times it exhibits small consistent differences. Systematic differences in the MC solution with increasing evolution time appear when LCG-F, ICG, EICG and MT-MN are used. Random "noise" in the MC solution is observed when the CLCG-PL generator is used, which, however, is unbiased. When $k^2 > 2000$ the results using any prng's disagree in the first peak of the distribution. This can be explained because the product $kf(k,t)$ for bigger values of k is sensitive to even small errors in the MC solution. Table 2 shows the computational complexity (CPU time for all 60 points) of the algorithm using all the prng's. We see that the computational cost is the least when the MT-MN generator is used. The CPU time of the algorithm using the SNWS and SPRNG parallel prng's is closely to the CLCG-PL sequential generator and they are faster than the EICG and LCG-F sequential generators. The ICG generator is twiceas slow as the parallel generators with the MC algorithm and therefore it should not be employed for solving this problem. Also, the quantity $E(l_\varepsilon)$ very slowly increases with increasing evolution time.

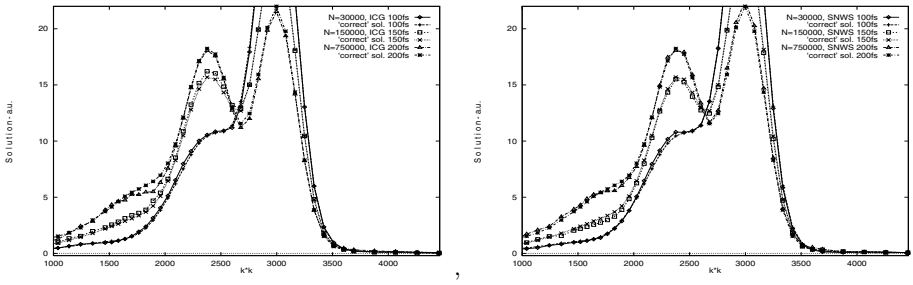


Fig. 3. Comparison of the electron energy distribution obtained by using of ICG and SNWS generators with the "correct" solution on the left and the right pictures, respectively.

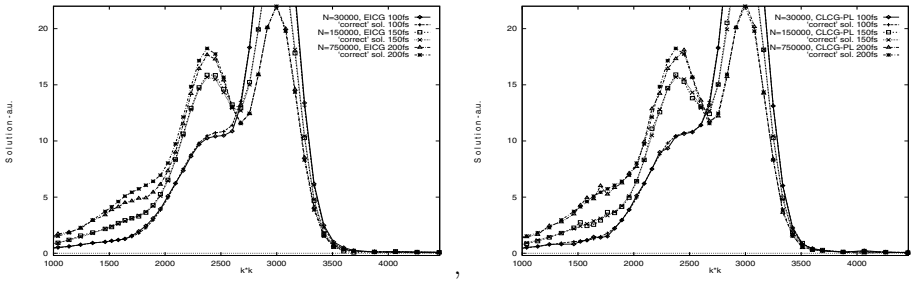


Fig. 4. Comparison of the electron energy distribution obtained by using of EICG and CLCG-PL generators with the "correct" solution on the left and the right pictures, respectively.

4 Summary

Statistically, the solution of the B-F equation would be expected to be noisier at $O(10^{-2})$ than at $O(10^{-3})$ mean square error. It is gratifying that the two parallel prng's used gave the same answer at $O(10^{-3})$ precision. However, even at $O(10^{-2})$ mean square error, if the solution was unbiased, we would expect random fluctuations about the more precise solution. This was only observed with the CLCG-PL prng. All the other sequential generators exhibited systematic rather than random differences.

Therefore we conclude that parallel prng's are preferable to solve this problem as the evolution time increases. In this case, the CPU time of the algorithm become crucial. Thus, to predict the solution we need parallel realizations of the algorithm and/or we have to estimate the solution with coarser stochastic error. To obtain a high parallel efficiency in the case of the parallel realization of the algorithm, the random sequences have to be produced with similar CPU times.

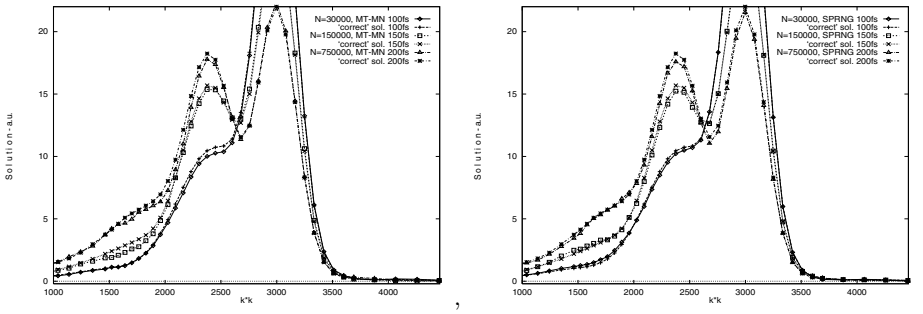


Fig. 5. Comparison of the electron energy distribution obtained by using of MT-MN and SPRNG generators with the "correct" solution on the left and the right pictures, respectively.

References

1. Barker, J., Ferry, D.: Self-scattering path-variable formulation of high field time-dependent quantum kinetic equations for semiconductor transport in the finite-collision-duration regime, *Phys. Rev. Lett.* **42**(26) (1979) 1779–1781.
2. Nedjalkov, M. et al: Statistical modeling of pulse excited electron quantum kinetics in a one-band semiconductor, *Math.& Comp. in Simul.* **47** (1998) 391–402.
3. Gurov, T.V., Whitlock, P.A.: An efficient backward Monte Carlo estimator for solving of a quantum kinetic equation with memory kernel, *Math. & Comp. in Simul.* **60** (2002) 85–105.
4. Sobol, I.M.: On quasi-Monte Carlo integration, *Math.& Comp. in Simul.* **47** (1998) 103–112.
5. L'Ecuyer, P.: Efficient and Portable Combined Random Number Generators, *Communications of the ACM* **31** (1988) 742–774.
6. Eichenauer-Hermann, J.: Statistical independence of a new class of inversive congruential pseudorandom numbers, *Math. Comp.* **60** (1993) 375–384.
7. Eichenauer, J., Lehn, J.: A non-linear congruential pseudo-random number generator, *Stat. Papers* **27** (1986) 315–326.
8. Fishman, G.S.: Multiplicative congruential random number generators with modulus 2^{β}: an exhaustive analysis for $\beta = 32$ and a partial analysis for $\beta = 48$, *Math. Comp.* **54** (1990) 331–344.
9. Park, S.K., Miller, K.W.: Random Number Generators: Good Ones Are Hard to Find, *Communications of the ACM* **31** (10) (1988) 1192–1201.
10. Matsumoto, M., Nishimura, T.: Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator, *ACM Trans. on Modeling & Comp. Simul.* **8** (1) (1998) 3–30.
11. Holian,B.L. et al: Pseudorandom number generator for massively parallel molecular-dynamics simulation, *Phys. Rev. E* **52**(2) (1994) 1607–1615.
12. *Scalable Parallel Random Number Generators Library for Parallel Monte Carlo Computations*, SPRNG 1.0 and SPRNG 2.0 – http://sprng.cs.fsu.edu.

Design and Distributed Computer Simulation of Thin p^+–i–n^+ Avalanche Photodiodes Using Monte Carlo Model

Mikhail Yakutovich

Belarussian State University

Abstract. The output current of an avalanche photodiodes (APD's) fluctuates in the absence of light as well as in its presence. The noise in APD's current arises from three sources: randomness in the number and in the positions at which dark carrier pairs are generated, randomness in the photon arrival number, and randomness in the carrier multiplication process. A Monte Carlo model has been used to estimate the excess noise factor in thin p^+–i–n^+ GaAs avalanche photodiodes. As this approach is computation intensive, simple parallel algorithm considering heterogeneous cluster based on MPICH was designed and implemented. Very good performance gain was achieved. It was found that APD model provides very good fits to the measured gain and noise and as such provides an accurate picture of the device operation. In this way, various device structures can be analyzed prior to their experimental realization. Through "computer experiments" like this outlined here, the effect of various geometries and material compositions on device performance can be assessed and optimal designs achieved.

1 Introduction

The avalanche photodiode (APD) is used in optical communications systems to convert a light signal into an electrical signal. The APD has internal gain, multiplying the signal current by the process of impact ionization in which a very energetic electron or hole creates a secondary electron-hole pair. A newly generated carrier must travel some distance (the dead space) in order to gain sufficient energy from the electric field to initiate an ionization event. The multiplication process introduces noise as a result of randomness in the ionization path length. The consequent variability in the generation of secondary carriers results in fluctuations in the total number of carriers produced per initial photocarrier, or multiplication. This noise component was first quantified by McIntyre [1,2] who found the following expression for the mean square noise current per unit bandwidth:

$$\langle i^2 \rangle = 2qI_p \langle M^2 \rangle F(\langle M \rangle) \tag{1}$$

where q is the electronic charge, I_p is the primary photo-current, $\langle M \rangle$ is the average current multiplication, and $F(\langle M \rangle)$ is the excess noise factor given by

$$F(\langle M \rangle) = k\langle M \rangle + (2 - 1/\langle M \rangle)(1 - k) \tag{2}$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 515–522, 2004.
© Springer-Verlag Berlin Heidelberg 2004

k is the ratio of the electron ionization coefficient α and the hole ionization coefficient β. If the primary photocarrier is a hole then $k = \alpha/\beta$ and if it is an electron then $k = \beta/\alpha$. However, recent experimental measurements on GaAs APD's [3,4,5,6] have shown a significant reduction in excess noise factor as i-region thickness decreases below one micron. A carrier starting with near zero energy, relative to the band edge, will have an almost zero chance of having an ionizing collision until it has gained sufficient energy from the electric field to attain the necessary energy to permit impact ionization [7,8]. Numerous analytical and numerical techniques have been proposed to address the nonlocal nature of impact ionization. Attempts to predict the ionization coefficients using Monte Carlo [9] and analytical techniques [10] have shown that, on average, carriers must travel a distance over which the potential drop is equivalent to $1.5 - 2$ times the ionization threshold energy before the probability of ionization, of a carrier which has not yet had an ionizing collision, rises to a steady-state, or "equilibrium," level.

Several techniques can potentially model the avalanche process while accounting for deadspace effects. These calculations would ideally be carried out using a Monte Carlo model with a full band structure (FBMC) calculated by the pseudopotential method, which provides the most realistic description of the transport. In recent years, full-band calculations have considerably advanced the understanding of impact ionization by showing that most carriers initiate events from higher lying bands producing secondary carriers with significant energy. The conventional Keldysh formula for the ionization rate, R_{ii} , which assumes a quadratic energy dependence, has also been shown to overestimate the ionization probability [11,12,13,14]. Stobbe [13] noted that different band structure assumptions can give different forms of the ionization rate which means that the accuracy of FBMC models for device simulations is questionable.

The lucky-drift (LD) model of Ridley [15,16] greatly simplifies the transport by using artificial trajectories based on the energy relaxation length which allows an analytic expression to be derived for the ionization coefficient. The complicated transport details are subsumed into a few material parameters which allows experimental results to be readily fitted and reveal a chemical trend. However, it was demonstrated in [17] that the use of energy relaxation path lengths to quantify phonon scattering in LD theory imposes a lower spatial limit of $0.1\mu m$. Furthermore, the model gives the incorrect spatial ionization probability which might lead to errors when calculating the excess noise factor associated with the avalanche process.

We used a simple Monte Carlo model (SMC) [18] for simulating thin APD's. It is an accurate, efficient and self-contained model for the avalanche process which can be used to predict both the multiplication and excess noise characteristics of all practical device geometries. Furthermore, this model allows experimental data to be fitted and interpreted with few parameters in a similar way to the LD model.

Since any Monte Carlo calculations are time consuming, which is especially restrictive when many ionization events need to be simulated to give reliable

statistics for the avalanche process, in this paper we present simple distribution algorithm. It takes into account the heterogeneous of cluster and allows achieving significant gain performance depending on contribution of any machine, independent of their relative speed.

Good results are shown between the calculated and measured multiplication and excess noise results from [3]. The simulation was tested on different heterogeneous clusters consisting of considerably different machines. Addition of relative "slower" machine leaded to achievement of gain, not deceleration. Our model can be used for simulation of complicated models utilizing relatively cheap clusters.

2 The Monte Carlo Model (MC)

We used a simple Monte Carlo model (SMC) [18] for simulating thin APD's. SMC uses single effective parabolic valleys and accurately accounts for deadspace effects. An effective parabolic valley is used for both electrons and holes which gives an energy independent mean-free path when phonon scattering is assumed to be solely by the deformation potential, which dominates at high electric fields.

To calculate the avalanche process in devices, the electric field profiles are generated in the depletion approximation assuming a built in voltage of $1.2V$. The values of the i-region thickness, w, the cladding doping, p^+ and n^+, and the unintentional p-type doping in the i-region, p^- were extracted from [6].

3 Estimation of Multiplication and Excess Noise Factor

The excess noise factor F is defined as the normalized second moment of the multiplication random variable M, when a single photocarrier initiates the multiplication. Thus,

$$F = \langle M^2 \rangle / \langle M \rangle^2 \tag{3}$$

where $\langle M \rangle$ is the mean multiplication and $\langle M^2 \rangle$ is the mean square multiplication.

The calculation uses an iterative scheme. The photo-generated electrons are first simulated yielding secondary electron and holes distributions. The secondary electrons are simply added to the electron simulation, and the calculation continues until all of the electrons are collected. The secondary hole distribution is then simulated based on the spatial of each particle's birth obtained from the initial electron simulation. Secondary holes arising from hole-initiated ionization effects are added to the hole simulation, and again, the calculation continues until all of the holes are collected. The electron simulation is then rerun with the secondary electrons. The total number of ionization events, N_t is recorded when all the carriers have left the multiplication region; the multiplication for that trial is then given by $M = N_t + 1$. By repeating the procedure for many trials, $\langle M \rangle$, and $\langle M^2 \rangle$ and F can be calculated. The number of trials is extended until successive values of differ by less than 0.0001. This usually requires at least 1000 trials for $\langle M \rangle = 2$ and $10000 - 50000$ trials for $\langle M \rangle$ larger than eight.

4 The Parallel MC Algorithm

Monte Carlo simulations of carrier transport in semiconductors are based on following the time evolution of an ensemble of particles through the material in both real and momentum space. The motion of each particle in the ensemble has to be simulated in turn, for the full duration of the simulation. It is assumed that these particles are effectively independent which makes the MC simulation well suited to parallel implementations to reduce computation time. The flow chart of a typical MC device simulation is shown in Fig. 1.

The parallel MC algorithm is based on a master-slave model [20]. The ensemble of particles is divided into subensembles, each of which is dedicated to a separate processor (slave). The slaves are solely responsible for simulating the particles' dynamics under the influence of the internal field distribution. The master processor updates the field distribution consistently with the port conditions enforced by the external circuitry. The master also serves as user interface. The MC algorithm will now be discussed with the aid of the flow chart (Fig. 2):

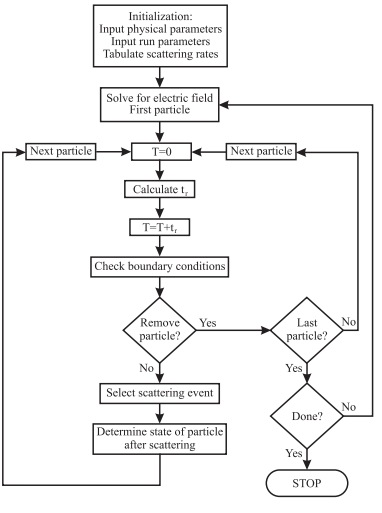


Fig. 1. Flow chart for a typical Monte Carlo device simulation algorithm

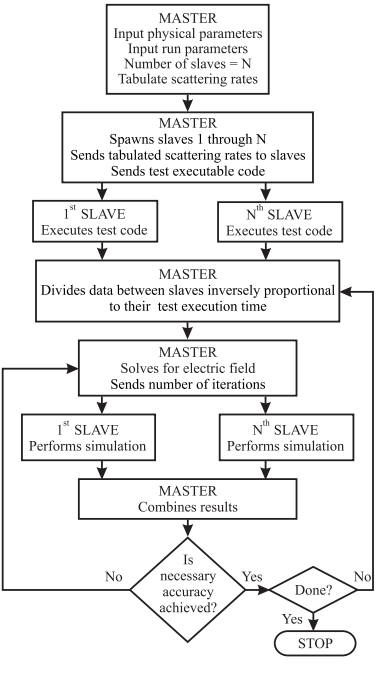


Fig. 2. Flow chart for a parallel MC device simulation algorithm

1. The master defines the physical device simulation problem and inputs the material and run parameters needed for the simulation. It also tabulates the various scattering rates as a function of particle energy.

2. The master spawns the slave executable code on N different slaves and sends the material parameters and tabulated scattering rates to each slave.
3. Each slave proceeds its own part until desired accuracy is attained.

To attain optimal efficiency, the computational load has to be shared among the processors in relation to their respective processing power. For a network of equally fast slave processors this implies that the number of particles in each subensemble must be kept equal throughout the simulation.

The main goal of dynamic load sharing is to equalize time τ_i of calculation on each slave machine, so the maximum theoretical performance gain will be achieved. Test computation performs on each machine before each simulation. Depending on test time τ_{ii}, the data is divided inversely proportional to τ_{ii}. So we can expand proposed algorithm:

1. The master defines the physical device simulation problem and inputs the material and run parameters needed for the simulation. It also tabulates the various scattering rates as a function of particle energy.
2. The master spawns the slave test executable code on N different slaves.
3. Each slave performs test code and sends execution time to master.
4. The master divides data between slaves inversely proportional to τ_{ii} and spawns the slave executable code on N different slaves and sends the material parameters and tabulated scattering rates to each slave.
5. Each slave proceeds its own part until desired accuracy is attained.

5 Results

The accuracy of the proposed parallel MC algorithm has been tested by comparing the results to those obtained by Plimmer [18]. The fit to the range of measured data for electrons and for holes is shown to be very good for M_e in Fig. 3 and for M_h in Fig. 4. In Fig. 5, the SMC-calculated excess noise factors are compared with the measured values from [3] for the range of p^+–i–n^+ structures with for the case of electron injection. This plot shows the calculated values to be in agreement with the measured results from [3] with the structure giving values close to those which would be predicted using the noise theory of McIntyre [1]. There is greater experimental uncertainties in measuring noise characteristics compared to the multiplication, but the MC predicts results close to experiment from all the diodes down to $0.1\mu m$ as shown on the plot.

The efficiency of the parallel SMC-algorithm have been quantified as the gain in computational speed achieved by employing multiple slaves relative to a master with single slave configuration and with different configurations. The modelling was executed on different homogeneous and heterogeneous clusters. The curves of the obtained speed-up as a function of the number of slaves are given in Fig. 6. In case of homogeneous cluster speed-up is close to ideal as our problem is perfectly suitable for parallelization.

Results of modelling in heterogeneous cluster are shown in Fig. 6 and Fig. 7. Every newly subsequent added computer had relatively lesser computing power.

The overall computing power of heterogeneous cluster was smaller then that of homogeneous cluster, hence, in general the overall speed-up in heterogeneous cluster reduced (Fig. 6). As it is shown in Fig. 7, our developed algorithm allowed efficiently distribute data between computers in a cluster, so the average idling time was greatly decreased. Addition of relative "slower" machine leaded to achievement of gain in any case. It is evident from the speed-up curves in Fig. 6 and Fig. 7 that the network communication plays a minor role in the efficiency of the algorithm, implying that a considerable improvement in computational speed is possible with the use of more powerful slave processors.

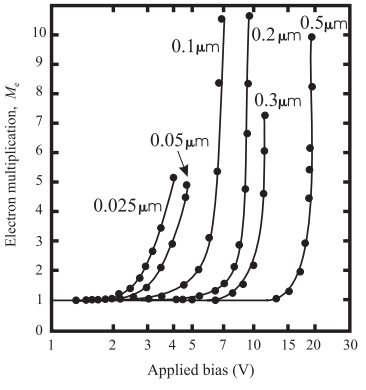


Fig. 3. Electron multiplication from MC calculations (•) compared against measured values (—) from pin's whose nominal i-region thicknesses are labelled on the plot

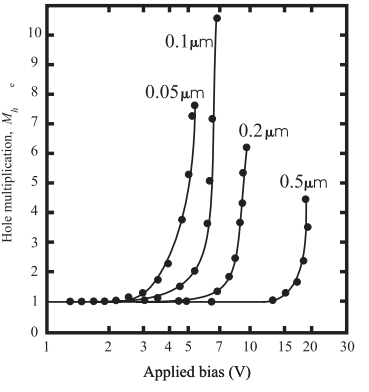


Fig. 4. Hole multiplication from MC calculations (•) compared against measured values (—) from pin's whose nominal i-region thicknesses are labelled on the plot

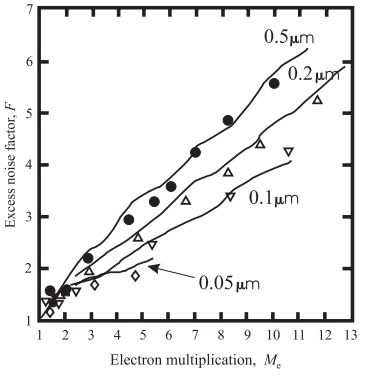


Fig. 5. SMC calculated excess noise, F, for electron injection from pin's with nominally $w = 0.5\mu m(\bullet)$, $0.2\mu m(\triangle)$, $0.1\mu m(\nabla)$ and $0.05\mu m(\diamondsuit)$ along with the measured results (—)

6 Conclusion

An efficient parallel implementation of the Monte Carlo particle simulation technique on a network of personal computers has been introduced. The parallel implementation have been successfully applied to the MC simulation of multiplication noise in GaAs p^+–i–n^+ avalanche photodiodes. These predict a decrease in excess noise factor as the multiplication length decreases from 1.0 to $0.05\mu m$ for both electron and hole injected multiplication. It was demonstrated that the excess noise factor depends strongly on the ionization path length distribution function.

Distributed computer simulation with dynamic load balancing greatly reduces computational time. Modelling was executed on different homogeneous and heterogeneous clusters. Addition of relative "slower" machine in heterogeneous cluster leaded to achievement of gain, not deceleration. Such algorithm can be widely used in different clusters.

Through "computer experiments" like this outlined here, the effect of various geometries and material compositions on device performance can be assessed and optimal designs achieved.

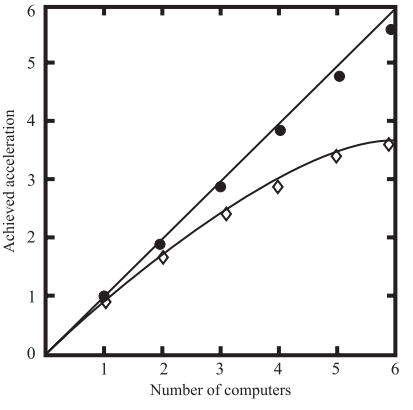


Fig. 6. The obtained speed-up curves for the MC-parallel algorithm in homogeneous (•) and heterogeneous (◇) clusters along with the ideal (—) speed-up curve

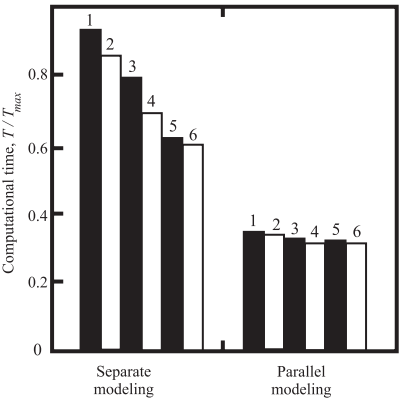


Fig. 7. Computational time for each processor working separately and in heterogeneous cluster

References

1. R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," *IEEE Trans. Electron Devices*, vol. ED-13, Jan. 1966.
2. R. J. McIntyre, "The distribution of gains in uniformly multiplying photodiodes: Theory," *IEEE Trans. Electron Devices*, vol. ED-19, pp.703-713, 1972.

3. K. F. Li, D. S. Ong, J. P. R. David, G. J. Rees, R. C. Tozer, P. N. Robson, and R. Grey, "Avalanche multiplication noise characteristics in thin GaAs p^+–i–n^+ diodes," *IEEE Trans. Electron Devices*, vol. 45, pp. 2102-2107, Oct. 1998.
4. C. Hu, K. A. Anselm, B. G. Streetman, and J. C. Campbell, "Noise characteristics of thin multiplication region GaAs avalanche photodiodes," *Appl. Phys. Lett.*, vol. 69, pp. 3734-3736, 1996.
5. K. F. Li, D. S. Ong, J. P. R. David, P. N. Robson, R. C. Tozer, G. J. Rees, and R. Grey, "Low excess noise characteristics in thin avalanche region GaAs diodes," *Electron. Lett.*, vol. 34, pp. 125-126, 1998.
6. S. A. Plimmer, J. P. R. David, D. C. Herbert, T.-W. Lee, G. J. Rees, P. A. Houston, R. Grey, P. N. Robson, A. W. Higgs, and D. R. Wight, "Investigation of impact ionization in thin GaAs diodes," *IEEE Trans. Electron Devices*, vol. 43, pp. 1066-1072, July 1996.
7. Y. Okuto and C. R. Crowell, "Energy-conservation considerations in the characterization of impact ionization in semiconductors," *Phys. Rev. B*, vol. 6, pp. 3076-3081, 1972.
8. Y. Okuto and C. R. Crowell, "Ionization coefficients in semiconductors: A nonlocal property," *Phys. Rev. B*, vol. 10, pp. 4284-4296, 1974.
9. K. F. Brennan, "Calculated electron and hole spatial ionization profiles in bulk GaAs and superlattice avalanche photodiodes," *IEEE J. Quantum Electron.*, vol. 24, pp. 2001-2006, 1988.
10. A. Spinelli, A. Pacelli, and A. L. Lacaita, "Dead space approximation for impact ionization in silicon," *Appl. Phys. Lett.*, vol. 69, no. 24, pp. 3707-3709, 1996.
11. J. Bude and K. Hess, "Thresholds of impact ionization in semiconductors," *J. Appl. Phys.*, vol. 72, pp. 3554-3561, 1992.
12. N. Sano, T. Aoki, and A. Yoshii, "Soft and hard thresholds in Si and GaAs," *Appl. Phys. Lett.*, vol. 55, pp. 1418-1420, 1989.
13. M. Stobbe, R. Redmer, and W. Schattke, "Impact ionization rate in GaAs," *Phys. Rev. B*, vol. 49, pp. 4494-4497, 1994.
14. H. K. Jung, K. Taniguchi, and C. Hamaguchi, "Impact ionization model for full band Monte Carlo simulation in GaAs," *J. Appl. Phys.*, vol. 59, pp. 2473-2480, 1996.
15. B. K. Ridley, "Lucky-drift mechanism for impact ionization in semiconductors," *J. Phys. C: Solid State Phys.*, vol. 16, pp. 3373-3388, 1983.
16. B. K. Ridley, "A model for impact ionization in wide-gap semiconductors," *J. Phys. C: Solid State Phys.*, vol. 16, pp. 4733-4751, 1983.
17. S. A. Plimmer, J. P. R. David, and G. M. Dunn, "Spatial limitations to the application of the Lucky-drift theory of impact ionization," *IEEE Trans. Electron Devices*, vol. 44, pp. 659-663, Apr. 1997.
18. S. A. Plimmer, J. P. R. David, D. S. Ong, K. F. Li, "A Simple Model for Avalanche Multiplication Including Deadspace Effects," *IEEE Trans. Electron Devices*, vol. 46, April 1999.
19. L. V. Keldysh, "Kinetic theory of impact ionization in semiconductors," *Sov. Phys.-JETP*, vol. 10, pp. 509-518, 1960.
20. Robert R. van Zyl, Willem J. Perold, Hans Grobler "A Parallel Implementation of the Monte Carlo Particle Simulation Technique on a Network of Personal Computers"

Convergence Proof for a Monte Carlo Method for Combinatorial Optimization Problems

Stefka Fidanova

IPP – BAS, Acad. G. Bonchev str. bl.25A, 1113 Sofia, Bulgaria
stefka@parallel.bas.bg

Abstract. In this paper we prove the convergence of a Monte Carlo (MC) method for Combinatorial Optimization Problems (COPs). The Ant Colony Optimization (ACO) is a MC method, created to solve efficiently COPs. The Ant Colony Optimization (ACO) algorithms are being applied successfully to diverse heavily problems. To show that ACO algorithms could be good alternatives to existing algorithms for hard combinatorial optimization problems, recent research in this area has mainly focused on the development of algorithmic variants which achieve better performance than previous one. In this paper we present ACO algorithm with Additional Reinforcement (ACO-AR) of the pheromone to the unused movements. ACO-AR algorithm differs from ACO algorithms in several important aspects. In this paper we prove the convergence of ACO-AR algorithm.

1 Introduction

Some time it is more important to find quickly good although not necessarily optimal solution. In this situation, the heuristic methods are with big efficient. For some difficult Combinatorial Optimization Problems (COPs) one or more months is needed to find an optimal solution on powerful computer and only some minutes to find solution by heuristic methods, which is very close to optimal one. Typical examples of practical COPs are the machine-scheduling problem, the net-partitioning problem, the traveling salesman problem, the assignment problem, etc.. Monte Carlo methods have been implemented to efficiently provide flexible and computerized procedures for solving many COPs.

ACO [1,2,3] is a MC method, created to solve COPs. It is a meta-heuristic procedure for quickly and efficiently obtaining high quality solutions to complex optimization problems [9]. ACO algorithm can be interpreted as parallel replicated Monte Carlo systems [11]. MC systems [10] are general stochastic simulation systems, that is, techniques performing repeated sampling experiments on the model of the system under consideration by making use of a stochastic component in the state sampling and/or transition rules. Experimental results are used to update some statistical knowledge about the problem, as well as the estimate of the variables the researcher is interested in. In turn, this knowledge can be also iteratively used to reduce the variance in the estimation of the described variables, directing the simulation process toward the most interesting state space

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 523–530, 2004.
© Springer-Verlag Berlin Heidelberg 2004

regions. Analogously, in ACO algorithms the ants sample the problem's solution space by repeatedly applying a stochastic decision policy until a feasible solution of the considered problem is built. The sampling is realized concurrently by a collection of differently instantiated replicas of the same ant type. Each ant "experiment" allows to adaptively modify the local statistical knowledge on the problem structure. The recursive retransmission of such knowledge determines a reduction in the variance of the whole search process the so far most interesting explored transitions probabilistically bias future search, preventing ants to waste resources in not promising regions of the search.

In this paper, the basic ACO algorithm has been modified and a convergence proof is presented. The ACO algorithms were inspired by the observation of real ant colonies [1,2,4]. Ants are social insects, they live in colonies and whose behavior is directed more to the survival of the colony as a whole than to that of a single individual component of the colony. An interesting behavior is how ants can find the shortest paths between food sources and their nest. While walking from a food source to the nest and vice-versa, ants deposit on the ground a substance called pheromone. Ants can smell pheromone and then they tend to choose, in probability, paths marked by strong pheromone concentrations. The pheromone trail allows the ants to find their way back to the food source (or to the nest).

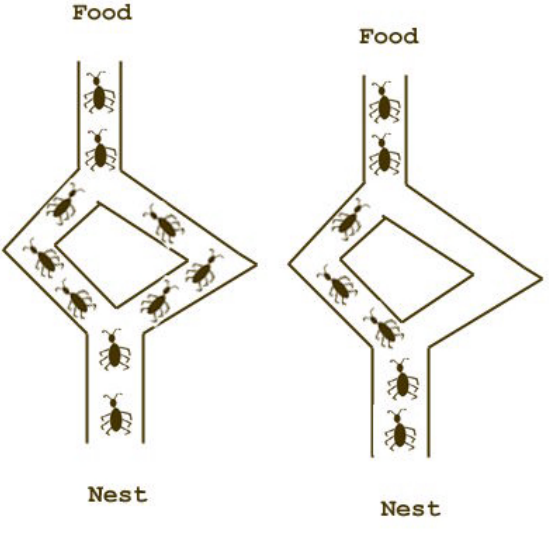


Fig. 1. Behavior of real ants at the beginning of the search and after some minutes

Figure 1 shows how ants can exploit pheromone to find a shortest path between two points. In this, figure ants arrive at a decision point in which they have to decide to turn on the left or on the right. The ants turning on the left first achieve the food sours. When they return back there is a pheromone only in a left side and they choose it and double the pheromone. Thus, after a short transition period the difference in the amount of pheromone on the two paths is sufficiently large and the new ants will prefer in probability to choose the left path, since at the decision point they receive a greater amount of pheromone on the left path. Very soon all ants will be using the shorter path.

The above behavior of real ants has inspired ACO algorithm. ACO algorithm, which is a population-based approach, has been successfully applied to many NP-hard problems [3,7,8]. One of its main ideas is the indirect communication among the individuals of ant colony. This mechanism is based on an analogy with trails of pheromone which real ants use for communication. The pheromone trails are a kind of distributed numerical information which is modified by the ants to reflect their experience accumulated while solving a particular problem.

The main purpose of this paper is to use additional reinforcement of the pheromone to the unused movements and thus to effectively avoid stagnation of the search and to prove the convergence of ACO-AR to the global optimum. The remainder of this paper is structured as follows. Section 2 describes the developed ACO-AR algorithm, while section 3 investigates its convergence. Section 4 shows parameter settings. The paper ends with conclusions and some remarks.

2 The ACO Algorithm

The ACO algorithms make use of simple agents called ants which iteratively construct candidate solutions to a COPs. The ants' solution construction is guided by pheromone trail and problem dependent heuristic information. The ACO algorithms can be applied to any COP by defining solution components which the ants use to iteratively construct candidate solutions and on which they may deposit a pheromone. An individual ant constructs a candidate solution by starting with a random partial solution and then iteratively adding new components to their partial solution until a complete candidate solution is generated. We will call each point at which an ant has to decide which solution component to add to its current partial solution a choice point. After the solution is completed, ants give feedback on their solutions by depositing pheromone on the components of their solutions. Typically, solution components which are part of the best solution, or are used by many ants, will receive a higher amount of pheromone and hence will be more attractive by the ants in following iterations. To avoid the search getting stuck before the pheromone trails get reinforced pheromone trails are decreased.

In general, all ACO algorithms adopt specific algorithmic scheme as follows. After the initialization of the pheromone trails and control parameters, a main loop is repeated until the stopping criteria are met. The stopping criteria can be a certain number of iterations or a given CPU-time limit. In the main loop, the

ants construct feasible solutions, then the pheromone trails are updated. More precisely, partial solutions are seen as follow: each ant moves from a state i to another state j of the partial solution. At each step, ant k computes a set of feasible expansions to its current state and moves to one of these expansions, according to a probability distribution specified as follows. For ant k, the probability p_{ij}^k of moving from state i to a state j depends on the combination of two values:

$$p_{ij}^k = \begin{cases} \frac{\tau_{ij}\eta_{ij}}{\sum_{l \in allowed_k} \tau_{il}\eta_{il}} & \text{if } j \in allowed_k \\ 0 & otherwise \end{cases} \quad (1)$$

where:

- η_{ij} is the attractiveness of the move as computed by some heuristic information indicating the a prior desirability of that move;
- τ_{ij} is the pheromone trail level of the move, indicating how profitable it has been in the past to make that particular move (it represents therefore a posterior indication of the desirability of that move).
- $allowed_k$ is the set of remaining feasible states.

Thus, the higher the value of the pheromone and the heuristic information, the more profitable it is to include state j in the partial solution. In the beginning, the initial pheromone level is set to τ_0, which is a small positive constant. While building a solution, ants change the pheromone level of the elements of the solutions by applying the following updating rule:

$$\tau_{ij} \leftarrow \rho\tau_{ij} + \Delta\tau_{ij} \quad (2)$$

where in the rule $0 < \rho < 1$ models evaporation and $\Delta\tau_{ij}$ is different for different ACO algorithms. In ant system [1], the first ant algorithm, all ants change the pheromone level depending to the quality of their solution. In ant colony system [2] extra pheromone is put on the elements of the best solution. In ACO algorithms with elitist ants [3] only small number of ants update the pheromone and so on.

Stagnation situation may occur when we perform the ACO algorithm. This can be happened when the pheromone trail is significantly higher for one choice than for all others. This means that one of the choices has a much higher pheromone level than the others and an ant will prefer this solution component over all alternatives. In this situation, ants construct the same solution over and over again and the exploration of the search space stops. The stagnation situation should be avoided by influencing the probabilities for choosing the next solution component which depend directly on the pheromone trails.

The aim of the paper is to develop the functionality of the ACO algorithms by adding some diversification such as additional reinforcement of the pheromone. This diversification guides the search to areas in the search space which have not been yet explored and forces ants to search for better solutions. We will call the

modified ACO algorithm with additional reinforcement [5,6]. If some movements are not used in the current iteration, additional pheromone reinforcement will be used as follows.

$$\tau_{ij} \leftarrow \alpha\tau_{ij} + q\tau_{max} \quad \alpha = \begin{cases} 1 \text{ if unused movements are evaporated} \\ \rho \text{ otherwise} \end{cases} \tag{3}$$

where $q \geq 0$, τ_{max} is the maximal value of the pheromone. Using ACO-AR algorithm the unused movements have the following features:

- they have great amount of the pheromone then the movements belonging to poor solutions.
- they have less amount of the pheromone then the movements belonging to the best solution

Thus the ants will be forced to choose new direction of search space without repeating the bad experience.

3 Convergence of the ACO-AR Algorithm

This section describes the convergence of the ACO-AR algorithm to the global optimum. We will use the Theorem 1 from [12], which proves that if the amount of the pheromone has a finite upper bound τ_{max} and a positive lower bound τ_{min}, then the ACO algorithm converges to the optimal solution. From the Proposition 1 (see [12]),the upper bound of the pheromone level is $g(s^*)/(1-\rho)$, where $g(s^*)$ is the maximum possible amount of a pheromone added after any iteration.

In some of ACO algorithms a pheromone is added on all used movements and in others ACO algorithms only on used movements which belong to the best solutions. The other possible movements are evaporated or stay unchanged. Thus, the lower bound of the pheromone level of some of them can be 0. After additional reinforcement of the unused movements a lower bound of their pheromone is greater then $q\tau_{max}$. The ACO algorithm for which a pheromone is added on used movements the lower bound of the pheromone value on used movements is greater or equal to τ_0. Thus after additional reinforcement of unused movements and by Theorem 1 from [12] the algorithm will converge to the optimal solution.

Let us consider an ACO algorithm with some elitist ants. In this case only a small number of ants update the pheromone belonging to their solutions. Thus, the big part of pheromone is only evaporated and its value decreases after every iteration. Assuming in the first iteration, the movement from a state i to a state j is unused and the movement from a state i to a state k is used and does not belong to the best solution. The probability to choose the state j and the state k are respectively:

$$p_{ij} = \tau_0\eta_{ij} / \sum_l \tau_{il}\eta_{il} \tag{4}$$

and

$$p_{ik} = \tau_0 \eta_{ik} / \sum_l \tau_{il} \eta_{il} \tag{5}$$

If the movement from i to k is used it means that $\eta_{ij} < \eta_{ik}$. After additional reinforcement, the pheromone level of the movement from i to j will increase and the pheromone level of the movement from i to k will decrease. Thus, after a transition period t_{0i} the probability to choose the movement from i to j will be greater than the probability to choose the movement from i to k. Also the movement from i to k will become unused and will receive additional reinforcement. Therefore, $\rho^{t_0} \tau_0 > 0$ is a lower bound of the pheromone value, where $t_0 = \max t_{0i}$.

Independently of used ACO algorithm, after additional reinforcement of unused movements the lower bound of the pheromone is greater than 0 and the Theorem 1 can be used. Thus, the convergence of ACO-AR algorithm to optimal solution have been proved.

We will estimate the length of the transition period t_{0i}. Let $\eta_j = \min_s(\eta_{is})$ and $\eta_k = \max_s(\eta_{is})$. At first iteration the pheromone level for all movements from arbitrary state i to any other state is equal to τ_0 and therefore the ants choose the state with greater heuristic. After number of iterations t_{0i} the pheromone of movements from the state i to a state with less heuristic information (i.e. unused movement) is:

$$\rho^{t_{0i}} \tau_0 + q\tau_{max}(1 - \rho^{t_{0i}})/(1 - \rho). \tag{6}$$

While the pheromone of the movement from the state i to a state with greater heuristic information (i.e. used movement) is $\rho^{t_{0i}} \tau_0$. From the above discussion it can be seen the used movements become unused if they have less probability as follows:

$$\rho^{t_{0i}} \tau_0 \eta_k < \rho^{t_{0i}} \tau_0 \eta_j + q\tau_{max} \eta_j (1 - \rho^{t_{oi}})/(1 - \rho) \tag{7}$$

The value of t_{0i} can be calculated from upper inequality.

4 Parameter Value for *q*

In this section we discus the value of the parameter q. Our aim is the diversification and exploration of the search space while keeping the best found solution. Let the movement from a state i to a state k belong to the best solution and the movement from a state i to a state j is unused. The aim is the pheromone level (τ_{ij}) of unused movements to be less than the pheromone level (τ_{ik}) of the movements that belong to the best solution (i.e. $\tau_{ij} \leq \tau_{ik}$). The values of τ_{ij} and τ_{ik} are as follows:

$$\tau_{ij} = \rho^{k_1+k_2} \tau_0 + \frac{1 - \rho^{k_2}}{1 - \rho} \rho^{k_1} g_1 + \frac{1 - \rho^{k_1}}{1 - \rho} g(s^*) \tag{8}$$

$$\tau_{ik} = \rho^{k_1+k_2}\tau_0 + \frac{1-\rho^{k_1+k_2}}{1-\rho} g(s^*) \tag{9}$$

where:

- k_1 is the number of iterations for which the movement from i to j belongs to poor solutions;
- k_2 is the number of iterations for which the movement from i to j is unused;
- g_1 is the maximal added pheromone to a movement that belong to poor solution;

From equations (8) and (9) and $0 < g_1 < g(s^*)$ follows that $q \leq \rho$. Evaporation parameter ρ depends of the problem.

5 Conclusion

Recent research has strongly focused on improving the performance of ACO algorithms. In this paper we have presented the ACO-AR algorithm to exploit the search space, which have not been exploited yet, and to avoid premature stagnation of the algorithm. We have shown that ACO-AR algorithm converges to the optimal solution when the algorithm run for a sufficiently large number of iterations. The main idea introduced by ACO-AR, the additional reinforcement of the unused movements, can be apply in a variety ACO algorithms. Our future work will be to apply ACO-AR to other NP-hard COPs and to investigate the search space exploration.

Acknowledgments. Stefka Fidanova was supported by the CONNEX program of the Austrian federal ministry for education, science and culture, and Center of Excellence BIS-21 grant ICA1-2000-70016.

References

1. M. Dorigo and G. Di Caro: The ant colony optimization metaheuristic, in: D. Corne,M. Dorigo and F. Glover,eds., New Idea in Optimization, McGrow-Hill (1999) 11–32.
2. M. Dorigo and L. M. Gambardella: Ant colony system: A cooperative learning approach to the traveling salesman problem, IEEE Transactions on Evolutionary Computation 1. (1999) 53–66.
3. M. Dorigo, G. Di Caro and Gambardella: Ant algorithms for distributed discrete optimization, Artificial Life 5. (1999) 137–172.
4. M. Dorigo,V. Maniezzo and A. Colorni: The ant system: Optimization by a colony of cooperating agents, IEEE Transaction on Systems, Man. and Cybernetics - Part B 26. (1996) 29–41.
5. S. Fidanova: ACO Algorithm with Additional reinforcement, In:M. Dorigo, G. Di Carro eds., From Ant Colonies to Artificial Ant, Lecture Notes in Computer Science 2542, Springer (2002) 292–293.

6. S. Fidanova: Ant Colony optimization and Pheromone Model, Int. conf. on Real-Life Applications of Metaheuristics, http://www.ruca.ua.ac.be/eume/workshops/reallife/programme.php (2003)
7. M. L. Gambardella, E. D. Taillard and G. Agazzi: A multiple ant colony system for vehicle routing problems with time windows, in: D. Corne, M. Dorigo and F. Glover, eds., New Ideas in Optimization, McGraw-Hill (1999) 63–76.
8. L. M. Gambardella, E. D. Taillard and M. Dorigo: Ant colonies for the QAP, J. of Oper. Res. Soc. 50. (1999) 167–176.
9. I.H. Osman and J. P. Kelley: Methaheuristic:An Overview, In: I. H. Osman and J. P. Kelley eds., Mathematics: Theory and Applications, Kluwer Academic Publishers (1996).
10. R. Y. Rubinstein: Simulation and the Monte Carlo Method John Wiley& Sons. (1981).
11. S. Streltsov and P. Vakili: Variance Reduction Algorithms for Parallel Replicated Simulation of Uniformied Markov Chains, J. of Discrete Event Dynamic Systems: Theory and Applications 6. (1996) 159–180.
12. T. Stützle and M. Dorigo: A Short Convergence Proof for a Class of Ant Colony Optimization Algorithms, IEEE Transactions on Evolutionary Computation 6(4). (2002) 358–365.

Monte Carlo Algorithm for Maneuvering Target Tracking and Classification*

Donka Angelova[1], Lyudmila Mihaylova[2], and Tzvetan Semerdjiev[1]

[1] Institute for Parallel Processing, Bulgarian Academy of Sciences
25A Acad. G. Bonchev St, 1113 Sofia, Bulgaria
donka@bas.bg
[2] Department of Electrical and Electronic Engineering, University of Bristol,
Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK
mila.mihaylova@bristol.ac.uk

Abstract. This paper considers the problem of joint maneuvering target tracking and classification. Based on the recently proposed particle filtering approach, a multiple model particle filter is designed for two-class identification of air targets: commercial and military aircraft. The classification task is implemented by processing radar (kinematic) measurements only, no class (feature) measurements are used. A speed likelihood function for each class is defined using a priori information about speed constraints. Class-dependent speed likelihoods are calculated through the state estimates of each class-dependent tracker. They are combined with the kinematic measurement likelihoods in order to improve the process of classification. The performance of the suggested multiple model particle filter is evaluated by Monte Carlo simulations.

1 Introduction

A lot of research has been performed recently concerning the problem of joint target tracking and classification. Actually, the simultaneous implementation of these two important tasks in the surveillance systems facilitates the situation assessment, resource allocation and decision-making [1,2]. *Classification* (or *identification*) usually includes target allegiance determination and/or target profile assessment such as *vehicle*, *ship* or *aircraft* type. Target class information could be obtained from an *electronic support measure* (*ESM*) sensor, friend-and-foe identification system, high resolution radar or other identity sensors. It could be inferred from a tracker, using kinematic measurements only or in a combination with identity sensors. On the other hand, target type knowledge applied to the tracker can improve tracking performance by the possibility of selecting appropriate target models. Classification information can assist in correct data association and false tracks elimination in multiple target tracking systems.
Two basic alternative approaches to classification exist based on *Bayesian* and *Dempster-Shafer theories*. Comparative studies [1] of these inferring techniques come to the conclusion that Dempster-Shafer is more robust than the Bayesian technique, but this is

* Research supported in part by Center of Excellence BIS21 grant ICA1-2000-70016, by the Bulgarian Foundation for Scientific Investigations under grants I-1202/02 and I-1205/02, and in part by the UK MOD Data and Information Fusion Defence Technology Center.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 531–539, 2004.
© Springer-Verlag Berlin Heidelberg 2004

achieved at the expense of delay in time [2]. The Bayesian approach is more certain regarding the correct decisions. Efficient approximations to the optimal Bayesian solution of *joint tracking and classification (JTC)* are provided by Monte Carlo methods. Feasible implementations of the JTC task via particle filtering are reported in [3,4]. A unified algorithm for JTC is suggested in [2] within the framework of the Bayesian theory. A bank of filters, covering the state and feature space are run in parallel with each filter matched to different target class. A robust particle filter is proposed as a concrete realization of the algorithm.

The accurate representation of joint state and class probability distributions is an important advantage of the particle filtering approach. The highly non-linear relationships between state and class measurements and non-Gaussian noise processes can be easily processed by the particle filter. In addition, flight envelope constraints, which are essential part of the JTC task, can be incorporated into the filtering algorithm in a natural and consistent way [5].

The objective of the present paper is to explore the capabilities of this particle filtering technique to track and classify *a maneuvering target*. Two air target classes are considered: commercial aircraft and military aircraft. A bank of two *interacting multiple model* class dependent particle filters is designed and implemented. The novelty of the paper relies also on accounting for two kinds of constraints : both on the acceleration and on the speed. Two speed likelihood functions are defined based on a priori information about speed constraints of each class. Such kind of constraints are used in other approaches for decision making (see e.g. [6]). At each filtering step, the estimated speed from each class dependent filter is used to calculate a class dependent speed likelihood. The speed likelihoods are combined with kinematic likelihoods in order to improve the process of classification.

The remaining part of the paper is organized as follows. Section 2 summarizes the Bayesian formulation of the JTC problem according to [2,4,7,8]. Section 3 presents the developed multiple model particle filter using both speed and acceleration constraints. Simulation results are given in Section 4, with conclusions generalized in Section 5.

2 Problem Formulation

Consider the following model of a discrete-time jump Markov system, describing the target dynamics and sensor measurement

$$x_k = F(m_k)\, x_{k-1} + G(m_k)\, u_k + B(m_k)\, w_k, \tag{1}$$

$$z_k = h(m_k, x_k) + D(m_k)\, v_k,\ k = 1, 2, \dots, \tag{2}$$

where $x_k \in \mathbb{R}^{n_x}$ is the *base (continuous) state* vector with transition matrix F, $z_k \in \mathbb{R}^{n_z}$ is the measurement vector with measurement function h, and $u_k \in \mathbb{U}$ is a known control input. The noises w_k and v_k are *independent identically distributed* (*i.i.d.*) Gaussian processes having characteristics $w_k \sim N(0, Q)$ and $v_k \sim N(0, R)$, respectively. All vectors and matrices are assumed of appropriate dimensions. The *modal (discrete) state* $m_k \in \mathbb{S} \triangleq \{1,\ 2,\ \dots,\ s\}$ is a time-homogeneous first-order Markov chain with transition probabilities $p_{ij} \triangleq Pr\{m_k = j \mid m_{k-1} = i\}$, $(i, j \in S)$ and initial probability distribution $P_0(i) \triangleq Pr\{m_0 = i\}$ for $i \in S$, such that $P_0(i) \geq 0$, and

$\sum_{i=1}^{s} P_0(i) = 1$. We assume that the target belongs to one of the M classes $c \in C$ where $C = \{c_1, c_2, \ldots, c_M\}$ represents the set of the target classes. Generally, the number of the discrete states $s = s(c)$, the initial probability distribution $P_0^c(i)$ and the transition probability matrix $[p_{ij}]^c$, $i, j \in S$ are different for each target class.

The joint state and class is time varying with respect to the state and time invariant with respect to the class [2]. Let $\left\{Z^k, Y^k\right\} = \{z_i, y_i\}_{i=1}^k$ be the cumulative set of *kinematic* (Z^k) and *class (feature)* measurements (Y^k) up to time k.

The *goal* of the joint tracking and classification task is to estimate the *state* x_k and the *posterior classification probabilities* $P\left(c \mid \left\{Z^k, Y^k\right\}\right), c \in C$ based on all available measurement information $\left\{Z^k, Y^k\right\}$. If we can construct the *posterior joint state-class probability density function* (*pdf*) $p\left(x_k, c \mid \left\{Z^k, Y^k\right\}\right)$, then the posterior classification probabilities can be obtained by marginalisation over x_k:

$$P\left(c \mid \left\{Z^k, Y^k\right\}\right) = \int_{x_k} p\left(x_k, c \mid \left\{Z^k, Y^k\right\}\right) dx_k. \tag{3}$$

Suppose that we know the posterior joint state-class pdf $p\left(x_{k-1}, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right)$ at time instant $k-1$. According to the Bayesian philosophy, $p\left(x_k, c \mid \left\{Z^k, Y^k\right\}\right)$ can be computed recursively from $p\left(x_{k-1}, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right)$ within the framework of two steps – *prediction* and *measurement update* [2,4].

The predicted state-class pdf $p\left(x_k, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right)$ at time k is given by

$$p\left(x_k, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right) = \int_{x_{k-1}} p\left(x_k \mid x_{k-1}, c\right) p\left(x_{k-1}, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right) dx_{k-1}, \tag{4}$$

where the conditional state prediction pdf $p\left(x_k \mid x_{k-1}, c, \left\{Z^{k-1}, Y^{k-1}\right\}\right)$ is obtained from the state transition equation (1).

The conditional pdf of the measurements $p\left(\{z_k, y_k\} \mid x_k, c\right) = \lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right)$ is usually known. This is the likelihood $\lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right)$ of the joint state and feature. When the measurements $\{z_k, y_k\}$ arrive, the update step can be completed

$$p\left(x_k, c \mid \left\{Z^k, Y^k\right\}\right) = \frac{1}{\bar{d}_k} \lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right) p\left(x_k, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right), \tag{5}$$

where $\bar{d}_k = \sum_{c \in C} \int_{x_k} \lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right) p\left(x_k, c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right) dx_k$ is a normalizing constant. The recursion (4)-(5) begins with the prior density $P\{x_0, c\}, x_0 \in \mathbb{R}^{n_x}, c \in C$, which is assumed known. Then using Bayes' theorem, the target classification probability is calculated by the equation

$$P\left(c \mid \left\{Z^k, Y^k\right\}\right) = \frac{p\left(\{z_k, y_k\} \mid c, \left\{Z^{k-1}, Y^{k-1}\right\}\right) P\left(c \mid \left\{Z^{k-1}, Y^{k-1}\right\}\right)}{\sum_{c \in C} p\left(\{z_k, y_k\} \mid c, \{Z^{k-1}, Y^{k-1}\}\right) P\left(c \mid \{Z^{k-1}, Y^{k-1}\}\right)}$$

with an initial prior target classification probability $P_0(c)$, and $\sum_{c \in C} P_0(c) = 1$.

The class-dependent state estimate $\hat{x}_k^c$, $c \in C$ takes part in the calculation of the *combined* state estimate $\hat{x}_k$

$$\hat{x}_k^c = \int_{x_k} x_k p\left(x_k, c \mid \left\{Z^k, Y^k\right\}\right) dx_k, \quad \hat{x}_k = \sum_{c \in C} \hat{x}_k^c P\left(c \mid \left\{Z^k, Y^k\right\}\right). \tag{6}$$

It can be seen from (4)-(6) that the estimates, needed for each class, can be calculated independently from the other classes. Therefore, the JTC task can be accomplished by the simultaneous work of M independent filters [2,4]. The scheme of the particle filter bank, implemented in the present paper is described in Section 3.

It should be noted that because in our case we don't have feature measurements, the set $\{Y^k\}$ is replaced in the particle filter by the speed estimates from the M classes. Together with a speed envelope which form is given in Section 3, they form a virtual "feature measurement".

3 Particle Filter for Maneuvering Target Tracking and Classification

Maneuvering target model. The two-dimensional target dynamics is given by

$$x_k = Fx_{k-1} + G\left[u_k + w_k\right], \ k = 1, 2, \ldots \tag{7}$$

where the state vector $x = (x, \dot{x}, y, \dot{y})'$ contains target positions and velocities in the horizontal (Oxy) Cartesian coordinate frame. The control input vector $u = (a_x, a_y)'$ includes target accelerations along x and y coordinates. The process noise $w = (w_x, w_y)'$ models perturbations in the accelerations. The transition matrices F and G are [9]

$$F = diag\left[F_1, F_1\right], \text{ for } F_1 = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}; \quad G = diag\left[g_1, g_1\right], \text{ for } g_1 = \left[\frac{T^2}{2} \ T\right]',$$

where T is the sampling interval. The target is assumed to belong to one of two classes ($M = 2$), representing either a lower speed *commercial aircraft* with limited maneuvering capability (c_1) or a highly maneuvering *military aircraft* (c_2) [7]. The flight envelope information comprises speed and acceleration constrains, characterizing each class. The speed $v = \sqrt{\dot{x}^2 + \dot{y}^2}$ of each class is limited respectively to the interval:

$$\{c_1 : v \in (100, 300)\} \ [m/s] \quad \text{and} \quad \{c_2 : v \in (150, 650)\} \ [m/s].$$

The range of the speed overlap section is $[150, 300]$. The control inputs are restricted to the following sets of accelerations:

$$\{c_1 : u \in (0, +2g, -2g)\} \quad \text{and} \quad \{c_2 : u \in (0, +5g, -5g)\} \ ,$$

where $g\,[m/s^2]$ is the gravity acceleration. The acceleration process u_k is a Markov chain with five states $s(c_1) = s(c_2) = 5$:

1. $a_x = 0, \quad a_y = 0, \quad$ 2. $a_x = A, \quad a_y = A, \quad$ 3. $a_x = A, \quad a_y = -A,$
4. $a_x = -A, \quad a_y = A, \quad$ 5. $a_x = -A, \quad a_y = -A$

where $A = 2g$ stands for class c_1 target and $A = 5g$ refers to the class c_2. The two target types have equal transition probabilities $p_{ij}, i, j \in S$: $p_{ij} = 0.7$, $i = j$; $p_{1j} = 0.15$, $j = 2, \ldots, 5$; $p_{ij} = 0.05$, $j \neq i$, $i, j = 2, \ldots, 5$. The initial probabilities are selected as follows: $P_0(1) = 0.6$, $P_0(2) = P_0(3) = P_0(4) = P_0(5) = 0.1$. The standard deviations of the process noise $w \sim N(0, diag(\sigma_{wx}^2, \sigma_{wy}^2))$ are different for each mode and class:

$\left\{c_1 : \sigma_w^j = 5.5 \ [m/s^2], j = 1, \ldots, 5\right\}$ and
$\left\{c_2 : \sigma_w^1 = 7.5, \sigma_w^j = 17.5 \ [m/s^2], j = 2, \ldots, 5\right\}$, where ($\sigma_w = \sigma_{wx} = \sigma_{wy}$).

Measurement model. The measurement model at time k is described by

$$z_k = h(x_k) + v_k, \qquad h(x) = \left(\sqrt{x^2+y^2}, \arctan\frac{x}{y}\right)', \tag{8}$$

where the measurement vector $z = (D, \beta)'$ contains the distance to the target D and bearing β, measured by the radar. The parameters of the measurement error vector $v \sim N(0, R)$, $R = diag(\sigma_D^2, \sigma_\beta^2)$ are as follows: $\sigma_D = 100.0$ $[m]$; $\sigma_\beta = 0.15$ [deg]. The sampling interval of $T = 5$ $[s]$ is chosen in the simulations below.

Speed constraints. Acceleration constraints are imposed on the filter operation by an appropriate choice of the control input in the target model. The speed constraints are enforced through speed likelihood functions. Using the speed envelope information, the speed likelihood functions are defined by the following relationships

$$g_1\left(v_k^{c_1}\right) = \begin{cases} 0.8, & \text{if } v_k^{c_1} \le 100 \text{ [m/s]} \\ 0.8 + \kappa_1\left(v_k^{c_1} - 100\right), & \text{if } \left(100 < v_k^{c_1} \le 300\right) \text{ for } \kappa_1 = -0.7/200 \\ 0.1, & \text{if } v_k^{c_1} > 300 \text{ [m/s]} \end{cases}$$

$$g_2\left(v_k^{c_2}\right) = \begin{cases} 0.1, & \text{if } v_k^{c_1} \le 150 \text{ [m/s]} \\ 0.1 + \kappa_2\left(v_k^{c_2} - 150\right), & \text{if } \left(150 < v_k^{c_1} \le 650\right) \text{ for } \kappa_2 = 0.85/500 \\ 0.95 & \text{if } v_k^{c_1} > 650 \text{ [m/s].} \end{cases}$$

According to the problem formulation, presented in Section 2, two class-dependent filters are working in parallel. At time step k, each filter gives state estimate $\{\hat{x}_k^c,\ c = 1, 2\}$. Let us assume, that the estimated speed from the previous time step, $\left\{\hat{v}_{k-1}^c,\ c = 1, 2\right\}$, is a kind of "feature measurement". The likelihood $\lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right)$ is factorized [2]

$$\lambda_{\{x_k, c\}}\left(\{z_k, y_k\}\right) = f_{x_k}\left(z_k\right) g_c\left(y_k^c\right), \tag{9}$$

where $y_k^c = \hat{v}_{k-1}^c$. Practically, the normalized speed likelihoods represent estimated by the filters speed-based class probabilities. The posterior class probabilities are modified by this additional speed information at each time step k. The inclusion of the speed likelihoods is done after some "warming-up" interval, including filter initialization.

Particle Filter Algorithm. Consider the hybrid particle $\mathbf{x} = \{x, m, c\}$, containing all necessary state, mode and class information. Let N_c the number of particles for class c. Then the detailed scheme of the proposed particle filter comprises the following steps:

1. Initialization, $k = 0$.
   For class $c = 1, 2, \ldots, M$, set $P(c) = P_0(c)$
   * For $j = 1, \ldots, N_c$, sample
   $\left\{x_0^{(j)} \sim p_0(x_0, c),\ m_0^{(j)} \sim \{P_0^c(m)\}_{m=1}^s(c),\ c^{(j)} = c\right\}$ and set $k = 1$.
   End for c

2. For $c = 1, \ldots, M$ (possibly in parallel) execute

   * Prediction step

   For $j = 1, \ldots, N_c$ generate samples
   $m_k^{(j)} \sim \{p_{lm}^c\}_{m=1}^{s(c)}$ for $l = m_{k-1}^{(j)}$,
   $w_k \sim N(0, Q(m_k^{(j)}, c))$, $x_k^{(j)} = Fx_{k-1}^{(j)} + Gu_k(m_k^{(j)}, c) + Gw_k$

   * Measurement processing step : on receipt of a new measurement $\{z_k, y_k\}$:

   For $j = 1, \ldots, N_c$ evaluate the weights $w_k^{(j)} = f(z_k \mid x_k^{(j)}) g_c(y_k^c)$,
   where $f(z_k \mid x_k^{(j)}) = N(z_k; h(x_k^{(j)}), R)$ and $g_c(y_k^c) = g_c\left(\hat{v}_{k-1}^{c(j)}\right)$;
   calculate
   $p\left(\{z_k, y_k\} \mid c, \left\{Z^{k-1}, Y^{k-1}\right\}\right) = \sum_{j=1}^{N_c} w_k^{(j)}$ and set $L(c) = \sum_{j=1}^{N_c} w_k^{(j)}$

   * Selection step

   normalize the weights $w_k^{(j)} = w_k^{(j)} / \sum_{j=1}^{N_c} w_k^{(j)}$
   resample with replacement N_c particles $(x_k^{(j)}; j = 1, \ldots, N_c)$
   from the set $(x_k^{(l)}); l = 1, \ldots, N_c)$, according to the importance weights

   * Compute updated state estimate and posterior model probability

   $\hat{x}_k^c = \frac{1}{N_c} \sum_{j=1}^{N_c} x_k^{(j)}$; $P^c(m_k = l) = \frac{\sum_{j=1}^{N_c} m_k^{(j)} = l}{\sum_{j=1}^{N_c} m_k^{(j)}}$, $l = 1, \ldots, s(c)$

   End for c
3. Output: Compute posterior class probabilities and combined output estimate
   $P\left(c \mid \left\{Z^k, Y^k\right\}\right) = L(c) P\left(c \mid Z^{k-1}, Y^{k-1}\right) / \sum_{c=1}^{M} L(c) P\left(c \mid Z^{k-1}, Y^{k-1}\right)$
   $\hat{x}_k = \sum_{c=1}^{M} P\left(c \mid \left\{Z^k, Y^k\right\}\right) \hat{x}_k^c$
4. Set $k \longleftarrow k + 1$ and go to step 2.

4 Simulation Results

The performance of the implemented tracking filter is evaluated by simulations over a representative test trajectory, depicted in Fig. 1.
Measures of performance. Root-Mean Squared Errors (RMSE) [9]: on position (both coordinates combined) and speed (magnitude of the velocity vector), *average probability of correct discrete state identification*, *average probability of correct class identification* and *average time per update* are used to evaluate the filter performance.
The results presented below are based on 100 Monte Carlo runs. The number of particles for each class is $N_c = 3000$. The prior class probabilities are chosen as follows: $P_0(1) = P_0(2) = 0.5$. The parameters of base state vector initial distribution $x_0 \sim N[x_0; m_0, P_0]$ are selected as follows: $P_0 = diag\{150^2, 20.0^2, 150^2, 20.0^2\}$; m_0 contains the exact initial target parameters.

Test trajectory. The target performs two coordinated turn maneuvers with normal accelerations $2g$ and $-5g$, respectively, within scan intervals $(16 \div 23)$ and $(35 \div 37)$. The

selected speed value is $v = 250\,[m/s]$. Then a maneuver is performed with longitudinal acceleration of $1g$ and normal acceleration of $2g$ in the frame of 3 scans $(45 \div 47)$. The longitudinal acceleration increases the speed up to $v = 400\,[m/s]$. These maneuvers are followed by another two maneuvers (during scans $(48 \div 56)$ and $(63 \div 70)$) with normal accelerations of $2g$ and $-2g$, respectively. The speed values (from 250

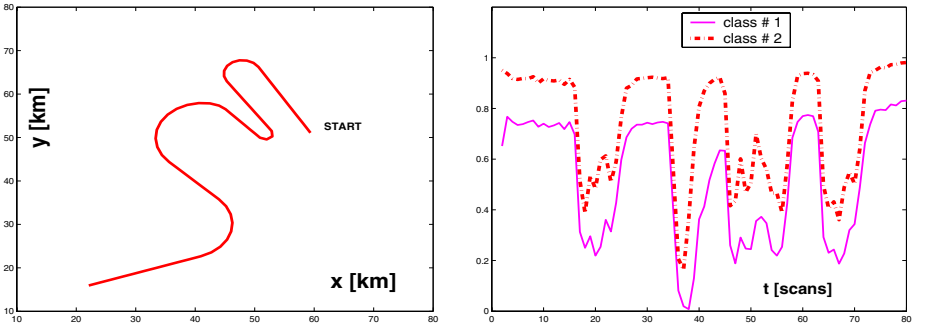


Fig. 1. Test trajectory (a) and (b) Posterior probability of model 1

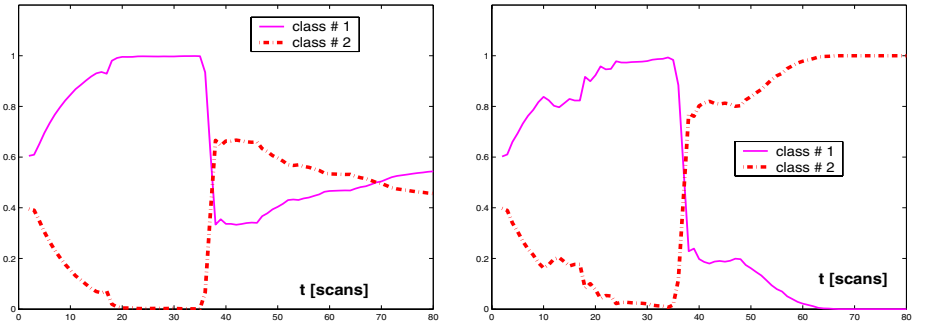


Fig. 2. Class probabilities without (a) and (b) with speed constraints

to $400\,[m/s]$) and the normal $5g$ acceleration are typical for the second class target. After the $5g$ maneuver, the filter correctly recognizes the real second class, but after the subsequent maneuvers of $2g$, a tendency for misclassification is present (Fig. 2(a)) in the filter without speed constraints. It is due to the fact, that the filtering system can "change its mind" regarding the class identification, if the second class target performs maneuvers, natural for the first class. The advantage of incorporation of speed constraints is illustrated in Fig. 2(b). According to the results from the RMSEs (Fig. 3) the developed particle filter with acceleration and speed constraints can reliably track maneuvering targets. The discrete (mode) states are determined correctly by the filter (Fig.1(b)). It

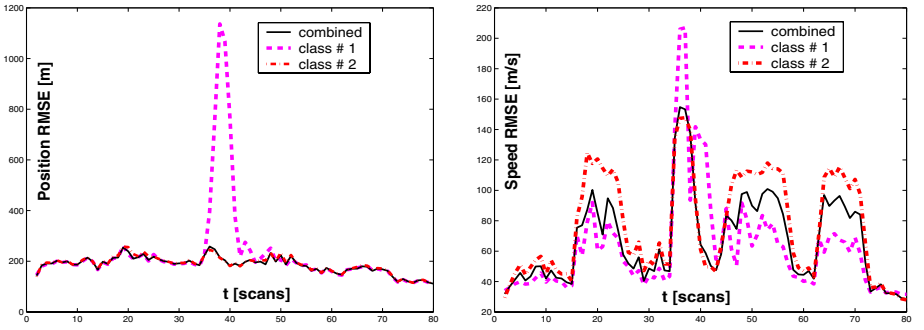


Fig. 3. Position RMSE [m] (a) and (b) Speed RMSE [m/s]

should be noted that the filter's computational complexity allows for an on-line processing. The average time per update, estimated in the MATLAB environment (on PC with AMD Athlon processor) is $1.52\ [s]$. It is less than the sampling interval of $5\ [s]$.

5 Conclusions

A Bayesian joint tracking and classification algorithm has been proposed recently in the work [2]. Based on this approach, a particle filter is developed in the present paper for *maneuvering* target tracking and classification. A bank of *interacting multiple model* class dependent particle filters is designed and implemented in the presence of speed and acceleration constraints. The acceleration constraints for each class are imposed by using different control inputs in the target model. The speed constraints are enforced by constructing class dependent speed likelihood functions. Speed likelihoods are calculated at each filtering step and assist in the process of classification.
The filter performance is analyzed by simulation over a typical $2-D$ target scenario. The results show a reliable tracking and correct target type classification.

References

1. Leung, H., Wu, J.: Bayesian and Dempster-Shafer target identification for radar surveillance. IEEE Trans. Aerospace and Electr. Systems, Vol. 36 (2000) 432–447
2. Gordon, N., Maskell, S., Kirubarajan, T.: Efficient particle filters for joint tracking and classification. Proc. SPIE Signal and Data Proc. of Small Targets, Vol. 4728, USA, (2002) 1–5
3. Herman, S., Moulin, P.: A Particle Filtering Appropach to FM-Band Passive Radar Tracking and Automatic Target Recognition. Proc. IEEE Aerospace Conf., Big Sky, Montana (2002)
4. Mallick, M., Maskell, S., Kirubarajan, T., Gordon, N.: Littoral Tracking Using Particle Filter. Proc. Fifth Int. Conf. Information Fusion, Annapolis, MD, USA (2002) 935–942
5. Challa, S., Bergman, N.: Target Tracking Incorporating Flight Envelope Information. Proc. Third International Conf. on Information Fusion, Paris, France (2000) ThC2-22–27
6. Tchamova, A., Semerdjiev, Tz., Dezert, J.: Estimation of Target Behaviour Tendencies using Dezert-Smarandache Theory. Proc. Sixth International Conf. on Information Fusion, Australia (2003) 1349–1356

7. Challa, S., Pulford, G.: Joint target tracking and classification using radar and ESM sensors. IEEE Trans. Aerospace and Electr. Systems, Vol. 37 (2001) 1039–1055
8. Doucet,A., de Freitas, N., Gordon, N.(ed.): Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York (2001)
9. Bar-Shalom, Y., Li, X.R.: Multitarget–Multisensor Tracking: Principles and Techniques. YBS Publishing (1995)

Coupling a Lattice Boltzmann and a Finite Difference Scheme

Paul Albuquerque[1,2], Davide Alemani[3], Bastien Chopard[1], and Pierre Leone[1,2]

[1] Computer Science Department, University of Geneva, 1211 Geneva 4, Switzerland
[2] LII, Ecole d'Ingénieurs de Genève, HES-SO, 1202 Geneva, Switzerland
[3] CABE, University of Geneva, 1211 Geneva 4, Switzerland

Abstract. We show how a lattice Boltzmann (LB) scheme can be spatially coupled with a finite difference (FD) scheme in order to solve the same problem. The typical situation we consider is a computational domain which is partitioned in two regions. The same spatio-temporal physical process extends over the full domain but a different numerical method is used over each region. At the interface of the subdomains, the LB and FD must be connected so as to ensure a perfect continuity of the physical quantities. We derive the theoretical concepts, which allow us to link both methods in the case of a diffusion process, and validate them with numerical simulations on a 2D domain.

1 Introduction

Many physical systems include phenomena at different time and space scales. Their description in terms of a numerical model is therefore a difficult task as often the same numerical scheme is not efficient over the full range of scales. In addition, it occurs that various physical processes take place and couple different parts of the system. Again, it is challenging to devise a numerical method which is able to efficiently deal with such constraints.

Here we consider the case of a spatially extended system in which separate spatial regions are treated with different numerical schemes. The motivation is that, depending on the nature of each region, optimal efficiency is obtained with different numerical methods.

For instance the lattice Boltzmann (LB) method [1] has a more detailed microscopic description than a finite difference (FD) scheme because the LB approach includes the molecular velocity of the particles. In addition, important physical quantities, such as the stress tensor, or particle current, are directly obtained from the local information. However, the LB scheme may require more memory than the corresponding FD scheme. Another motivation is that boundary conditions are more or less naturally imposed on a given numerical scheme. Therefore, in order to improve the global quality of the numerical solution, it may be quite efficient to assume that several solvers are coupled.

Obviously, this coupling should not produce any discontinuities at the interface between regions that are treated with the LB or the FD method. Since each

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 540–547, 2004.
© Springer-Verlag Berlin Heidelberg 2004

scheme uses different variables, it is crucial to understand how the LB set of variables is related to the FD set, and conversely.

Our solution follows the same argument as developed in [2] for a multigrid LB scheme: the connection between the full set of LB variables and the standard macroscopic physical quantities is based on the splitting of the particle distribution functions in an equilibrium and a nonequilibrium part. The former is built on the physical quantities and the latter on the gradient of these quantities. This observation is quite general and could be used to couple any scheme (e.g. finite volume or finite element) with any LB method. Indeed, if the two schemes use different variables, one must compute the particle distribution functions starting from physical quantities known from another solver or from an analytical expression. Here we will only consider the case of a diffusion process on a given domain which is solved by the LB approach on a chosen subdomain and with a FD solver on the rest of the domain.

Hybrid methods have already been proposed in the literature. In [3] a finite volume and a finite element method are combined and then coupled with a finite difference time domain solver for the wave equation, allowing the simulation of wave propagation in complex 3D geometry. In [4], the LB method is coupled with a molecular dynamics simulation of polymers. However, to our knowledge, the FD and LB schemes have never been coupled across adjacent regions.

The paper is organized as follows. In section 2 we briefly introduce the LB approach (we assume that the reader is familiar with the FD method). The special case of a diffusion process is discussed in section 3. In particular, it is shown how the Chapman-Enskog expansion offers a mapping between the LB variables and the macroscopic quantities and their spatial derivatives. In section 4 we describe the coupling algorithm. Then, in section 5 we perform a numerical simulation to demonstrate that the proposed coupling is correct. Finally, section 6 summarizes our findings and draws some conclusions.

2 The Lattice Boltzmann Approach

A lattice Boltzmann (LB) model [1,5] describes a physical system in terms of a mesoscopic dynamics: fictitious particles move on a regular lattice, synchronously at discrete time steps. An interaction is defined between the particles that meet simultaneously at the same lattice site. Particles obey collision rules which reproduce, in the macroscopic limit, an equation of physics.

A LB model can be interpreted as a discretization of the Boltzmann transport equation on a regular lattice of spacing Δr and with discrete time step Δt. The possible velocities for the pseudo-particles are the vectors $\boldsymbol{v}_i$. They are chosen so as to match the lattice direction: if $\boldsymbol{r}$ is a lattice site, $\boldsymbol{r} + \boldsymbol{v}_i \Delta t$ is also a lattice point. We thus consider a dynamics with $z+1$ possible velocities, where z is the lattice coordination number and $\boldsymbol{v}_0 = 0$ describe the population of rest particles.

For isotropy reasons the lattice topology must satisfy the conditions

$$\sum_i v_{i\alpha} = 0 \quad \text{and} \quad \sum_i v_{i\alpha} v_{i\beta} = v^2 C_2 \delta_{\alpha\beta} \tag{1}$$

where C_2 is a numerical coefficient which depends on the lattice topology. The greek indices label the spatial dimensions and $v = \Delta t/\Delta r$. The first condition follows from the fact that if $\boldsymbol{v}_i$ is a possible velocity, then so is $-\boldsymbol{v}_i$.

In the LB approach a physical system is described through density distribution functions $f_i(\boldsymbol{r}, t)$. For hydrodynamics and diffusion processes, $f_i(\boldsymbol{r}, t)$ represents the distribution of particles entering site $\boldsymbol{r}$ at time t and moving in direction $\boldsymbol{v}_i$. For the wave model, the interpretation is less obvious as f_i can be positive as well as negative. Physical quantities can be defined from moments of these distributions. For instance, the local density is obtained by $\rho = \sum_{i=0}^{z} f_i$.

A LB model can be determined by specifying a lattice, a kinetic equation and an equilibrium distribution. In its simplest form (BGK model), the dynamics can be written as a relaxation to a given local equilibrium

$$f_i(\boldsymbol{r} + \boldsymbol{v}_i \Delta t, t + \Delta t) - f_i(\boldsymbol{r}, t) = \frac{1}{\tau} \left(f_i^{eq}(\boldsymbol{r}, t) - f_i(\boldsymbol{r}, t) \right) \tag{2}$$

where τ is a relaxation time, which is a free parameter of the model. The local equilibrium solution f_i^{eq} contains all the information concerning the physical process investigated. It changes according to whether we consider hydrodynamics, diffusion or wave propagation.

3 The Lattice Boltzmann Diffusion Model

Diffusion processes were first modeled using a cellular automata approach [6]. The scheme was then extended to a LB dynamics with a BGK collision term [1, 5,7].

From now on, we assume that the diffusion process takes place on a d-dimensionnal square lattice. The key point is to choose correctly the local equilibrium f_i^{eq} so that the diffusion equation for the particle density $\rho = \sum_{i=0}^{2d} f_i$ can be derived from eq. (2). Since the particle density ρ is the only conserved quantity in a diffusive process, the local equilibrium f_i^{eq} is taken to be $f_i^{eq}(\boldsymbol{r}, t) = \rho(\boldsymbol{r}, t)/2d$ so that ρ is indeed conserved and f_i^{eq} depends on $\boldsymbol{r}$ and t only through the conserved quantities [1,5].

Let us now assume that Δt is small (and constant in all our computations). Taylor expanding the left hand side of eq. (2) up to second order, we get

$$\begin{aligned} \Delta t (\boldsymbol{v}_i \cdot \nabla) f_i(\boldsymbol{r}, t) + \Delta t\, \partial_t f_i(\boldsymbol{r}, t) + \frac{\Delta t^2}{2} \partial_t^2 f_i(\boldsymbol{r}, t) + \\ \Delta t^2 (\boldsymbol{v}_i \cdot \nabla) \partial_t f_i(\boldsymbol{r}, t) + \frac{\Delta t^2}{2} (\boldsymbol{v}_i \cdot \nabla)^2 f_i(\boldsymbol{r}, t) = \frac{1}{\tau} \left(f_i^{eq}(\boldsymbol{r}, t) - f_i(\boldsymbol{r}, t) \right). \end{aligned} \tag{3}$$

We then use the multiscale Chapman-Enskog expansion to solve eq. (3). Thus, we set $f_i = f_i^{(0)} + f_i^{(1)} + f_i^{(2)} + \ldots$ and introduce next a small parameter ϵ along with the change of coordinates $(\boldsymbol{r}, t) \mapsto (\boldsymbol{r}_1, t_1, t_2) = (\epsilon \boldsymbol{r}, \epsilon t, \epsilon^2 t)$. We also consider a new function $f_i^{\epsilon}(\boldsymbol{r}_1, t_1, t_2)$ which satisfies $f_i(\boldsymbol{r}, t) = f_i^{\epsilon}(\epsilon \boldsymbol{r}, \epsilon t, \epsilon^2 t)$. After formally substituting $f_i \longrightarrow f_i^{\epsilon}$, $\nabla_r \longrightarrow \epsilon \nabla_{r_1}$, $\partial_t \longrightarrow \epsilon \partial_{t_1} + \epsilon^2 \partial_{t_2}$, into eq. (3), we obtain a new equation for f_i^{ϵ}.

To find an asymptotic series solution, we expand f_i^ϵ in powers of ϵ

$$f_i^\epsilon(\boldsymbol{r}_1, t_1, t_2) = f_i^{\epsilon(0)}(\boldsymbol{r}_1, t_1, t_2) + \epsilon f_i^{\epsilon(1)}(\boldsymbol{r}_1, t_1, t_2) + \epsilon^2 f_i^{\epsilon(2)}(\boldsymbol{r}_1, t_1, t_2) + \dots \quad (4)$$

By introducing (4) into the equation for f_i^ϵ, we can recursively determine the $f_i^{\epsilon(j)}$s. Using particle density conservation and the first equation in (1), we get at zero and first order

$$f_i^{(0)}(\boldsymbol{r}, t) = \frac{\rho(\boldsymbol{r}, t)}{2d} = f_i^{eq}(\boldsymbol{r}, t)\,, \quad f_i^{(1)} = -\frac{\tau \Delta t}{2d} \boldsymbol{v}_i \cdot \nabla \rho(r, t)\,.$$

It is still worth noticing, with respect to the macrocospic behaviour of our model, that the particle density $\rho(\boldsymbol{r}, t)$ satisfies the diffusion equation. Indeed, using the second equation in (1), we get

$$\partial_t \rho(\boldsymbol{r}, t) = \frac{C_2 v^2 \Delta t}{2d} \big(\tau - \frac{1}{2}\big) \nabla^2 \rho(\boldsymbol{r}, t),$$

with diffusion coefficient $D = (\tau - 1/2) C_2 v^2 \Delta t/(2d)$.

4 The Coupling Algorithm

The result of the previous section shows that the LB variables can be written as $f_i = f_i^{eq} + f_i^{neq}$, where

$$f_i^{eq} = f_i^{(0)} = \frac{\rho(\boldsymbol{r}, t)}{2d}\,, \quad f_i^{neq} \approx f_i^{(1)} = -\frac{\tau \Delta t}{2d} \boldsymbol{v}_i \cdot \nabla \rho(r, t)\,.$$

These relations give a dictionary to go from the standard description of diffusion, where only ρ is considered, to the LB description. Note that the inverse connection is straightforward since we always have, by definition, $\rho = \sum_i f_i$. It must be clear that the LB scheme requires more information on the physical variables because it contains more degrees of freedom.

To make the coupling between a FD and a LB scheme more explicit, we now consider the situation of a two-dimensional diffusion problem $\partial_t \rho = D \nabla^2 \rho$ on a square domain $\Omega \subset \mathbb{R}^2$ of size $L_x \times L_y$. We cut Ω into two parts Ω_1 and Ω_2 such that $\Omega = \Omega_1 \bigcup \Omega_2$. We apply in Ω_1 the FD method and in Ω_2 the LB method (see fig. 1). We assume here that the same mesh is used for both methods.

We define the quantities $\boldsymbol{v}_i$, $i = 1, ..., 4$, as vectors pointing in the four lattice directions: right, up, left and down. Their lengths are such that $\boldsymbol{v}_i \Delta t$ connects the neighboring sites along direction i.

For the points $\boldsymbol{r} \in \Omega_1$, we apply the FD scheme to the diffusion equation

$$\rho(\boldsymbol{r}, t + \Delta t) = \rho(\boldsymbol{r}, t) + \frac{\Delta t}{\Delta r^2} D \sum_{i=1}^{4} \left(\rho(\boldsymbol{r} + \boldsymbol{v}_i \Delta t, t) - \rho(\boldsymbol{r}, t)\right), \quad (5)$$

whereas the points $\boldsymbol{r} \in \Omega_2$ are treated with the LB scheme given by eq. (2). There are several ways to define the interface between the two regions, Ω_1 and

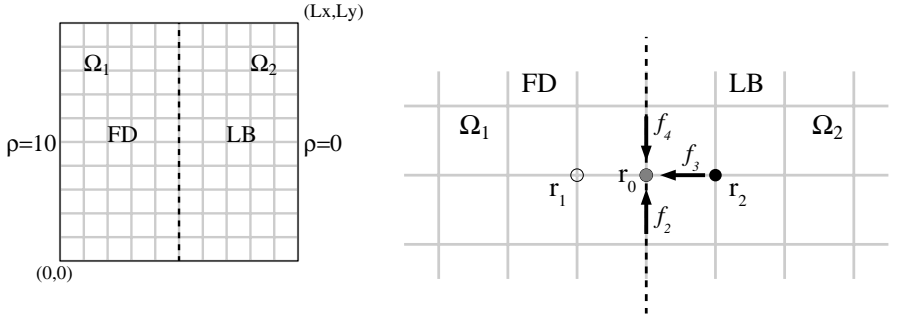


Fig. 1. Left: The computational grid and its partioning into a subdomain Ω_1, where the FD scheme is applied, and a subdomain Ω_2, where the LB scheme is used. The boundary conditions we impose for the diffusion process are $\rho = 10$ on the left wall, $\rho = 0$ on the right one and periodic along the vertical axis. Right: The point $\boldsymbol{r}_1$ (white circle), resp. $\boldsymbol{r}_2$ (black circle), is completely treated with the FD, resp. LB, scheme, and $\boldsymbol{r}_0$ (gray circle) is updated with both methods. The figure shows that for $\boldsymbol{r}_0$, the distribution f_1 is unknown (because it should come out of $\boldsymbol{r}_1$). Thus, we use our coupling algorithm to compute it

Ω_2. Here, we assume an overlap so that a point $\boldsymbol{r}_0$ at the interface belongs to both Ω_1 and Ω_2 (see fig. 1). On such points $\boldsymbol{r}_0$, the LB variables as well as the FD variables are computed. We denote by $\boldsymbol{r}_1 \in \Omega_1$ and $\boldsymbol{r}_2 \in \Omega_2$, two neighbors of a site $\boldsymbol{r}_0$ of the interface. According to fig. 1 and eq. (5), the calculation of $\rho(\boldsymbol{r}_0, t + \Delta t)$ requires the knowledge of $\rho(\boldsymbol{r}_2, t)$, where $\boldsymbol{r}_2$ is only subject to the LB scheme. However, the LB scheme offers naturally this quantity as $\rho(\boldsymbol{r}_2, t) = \sum_{i=1}^{4} f_i(\boldsymbol{r}_2, t)$. Therefore, the coupling of a FD site to a LB site is straightforward. The reverse is a bit more involved. In order to update the LB scheme at $\boldsymbol{r}_0$ and time $t + \Delta t$, we need $f_1(\boldsymbol{r}_0, t)$. This quantity is not known because the site $\boldsymbol{r}_1$ is only treated with the FD scheme. Indeed, if the LB scheme had been applied to the full domain, then $f_1(\boldsymbol{r}_0, t)$ would have been propagated from lattice site $\boldsymbol{r}_1$. However, the value of $f_1(\boldsymbol{r}_0, t)$ can be computed from our dictionary $f_1(\boldsymbol{r}_0, t) = f_1^{(0)}(\boldsymbol{r}_0, t) + f_1^{(1)}(\boldsymbol{r}_0, t)$ with $f_1^{(0)}(\boldsymbol{r}_0, t) = \rho(\boldsymbol{r}_0, t)/4$ and $f_1^{(1)}(\boldsymbol{r}_0, t) = -(\tau \Delta t/4)\, \boldsymbol{v}_i \cdot \nabla\rho(\boldsymbol{r}_0, t)$. In order to obtain $\nabla\rho(\boldsymbol{r}_0, t)$, we use a second order finite difference approximation of ρ over the nearest neighbors

$$\nabla\rho(\boldsymbol{r}_0, t) = \frac{\rho(\boldsymbol{r}_2, t) - \rho(\boldsymbol{r}_1, t)}{2\Delta r}\,. \tag{6}$$

Note that in the particular case where only one f_i is missing, the connection can be made in a simpler way. As $\rho = \sum_i f_i$ is known from the FD calculation, and f_2, f_3 and f_4 are known from the LB calculation, one has $f_1 = \rho - f_2 - f_3 - f_4$. In the simple case described here, this approach gives a correct coupling. However, it no longer works if the interface between Ω_1 and Ω_2 is irregular because the previous expression is not sufficient to determine more than one f_i.

5 Numerical Validation

To validate the coupling algorithm proposed in the previous section, we performed the simulation of a source-sink problem on the geometry defined in fig. 1. The domain size is $L_x = L_y = 30$ in lattice units. The boundary conditions are $\rho((0, y), t) = 10$ and $\rho((L_x, y), t) = 0$ on the left and right sides of Ω and the initial conditions are $\rho((0, y), 0) = 10$ and $\rho((x, y), 0) = 0$ for $x \neq 0$. On the lower and upper walls, we apply periodic conditions $\rho((x, L_y + 1), t) = \rho((x, 0), t)$ and $\rho((x, -1), t) = \rho((x, L_y), t)$. The vertical interface between the subdomains is positioned at $x = 6$, with the FD scheme over the left region and the LB over the right one. Hence, the geometry is actually 1-dimensional.

This simple set-up with the above boundary and initial conditions allows for an analytic solution to the diffusion equation for the time-dependent dynamics,

$$\begin{aligned}\rho((x,y),t) =& \rho_0 \sum_{i=0}^{\infty} \mathrm{erfc}\left(\frac{\mathrm{L_x}(1-(1-\mathrm{x/L_x})+2\mathrm{i})}{2\sqrt{\mathrm{Dt}}}\right) \\ &- \rho_0 \sum_{i=0}^{\infty} \mathrm{erfc}\left(\frac{\mathrm{L_x}(1-(1+\mathrm{x/L_x})+2\mathrm{i})}{2\sqrt{\mathrm{Dt}}}\right)\end{aligned}$$

where $\mathrm{erfc(v)} = \int_{\mathrm{v}}^{\infty} \mathrm{e}^{-\mathrm{u}^2}\mathrm{du}$ and D is the diffusion coefficient (here $\rho_0 = 10$).

After several iterations, the system converges to a steady state in which the density profile is expected to be a linear function of x satisfying the boundary conditions. To illustrate the importance of a correct coupling at the subdomain interface, we plot in figs. 2a, 2b, 2c (left) the density profile obtained from the numerical solution with the full coupling $f_1 = f_1^{(0)} + f_1^{(1)}$, the one without the gradient correction, i.e. with the approximation $f_1 \approx f_1^{(0)}$, and the theoretical profile. We also plot in figs. 2a, 2b, 2c (right) the error with respect to the analytic solution. From the breaking of the slope observed in fig. 2c (left) we conclude that the gradient correction is necessary to obtain the correct density profile and continuity of the particle current.

We computed the relative error in the L^2-norm for the lattice resolution used in the simulation. For 100 (resp. 500, 5000) time steps, we get a 1.2% (resp. 0.6%, 0.3%) relative error for the full coupling. We expect the level of accuracy of our mapping to be the same as the LB itself: second order in the lattice spacing.

6 Conclusion

In this work, a LB scheme is spatially coupled to a FD scheme on a computational domain partitioned in two regions. We propose a way to relate the LB distribution functions f_i with the classical physical quantities and their derivatives. This is a first step towards coupling correctly a LB scheme with another method. Hence, to focus on the methodology only, we consider the simplest possible situation: a diffusion process solved by the LB approach on one region and

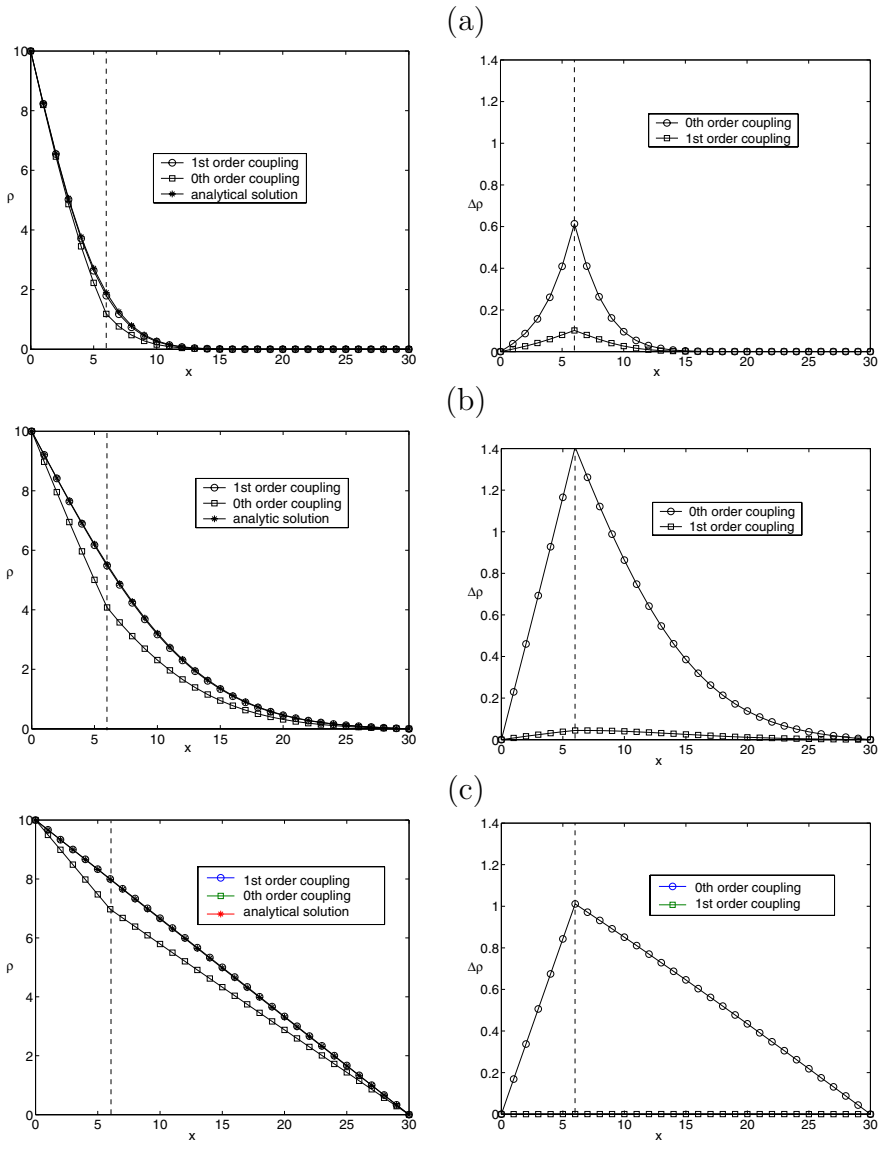


Fig. 2. The density profile along the horizontal axis of the domain Ω is plotted on the left side and the error with respect to the analytic solution on the right side. The squares correspond to a simulation in which the coupling algorithm does not include the gradient correction, whereas the circles represent the case where the full coupling is considered. The asterisks of the analytic solution are diffcult to distinguish from the circles of the full coupling. The total grid size is 31×31, and there are respectively (a)100, (b)500 and (c)5000 time steps. The vertical interface (dashed line) is located at $x = 6$. The relaxation time is $\tau = 0.7$, the lattice spacing $\Delta r = 1$ and the time step $\Delta t = 10^{-1}$

with a FD solver on the other. At the interface, the LB and FD are connected so as to preserve continuity of the physical quantities. The connection between the f_is and the standard macroscopic physical quantities is obtained using a multiscale Chapman-Enskog expansion to split the f_is in an equilibrium and a nonequilibrium part. The former is related to the particle density and the latter to its gradient. Our coupling methodology is indeed an approximation since we neglect higher-order derivatives in the nonequilibrium distributions. A validation was performed by running numerical simulations on a 2D domain and comparing them with an analytic solution. Our mapping has the same level of accuracy as the LB itself: second order in the lattice spacing.

Regarding future work, it seems natural to try to couple two domains with different meshes. A good enough interpolation is needed to keep the second-order accuracy of the coupling scheme (see [2]). This as well as the case of an irregular boundary is under investigation. Other perspectives include applying the same methodology to the LB reaction-diffusion, wave and fluid models. We also wish to couple different LB dynamics: a diffusive region with a convective one. With respect to coupling an incompressible Navier-Stokes solver with a "compressible" LB fluid model, we believe that the pressure obtained from the Navier-Stokes solver can be used to specify the density of the LB scheme, and conversely. We plan to examine this on a Poiseuille flow and on Blasius' problem.

References

1. B. Chopard, P.O. Luthi, A. Masselot, and A. Dupuis. Cellular automata and lattice Boltzmann techniques: An approach to model and simulate complex systems. *Advances in Complex Systems*, 5(2), 2002. http://cui.unige.ch/∼chopard/FTP/CA/advp.ps.gz.
2. A. Dupuis and B. Chopard. Theory and applications of alternative lattice Boltzmann refinement algorithm. *Phys. Rev. E*, 67:066707, 2003. In press.
3. L. Beilina. A hybrid method for elastic waves. Technical report, Finite Element Center, Chalmers University of Technology, 2003. Preprint 2003-01.
4. A. Malevanets and R. Kapral. Mesoscopic model for solvent dynamics. *J. of Chem. Phys.*, 110(17):8605–8613, May 1999.
5. D.A. Wolf-Gladrow. *Lattice-Gas Cellular Automata and Lattice Boltzmann Models: an Introduction*. LNM, 1725. Springer, Berlin, 2000.
6. B. Chopard and M. Droz. Cellular automata model for diffusion processes. *J. Stat. Phys.*, 64:859–892, 1991.
7. R. G. M. Van der Sman and M. H. Ernst. Convection-diffusion lattice boltzmann scheme for irregular lattices. *J. Comp. Phys.*, 160:766–782, 2000.

Accuracy versus Performance in Lattice Boltzmann BGK Simulations of Systolic Flows

Abdel Monim Artoli, Lilit Abrahamyan*, and Alfons G. Hoekstra

Section Computational Science, Laboratory for Computing, System Architecture and Programming,Faculty of Science,
University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{artoli,labraham,alfons}@science.uva.nl
http://www.science.uva.nl/research/scs/

Abstract. The aim of this work is to tune the lattice Boltzmann BGK (LBGK) simulation parameters in order to achieve optimum accuracy and performance for time dependent flows. We present detailed analysis of the accuracy and performance of LBGK in simulating pulsatile Newtonian flow in a straight rigid 3D tube. We compare the obtained velocity profiles and shear stress to the analytic Womersley solutions. A curved boundary condition is used for the walls and the accuracy and performance are compared to that obtained by using the bounce-back on the links. A technique to reduce compressibility errors during simulations based on reducing the Mach number is presented.

Keywords: Lattice Boltzmann Method, Systolic Cycle, Abdominal Aorta.

1 Introduction

Suitability and accuracy of the newly established lattice Boltzmann method in simulating time dependent fluid flows is demonstrated in the literature [1,2,3]. It is shown that use of curved boundary conditions noticeably enhances the accuracy as compared to using the simple bounce-back on the links [4,5].
The aim of this study is to end up with optimal simulation parameters for a desired accuracy with minimum simulation time. Simulation parameters for fixed Reynolds and Womersley parameters are studied. The paper is organized as follows: First, we briefly review the LBGK model we are using and then, we discuss the convergence behavior under different simulation choices and set up the optimal conditions for best performance.

2 The Lattice Boltzmann BGK Method

The method is based on a discretized Boltzmann equation with simplified collision operator via the single particle relaxation time approximation proposed by

* Corresponding author.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 548–555, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Bhatnagar, Gross and Krook in 1954. [6] The LBGK scheme is based on two steps; streaming to the neighboring nodes and colliding with local node populations which are represented by the probability f_i of a particle moving with a velocity $\mathbf{e}_i$ per unit time step δt. Populations, which are assumed Maxwellians, relax towards equilibrium during a collision process. The equilibrium distribution function

$$f_i = w_i \rho \left(1 + \frac{3}{v^2}\mathbf{e}_i \cdot \mathbf{u} + \frac{9}{2v^4}(\mathbf{e}_i \cdot \mathbf{u})^2 - \frac{3}{2v^2}\mathbf{u} \cdot \mathbf{u}\right), \quad (1)$$

is a low Mach number approximation to the expansion of the Maxwellian distribution. Here, w_i is a weighting factor, $v = \delta x/\delta t$ is the lattice speed, and δx and δt are the lattice spacing and the time step, respectively.[7]The well-known lattice BGK equation

$$f_i(\mathbf{x} + \mathbf{e}_i\delta_t, \mathbf{e}_i, t + \delta_t) - f_i(\mathbf{x}, \mathbf{e}_i, t) = -\frac{1}{\tau}[f_i(\mathbf{x}, \mathbf{e}_i, t) - f_i^{(0)}(\mathbf{x}, \mathbf{e}_i, t)] \quad (2)$$

can be obtained by discretizing the evolution equation of the distribution functions in the velocity space using a finite set of velocities $\mathbf{e}_i$. In this equation, τ is the dimensionless relaxation time.In this article, we used the standard quasi incompressible D3Q19 model which has three types of particles on each node; a rest particle, six particles moving along x, y and z principal directions. The hydrodynamic density, ρ, and the macroscopic velocity, $\mathbf{u}$, are determined in terms of the particle distribution functions from $\rho = \sum_i f_i = \sum_i f_i^{(eq)}$ and $\rho\mathbf{u} = \sum_i \mathbf{e}_i f_i = \sum_i \mathbf{e}_i f_i^{(eq)}$.

3 Simulations

In all simulations the flow rate in the tube is computed from a measured aortic pressure at the entrance. Its Fourier terms, up to the 8th harmonic, are used to set a suitable pressure gradient for obtaining an average Reynolds number $R_e = \frac{UD}{\nu} = 590$ and a Womersley parameter $\alpha = R\sqrt{\frac{\omega}{\nu}} = 16$, where $R = D/2$ is the radius of the tube, $\omega = 2\pi/T$ is the angular frequency and $T = 1/f$ is the period, with f being the number of heart beats per second. For details see [3]. Pressure boundary conditions are used for the inlet and the outlet boundaries and, for the walls, either the bounce-back on the links (BBL) or the Bouzidi boundary condition (BBC) [5] is used. We have performed three different categories of simulations of systolic flow in a 3D rigid tube benchmark and one simulation for the aortic bifurcation. The first set of simulations compares BBL with BBC. The second set deals with error behavior and the third set investigates the convergence behavior.[7]

3.1 Influence of the Wall Boundary Conditions

A wall boundary condition is needed to evaluate distributions coming from solid boundaries, while inlet/outlet conditions are needed to drive the flow. In this work, we investigate the error behavior for a complete systolic cycle which contains at least 16 harmonic terms. Moreover, we compare the error behavior for

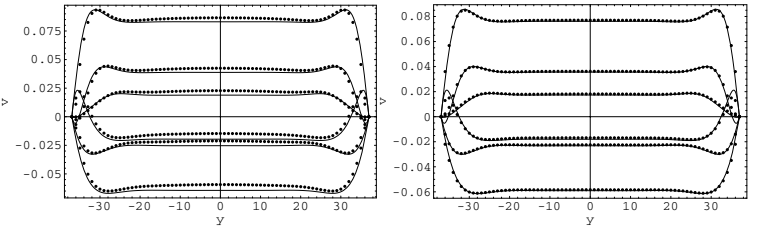
the two boundary conditions at a fixed Mach number in order to know how much gain we get from using a more accurate, rather sophisticated boundary condition, such as the BBC, over the less accurate but simple BBL.
The diameter of the tube is represented by 74 lattice nodes and the tube length is $L = 148$ lattice nodes. First, BBL is used to simulate systolic flow in the tube. The simulation parameters are set to yield the required Womersley and Reynolds numbers which are kept fixed to the values mentioned above. For this simulation, $T = 2000$, pressure gradient $G = 1.1 \times 10^{-5}$, $\tau = 0.55$ and A is amplitude of Furie series. Samples of obtained velocity profiles at different times of the systolic cycle are shown in Fig. 1(a) compared to the real part of the analytical Womersley solutions [8]. The average Mach number is 0.05 for this simulation. The relative error in velocity at each time step is defined by

$$Ev = \frac{\sum_{i=0}^{n} |\mathbf{u}_{th}(\mathbf{x}_i) - \mathbf{u}_{lb}(\mathbf{x}_i)|}{\sum_{i=1}^{n} |\mathbf{u}_{th}(\mathbf{x}_i)|} \tag{3}$$

where $\mathbf{u}_{th}(\mathbf{x}_i)$ is the analytical solution for the axial velocity and $\mathbf{u}_{lb}(\mathbf{x}_i)$ is the velocity obtained from the LBGK simulations. The bounce back on the links yields a time-averaged error of 0.11 at a Mach number of 0.05 for this specific simulation. The set of simulations is conducted for the same simulation parameters but using the BBC boundary condition. The agreement with analytical solutions enhances significantly, as shown in Fig. 1(b) and the average error reduces to approximately 0.03.

4 Grid Refinement

For pulsatile flow in rigid tubes, it is more convenient to tune the lattice viscosity ν, the Mach number M , the diameter D of the tube and the period T of the



(a) (b)

Fig. 1. Obtained samples of velocity profiles (dots) in lattice units during the systolic cycle in a 3D tube, compared to the analytical Womersley solution (lines) with: (a) BBL and (b)BBC wall boundary conditions

pulse. Changes in any of these parameters will result in changes in the space and time resolutions of the grid, and accordingly, the Mach and the Knudsen numbers. In LBGK simulations the error behavior is influenced by the Mach number, $M = \frac{U}{c_s}$, and the Knudsen number $\epsilon \sim (2\tau - 1)/D$ which approximates the mean free path in the lattice BGK method. The compressibility error of lattice BGK is given by [9]

$$\phi = \delta_x^2 \; R_e^2 \; \nu^2 \; c^2 \partial_t \rho \tag{4}$$

increases with increasing Reynolds and Mach numbers, since $R_e = MDc_s/\nu$ and goes to zero as $\tau = 1$ where the method behaves the same as finite difference methods. As a function of Womersley number, the compressibility error can be written as

$$\phi = \frac{2\delta_x^2 c_s c^2}{\pi} \; \frac{\alpha^2 MD}{St} \; \partial_t \rho \tag{5}$$

where $St = Df/U$ is the Strouhal number. In our simulations, the dimensionless hydrodynamic numbers (Re, α and St) and the Mach or Knudsen numbers are all fixed. This implies that the grid must be refined. There are three different ways to do this;

1. **fixed M method**: in which the the diameter D, the period T and the viscosity ν are changed while keeping fixed the Mach number.
2. **fixed τ method**: by changing the diameter D, the period T and the Mach number M while keeping fixed the lattice viscosity ν.
3. **fixed D method** : by keeping fixed the diameter D while changing the viscosity, period and the Mach number.

Table 1. Relative changes in simulation parameters under fixed Reynolds and Womersley numbers with respect to an n times change in one of the parameters of a reference simulation

Lattice Parameter	D'/D	ν'/ν	T'/T	U'/U	$\delta x'/\delta x$	$\delta t'/\delta t$	M'/M	ϵ'/ϵ
Fixed D	1	$1/n$	n	$1/n$	1	$1/n$	$1/n$	1/n
Fixed τ	n	1	n^2	$1/n$	$1/n$	$1/n^2$	$1/n$	$1/n$
Fixed M	n	n	n	1	$1/n$	$1/n$	1	1

The effects of these changes on the grid resolution are shown in Table 1, in which we assume an n times change in one of the parameters and compute the corresponding changes in the other parameters to return the fixed R_e and α. The fixed M method does not involve reduction of the Mach number, which is a major contributer to the error when considering time dependent flows and, therefore, it is not attractive in this study.

Table 2. Simulation parameters with respect to the reference simulation for which $\tau = 1, M = 0.50$ and $\epsilon = 1/74$, . The average error Eav associated with each set is shown for BBL, BBC1 and BBC2 boundary conditions

T	200	2000	20000
n	1	10	100
τ	1	0.55	0.505
$\frac{G'}{G}$	1	1/100	1/10000
$\frac{M'}{M}$	1	1/10	1/100
$\frac{\epsilon'}{\epsilon}$	1	1/10	1/100
$\frac{\epsilon' M'}{\epsilon M}$	1	1/100	1/10000
Eav, BBL	instable	0.120	0.027
Eav, BBC1	0.0627	0.0352	0.0253
Eav, BBC2	0.0615	0.0102	instable

Accuracy and Performance with the Fixed *D* Method. In all simulations, the system is initialized from rest and the simulation ends after 40-50 complete periods. The BBL, BBC1 (first order Bouzidi) and BBC2 (second order Bouzidi) are used separately for each parameter set to end up with 9 simulations. The simulations are performed on 8 nodes of a Beowulf cluster using slice decomposition.

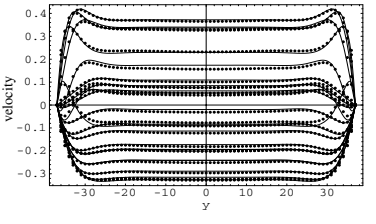


Fig. 2. Velocity profiles at $M = 0.50$ using the BBC1 boundary condition with overall average error of about 0.07, still less than the BBL results at a 10 times lower Mach number. The system is instable with the BBL at this Mach number

The mean time per iteration is 0.2 seconds. Samples of BBC1 obtained velocities and shear stress profiles during the systolic cycle are shown in Fig. 2 for $M = 0.5$, compared to the analytical Womersley solutions.Although M is relatively high, the agreement with the analytical solution is still better than those obtained by a 10 times smaller Mach number with BBL shown in Fig. 1(a). The last three rows in Table 2 list the average error associated with the three wall boundary conditions.

Accuracy and Performance with the Fixed τ Method. In order to reduce simulation time, it is necessary to have a large time-step in a coarse grid at a high Mach number.We use the fixed τ method to perform a set of simulations in which the period is set to the smallest possible value that leads to a stable solution on the coarsest grid. Then the corresponding values for the pressure gradient and

Table 3. Simulation parameters used to enhance the spatial resolution. The mean relative error, Eav, is listed for each case

D	21	65	105	21 n
n	1	3	5	n
T	40	360	1000	40 n^2
$\frac{G'}{G}$	1	1/27	1/125	$1/n^3$
$\frac{M'}{M}$	1	1/3	1/5	$1/n$
δt	1	1/9	1/25	$1/n^2$
Eav, BBL	0.2412	0.1189	0.0262	
Eav, BBC	0.2301	0.0557	0.0262	
Eav, BBC2	instable	0.0560	0.0266	

the relaxation parameter are set to yield the desired Womersley and Reynolds numbers. The convergence behavior is studied by grid refinement in both δ_x and δ_t, as explained in Table 1. The simulation parameters are listed in Table 3 together with obtained average errors associated with the three used boundary conditions. As it is shown in Fig. 3, at least second order convergence behavior is guaranteed with this method. Moreover, solutions with periods smaller than the fixed D method are stable and therefore the simulation time is less, but it scales as n^2. The convergence behavior as a function of time for this method is shown in Fig. 4, which shows the difference between the analytical and obtained velocity profiles at different simulation times. In this figure, the method converges to a reasonable accuracy after 40 complete periods, similar to the fixed D method, but with a major computational gain, since the length of the period is smaller (i.e. δ_t is larger). This figure also illustrates that the error is localized near to the walls, where large gradients exist, and it does not enhance noticeably near to the walls on the same grid. Table 4 lists the error dependence as a function of simulation times for BBL, BBC1 and BBC2 boundary conditions for a tube with $D = 65$ lattice nodes. In conclusion, this method is computationally more feasible than the fixed D method.

4.1 Convergence Behavior

Convergence to steady state of the lattice Boltzmann method is significantly affected by two local processes; initialization and boundary conditions. In this section, we focus on the influence of initial and boundary conditions on the convergence behavior.

Table 4. Mean, variance and mean deviation of the Relative error for BBL, BBC boundary conditions with $D = 65$ lattice nodes

Boundary condition	Mean error (Eav)	Variance	Mean deviation
BBL	0.1189	0.0013	0.0219
BBC	0.0557	0.0027	0.0352
BBC2	0.0560	0.0027	0.0350

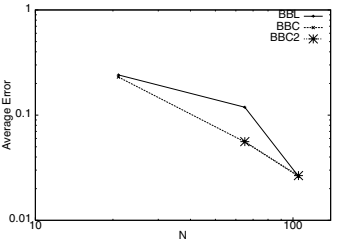


Fig. 3. Convergence behavior obtained by reducing the grid spacing n times, time-step n^2 times and increasing the period n^2 times, for the BBL, BBC and BBC2 boundary conditions as a function of grid points.The relaxation parameter is kept constant and the body force is reduced n^3 times to return the same Reynolds and Womersley parameters at $Re =$ and $\alpha = 16$

Convergence and Walls Boundary Conditions. For the walls, bounce-backs and curved boundaries can be recognized. The bounce back is a very efficient boundary condition since it only involves a single memory swapping process for each relevant distribution on each node on the surface of the simulated object. For all curved boundaries, the exact position of the walls is determined at least once if the boundary is fixed and needs to be computed dynamically for moving boundaries. This is more costly than using the bounce back rule.[7]

Inlet and Outlet Conditions. For non-periodic geometry, inlets and outlets need to be treated differently in the following manner:

- Velocity and pressure : assign one and compute the other [10], assign both (only for inlets) extrapolate or no flux normal to the walls (only for outlets).
- Unknown distributions: compute explicitly [10], set to their equilibrium, copy from nearest neighbors, interpolate or extrapolate.

For the first item at least 15 additions and two multiplications are needed per node on the boundary and therefore is at least 15 times more expensive than periodic boundary conditions. A reasonable choice is then to assign pressure and compute velocity at the inlet, no-flux at the outlets and set the unknown distributions to their equilibrium values. If the outlets are far enough from inflow, copying from upstream would be the most efficient outlet condition.

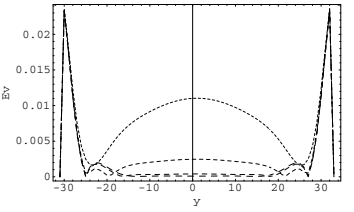


Fig. 4. Local deviations from Analytical solutions, δE, computed for the velocity field at $t = 20T$ (top Curve), $30T, 40T$ and $50T$ (bottom curve). The diameter of the tube is represented by 65 nodes and the period is $T = 360$ sampling points

5 Conclusions

We have shown that the lattice Boltzmann BGK method is an accurate and efficient method as a solver for time-dependent flows. Different methods for performing time dependent flows at fixed simulation parameters are tested in terms of accuracy and performance. An aortic pressure is used as an inlet condition to drive the flow in a 3d rigid tube and the Womersley solution is recovered to an acceptable accuracy. The influence of walls, inlet and outlet boundary conditions on accuracy and performance is studied in details as a function of Mach and Knudsen numbers.

Acknowledgments. We acknowledge Prof. Peter M.A. Sloot for his support and helpfull discussions. This work was partially funded by the Steunfonds Soedanese Studenten, Leiden, The Netherlands, and by the Dutch National Science Foundation, NWO, through the DIME-Token2000 project.

References

1. Xiaoyi He and Li-Shi Luo, *J. Stat. Phys.* **88**, 927 (1997); Q. Zou, S. Hou, S.Chen and G. D. Doolen, *J. Stat. Phys.* **81**, 35 (1995); Guo, Shi and Wang, *J. Comp. Phys.* **165**, 288 (2000).
2. M. Krafczyk, M. Cerrolaza, M. Schulz, E. Rank, *J. Biomechanics* **31**, 453 (1998).
3. A. M. Artoli, A. G. Hoekstra, and P. M. A. Sloot, *Int. J. Mod. Phys.C*, **13** No. 8, 1119 (2002).
4. O. Fillippova and D. Hänel, *J. Comp. Phys.* **147**, 219 (1998).
5. M. Bouzidi, M. Fidouss and P. Lallemand, *Phys. Fluids* **13**, 3452 (2001).
6. P. L. Bhatnagar, E. P. Gross, and M. Krook, *Phys. Rev. A*, **94**, 511 (1954).
7. A. M. Artoli, *Mesoscopic Computational Haemodinamics*, PhD thesis, University van Amsterdam, The Netherlands (2003).
8. C. Pozrikidis, *Introduction to Theoretical and Computational Fluid Dynamics*, OUP, (1997).
9. D. Holdych, D. Noble, J. G. Georgiadis, and R. O. Buckius, Proceedings of the 11th international conference of Discrete simulation of fluid dynamics and soft condensed matter, Shanghai, China, August 5-9 (2002).
10. Q. Zao and X. He, *Phys. fluids* **9**, 1591(1997).

Mesoscopic Modelling of Droplets on Topologically Patterned Substrates

A. Dupuis and J.M. Yeomans

Theoretical Physics,University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK.

Abstract. We present a lattice Boltzmann model to describe the spreading of droplets on topologically patterned substrates. We apply it to model superhydrophobic behaviour on surfaces covered by an array of micron-scale posts. We find that the patterning results in a substantial increase in contact angle, from 110^o to 156^o.

1 Introduction

A droplet in contact with a substrate will try to spread to an equilibrium shape determined by Young's law which describes the balance of surface tensions. There are many parameters which affect this process. For example surface disorder in the form of chemical or topological heterogeneities can pin a droplet or change its final shape. This has usually been viewed as a nuisance in experiments and applications. However with the advent of microfabrication techniques it is becoming possible to harness controlled surface topologies to explore new physical phenomena.

A beautiful example of this, inspired by the leaves of the lotus plant, is a superhydrophobic substrate. The angle θ between the tangent plane and the droplet is usually known as the contact angle. The higher the contact angle the more repellent the surface. There are applications, for example raincoats and windscreens, where repellent surfaces are highly desirable. Surface coatings and chemical modifications of the substrate are common ways to increase the contact angle but it is difficult to achieve an angle of more than 120^o. However surfaces patterned with posts on a micron length scale allow contact angles of 160^o to be reached [1,2].

The aim of this paper is to present a lattice Boltzmann algorithm which can be used to investigate the behaviour of droplets on topologically patterned substrates. Lattice Boltzmann is a particularly appropriate approach in that it solves the Navier Stokes equations but also inputs the thermodynamic information such as surface tensions needed to describe the behaviour of droplets. Moreover its natural length scale, for fluids such as water, is of order microns where much of the exciting new physics is expected to appear. The method has already shown its capability in dealing with spreading on surfaces with chemical patterning [3].

In section 2 we summarise the algorithm and, particularly, describe the new thermodynamic and velocity boundary conditions needed to treat surfaces with

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 556–563, 2004.
© Springer-Verlag Berlin Heidelberg 2004

topological patterning. In section 3 we present results for a substrate patterned by an array of posts. The patterning leads to a considerable increase in contact angle. Finally we discuss directions for future work using this approach.

2 The Mesoscopic Model

We consider a liquid-gas system of density $n(\mathbf{r})$ and volume V. The surface of the substrate is denoted by S. The equilibrium properties are described by the free energy

$$\Psi = \int_V \left(\psi_b(n) + \frac{\kappa}{2} \left(\partial_\alpha n \right)^2 \right) dV + \int_S \psi_c(n)\, dS. \tag{1}$$

$\psi_b(n)$ is the free energy in the bulk. We choose a Van der Waals form

$$\psi_b(n) = p_c \left(\nu_n + 1 \right)^2 \left(\nu_n^2 - 2\nu_n + 3 - 2\beta\tau_w \right) \tag{2}$$

where $\nu_n = (n - n_c)/n_c$, $\tau_w = (T_c - T)/T_c$ and $p_c = 1/8$, $n_c = 7/2$ and $T_c = 4/7$ are the critical pressure, density and temperature respectively and β is a constant typically equal to 0.1. The bulk pressure

$$p_b = p_c(\nu_n + 1)^2(3\nu_n^2 - 2\nu_n + 1 - 2\beta\tau_w). \tag{3}$$

The derivative term in equation (1) models the free energy associated with an interface. κ is related to the surface tension. $\psi_c(n_s) = \phi_0 - \phi_1 n_s + \cdots$ is the Cahn surface free energy [4] which controls the wetting properties of the fluid.

The lattice Boltzmann algorithm solves the Navier-Stokes equations for this system. Because interfaces appear naturally within the model it is particularly well suited to looking at the behaviour of moving drops.

2.1 The Lattice Boltzmann Algorithm

The lattice Boltzmann approach follows the evolution of partial distribution functions f_i on a regular, d-dimensional lattice formed of sites $\mathbf{r}$. The label i denotes velocity directions and runs between 0 and z. $DdQz + 1$ is a standard lattice topology classification. The $D3Q15$ lattice we use here has the following velocity vectors $\mathbf{v}_i$: $(0,0,0)$, $(\pm1,\pm1,\pm1)$, $(\pm1,0,0)$, $(0,\pm1,0)$, $(0,0,\pm1)$ in lattice units as shown in fig. 1.

The lattice Boltzmann dynamics are given by

$$f_i(\mathbf{r} + \Delta t \mathbf{v}_i, t + \Delta t) = f_i(\mathbf{r}, t) + \frac{1}{\tau} \left(f_i^{eq}(\mathbf{r}, t) - f_i(\mathbf{r}, t) \right) \tag{4}$$

where Δt is the time step of the simulation, τ the relaxation time and f_i^{eq} the equilibrium distribution function which is a function of the density $n = \sum_{i=0}^{z} f_i$ and the fluid velocity $\mathbf{u}$, defined through the relation

$$n\mathbf{u} = \sum_{i=0}^{z} f_i \mathbf{v}_i. \tag{5}$$

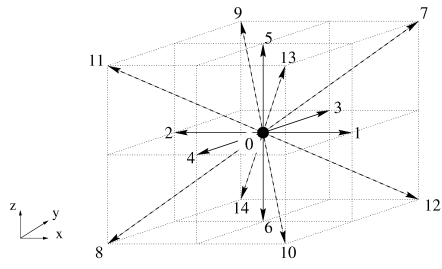


Fig. 1. Topology of a $D3Q15$ lattice. The directions i are numbered and correspond to the velocity vectors $\mathbf{v}_i$

The relaxation time tunes the kinematic viscosity as [5]

$$\nu = \frac{\Delta \mathbf{r}^2}{\Delta t} \frac{C_4}{C_2} (\tau - \frac{1}{2}) \tag{6}$$

where $\Delta \mathbf{r}$ is the lattice spacing and C_2 and C_4 are coefficients related to the topology of the lattice. These are equal to 3 and 1 respectively when one considers a $D3Q15$ lattice (see [6] for more details).

It can be shown [7] that equation (4) reproduces the Navier-Stokes equations of a non-ideal gas if the local equilibrium functions are chosen as

$$\begin{aligned} f_i^{eq} &= A_\sigma + B_\sigma u_\alpha v_{i\alpha} + C_\sigma \mathbf{u}^2 + D_\sigma u_\alpha u_\beta v_{i\alpha} v_{i\beta} + G_{\sigma\alpha\beta} v_{i\alpha} v_{i\beta}, \quad i > 0, \\ f_0^{eq} &= n - \sum_{i=1}^{z} f_i^{eq} \end{aligned} \tag{7}$$

where Einstein notation is understood for the Cartesian labels α and β (i.e. $v_{i\alpha} u_\alpha = \sum_\alpha v_{i\alpha} u_\alpha$) and where σ labels velocities of different magnitude. A possible choice of the coefficients is [6]

$$\begin{aligned} A_\sigma &= \frac{w_\sigma}{c^2} \left(p_b - \frac{\kappa}{2} (\partial_\alpha n)^2 - \kappa n \partial_{\alpha\alpha} n + \nu u_\alpha \partial_\alpha n \right), \\ B_\sigma &= \frac{w_\sigma n}{c^2}, \quad C_\sigma = -\frac{w_\sigma n}{2c^2}, \quad D_\sigma = \frac{3 w_\sigma n}{2c^4}, \\ G_{1\gamma\gamma} &= \frac{1}{2c^4} \left(\kappa (\partial_\gamma n)^2 + 2\nu u_\gamma \partial_\gamma n \right), \quad G_{2\gamma\gamma} = 0, \\ G_{2\gamma\delta} &= \frac{1}{16 c^4} \left(\kappa (\partial_\gamma n)(\partial_\delta n) + \nu (u_\gamma \partial_\delta n + u_\delta \partial_\gamma n) \right) \end{aligned} \tag{8}$$

where $w_1 = 1/3$, $w_2 = 1/24$ and $c = \Delta \mathbf{r} / \Delta t$.

2.2 Wetting Boundary Conditions

The major challenge in dealing with patterned substrates is to handle the boundary conditions correctly. We consider first wetting boundary conditions which

control the value of the density derivative and hence the contact angle. For flat substrates a boundary condition can be set by minimising the free energy (1) [4]

$$\hat{\mathbf{s}} \cdot \nabla n = -\frac{\phi_1}{\kappa} \tag{9}$$

where $\hat{\mathbf{s}}$ is the unit vector normal to the substrate. It is possible to obtain an expression relating ϕ_1 to the contact angle θ as [6]

$$\phi_1 = 2\beta\tau_w\sqrt{2p_c\kappa}\,\mathrm{sign}\left(\frac{\pi}{2}-\theta\right)\sqrt{\cos\frac{\alpha}{3}\left(1-\cos\frac{\alpha}{3}\right)} \tag{10}$$

where $\alpha = \cos^{-1}(\sin^2\theta)$ and the function sign returns the sign of its argument.

Equation (9) is used to constrain the density derivative for sites on a flat part of the substrate. However, no such exact results are available for sites at edges or corners. We work on the principle that the wetting angle at such sites should be constrained as little as possible so that, in the limit of an increasingly fine mesh, it is determined by the contact angle of the neighbouring flat surfaces.

For edges (labels $9-12$ in fig. 2) and corners (labels $1-4$) at the top of the post each site has 6 neighbours on the computational mesh. Therefore these sites can be treated as bulk sites.

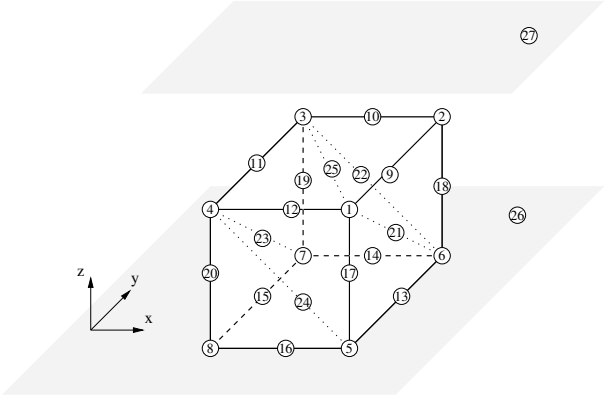


Fig. 2. Sketch of a post on a substrate. Encircled numbers label sites in different topological positions. Labels 26 and 27 denote sites on the bottom ($z = z_{min}$) and the top ($z = z_{max}$) of the domain respectively

At bottom edges where the post abuts the surface (labels $13-16$ in fig. 2) density derivatives in the two directions normal to the surface (e.g. x and z for sites labeled 13) are calculated using

$$\partial_z n = \partial_{x/y} n = -\frac{1}{\sqrt{2}}\frac{\phi_1}{\kappa} \tag{11}$$

where the middle term constrains the density derivative in the appropriate direction x or y.

At bottom corners where the post joins the surface (labels $5-8$ in fig. 2) density derivatives in both the x and y directions are known. Therefore these sites are treated as planar sites.

2.3 Velocity Boundary Conditions

We impose a no-slip boundary condition on the velocity. As the collision operator (the right hand side of equation (4)) is applied at the boundary the usual bounce-back condition is not appropriate as it would not ensure mass conservation [8].

Indeed after applying equation (4) there are missing fields on the substrate sites because no fluid has been propagated from the solid. Missing fields are

Table 1. Velocity boundary conditions

Label	Conditions
1	$f_{13} = f_{14}$
2	$f_7 = f_8$
3	$f_9 = f_{10}$
4	$f_{11} = f_{12}$
5	$f_5 = f_6$ $f_{13} = (f_3 - f_4 - f_1 + f_2)/2 + f_9 + f_{14} - f_{10}$ $f_{11} = (f_1 - f_2)/2 - f_9 + f_{10} + f_{12}$ $f_7 = (-f_3 + f_4)/2 + f_8 - f_9 + f_{10}$
6	$f_5 = f_6$ $f_{13} = (f_3 - f_4)/2 - f_{11} + f_{12} + f_{14}$ $f_9 = (f_1 - f_2)/2 - f_{11} + f_{10} + f_{12}$ $f_7 = (-f_1 + f_2 - f_3 + f_4)/2 + f_8 + f_{11} - f_{12}$
7	$f_5 = f_6$ $f_{11} = (f_3 - f_4)/2 - f_{13} + f_{12} + f_{14}$ $f_9 = (f_1 - f_2 - f_3 + f_4)/2 + f_{13} - f_{14} + f_{10}$ $f_7 = (-f_1 + f_2)/2 + f_8 - f_{13} + f_{14}$
8	$f_5 = f_6$ $f_{11} = (f_3 - f_4 + f_1 - f_2)/2 + f_7 - f_8 + f_{12}$ $f_9 = (-f_3 + f_4)/2 - f_7 + f_8 + f_{10}$ $f_{13} = (-f_1 + f_2)/2 - f_7 + f_8 + f_{14}$
9	$f_{13} = f_{14}$; $f_7 = f_8$
10	$f_9 = f_{10}$; $f_7 = f_8$
11	$f_9 = f_{10}$; $f_{11} = f_{12}$
12	$f_{13} = f_{14}$; $f_{11} = f_{12}$
13	$f_5 = f_6$ $f_1 = 2(-f_{10} + f_9 + f_{11} - f_{12}) + f_2$ $f_{13} = (f_3 - f_4)/2 - f_{11} + f_{12} + f_{14}$ $f_7 = (-f_3 + f_4)/2 + f_8 - f_9 + f_{10}$
14	$f_5 = f_6$ $f_9 = (f_1 - f_2)/2 - f_{11} + f_{10} + f_{12}$ $f_7 = (-f_1 + f_2)/2 + f_8 - f_{13} + f_{14}$ $f_3 = 2(-f_{12} + f_{11} + f_{13} - f_{14}) + f_4$
15	$f_5 = f_6$ $f_2 = 2(-f_{14} + f_7 + f_{13} - f_8) + f_1$ $f_{11} = (f_3 - f_4)/2 - f_{13} + f_{12} + f_{14}$ $f_9 = (-f_3 + f_4)/2 + f_8 - f_7 + f_{10}$
16	$f_5 = f_6$ $f_{11} = (f_1 - f_2)/2 - f_9 + f_{10} + f_{12}$ $f_{13} = (-f_1 + f_2)/2 - f_7 + f_8 + f_{14}$ $f_4 = 2(-f_{10} + f_7 + f_9 - f_8) + f_3$
17	$f_{10} = f_9$; $f_{13} = f_{14}$
18	$f_7 = f_8$; $f_{12} = f_{11}$
19	$f_9 = f_{10}$; $f_{14} = f_{13}$
20	$f_8 = f_7$; $f_{11} = f_{12}$
21	$f_1 = f_2$; $f_7 = f_8$ $f_{12} = (-f_3 + f_4)/2 + f_{11}$ $f_{13} = (-f_5 + f_6)/2 + f_{14}$ $f_{10} = (f_3 - f_4 + f_5 - f_6)/2 + f_9$
22	$f_3 = f_4$; $f_7 = f_8$ $f_9 = (-f_5 + f_6)/2 + f_{10}$ $f_{14} = (f_1 - f_2 + f_5 - f_6)/2 + f_{13}$ $f_{12} = (-f_1 + f_2)/2 + f_{11}$
23	$f_2 = f_1$; $f_8 = f_7$ $f_{11} = (f_3 - f_4)/2 + f_{12}$ $f_{14} = (f_5 - f_6)/2 + f_{13}$ $f_9 = f_{10} + (-f_3 + f_4 - f_5 + f_6)/2$
24	$f_8 = f_7$; $f_4 = f_3$ $f_{11} = (f_1 - f_2)/2 + f_{12}$ $f_{10} = (f_5 - f_6)/2 + f_9$ $f_{13} = (-f_5 + f_6 - f_1 + f_2)/2 + f_{14}$
25	$f_7 = f_8$; $f_5 = f_6$ $f_{11} = (f_1 - f_2 + f_3 - f_4)/2 + f_{12}$ $f_9 = (-f_3 + f_4)/2 + f_{10}$ $f_{13} = (-f_1 + f_2)/2 + f_{14}$
26	$f_7 = f_8$; $f_5 = f_6$ $f_{11} = (f_1 - f_2 + f_3 - f_4)/2 + f_{12}$ $f_9 = (-f_3 + f_4)/2 + f_{10}$ $f_{13} = (-f_1 + f_2)/2 + f_{14}$
27	$f_6 = f_5$; $f_8 = f_7$ $f_{10} = (f_3 - f_4)/2 + f_9$ $f_{14} = (f_1 - f_2)/2 + f_{13}$ $f_{12} = (-f_3 + f_4 - f_1 + f_2)/2 + f_{11}$

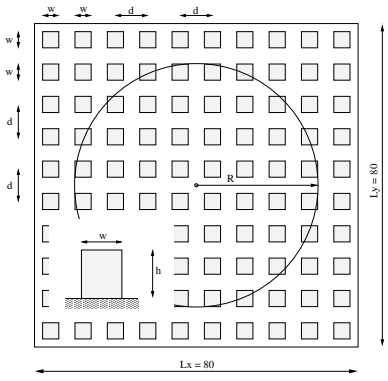


Fig. 3. Sketch of the substrate. Dimensions are expressed in simulation units. Shaded areas are posts

determined to fulfill the no-slip condition given by equation (5) with $\mathbf{u} = 0$. This does not uniquely determine the f_i's. For most of the cases (i.e. $1-20$) arbitrary choices guided by symmetry are used to close the system. This is no longer possible for sites $21-27$ where four asymmetrical choices are available. Selecting one of those solutions or using a simple algorithm which chooses one of them at random each time step leads to very comparable and symmetrical results. Hence we argue that an asymmetrical choice can be used. Possible conditions, which are used in the results reported here, are listed in table 1.

The conservation of mass is ensured by setting a suitable rest field, f_0, equal to the difference between the density of the missing fields and the one of the fields entering the solid after collision.

3 Results

As an example we consider here the superhydrophobic behaviour of droplet spreading on a substrate patterned by square posts arranged as in fig. 3.

The size of the domain is $L_x \times L_y \times L_z = 80 \times 80 \times 80$ and the height, spacing and width of posts are $h = 5$, $d = 8$ and $w = 4$ respectively. A spherical droplet of radius $R = 30$ is initially centered around the point $(x; y; z) = (41; 41; 36)$. The contact angle $\theta_{input} = 110^o$ is set on every substrate site. The surface tension and the viscosity are tuned by choosing parameters $\kappa = 0.002$ and $\tau = 0.8$ respectively. The liquid density n_l and gas density n_g are set to $n_l = 4.128$ and $n_g = 2.913$ and the temperature $T = 0.4$.

Fig. 4 shows the final state attained by the droplet for different substrates and initial conditions. For comparison fig. 4(a) shows a planar substrate. The equilibrium contact angle is $\theta_a = 110^o = \theta_{input}$ as expected [6]. In fig. 4(b) the substrate is patterned and the initial velocity of the drop is zero. Now the contact angle is $\theta_b = 156^o$, a demonstration of superhydrophobic behaviour. Fig. 4(c)

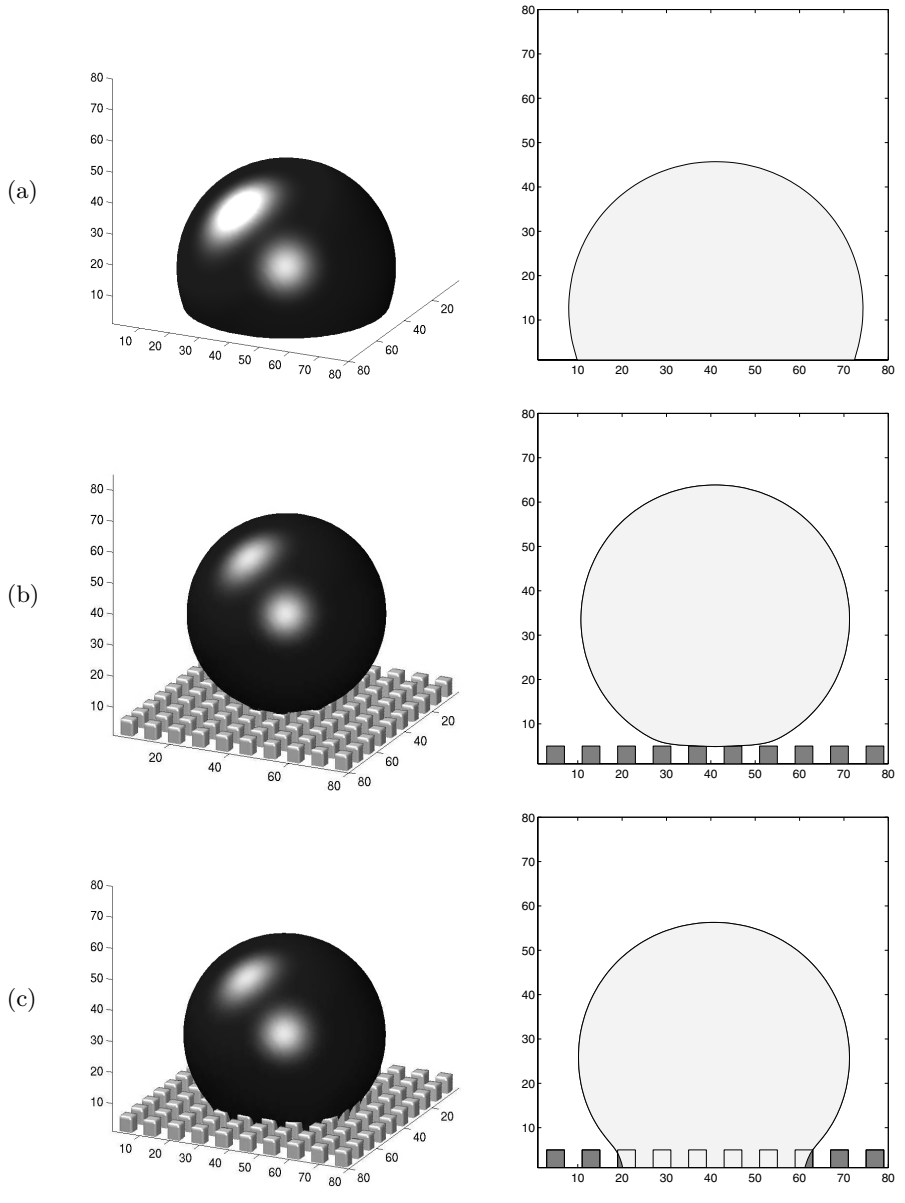


Fig. 4. Final states of a spreading droplet. The right column reports cuts at $y = 41$. (a) The substrate is flat and homogeneous. (b) The substrate is decorated with posts and the initial velocity of the droplet is 0. (c) Same geometry as (b) but the droplet reaches the substrate with a velocity $0.01\Delta\mathbf{r}/\Delta t$. Each of these simulations ran for approximately 8 hours on 8 processors on a PC cluster

reports an identical geometry but a drop with an initial impact velocity. Now the drop is able to collapse onto the substrate and the final angle is $\theta_b = 130^o$. These angles are compatible with those in [2] where similar parameters are considered.

For the parameter values used in these simulations the state with the droplet suspended on the posts has a slightly higher free energy than the collapsed state. It is a metastable state and the droplet needs an impact velocity to reach the true thermodynamic ground state. For macroscopic drops gravity will also be important in determining whether the drop remains suspended on top of the posts. Extrand has predicted the minimum post perimeter density necessary for a droplet to be suspended [9]. A next step will be to add gravity to the simulation to compare to his prediction.

Superhydrophobicity occurs over a wide range of d, the distance between the posts. For suspended drops of this size and $d \geq 12$ the drop resides on a single post and the contact angle is 170^o. For $d < 12$ the contact angle lies between 148^o and 156^o with the range primarly due to the commensurability between drop radius and post spacing.

It is of course also of interest to look further at the dynamics of the spreading. The droplet random motion reported in [2] and the bouncing back of droplets on nanotubes [10] pose many interesting directions for future research.

References

1. J. Bico, C. Marzolin, and D. Quéré. Pearl drops. *Eur. Phys. Lett.*, 47(2):220–226, 1999.
2. D. Öner and T.J. McCarthy. Ultrahydrophobic surfaces. Effects of topography length scales on wettability. *Langmuir*, 16:7777–7782, 2000.
3. J. Léopoldès, A. Dupuis, D.G. Bucknall, and J.M. Yeomans. Jetting micron-scale droplets onto chemically heterogeneous surfaces. *Langmuir*, 19(23):9818–9822, 2003.
4. J.W. Cahn. Critical point wetting. *J. Chem. Phys.*, 66:3667–3672, 1977.
5. S. Succi. *The Lattice Boltzmann Equation, For Fluid Dynamics and Beyond.* Oxford University Press, 2001.
6. A. Dupuis and J.M. Yeomans. Lattice Boltzmann modelling of droplets on chemically heterogeneous surfaces. *Fut. Gen. Comp. Syst.*, in press.
7. M.R. Swift, E. Orlandini, W.R. Osborn, and J.M. Yeomans. Lattice Boltzmann simulations of liquid-gas and binary fluid systems. *Phys. Rev. E*, 54:5051–5052, 1996.
8. B. Chopard and A. Dupuis. A mass conserving boundary condition for lattice Boltzmann models. *Int. J. Mod. Phys. B*, 17:103–106, 2002.
9. C.W. Extrand. Model for contact angles and hysteresis on rough and ultraphobic surfaces. *Langmuir*, 18:7991–7999, 2002.
10. K.K.S. Lau, J. Bico, K.B.K. Teo, M. Chhowalla, G.A.J. Amaratunga, W.I. Milne, G.H. McKinley, and K.K. Gleason. Superhydrophobic carbon nanotube forests. *Nano Lett.*, in press.

Soot Particle Deposition within Porous Structures Using a Method of Moments – Lattice Boltzmann Approach

Bernhard F.W. Gschaider, Claudia C. Honeger, and Christian E.P. Redl

Christian-Doppler-Laboratory for Applied Computational Thermofluiddynamics,
Mining University Leoben,
Franz-Josef-Strasse 18, 8700 Leoben, Austria

Abstract. This paper deals with the combination of two computational methods to simulate the flow of particle laden fluids through porous structures: the Lattice Boltzmann Method (LBM) which is a method to solve the Navier-Stokes equation in complex geometries and the Method of Moments (MoM) which describes the time evolution of non-homogeneous particle distributions. The combination of these methods makes it possible to take phenomena into account that depend on particle size and size distribution of the transported material. It is also possible to simulate changes in the size distribution.

1 Introduction

The simulation of particle laden flows covers a large area of useful applications e.g. emissions of internal combustion engines, transport of pollutants in the ground-water, aerosols in the atmosphere.

Most of these applications involve the need to simulate a large number of particles with a non-uniform size distribution. This size distribution is described by a Probability Density Function (PDF). The shape of this PDF may change due to coagulation, nucleation or the breaking-up of particles.

A typical example for combined fluid-particle flow is the exhaust gas flow of combustion engines. A variety of problems arise within this field of application. One of these is the soot particle laden gas flow through particulate filter materials. The purpose of these simulations is to judge the efficiency of the filter material with respect to deposition of soot particles on the pore walls.

One difficulty with the simulation of flows inside porous materials is the complexity of the pore geometry. This makes it hard to generate a body fitted computational grid essential for traditional methods of flow simulation like the Finite Volume Method. Therefore a Lattice Boltzmann Method (LBM), which is especially well suited for complex geometries, is used. The computational domains for the LBM can be generated automatically from computer tomography or reconstructed using statistical correlations.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 564–571, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Lattice Boltzmann Method

A standard incompressible formulation of the Lattice Boltzmann equation is used for simulating the fluid flow [1] [2].

The incompressible Lattice Boltzmann equation can be written as :

$$f_i(t+1, \mathbf{x}+\mathbf{e}_i) = f_i(t, \mathbf{x}) - \frac{1}{\tau}(f_i(t, \mathbf{x}) - {f_i}^{eq}(t, \mathbf{x})) \tag{1}$$

where the equilibrium distributions are given by

$${f_i}^{eq} = t_i\{p + p_0(\frac{\mathbf{e}_{i\alpha}\mathbf{u}_\alpha}{{c_s}^2} + \frac{\mathbf{u}_\alpha\mathbf{u}_\beta}{2{c_s}^2}(\frac{\mathbf{e}_{i\alpha}\mathbf{e}_{i\beta}}{{c_s}^2} - \delta_{\alpha\beta}))\} + S_i \tag{2}$$

In the above equations, p_0 is a reference pressure, f_i are the distribution functions and $\mathbf{e}_i$ are the lattice vectors. S_i is a volumetric source term which may result from the particle transport. The lattice dependent weighting factors for the D2Q9 model are $t_0 = 4/9$ for the rest particle distribution, $t_{1-4} = 1/9$ for the cartesian directions and $t_{5-8} = 1/36$ for the diagonal directed lattice vectors. The lattice speed of sound is $c_s = 1/\sqrt{3}$.

The hydrodynamic quantities can be obtained by computing the moments of the particle distribution functions :

$$p = \sum_i f_i \tag{3}$$

$$p_0\mathbf{u} = \sum_i f_i\mathbf{e}_i \tag{4}$$

The relaxation parameter τ is related to the viscosity as follows:

$$\tau = \frac{6\nu + 1}{2} \tag{5}$$

Due to the tremendous computational overhead transport equations for species are not solved within the LBM framework but by other means, e. g. finite differences. Another limiting factor is the poor numerical characteristics of LB equations for scalar transport. As it is essential to have a minimum of numerical diffusion for the transport of the moments in the MoM in this works a Van-Leer flux splitting scheme [3] is adopted to solve the transport equations for these moments. The fluid flow is solved by the LBM which accuracy for laminar flows is widely accepted (see e.g. [4]).

3 Particle Models

The appropriate particle model for the presented approach under discussion must include the possibility to simulate a large number of particles of different sizes and also needs to account for their change in size.

3.1 Comparison of Particle Models

Traditionally there are two approaches to model particles: The *Lagrangian* approach, where each particle is simulated individually by calculating its motion due to inertia and the drag force caused by the fluid. This approach can simulate the size changes in the particle population by detecting collisions between individual particles. The disadvantage of this approach is that only a limited number of particles can be treated with the available computational power. An example of a Lagrangian particle model in combination with a Lattice Boltzmann solver can be found in the papers of Redl et al [5] or Filippova et al [6].

The other approach is the *Eulerian* particle model. In this model the particles are represented by a density ϕ of particles of uniform size and properties. These particles are transported by the fluid without deviating from the streamlines using the usual transport equation:

$$\frac{\partial \rho\phi}{\partial t} = \frac{\partial}{\partial x_i}\left(\rho u_i \phi - D\frac{\partial \phi}{\partial x_i}\right) = S \tag{6}$$

In order for this approach to be valid the particles have to be small enough:

$$\frac{F_{drag}}{F_{inertial}} \propto \frac{r\mu v}{r^3 \rho v^2} \gg 1 \tag{7}$$

This approach allows an unlimited number of particles, but it can not model changes in the particle size distribution.

An approach, that is widely used together with the LBM, is the usage of a cellular automata formulation where the number of particles in each computational cell is stored. Particle transport and deposition are updated at each time-step by calculating the probability that a particle moves to a neighboring cell (using the velocities from the flow solution) and moving the corresponding number of particles. These models allow for a much larger number of particles than the Lagrangian particle model, but they assume a homogeneous particle population and they don't take particle-particle interactions into account. They have been successfully used for the calculation of snow and sand transport ([7],[8]).

3.2 Method of Moments (MoM)

In order to take account for the change in particle size distribution, particles are separated into classes of mass $m_i = i\Delta m$ where the particle class i contains N_i particles. The evolution of these particle classes can be described by the Smoluchowski model [9]:

$$\frac{dN_1}{dt} = -\sum_{j=1}^{\infty} \beta_{1,j} N_1 N_j \tag{8}$$

$$\frac{dN_i}{dt} = \frac{1}{2}\sum_{j=1}^{i-1} \beta_{1,i-j} N_j N_{i-j} - \sum_{j=1}^{\infty} \beta_{i,j} N_i N_j \tag{9}$$

$$i = 2, \ldots, \infty$$

with the collision operator $\beta_{i,j}$ describing the probability that two particles from the particle classes i and j coagulate. To simulate the complete evolution of the particle population, all these equations would have to be solved simultaneously which of course is not feasible for non-trivial PDFs.

An alternative approach is the *Method of Moments* (MoM). The MoM is based on the moments of the particle population:

$$M_r = \sum_i m_i^r N_i \tag{10}$$

Moment M_0 equals the total number of particles, moment M_1 is equivalent to the total mass of the particles. The higher order moments correspond to statistical measures of the PDF and consequently its shape.

Knowing all the moments M_r is equivalent to knowing the exact size distribution of the particles. However, to reconstruct an approximation of the PDF only a small number of moments is needed (according to Frenklach [10] a number of 6 moments is sufficient for most applications).

The changes of the moments can be calculated from the changes of the particle classes (8):

$$S_0 = \tfrac{dM_0}{dt} = -\frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\beta_{i,j}N_iN_j \tag{11}$$

$$S_1 = \tfrac{dM_1}{dt} = 0 \tag{12}$$

$$S_2 = \tfrac{dM_2}{dt} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ij\beta_{i,j}N_iN_j \tag{13}$$

$$S_3 = \tfrac{dM_3}{dt} = 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ij^2\beta_{i,j}N_iN_j \ldots \tag{14}$$

For specific collision terms $\beta_{i,j}$ (for an example see [11]) the sums in these formulas can be reduced to expressions that involve only a finite number of terms that are expressed by moments M_r (where r isn't necessarily an integer number). These moments can be approximated from the known moments by interpolation.

The transport of the PDF is implemented by the usual transport equation (6) using the time evolutions of the moments (11) - (14) as source terms:

$$\frac{\partial \rho M_r}{\partial t} = \frac{\partial}{\partial x_i}\left(\rho u_i M_r - D\frac{\partial M_r}{\partial x_i}\right) = S_r \tag{15}$$

Equation (15) is the transport equation for the moment M_r. The source term S_r is the source term as calculated in equations (11)-(14). In the general case for the calculation of these source terms the values of the other moments are needed (coupling the calculation of the moments).

Solving equation (15) for the moments is equivalent to solving (6) for each particle class: each particle with a mass m_i carries a contribution to M_r of m_i^r. All N_i particles in class i have a moment $N_i m_i^r$ for which we could solve (6). If these equations are added for all particle classes and we use (10), we get to (15).

4 Implementation

The Method of Moments has been integrated into the existing LBM solver. Due to the explicit nature of the LBM the calculation of the flow solution and the particle transport can be easily decoupled: at each time-step the new solution of the Navier-Stokes equations is calculated by the LBM. Then the transport equations (15) for the moments are solved using that flow solution.

Then the source terms S_r are calculated from equations (11) - (14). For general collision terms $\beta_{i,j}$ fractional moments M_f have to be used, which are interpolated from the (known) integer moments. Because there are several orders of magnitude between the moments, which leads to numerical instabilities in the interpolation, the interpolation is performed with the logarithms of the moments making the computation of the sources much stabler (this is possible because the moments (10) are positive).

In addition to the source terms from particle-particle interactions, in the computational cells next to walls, a source term for the particle deposition on the walls is calculated (these are obviously the only source terms S_1 for the mass density M_1). For each computational cell a separate variable ρ_{dep} for the density of the deposited soot is introduced. The equation

$$\frac{\partial \rho_{dep}}{\partial t} = -S_1 \tag{16}$$

for the deposited soot is solved. As soon as ρ_{dep} reaches a critical density ρ_{full} (which is known from weight measurements of deposited soot) the computational cell is marked as *full*. It is then treated like a wall boundary cell by the LBM and soot can deposit in the neighboring cells.

An alternate approach is to treat the deposited soot as a permeable medium and allow the flow of air through through the soot modified by the Darcy-Forchheimer term. This method also allows for the treatment of partially filled cells and therefor a more accurate flow solution. The disadvantage is that the additional source term for the Navier-Stokes equations degrades the stability of the LBM.

The necessity to store seven additional scalars (6 moments and the deposition density) per volume cell increases the memory consumption of the simulation moderately. More significant is the increase in CPU-time, as the calculation of the particle transport and the source terms have to be performed at every time-step and are computationally intensive.

5 Results

Two simulation results are shown to illustrate the capabilities of the Method of Moments: the first simulation is a simple case that illustrates the effect of the particle coagulation. The second case shows particle flow in a more complex geometry.

5.1 Deposition in a Channel

The first simulation shows laminar flow through a channel at a constant pressure gradient. Soot particles are inserted at the left end of the channel, and deposit eventually on the channel walls and on the rectangular obstacle located at the right end of the channel. The results are shown in figure 1. The obstacles are shown in light grey, the deposited soot is shown in black. The values of the moments are depicted by isolines.

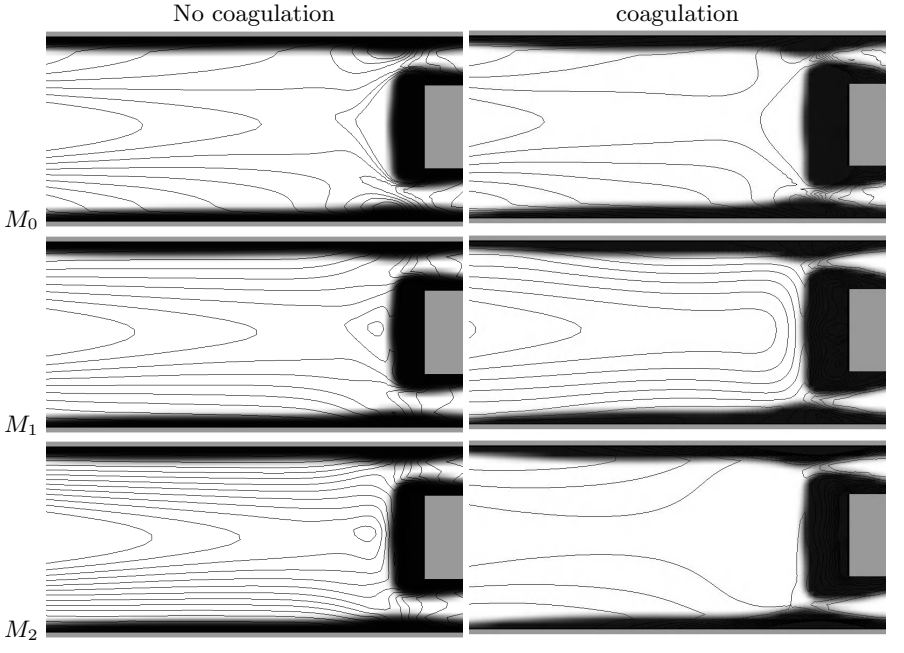


Fig. 1. Soot deposition in a channel with obstacle. The same simulation is shown with and without coagulation

In the first column the simulation without coagulation is shown. The second column shows the same simulation, but the PDF changes due to coagulation. For both simulations the result at the same time is given.

The images in the first line show the particle density, M_0. Coagulation causes a reduction in particle density in flow direction. The second line shows the material density M_1, which is nearly the same in both cases (differences near the obstacles are due to the different flow fields caused by the different soot depositions). The last line shows the moment M_2 which corresponds to the width of the PDF which is much smaller in the coagulation case.

The amount of deposited soot is larger in the case of coagulation, because the larger particles tend to deposit.more easily.

5.2 Deposition on a Pore Structure

The pictures in figure 2 show fluid-particle flow through a more realistic structure, reconstructed from existing porous material. The average pore radius is 0.1 millimeters.

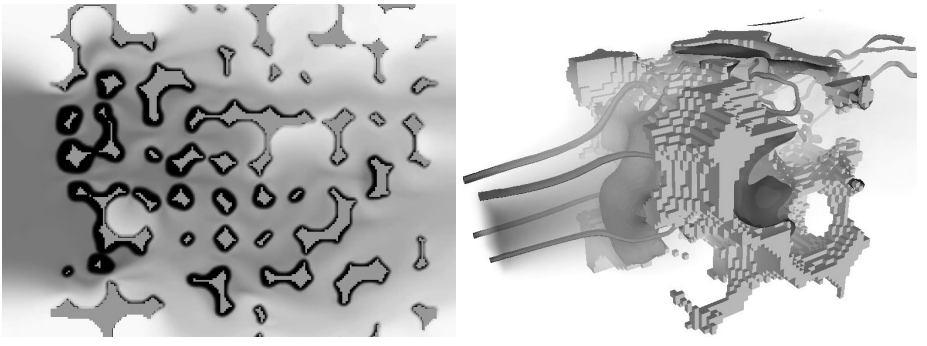


Fig. 2. Deposition in porous material. The left picture shows a two-dimensional reconstruction of a pore with deposited soot and soot density. The right picture shows a three-dimensional simulation of soot deposition on a pore structure. Stream line of the velocity, soot density in the air and soot deposited on the pore are visible

The left picture in figure 2 was obtained from a two dimensional simulation. The pore itself is shown in light gray, deposited soot in black. Flow through the pore structure goes from the left to the right and is driven by a constant pressure difference, that was applied at the boundaries. A soot distribution is inserted at the inlet. The density M_1 of the soot in the air is shown in shades of gray.

It can be clearly seen, that the density of the soot in the air drops due to the deposition of the soot on the pore structure.

The deposited soot changes the flow distribution through the pore and therefore the permeability of the filter and the filtration efficiency.

The picture on the right of figure 2 is taken from a 3D simulation of a similar pore. The pore walls are represented by the the cubes. The air flows through the pore structure from the left to the right (four "stream tubes" depict the fluid velocity). Soot is inserted at a small fraction of the inflow (soot density is shown by the gray fog). Soot deposition on the pore is shown by the dark gray isosurfaces.

6 Conclusion

The Method of Moments for soot particle tracking and deposition is successfully combined with the Lattice Boltzmann Method. Particle coagulation and deposition can be easily implemented by this approach.

Application of the suggested approach has been shown for two cases: Soot deposition in simple geometries and realistic porous structures.

The combination of these two methods makes it possible to perform simulations in a wide range of applications, where complex geometries and changing particle density functions are encountered.

References

1. Chen, S., Doolen, G.D.: Lattice boltzmann method for fluid flos. Ann. Rev. Fluid Mech. (1998)
2. Yu, D., Mei, R., Luo, L.S., Shyy, W.: Viscous flow computations with the method of lattice boltzmann equation. Progress in Aerospace Sciences (2003)
3. Leer, B.V.: Toward the ultimate convervative difference scheme v. Journal of Computational Physics (1979)
4. Luo, L.S.: The lattice gas and lattice boltzmann methods: Past, present and future. In: Proceedings "International Conference on Applied Computational Fluid Dynamics", Beijing, China. (2000)
5. Redl, C., Wassermayr, C., Leixnering, J.: On the numerical simulation in foam filter design for diesel exhaus gas systems. In: Proceedings "International Congress for Engine Combustion Process", Munich, Germany. (2003)
6. Filippova, O., Hänel, D.: Lattice-boltzmann simlation of gas-particle flow in filters. Computers & Fluids (1997)
7. Chopard, B., Masselot, A., Dupuis, A.: A lattice gas model for erosion and particle transport in a fluid. Computer Physics Communications (2000)
8. Dupuis, A., Chopard, B.: Lattice gas modeling of scour formation under submarine pipelines. Journal of Computational Physics (2002)
9. Smoluchowski, M.: Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. Zeitschrift für physikalische Chemie (1917)
10. Frenklach, M.: Method of moments with interpolative closure. Chemical Engineering Science (2002)
11. Kazakov, A., Frenklach, M.: Dynamic modeling of soot particle coagulation and aggregation: Implementation with the method of moments and application to high-pressure laminar premixed flames. Combustion and Flame (1998)
12. McGraw, R., Wright, D.L.: Chemically resolvedaerosol dynamics for internal mixtures by the quadrature method of moments. Journal of Aerosol Science (2003)
13. Wu, J., Hu, B.X., Zhang, D., Shirley, C.: A three-dimensional numerical method of moments for groundwater flow and solute transport in a nonstationary conductivity field. Advances in Water Resources (2003)

Numerical Bifurcation Analysis of Lattice Boltzmann Models: A Reaction-Diffusion Example

Pieter Van Leemput and Kurt Lust

Department of Computer Science, K.U.Leuven,
Celestijnenlaan 200A, B-3001 Heverlee, Belgium.
{pieter.vanleemput,kurt.lust}@cs.kuleuven.ac.be

Abstract. We study two strategies to perform a time stepper based numerical bifurcation analysis of systems modeled by lattice Boltzmann methods, one using the lattice Boltzmann model as the time stepper and the other the *coarse-grained* time stepper proposed in Kevrekidis *et al.,* CMS 1(4). We show that techniques developed for time stepper based numerical bifurcation analysis of partial differential equations (PDEs) can be used for lattice Boltzmann models as well. The results for both approaches are also compared with an equivalent PDE description.

Keywords: numerical bifurcation analysis, lattice Boltzmann method, Newton-Picard method, reaction-diffusion systems.

1 Introduction

Time simulation is a popular method to study the influence of varying the parameters in a parameter-dependent dynamical system, but comparing simulation results for many different parameter values is tedious. Also, only stable solutions can be explored in this way. Bifurcation theory offers a more systematic way to study the asymptotic behavior of a parameter-dependent system.

In a numerical bifurcation analysis, one computes branches of solutions of some type by varying one or more parameters. Along the branch, the stability information is monitored and the critical (or *bifurcation*) points where stability changes occur are detected and often computed. Several techniques and codes have been developed for partial differential equations (PDEs), some of which work on top of an existing time simulation code. In this paper, we show that these techniques can be used for bifurcation analysis of lattice Boltzmann (LB) models also. Examples of such *time stepper* or *map based* techniques are the Recursive Projection Method [1] and the Newton-Picard method [2]. These methods exploit a spectral property of the time integration operator that is typical of systems with a dissipative operator: the time integration operator has only few eigenvalues close to or outside the unit circle. In other words, the system's long-term behavior is governed by only a small number of slow modes. This is an

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 572–579, 2004.
© Springer-Verlag Berlin Heidelberg 2004

inherent property of the physics of the system. It is also the starting assumption for techniques such as inertial manifolds and reduced-order modeling.

We will discuss two strategies for the bifurcation analysis of LB models. One approach uses the LB time stepper as the time stepper in the bifurcation code. The state can be expressed either as distribution functions or as a full set of velocity moments. We will also make a slight extension of the time stepper to compute periodic solutions. The second approach uses the *coarse-grained* time stepper proposed by Kevrekidis *et al.* [3,4] which performs a time integration for the macroscopic variables (a subset of the velocity moments in a LB model) using only microscopic or mesoscopic simulations. Their coarse-grained time stepper is developed for cases where a macroscopic model is *conceptually* possible, but not yet known in a closed analytical form. Whereas our numerical bifurcation analysis techniques rely on a separation of time scales at the macroscopic level, the coarse-grained time stepper relies on a time scale separation between the macroscopically relevant variables, i.e., the lower-order moments of the distribution functions in a LB model, and the other variables that are only relevant at the microscopic or mesoscopic scales.

We will use a FitzHugh-Nagumo (FHN) reaction-diffusion system on a one-dimensional domain as our test case. This system is usually specified as a PDE system, but it is possible to develop an equivalent LB-BGK model. We compare the results for both approaches based on the LB model with each other and with the results for the equivalent PDE model. Section 2 focusses on the numerical bifurcation analysis. In Sect. 3, we present the three models. Numerical results for the FHN system are shown in Sect. 4. Section 5 repeats the main conclusions.

2 Numerical Bifurcation Analysis

For numerical bifurcation analysis of autonomous parameter-dependent PDEs, the PDEs are first space-discretized to a large set of ordinary differential equations (ODEs)

$$y_t = f(y, \gamma) \ , \quad f : \mathbb{R}^N \times \mathbb{R} \mapsto \mathbb{R}^N \tag{1}$$

with y the state vector and γ the bifurcation parameter. Let $\varphi_T(y(0), \gamma)$ denote the result $y(T)$ of a time integration of (1) at time T with initial condition $y(0)$ and parameter γ. A steady state (y^*, γ^*) of (1) is also a fixed point of the map

$$y \mapsto \varphi_T(y, \gamma) \tag{2}$$

for any value of T. A periodic solution of (1) is a fixed point of (2) only when T is a multiple of the (unknown) period.

A steady state of (1) is asymptotically stable if all eigenvalues λ_j of the Jacobian matrix $(\partial f/\partial y)(y^*, \gamma^*)$ have a negative real part. Hence the rightmost eigenvalues determine the asymptotic stability. The equivalent fixed point of (2) is stable if all eigenvalues μ_j of the matrix $M := (\partial \varphi_T/\partial y)(y^*, \gamma^*)$ are smaller than one in modulus. Both sets of eigenvalues are related by

$$\mu_j = \exp(\lambda_j \, T) \tag{3}$$

and thus $|\mu_j| < 1 \Leftrightarrow \mathrm{Re}(\lambda_j) < 0$. Hence they provide exactly the same stability information.

In time stepper based bifurcation analysis, a continuous-time dynamical system is analyzed through the (almost) equivalent map (2). In fact, the time integrator can be any time integrator, including a LB simulation or the coarse-grained time integrator. A branch of fixed points is computed with a continuation method. Given a few points on the branch, a prediction is made for the next point and that point is then computed by solving the nonlinear system

$$\begin{cases} \varphi_T(y,\gamma) - y = 0 \ , \\ n(y,\gamma,\eta) = 0 \end{cases} \tag{4}$$

for y and γ. The scalar *parameterizing equation* $n(y,\gamma,\eta) = 0$ determines the position of the point along the branch through a reparameterization with parameter η. In our experiments, we used pseudo-arclength parameterization [5].

When computing a branch of periodic solutions, the period T becomes a variable as well. The system (4) is augmented with a *phase condition* $s(y,T,\gamma) = 0$, which ensures isolated solutions of the system. M is then called the *monodromy matrix* and its stability-determining eigenvalues the *Floquet multipliers*.

A popular way of solving (4) is Newton's method with Gaussian elimination. However, M and thus also the Jacobian matrix of (4) are in general dense matrices (even though $(\partial f/\partial y)(y,\gamma)$ is usually a sparse matrix). It is not feasible to compute and store M. However, since computing a matrix-vector product with M is equivalent to computing a directional derivative of the time integrator, these matrix-vector products can be computed using numerical differentiation.

The fact that M often has only a few eigenvalues close to or outside the unit circle is exploited by several numerical methods. One of the earliest is the Recursive Projection Method [1]. Although this method has only been derived for the computation of steady state solutions, it can be extended to compute periodic solutions also. However, robustness problems are fairly common with the original implementation of [1]. We will use the more robust Newton-Picard method [2] instead. This method was originally developed for the computation of periodic solutions but is equally well suited for the computation of steady states. First, an approximation to the low-dimensional generalized eigenspace $\mathcal{U}$ of all dominant eigenvalues of M is computed by orthogonal subspace iterations, requiring only matrix-vector products with M. Then, the linearized system is projected onto $\mathcal{U}$ and its high-dimensional orthogonal complement $\mathcal{U}^\perp$. The large subsystem in $\mathcal{U}^\perp$ is solved approximately using a fixed point (or *Picard*) iteration, which corresponds essentially to time integration in the high-dimensional subspace $\mathcal{U}^\perp$. The small subsystem in $\mathcal{U}$ is solved using Gaussian elimination. The total number of time integrations needed to converge to a given accuracy is mostly determined by the dominant eigenvalues and thus by the physics of the problem and not by the particular mathematical model or discretization. Hence we expect that a similar number of time integrations will be needed for the PDE model, the LB model and the coarse-grained description, though some differen-

ces are possible, e.g., because of a different number of variables or convergence problems of Newton's method when the nonlinearity becomes too strong.

3 System Descriptions

3.1 Macroscopic Description: A PDE System

The one-dimensional macroscopic FHN PDE model is given by

$$\begin{cases} \rho_t^{ac} = \rho_{xx}^{ac} + \rho^{ac} - (\rho^{ac})^3 - \rho^{in} \ , \\ \rho_t^{in} = \delta\rho_{xx}^{in} + \varepsilon(\rho^{ac} - a_1\rho^{in} - a_0) \ , \end{cases} \tag{5}$$

with homogeneous Neumann boundary conditions on the domain $[0, 20]$. The variables $\rho^{ac}(x,t)$ and $\rho^{in}(x,t)$ are the activator and inhibitor concentration respectively. We set $\delta = 4$, $a_0 = -0.03$ and $a_1 = 2$ and vary $\varepsilon \in [0, 1]$. We used a method of lines discretization with central differences in space at the midpoints of 200 lattice intervals and the trapezoidal rule for time integration.

3.2 Mesoscopic Description: Lattice Boltzmann Model

The variables in a LB model are the distribution functions $f_i^s(x,t)$ associated with a species $s \in \{ac, in\}$ and a discrete "particle" velocity v_i. They are defined on a space-time lattice with spacing Δx in space and Δt in time. We use only three discrete values for the velocity: $v_i = i\,\Delta x/\Delta t$ with $i \in I := \{-1, 0, 1\}$ (D1Q3 scheme). The concentration – the only relevant macroscopic variable – is the zeroth order velocity moment of the distribution functions, i.e.,

$$\rho^s(x,t) = \sum_{i=-1}^{1} f_i^s(x,t) \ . \tag{6}$$

The discrete time evolution of the distribution functions is described by the standard LB time stepper equation

$$f_i^s(x + v_i\Delta t, t + \Delta t) - f_i^s(x,t) = -\omega^s[f_i^s(x,t) - f_i^{s,eq}(x,t)] + R_i^s \ \ , i \in I \ . \tag{7}$$

We used halfway bounce-back boundary conditions [6] and the approach in [7] for the BGK collision term $-\omega^s[f_i^s(x,t) - f_i^{s,eq}(x,t)]$ and reaction term R_i^s. Note that the equilibrium distribution $f_i^{s,eq}(x,t) = \frac{1}{3}\rho^s(x,t), i \in I$ which will be important in the remaining of the paper.

Note that the single *local* time scale in the LB model should not be confused with the *global time scales* which are important in our numerical algorithms. The large range of global time scales results from the coupling of all lattice points. The slowest components in the global behavior of the system can be much slower than the local relaxation time.

When computing periodic solutions, the integration time T is continuous and not always an exact multiple of the LB time step Δt. We then determine a positive integer k such that $(k-1)\Delta t < T \leq k\Delta t$ and compute the state at time T by a linear interpolation of the states at the discrete time steps $k-1$ and k.

3.3 Coarse-Grained Description

The coarse-grained LB (CGLB) time integration procedure proposed in [3,4] is a time stepper for the macroscopic variables. A macroscopic time step ΔT consists of three basic steps. The first step is the *reconstruction* or *lifting*. Here, meaningful mesoscopic initial values $f_i^s(x,0)$ are derived, which are consistent with the governing macroscopic variable $\rho^s(x,0)$, i.e.,

$$f_i^s(x,0) = w_i\,\rho^s(x,0)\ \ , i \in I \quad \text{with} \quad \sum_{i=-1}^{1} w_i = 1\ . \tag{8}$$

We choose the weights w_i all equal to 1/3, i.e., equal to the weights for the diffusive equilibrium. Next, the mesoscopic variables $f_i^s(x,t)$ are evolved through the LB time stepper for a time interval ΔT. If ΔT is not an integer multiple of the LB time step Δt, we interpolate as in Sect. 3.2. In the final *restriction* or *coarse-graining* step, the macroscopic variable $\rho^s(x,\Delta T)$ is computed using (6). This procedure is repeated until time T. In our experiments, we use $\Delta T = 5$. Since ΔT is fairly large, we cannot interpolate between two time steps to obtain the result at an arbitrary time T. Instead we change ΔT slightly from the target value $\Delta T = 5$, such that it fits an integer number of times in T. As long as ΔT is not changed too much, this has no significant influence on the result. The choice of the reconstruction weights w_i and macroscopic time step ΔT is a topic of our current research and will be discussed in a future publication.

4 Numerical Results and Discussion

4.1 Bifurcation Diagrams

Figure 1 shows the bifurcation diagram for steady state (upper diagram) and periodic solutions (lower diagram), computed using the Newton-Picard method. We used $T = 5$ for the steady state solutions, a good value for the Newton-Picard method in this case. For the LB model, we used $\Delta t = 0.001$, which produced the best results. The steady state branch contains two bifurcation points: a fold point at $\varepsilon \approx 0.945$ and a supercritical Hopf point at $\varepsilon \approx 0.0183$ where the branch of periodic solutions in the lower diagram meets. The periodic solution branch has a fold point at $\varepsilon \approx 0.00087$. Though the unstable part of the branch has almost the same (ε, T)-projection as the stable one, the corresponding orbits are different. Computing these unstable solutions proved difficult, which is a well-known problem of single shooting based numerical methods. The bifurcation diagrams for the two LB approaches correspond very well with differences on the order of the discretization error. Moreover, we do obtain the same bifurcation information as from the equivalent PDE.

4.2 Eigenvalue Spectra

The Newton-Picard code computes the stability-determining eigenvalues through orthogonal subspace iterations. Table 1 lists the dominant eigenvalues λ_j for an

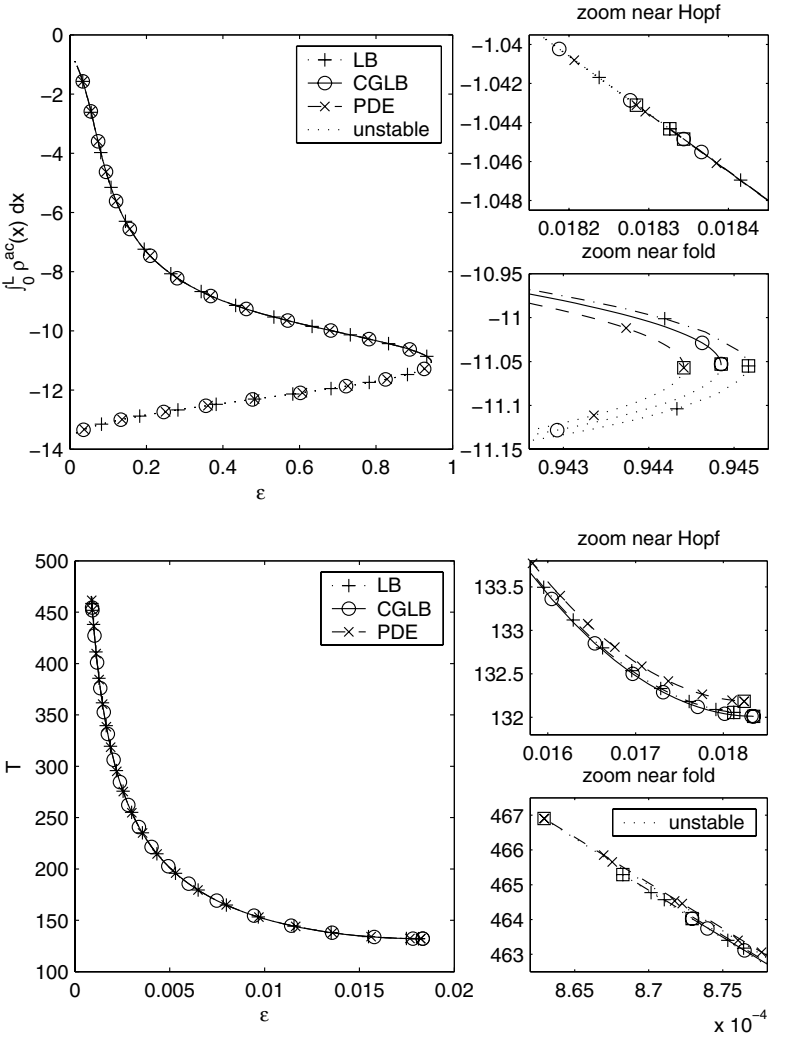


Fig. 1. Bifurcation diagram for the steady state (upper diagram) and periodic solutions. Unstable solutions are indicated by dotted lines and bifurcation points by boxed markers. The markers represent only a subset of the computed points

Table 1. Dominant eigenvalues for the unstable steady state on the upper part of the branch and stable periodic solution at $\varepsilon = 0.01$ (using $\Delta T \approx 5$ in the CGLB integrator)

	steady state			periodic solution	
	$\lambda_{1,2}$	λ_3	λ_4	trivial μ_1	μ_2
LB	$0.002010 \pm 0.039461i$	-0.124867	-0.411364	1.000000	0.514888
CGLB	$0.002012 \pm 0.039463i$	-0.124863	-0.411348	1.000000	0.514452
PDE	$0.001999 \pm 0.039446i$	-0.124861	-0.411288	1.000000	0.516712

unstable steady state at $\varepsilon = 0.01$ and Floquet multipliers μ_j for the stable periodic solution at the same parameter value. Again, the eigenvalues for both LB based approaches correspond very well with one another and give the same information as the equivalent PDE model. Note that periodic solutions of an autonomous system always have a trivial multiplier at one. The great accuracy of the computed value indicates that the time integration and eigenvalue computation is very accurate.

4.3 Time Stepper Calls

Table 2 lists the number of time integrations needed to continue a part of the branches of stable steady state and stable periodic solutions, not including the integrations for the accurate computation of the stability information at the end. For the steady states, we again used $T = 5$. Note that the LB model uses a different set of variables from the other two approaches and hence another measure for the distance. We changed the maximal step size in our variable step size continuation code for each approach such that the number of points computed along the branch is approximately the same. As expected, the average number of time integrations needed for each point is almost the same. The computing time however is much lower for the PDE based model since the implicit time stepper uses much larger steps than the LB model. The computing time for both LB based approaches is dominated by the LB simulations and virtually the same for both approaches.

Table 2. Total number of time integrations, the number of points computed and the average number of time integrations for the continuation of a part of the branches of stable steady state and periodic solutions

number of time integrations	steady state, $\varepsilon \in [0.05, 0.55]$			periodic, $\varepsilon \in [0.006, 0.016]$		
	total	nb. pts	avg.	total	nb. pts	avg.
LB	1352	20	67.600	1611	25	64.440
CGLB ($\Delta T \approx 5$)	1419	21	67.571	1623	25	64.920
PDE	1420	21	67.619	1621	25	64.840

5 Conclusions

In this paper, we have shown that algorithms developed for time stepper based numerical bifurcation analysis of PDEs can be used for lattice Boltzmann models as well in two different ways. It is possible to use either the LB model itself or the coarse-grained time integrator as the time stepper in a time stepper based bifurcation code. For our test case, the accuracy and computing time for both LB based approaches are comparable which was shown to be not unexpected. We

have also compared the results to those for an equivalent PDE model, demonstrating that the same bifurcation information can be obtained from a lattice Boltzmann model as from a PDE model. However, time integration of the PDE was much faster, resulting in much lower computing times for the bifurcation diagram of the PDE model.

The two approaches for bifurcation analysis of LB models have distinct advantages and disadvantages. Several parameters in the coarse-grained integrator need to be chosen carefully. This is currently a subject of our research and will be discussed in a future publication. On the other hand, the dimension of the state vector is much lower (only one third of the number of components in our case), resulting in a proportionally lower memory consumption of the Newton-Picard method for the coarse-grained time integrator. In some cases, this can be an issue and thus possibly an advantage of the coarse-grained approach.

Acknowledgements. Kurt Lust is a postdoctoral fellow of the Fund for Scientific Research - Flanders which also provided further funding through project G.0130.03. This paper presents research results of the Interuniversity Attraction Poles Programme – Belgian Science Policy. The scientific responsibility rests with its authors. We also acknowledge many helpful discussions with Prof. I.G. Kevrekidis.

References

1. Shroff, G.M., Keller, H.B.: Stabilization of unstable procedures: The Recursive Projection Method. SIAM Journal on Numerical Analysis **30** (1993) 1099–1120
2. Lust, K., Roose, D., Spence, A., Champneys, A.: An adaptive Newton-Picard algorithm with subspace iteration for computing periodic solutions. SIAM Journal on Scientific Computing **19** (1998) 1188–1209
3. Kevrekidis, I.G., Gear, C.W., Hyman, J.M., Kevrekidis, P.G., Runborg, O., Theodoropoulos, C.: Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. Communications in Mathematical Sciences **1** (2003) 715–762
4. Theodoropoulos, C., Qian, Y.H., Kevrekidis, I.G.: "Coarse" stability and bifurcation analysis using time-steppers: a reaction-diffusion example. Proceedings of the National Academy of Sciences **97** (2000) 9840–9843
5. Keller, H.B.: Numerical solution of bifurcation and nonlinear eigenvalue problems. In Rabinowitz, P.H., ed.: Applications of Bifurcation Theory, New York, Academic Press (1977)
6. Ginzbourg, I., Adler, P.M.: Boundary flow condition analysis for the three-dimensional lattice Boltzmann model. Journal of Physics II France **4** (1994) 191–214
7. Qian, Y.H., Orszag, S.A.: Scalings in diffusion-driven reaction $A + B \rightarrow C$: Numerical simulations by Lattice BGK Models. Journal of Statistical Physics **81** (1995) 237–253

Particle Models of Discharge Plasmas in Molecular Gases

S. Longo, M. Capitelli, and P. Diomede

Dipartimento di Chimica dell'Universita' di Bari, Via Orabona 4, 70126 Bari, Italy
IMIP/CNR, Via Orabona 4, 70126 Bari, Italy
cscpsl17@area.ba.cnr.it

Abstract. We describe a mixed particle/continuum model for discharge plasmas in molecular gases developed by our group, which couples a particle description of the plasma phase with the diffusion/reaction kinetics of atoms and molecules in the gas phase. The model includes an improved treatment of ion kinetics, which fits some serious problems of multi time scale physical chemistry. The hydrogen plasma is considered as a test case. Results and computational costs are shortly discussed also in comparison with a different code.

1 Introduction

The modelling of the discharge plasma in molecular gases [1,2] represents an important benchmark for the computational techniques of the plasma discharge community. Most of the models developed until now are 'fluid' models based on a macroscopic description of the charged particle kinetics in the discharge [3]. These models assume for the ionization coefficient, and in general for the rate coefficients of the electron/molecule reactions, values which are at most calculated by solving the so-called quasi-isotropic Boltzmann equation assuming a uniform plasma. These are later interpolated in terms of the electron energy and applied locally depending on the calculated value of the electron temperature. This approach cannot handle cases where the electrons have a Knudsen number higher than about 0.01 and behave like a rarefied gas. For a typical plasma dimension of 6cm, at room temperature, assuming an electron/neutral elastic cross section of 10^{-19} m^2, the borderline pressure at which Kn=0.01 is in the order of 0.5 torr. When the pressure is significantly lower than this, only methods based on numerical solutions of kinetic equations for the translational distribution function f(**r**,**v**,t) of the charged particles, can be applied in order to get accurate values for the rate coefficients of electron/molecule reactions.

In general terms, the coupling of the neutral gas and plasma phase kinetics is always to be taken into account, especially for molecular gases: on one side in fact, the rate coefficients for electron/molecule reactions are functionals of the local electron energy distribution function; on the other side, the Boltzmann collision integral is also a function of the gas composition, including the vibrational excitation.

The necessity to take into account chemical kinetics as well as charged particle kinetics in the plasma phase at the same time *creates an interesting problem of multiple time scales*. In fact:

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 580–587, 2004.
© Springer-Verlag Berlin Heidelberg 2004

1. the shortest electron particle kinetics time scale is the reverse of the electron plasma frequency, which in typical applicative cases is in the nanosecond range.
2. the time scale for charged particle relaxation to the steady state is the ambipolar diffusion time, which in typical applicative cases is of the order of 0.1-1 milliseconds
3. the time scale for the stabilization of a steady state chemical composition of the negative ions can reach (especially for electronegative plasmas) values of 0.01 seconds or higher.

So we have a time scale span of several orders of magnitude to be tackled by an appropriate technique, which must at the same time merge two different kinetic descriptions, one (for charged particles) based on transport kinetic equations, and the other (for neutral particles) based on chemical kinetics rate-diffusion equations.

Several particle models of these kind of plasmas already appeared in the literature [11], but as a rule they do not include the vibrational/chemical kinetics of gas phase molecules. A class of particle models including the kinetics of vibrationally excited molecules was considered in the past few years [5,6]. These models include a state-to-state chemical kinetics for molecules, this last term meaning that molecules in different vibrational levels are treated as separate species. Another important problem not yet addressed for this system is an accurate, kinetic-level description of the several kinds of ions existing in the discharge, which should take into account also the thermal distribution of neutral particles in the bulk region of the plasma, where it cannot be neglected because of the thermal energy of ions.

In this paper we present a fully self-consistent 1D model of the discharge plasma based on substantial improvements (to be described in details in the following) of previous models [5,6].

The algorithm presented in this paper includes several techniques to treat the different time scales described in the introduction, in particular the self-consistent treatment of neutral chemistry and plasma kinetics and a special scaling for the negative ion time scale.

2 Mathematical Statement of the Problem

In reactive plasmas usually the relaxation times of the chemical kinetics are by far higher than the corresponding ones for the electron kinetics, therefore adiabatic elimination of the time dependence of the neutral component densities can be applied.

In the 1D×3D (meaning 1 space dimension and 3 velocity components) case, which can be reduced to 1D×2D unless transversal fields are applied, we can formalize our problem as follows:

$$\left(\frac{\partial}{\partial t} + v_x \frac{\partial}{\partial x} - \frac{q_s}{m_s} \frac{\partial \varphi(x,t)}{\partial x} \frac{\partial}{\partial v_x} \right) f_s(x, \mathbf{v}, t) = C_s\left(\{F_c\}\right) \tag{1a}$$

$$\frac{\partial^2 \varphi(x,t)}{\partial x^2} = -\frac{1}{\varepsilon_0}\sum_s q_s \int d^3 v f_s(x,\mathbf{v},t) \tag{1b}$$

$$-D_c \frac{\partial^2 n_c(x)}{\partial x^2} = \sum_r \left(\nu'_{rc} - \nu_{rc}\right) k_r \left(\left\langle f_e \right\rangle_t \right) \prod_{c'} n_{c'}^{\nu_{rc}} \tag{1c}$$

where f_s and F_c are the kinetic distribution functions for the s-th charged species and the c-th neutral species respectively, f_e is the defined below (Eq.5), q_s and m_s are the s-th species electric charge and mass, φ is the electric potential, n_c is the number density of the c-th neutral component, D_c is its diffusion coefficient, k and ν are, respectively, the rate coefficient and the molecularity of the c-th species in the r-th elementary process in a set of N_r

$$\left\{ \sum_{c=1}^{N_c} \nu_{rc} X_c \rightarrow \sum_{c=1}^{N_c} \nu'_{rc} X_c \right\}_{1 \le r \le N_r} \tag{2}$$

where X_c is a generic neutral species out of a set of N_c.

We also include surface processes by an appropriate implementation of the boundary conditions [6].

In eq.(1a), C_s is the Boltzmann collision integral for charged/neutral particle collisions:

$$\begin{aligned} C_s\left(\{F_c\}\right) &= -f_s(v)\int d^3 v' p_{v\to v'} + \int d^3 v' p_{v'\to v} f_s(v') \\ p_{v'\to v} &= \int d^3 w d^3 w' \,|\, v' - w' \,|\, \sum_k \sigma_k(v', w', v, w) F_{c(k)}(r, w') \end{aligned} \tag{3}$$

where k is an index addressing a specific collision process, σ_k and c(k) are the differential cross section and the neutral collision partner of the k-th process.

In the case of reactions including the electron as a reactant, the rate coefficient of the process must be written as (under the approximation of static neutral target):

$$k(x) = \sqrt{2/m_e} \int_0^\infty \varepsilon f_e(\varepsilon, x)\sigma(\varepsilon) d\varepsilon \tag{4}$$

where σ is the related total cross section and $f_e(\varepsilon,x)$ is the so-called *electron energy distribution function* (*eedf*) defined as

$$\begin{aligned} f_e(\varepsilon, x) &= \varepsilon^{-1/2} \left\langle f_e(x,\mathbf{v},t)/n_e \right\rangle_{t,\Omega(\varepsilon)} \\ \Omega(\varepsilon) &= \left\{ \mathbf{v} \in R^3, m_e v^2 = 2\varepsilon \right\} \end{aligned} \tag{5}$$

this function is normalized to 1 for any position x, and is usually measured in $eV^{-3/2}$. The rate coefficient (eq.4) is a functional of the eedf and cannot be simply written as a

function of the electron mean energy; this can only be done in the quasi-equilibrium case, where the eedf is maxwellian.

To fix ideas, in the following we will consider a code implementation for molecular hydrogen: we have selected this test case in view of the special status of hydrogen in physics and chemistry, and the several important applications of gas discharges in pure hydrogen and mixture of hydrogen with other gases. The code version for hydrogen includes several reactions and plasma phase processes. The species considered are five particle species in the plasma phase, i.e. electrons, H_3^+, H_2^+, H^+, H^-; and sixteen neutral components, i.e. H_2(v=0 to 14) and hydrogen atoms.

A list of the collision and chemical processes is in ref.[6,(2001)]. The only changes are: the vibrational deactivation probability on the walls is here $\gamma_V = 0.1$, a mean between those used in the two refs.[6], the H^-/H_3^+ and H^-/H reaction rates are after Gorse et al. [8], and the H_3^+/H_2 cross section is from ref.[9].

3 Numerical Method

To solve the problem described in the previous sections we use a Particle in Cell/ Monte Carlo method (PIC/MC), which is a lagrangian/stochastic solution technique, for the transport equation and a grid-discretized relaxation technique for the reaction-diffusion part. A self-consistent approach involving joint solution of the two problems is necessary, for the reasons explained in the introduction. The procedure is repeated up to full relaxation.

In the PIC/MC approach the Newton equation for a large ensemble (10^4 - 10^5 particles) of mathematical point particles (assumed to represent a statistically representative sample of electrons and positive ions) is solved taking into account the local electric field as it results from local interpolation within a cell of a mathematical mesh (from which follows the name 'Particle in Cell'). The electric charge is determined by sampling the particles on the mesh. The electric potential and field are determined by solving the Poisson equation on the same mesh (which in the 1D case is trivial [4]).

The PIC method delivers a solution of the Vlasov-Poisson plasma problem (1a,b) in the following form:

$$f_s(\mathbf{r},\mathbf{v},t) = \frac{W_s}{N_s}\sum_{p=1}^{N} S(\mathbf{r}-\mathbf{r}_p)\delta(\mathbf{v}-\mathbf{v}_p) \tag{6}$$

W_s is the ratio between real and simulated particles, and S(x) is the particle shape factor which describes the way particles are assigned to the mesh. A typical choice for S is the so-called 'Cloud-in-Cell' one, which in 1D is:

$$S(x) = \begin{cases} 1 - 2x/\Delta x, & 0 \le x < \Delta x/2 \\ 1 + 2x/\Delta x, & -\Delta x/2 \le x < 0 \end{cases} \tag{7}$$

The Newton equations in PIC are usually solved by using the Leapfrog method.

We use a modified time step technique [4], where the particle dynamics evolves in time with an appropriate time step Δt, while the time-to-next collision t_c is an independent variable for any particle, which decreases during the free-flight. The exact time when $t_c(i)=0$ marks a new collision event.

As regards the inclusion of the collision term C_s, as it has been demonstrated [7], a stochastic calculation of the r.h.s. of eq.(3) in the von Neumann sense delivers directly and formally an improved version of the null-collision Monte Carlo method including the thermal distribution of neutrals.

In the plasma model the particle ensemble is represented by the set of numbers:

$$\left\{ i,\ r(i),\ v(i),\ t_C(i),\ s(i) \right\}_{1 \le i \le N} \quad ; \quad N = \sum_s N_s \tag{8}$$

Here i identifies the i-th simulated particle, r(i) and v(i) are the position and velocity vectors assigned to the particle, s(i) is the particle species and $t_c(i)$ is the time left to the next Monte Carlo collision event assigned only after a collision event by

$$\begin{aligned} t_c(i) &= -\frac{1}{\alpha_{s(i),max}} ln\eta \quad ; \quad 0 < \eta \le 1 \quad ; \quad f(\eta) = 1 \\ \alpha_{s,max} &= \max_{\mathbf{r},\, g<g_{max}} g \sum_k n_{c(k)}(\mathbf{r}) \sigma_{k,tot}(g) \end{aligned} \tag{9}$$

where g is the relative speed, g_{max} is a physically grounded cutoff value for g and s is the particle species as above. Unphysical (null) events due to the fact that the collision frequency is $< \alpha_{s,max}$ are removed by the rejection method.

Disappearing particles are removed from the simulation by putting s(i)=0. The particle list is compressed at regular times by removing 'zero' particles.

If N exceeds 2Nin, where Nin is its initial value, particles with odd i are removed and all W_s are multiplied by 2. If, instead, N<0.5Nin, the particle ensemble is doubled and all W_s are divided by 2.

Recombination processes cannot fit the basic PIC/MC formalism since they involve two charged particles. We treat these processes as a combination of two first order ones, each including one of two particle species involved in the process, during a PIC time step Δt. Charge neutrality is not enforced, but obtained on the average.

Two different values for the different W_s are used by applying the technique introduced in ref. [6], in order to reduce the statistical error over macroscopic quantities calculated for minoritary ionic species, i.e. H_2^+, H^+, H^-. In the calculation presented hereafter $W(H^+)/W(e) = 0.1$ and the same for H_2^+ and H^-.

Negative ions reach a steady state density arising from the balance of pseudo zero order production channel α (attachment) and pseudo first-order consumption channels $\beta n(H^-)$ (detachment, recombination), i.e. $\alpha = \beta n(H^-)$. The relaxation time $1/\beta$ is often very long with respect to the other plasma time scales (about 0.01 seconds in the test case of the next section).

A special technique has been devised to speed up the convergence of the negative ion density: we scale both α and β to $k\alpha$ and $k\beta$ respectively, with $k > 1$, while keeping $n(H^-) = \alpha/\beta$ constant. This implies scaling the cross sections for attachment

and detachment, and the rate coefficients in the treatment of H^- recombination (see above). The technique works since electrons and other ion densities are not affected significantly by attachment or H^- detachment. This has been checked by direct comparison with code results with k=1.

As regards the neutral kinetic equation (1c) we preliminarily linearize the loss term and rewrite the equation in the following form:

$$\left(-D_c \frac{\partial^2}{\partial x^2} + \sum_r \nu_{rc} k_r \prod_{c'} n_{c'}^{\nu_{rc}-\delta_{cc'}}\right) n_c(x) = \sum_r \nu'_{rc} k_r \prod_{c'} n_{c'}^{\nu_{rc}} \qquad (1c')$$

This procedure is convenient since in gas phase chemistry loss terms have mostly reaction order one in the lost species. Following space discretization, a Gauss–Seidel relaxation is applied to the resulting system of equations.

4 Test Case: Radio Frequency Discharge in Hydrogen

As a test case, we consider a pure hydrogen RF discharge plasma produced by the parallel-plate configuration, i.e. one plane electrode surface ($x = 0$) is kept at electric potential $V = 0$ (grounded), while the opposite one ($x = d$) is assumed to be driven by an external generator to an oscillating potential.

The PIC mesh has 400 cells, while the neutral reaction/diffusion equation mesh has 50 cells. The PIC time step is 0.05 ns, fitting the Courant condition.

The code was run for 10^6 PIC time steps with 5 PIC/kinetics iterations in order to reach a steady state solution. The main bottleneck is the (albeit scaled) negative ion density relaxation.

The values for physical quantities are: gas temperature T_g = 300 K, voltage amplitude 200 V, gas pressure p = 0.1torr, discharge frequency = 13.56MHz, discharge gap d = 0.06m, DC voltage (bias) = 0 V.

Figs.1-4 show a snapshot of the results. In particular:

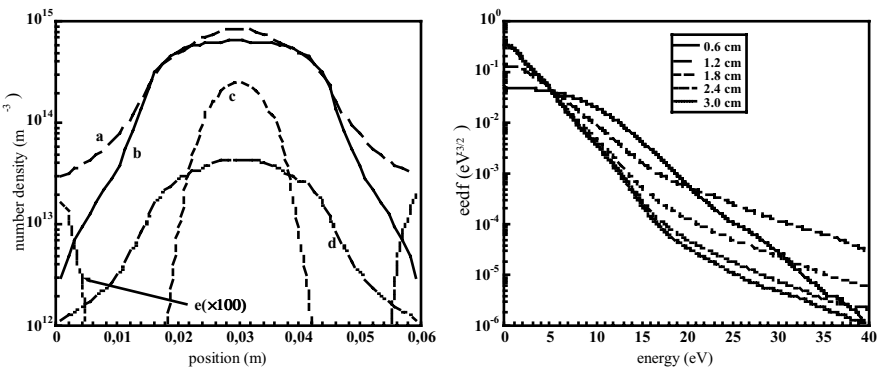


Fig. 1. Left: Number density of charged particles at the steady state for the test case: (a) H_3^+ ions, (b) electrons, (c) H^- ions, (d) H^+ ions, (e) H_2^+ ion; right: the eedf at different positions in the discharge.

Fig. 1 shows the results obtained for the number densities of charged particles. It can be seen that the role of negative ions in the central region of the plasma is not negligible, contrary to the implicit assumption of most of the literature in the field. A large difference is observed for what regards the negative ion density in Fig.1 with respect to ref.[6] due to the different value of γ_v.

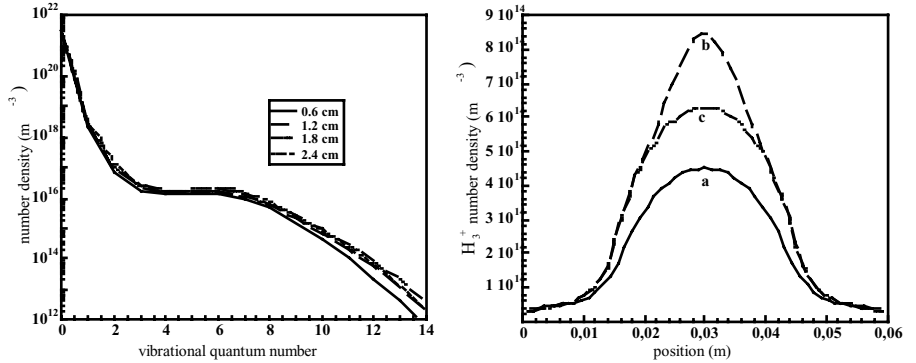


Fig. 2. Left: Plot of the vdf at different positions in the discharge; right: Comparison between H_3^+ number density, as a function of position, calculated by SIGLO-RF fluid model code (a) and our PIC/MC model taking into account (b) and neglecting (c) H^- ion production

Figs. 1-right/2-left show the eedf and the vibrational distribution function (vdf) in different positions in the discharge. For both it is confirmed a non equilibrium character. As already well known the vdf is characterised by a plateau due to EV excitation processes (i.e. high threshold processes with a singlet intermediate [1]) occurring in the sheath/bulk boundary region where the electron energy deposition is larger.

Fig.2-right shows a comparison between our PIC/MC code results and those obtained with the fluid modelling based SIGLO-RF [10] code in the same physical conditions and using mostly the same input data. This code, however, neglects the negative ion and the vibrational kinetics.

It can be noticed a good agreement for the density of the majority positive ion.

The comparison is better in the case of the code run which neglects negative ions, as expected.

These results could be obtained at the cost of 10.5 hours on a Pentium 4, 1.7 GHz PC. The most remarkable point on this respect is the success of the negative ion time scale technique, which allowed to reduce the CPU time by a factor of ten without affecting the overall results in any sensible way.

Since there is no need for a particle/mesh list, the approach is not only highly suitable for parallelization but possibly, with some modifications, for a GRID network implementation.

Of course the computational cost of a fluid model run is by far lower, but this last is less informative as regards the electron energy distribution, since it considers local field conditions based on the calculated local electron energy, and therefore cannot

reproduce kinetic tails, like in Fig.2, which are due to fast electron diffusion. Besides, for the same reason, the fluid approach is not applicable to more rarefied cases, which are easily accessed by particle models like the one presented here.

5 Conclusions

A one dimensional particle model for non equilibrium plasma developed in the past has been applied to hydrogen and improved as regards the treatment of ion kinetics.

In particular we have added a kinetic description of the interaction of charged particles with the thermal distribution of gas particles, a weighting method which enhances the statistics of minoritary ionic species and a scaling technique to speed up the negative ion stabilization. The new version of the method can deliver a very detailed view of the plasma kinetics at a reasonable computational cost.

Acknowledgments. This work has been partially supported by ASI (contr. I/R/055/02) and FISR "Modellistica Molecolare di Sistemi di complessita' Crescente"

References

1. M.Capitelli, C.M.Ferreira, F.Gordiets, A.I.Osipov, 'Plasma Kinetics in Atmospheric Gases', Springer Verlag 2000; M. Capitelli (Ed.), Non Equilibrium Vibrational Kinetics, Springer, New York, 1985
2. M.A.Liebermann and A.J.Lichtemberg, 'Principles of Plasma Discharges and Materials Processing', Wiley N.Y. 1994
3. A.Salabas, G.Gousset and L.L.Alves, Plasma Sources Sci.Techn. 11 448 (2002); J.P.Boeuf and Ph. Belenguer, J.Appl.Phys. 71 4751 (1992)
4. C. K. Birdsall, IEEE Trans. Plasma. Sci. 19 68-85 (1991); R. W. Hockney and J. W. Eastwood, 'Computer Simulation Using Particles', Adam Hilger 1991
5. S. Longo, K. Hassouni, D. Iasillo, and M. Capitelli, J. Phys. III 7 707 (1997); S. Longo, M. Capitelli and K. Hassouni, J. Phys. IV C4 271 (1997); S. Longo, M. Capitelli, K. Hassouni, Journal of Thermophysics and Heat Transfer 12 473 (1998)
6. S. Longo, I. D. Boyd, Chem. Phys. 238 445 (1998), S. Longo, A. Milella, Chem. Phys. 274 219 (2001).
7. S. Longo, Physica A 313 389 (2002); S. Longo, P. Diomede, Eur.Phys.J AP, in press
8. C. Gorse, R. Celiberto, M. Cacciatore, A. Laganà, M. Capitelli, Chem. Phys. 161 (1992) 211
9. T. Simko, V. Martisovits, J. Bretagne, G. Gousset, Phys. Rev. E 56 5908 (1997)
10. J.P. Boeuf, L.C. Pitchford, 1995, SIGLO-RF, PC version 1.0, a 1D User-Friendly Model for RF Discharges Simulation (Kinema Software, siglo@kinema.com, http://www.siglo-kinema.com/siglo-rf.htm, Monument, CO)
11. M. Yan, A. Bogaerts, W. J. Goedheer and R. Gijbels, Plasma Sources Sci. Techn. 9 583 (2000), K.Radouane et al., J.Appl.Phys. 90 4346 (2001).

Fully Kinetic Particle-in-Cell Simulation of a Hall Thruster

Francesco Taccogna[1], Savino Longo[1,2], Mario Capitelli[1,2], and Ralf Schneider[3]

[1]Dipartimento di Chimica dell'Universitài di Bari, via Orabona 4, 70126 Bari, Italy
cscpft61@area.ba.cnr.it
[2]IMIP-CNR, sect. Bari, via Orabona 4, 70126 Bari, Italy
[3]Max Planck Institute für Plasmaphysik, Wendelsteinstr. 1, D-17491 Greifswald, Germany

Abstract. A 2D axisymmetric fully kinetic Particle-in-Cell (PIC) model of the atom (Xe), ion (Xe^+) and electron dynamics of a stationary plasma thruster (SPT) is developed. Electron-neutral elastic scattering, excitation and ionization processes are modelled by Monte Carlo collision methodology. The interaction of the plasma discharge with the ceramic walls leads to plasma recombination, energy loss and secondary electron. These phenomena are included into the model by different approaches. The electric field is self-consistently solved from the Poisson equation, while the magnetostatic field is precomputed. The code is applied to a scaled SPT thruster geometry where fundamental physics parameters are kept constant. The model reproduces the discharge ignition dynamics. The numerical results will provide a better understanding of the experimentally observed enhanced axial electron current and high frequency oscillations.

1 Introduction

The modelling of electric thruster is a very important issue in view of the increasing importance of such propulsion in all space applications when specific impulse, and not just power, is important, i.e. for satellite guidance, orbit transfer and deep space exploration projects.

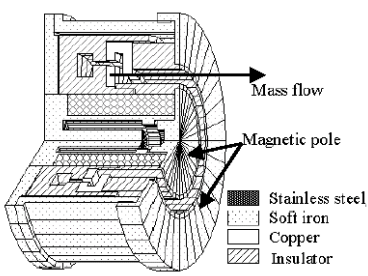


Fig. 1. Schematic representation of the discharge chamber in the SPT-100 thruster

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 588–595, 2004.
© Springer-Verlag Berlin Heidelberg 2004

A SPT can be schematically described (see Fig. 1) as an anode-cathode system, with a dielectric annular chamber where the propellant ionization and acceleration process occurs [1]. This thruster works using a perpendicular electric and magnetic fields configuration. A magnetic circuit generates an axisymmetric and quasi-radial magnetic field between the inner and outer poles. In operation, an electrical discharge is established between an anode (deep inside the channel), which is acting also as a gas distributor, and an external cathode, which is used also as an electron emitter. In this configuration, cathode electrons are drawn to the positively charged anode, but the radial magnetic field creates a strong impedance, trapping the electrons in cyclotron motion which follows a closed drift path inside the annular chamber. The trapped electrons act as a volumetric zone of ionization for neutral propellant atoms and as a virtual cathode to accelerate the ions which are not significantly affected by the magnetic field due to their larger Larmor radii. Generally, xenon is used as propellant. The quasi-radial magnetic field and the channel length L are chosen so that:

$$\begin{aligned} & r_{L,e} << L << r_{L,i} \\ & \omega_{c,i}\tau_i << 1 << \omega_{c,e}\tau_e \end{aligned} \tag{1}$$

where r_L is the Larmor radius, ω_c is the angular cyclotron frequency, τ is the mean time between collisions and the subscripts e and i means electrons and ions. Electrons are strongly magnetized, whereas ions are non-magnetized. The resulting external jet composed by the high speed ion beam is subsequently neutralized by part of electrons coming from the external cathode-compensator. The presence of an insulator as wall material has a profound effect on the plasma within a Hall type thruster. After an impact with dielectric walls, high energy electrons are absorbed and release less energetic secondary electrons that are more firmly confined by the magnetic field. The result is that the dielectric wall limits the temperature of the electrons confined into the channel. By limiting the electron temperature, a smooth continuous variation of the plasma potential results. These effects will be specially addressed within this work.

Models of Hall thrusters have been developed using hybrid fluid-particle approaches [2-7] to aid in the optimization of the performance of the thruster. In all these models, electrons are described as a fluid by the first three moments of the Boltzmann equation. They are 1D (axial or radial), quasi-1D (considering wall losses) or 2D (in the plane (r,z) or (r,θ)), and steady-state or transient (on the ion time scale). Most of the models are based on the quasi-neutrality (QN) assumption. Therefore, Poisson's equation is not solved and the constraints related to the explicit time integration of the transport equations and explicit space integration of the Poisson's equation were therefore eliminated. This assumption considerably simplified the numerical aspect of the simulation. Usually, in the electron momentum transport equation, Bohm and/or near-wall conductivity were included by means of empirical fitting parameters. A question which cannot be resolved by these models, and which in fact strongly limits the reliability of their results, is the electron transport in SPTs, and in particular the important role that the electron interactions with the channel walls play together with volume processes. As a result,

SPT performance is affected by both the state of the wall surface and the properties of plasma structures on Debye and electron-Larmor scales. Therefore, the construction of a self-consistent theory of SPT processes requires a kinetic description of not only heavy particles (atoms and ions), but of electrons as well. To accomplish this, we present a two-dimensional axisymmetric Particle-In-Cell (PIC) [8,9] model using Monte Carlo Collision (MCC) method to take into account electron-neutral interactions. Secondary electron emission from the wall is simulated by a probabilistic model [10]. The electric potential is calculated solving Poisson equation solution without assuming quasi-neutrality.

We first describe in Sec. 2 the numerical model. The results are presented and discussed in Sec. 3.

2 Numerical Method

2.1 Geometrical Scaling

To capture electron dynamics, we need a time-step on the order of the plasma oscillation. Unfortunately, using the real mass ratio, neutral particles will require millions of such time-steps to cross the simulation region. In order to speed-up the code we have reduced the dimension of the discharge while preserving the values of the relevant parameters that govern the physics [11]. All the basic plasma characteristics in gas-phase ionization devices rely heavily on the ionization process, whereby the neutrals are ionized in collisions with the electrons. As the size of the device is reduced, everything else remaining constant, the number of collisions that the electrons and the neutrals experience with each other as they traverse the effective length of the device is reduced. Thus, in order to maintain the effective collision probability, it is necessary to increase the number densities of all the species in proportion, that is the mass flow rate of propellant should scale as:

$$\dot{m} = M_{Xe} n_N c_s A \sim L \tag{2}$$

where M_{Xe}, n_N and c_s are the neutral mass, density and thermal velocity respectively, while A is the area of the anode. Moreover, to preserve the effectiveness of the electron confinement scheme (the ratio of the Larmor radius to the thruster dimension) under scaling, the strength of the magnetic field must vary inversely with length. The discharge current I_d, determined as the product of the current density times the area of the device, scales proportionally with length. Consequently the thrust T, that is the total force undergone by the SPT in reaction to the acceleration of the ions, scales linearly with the length while the specific impulse $I_{sp} = T / \dot{m}g$ remains invariant under geometrical scaling. Here we have assumed that the electron temperature is constant and independent of scale, as was shown in Ref. [11]. However, the benefits of this trick are limited. First, the plasma frequency must remain shorter than the electron gyrofrequency. Also, the Debye length must remain a small quantity with respect to overall thruster dimensions. If the sheaths become too large, they can interfere with the discharge. Under these constraints the geometrical reducing factor was ζ=0.02.

2.2 Magnetic Field

Because the self-induced magnetic field is negligible compared to the applied field from the Hall thruster's coils, the simulation uses only a constant magnetic field. Under this condition we can define the magnetic potential ϕ_B and use the corresponding Laplace equation to solve for the magnetic field. Boundary conditions are the same used by Fife [3]: at the right and left boundaries, one has zero flux, while on the outer magnetic pole, one has $\phi_B = 1$, and on the inner pole one has $\phi_B = 0$. We then solved for ϕ_B across the domain, and took the gradient to arrive at the magnetic field **B**. Finally, we specified a control point and a field strength at that control point, and used that to normalize the field.

2.3 Neutral Gas Particle Kinetics

The neutral density in this test case is two order of magnitude higher than the plasma density. Therefore we let neutral super-particles be some integral number s_N=100 times larger than plasma super-particles. As a neutral undergoes ionization events, s_N decreases by one unit until s_N=0 and the particle disappears. We initialize the simulation by expanding a plume of neutrals from the anode with a long time-step until we approach a steady state. The number of neutrals macroparticles created at the anode line (z=0) per time-step Δt_N, is a function of the scaled mass flow rate according to:

$$\Delta N_N = \frac{\dot{m}'}{M w_N} \Delta t_N \tag{3}$$

where w_N is the neutral statistical weight. They are injected with a radial position sampled from an uniform cylindrical density distribution and their initial velocity distribution is taken to be half-Maxwellian with a temperature of typically 500 K, by using the polar form of the Box-Muller transformation for velocity. Neutrals can disappear when they reach the right boundary of the geometry. Neutrals which hit the anode and walls are re-emitted according to an half-Maxwellian at a wall temperature (900 K) based on experimental data. One must consider the probability for a particle crossing the surface to have a given direction. This probability density actually follows a cosine law, due to the fact that particles with a large velocity component along the normal to the surface escape more frequently than others. The probability densities for axial and azimuthal velocity components are Gaussian distribution (the polar form of the Box-Muller transformation was used), while the probability density for the radial component is a Rayleigh distribution. Neutral-neutral collisions may be ignored, assuming a neutral free molecular regime because the mean free path is much longer than thruster dimension.

2.4 Plasma Phase

When neutral particles have filled the simulation region, electrons are introduced each time step Δt from the exit plane (the cathode is not included in the simulation region) with a steady state current control method of electron injection [12]:

$$n_e = \frac{I_d \ t}{qw} - \ n_i^c + \ n_e^c \tag{4}$$

where w is the statistical weight of electron super-particles, q is the elementary charge, Δn_i^c and Δn_e^c are the number of ions and electrons passing the free space boundary each iteration. Electron initial velocity distribution is taken to be half-Maxwellian with a temperature of typically about 15 eV. They can lose energy and change momentum by collisions (elastic scattering, excitation and ionization) with neutrals. These interactions are modelled by MCC technique and scattering cross sections are gathered from the literature [12]. For each electron we calculate the probability of an e-N scattering in a time Δt_{coll} short with respect to the mean free flight time:

$$P_{tot} = n_N \sigma_{tot}(v_e) v_e \Delta t_e = \sum_{k=1}^{3} n_N \sigma_k(v_e) v_e \Delta t_e = \sum_{k=1}^{3} P_k \tag{5}$$

where n_N is the neutral density, σ_{tot} is the total electron-neutral cross section, v_e is the electron velocity and P_1, P_2 and P_3 are the probability for the occurrence of collisional event 1, 2, and 3, respectively. P_{tot} is compared with a random number r_d sampled from an uniform distribution in the range [0,1] in order to decide if a collision event happens (in our case Δt_{coll} is chosen so that $P_{tot}<10^{-2}$). If $P_{tot}>r_d$, we compare another random number to the cross sections for elastic scattering, excitation, and ionization to determine which type of event occurs. We choose the collisional event j if

$$\sum_{k=1}^{j-1} \frac{P_k}{P_{tot}} < r_d \leq \sum_{k=1}^{j} \frac{P_k}{P_{tot}} \tag{6}$$

In all cases, the electron is scattered isotropically. If the collision is inelastic, energy (8.32 eV for the first excitation and 12.1 eV for the first ionization) is subtracted from the electrons. In the case of ionization, ions and secondary electrons are created at the primary electron's location. The energy of primary and secondary electrons is divided randomly.

When an electron strikes the dielectric wall (Boron Nitride BN), we decide the number of electrons emitted on the basis of the energy and the angle of impact of the incident electron implementing the probabilistic model of Furman and Pivi [10]. The energy spectrum of the emitted electrons is able to reproduce the three different type of secondary electrons, that is backscattered (high energy region), rediffused (middle energy region) and true secondary electrons (low energy region). For simplicity, we assume the same emission-angle distribution (~cosθ and uniform azimuthal angle) for all electrons, regardless of the physical mechanism by which they were generated.

Verboncoeur interpolation method [13] which guarantees charge density conservation on a radial metrics is used to weight particles to the grid nodes, where

the field equations are solved, and to weight the fields back to the particles. The electric potential is re-calculated each time-step using finite differences method for Poisson equation which is solved iteratively using successive over-relaxation (SOR) with Chebyscev acceleration technique [8]. As regards the boundary conditions, we keep the electric potential constant at the anode and at the exit plane and we also assume that the electric field changes its sign at the dielectric surface and its magnitude remains the same [14]. Then, the boundary condition at the walls can be written as:

$$E_w = -\frac{\rho}{2\varepsilon_0} \tag{7}$$

where ρ is the surface charge density and ε_0 is the free space permittivity constant; the possible surface conductivity of the dielectric is neglected. The potential at the wall surfaces is assumed to be zero. Electrons are moved discretizing the equations of motion by the leapfrog method of Boris [9]. All electrons which hit the anode boundary and the free space boundary are deleted. At the end of each electron loop, ions are created at the ionization locations. The initial ion velocity is set equal to the local neutral background velocity, but ions and neutrals are moved once each 100 electron loop. An ion which impacts the anode and the walls disappears and a neutral is created with one half of the ions initial kinetic energy, but with a random velocity direction. The particle is thus partially accommodated in energy, and fully accommodated in momentum. At the free space boundaries, all particles are deleted and a count of electron, ion and neutral fluxes is maintained.

3 Results and Discussions

In this section, we present the simulation results (scaled values) at the following operating conditions: channel length L=2.5 cm, inner radius R_{in}=3.5 cm, outer radius R_{out}=5 cm, discharge voltage ϕ_d=300 V, mass flow rate $\dot{m} = 3\,\mathrm{mg/s}$, discharge pressure P=2x10^{-5} mbar, discharge current I_d=3.2 A, maximum radial magnetic field $B_{r,max}$=150 Gauss.

Figure 2(a) shows long time history of electron, ion and neutral total number macroparticles. We can see that charged particles continue to oscillate (the simulated discharge is not stable), while the number of neutrals is not strongly changed due to the fact that one neutral transit time corresponding to the simulation time is not sufficient to see the low frequency phenomena related to the neutral scales.

The colour plots of Figs. 2(b)-3 demonstrate the main features of the discharge and show, respectively, the space variation of the electric potential, plasma density and electron temperature. As it can see in Fig. 4(b), most of the potential drop occurs in the exhaust region, where the magnetic field is large. This decrease compensates the low electron conductivity in this region and ensures current continuity. It is customary to allocate the acceleration region here. The large axial electric field resulting from this voltage drop is responsible for accelerating the ions from the ionization region to the exit plane and the electron from the outlet to the

anode. From the figure, also the little anode drop (~20 V) and the lateral wall sheaths, whose voltage drop decreases from the anode to the outlet, are evident. However, the computed potential vanishes at the channel exit, while observations [15] indicate that only one-half to one-third of the potential drop takes place downstream of the thruster exit. This difference is due imposing the zero potential boundary condition at the exit plane in the numerical simulation, i.e., the full potential drop is forced to occur inside the channel.

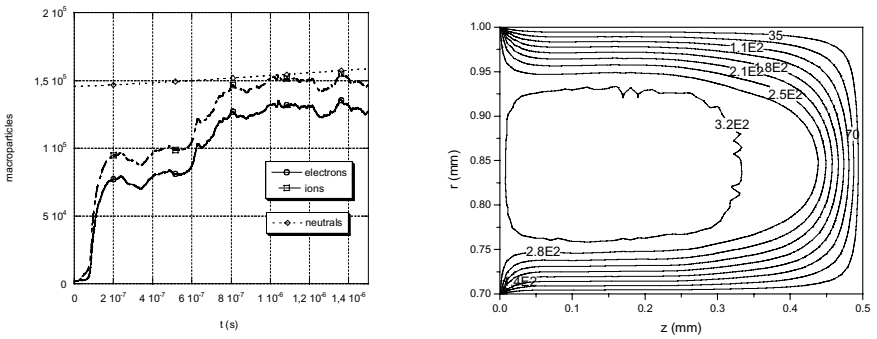


Fig. 2. (a) Long time history of electron, ion and neutral macroparticles. (b) Profile of the plasma potential

The space variation of the plasma density (Fig. 3.a) shows that the plasma reaches its maximum in the center of the channel and in the ionization region, while it decreases in the acceleration region due to the increasing ion velocity. The electron temperature (Fig. 3.b), calculated in each cell from the usual formula reaches a maximum (18 eV) close to the exhaust and decreases to about 6.7 eV in the low electric field region. The peak can be attributed to Ohmic heating due to the maximum azimuthal drift velocity in this region. Furthermore, the drop near the walls is a consequence of the emission of cold secondary electrons from the insulators. These numerical results are consistent with our global understanding of the stationary plasma thruster and reproduce quite well the experimental observations [15].

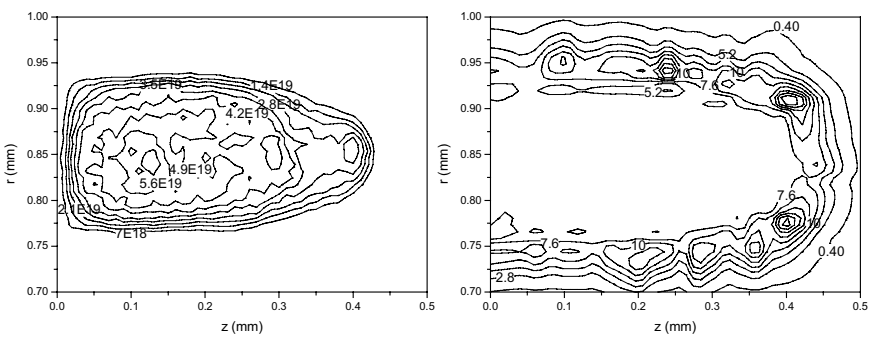


Fig. 3. Profile of the electron (a) density and temperature (b)

5 Conclusions

A 2D(r,z)-3V numerical model was developed to assess the effect of dielectric walls in stationary plasma thrusters. The model consists of a fully kinetic PIC-MCC for the plasma phase. The emission of secondary electron by electron impact from the walls is taken into account by a probabilistic model simulating the different kind of electrons created at the wall, backscattered, rediffused and true secondaries. In order to make the simulation possible and model the electron dynamics, a new kind of scaling-law is applied for the PIC model.

References

1. Zhurin, V. V., Kaufman, H. R., Robinson, R. S.: Plasma Sources Sci. Technol. 8 (1999) R1.
2. Komurasaki, K., Arakawa, Y.: J. Prop. Pow. 11(6) (1995) 1317.
3. Fife, J. M.: PhD thesis, Massachusetts Institute of Technologies (1998).
4. Hagelaar, G. J. M., Bareilles, J., Garrigues, L., Bœuf, J.-P.: J. Appl. Phys. 91(9) (2002) 5592.
5. Koo, J. W., Boyd, I. D.: AIAA paper 2003-10113 (2003).
6. Garrigues L. : PhD thesis, Université Paul Sabatier, Toulouse (1998).
7. Levchenko I., Keidar M.: IEPC-2003-162 (2003).
8. Eastwood, J. W., Hockney, R. W.: Computer Simulation using particles, McGraw-Hill, New York (1981).
9. Birdsall, C. K., Langdon, A. B.: Plasma Physics via Computer Simulation, Mc-Graw-Hill, New York (1985).
10. Furman, M. A., Pivi, M. T. F.: Phys. Rev. Special Topics – Accel. and Beams 5(12) (2002) 124404.
11. Khayms, V.: PhD thesis, Massachusetts Institute of Technology (2000).
12. Szabo, J. J. Jr.: PhD thesis, Massachusetts Institute of Technology (2002).
13. Verboncoeur, J. P. : J. Comp. Phys. 174 (2001) 421.
14. Morozov, A. I., Savel'ev, V. V.: Plasma Phys. Rep. 28(12) (2002) 1017.
15. Bishaev, A. M., and Kim, V. P.: Sov. Phys. Tech. Phis. 23 (1978) 1055.

Standard of Molecular Dynamics Modeling and Simulation of Relaxation in Dense Media

Alexey Y. Kuksin, Igor V. Morozov, Genri E. Norman, and Vladimir V. Stegailov

Institute for High Energy Densities, Russian Academy of Sciences
IHED-IVTAN, Izhorskaya 13/19, Moscow 125412, Russia
bogous@orc.ru

Abstract. Approaches to simulation and modeling of relaxation in dense media are developed which would be universal for some classes of relaxation processes. Three examples of relaxation in strongly dissimilar multi-scale systems are considered: equilibration of electrons and ions in a nonisothermal nonideal plasma, lifetime and nucleation of crystals at superheating, void formation in a crystal under negative pressures.

1 Introduction

Molecular dynamics (MD) method could be a powerful tool for studying nonequilibrium states and relaxation phenomena in dense many-particle systems. In this connection we would mention studies of melting front velocity [1], damping oscillation regime in one-component nonideal plasmas [2], relaxation to equilibrium in two-component two-temperature nonideal plasmas [3,4], recombination relaxation in ultracold plasmas [5], spontaneous decay of a superheated crystal [6], relaxation of SH radical in solid krypton [7], equilibration and sampling of a biomolecule [8], protein folding [9], relaxation in shock wave front [10].

Modeling and simulation of relaxation is a relatively new sector of computer science where the standard of approaches, models and numerics has not been established. This paper contribute to the development of this standard. In section 2 approaches are considered which are specific for the problem studied. They are different from those used in modeling and simulation of equilibrium dense media. Some examples of our results are presented in section 3 which validate the approaches developed. The examples chosen correspond to multi-scale systems: two timescales in plasma due to the electron-ion mass ratio, short and long relaxation stages in plasma due to dynamics and stochastic regimes; the lifetime and the decay duration in metastable crystals which are different by several orders of magnitude.

2 Standard

Any conventional MD simulation starts from more or less arbitrary initial conditions. Then different approaches are applied to equilibrate the system. Only

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 596–603, 2004.
© Springer-Verlag Berlin Heidelberg 2004

subsequent equilibrium run is used to obtain the useful information. The objective of the simulation of relaxation is of the opposite sense, i.e. to get the information from the non-equilibrium part of the MD run, from that part which is discarded and is not used in equilibrium MD simulations. It is evident, that the requirements to the simulation and modeling of the relaxation should differ from those for the equilibrium ones.

Initial state. The choice of the initial state cannot be an arbitrary one. The initial state should correspond to the physical problem we would like to study. The physical system is modeled with respect to the boundary conditions, the character of non-equilibrium and the minimum particle number which is required to reproduce in simulation the main features of the relaxation process studied.

Ensemble of initial states. One initial state is not sufficient as a rule to achieve the satisfactory accuracy. So an ensemble of initial states is needed to obtain a number of MD runs and perform the averaging of the results. Each state in the ensemble should differ from each other significantly but all of them are to be equivalent with respect to the non-equilibrium problem studied. The last requirement needs some art and can be checked by the following criteria. The increase of the number I of initial states increases the accuracy of averaging as $\sqrt{I}$ and does not shift the mean value.

Ensemble-dependence. Non-equilibrium states can be different for the same media. Different ensembles of initial state correspond to those cases in the simulation. The problem is if there are some features of the relaxation processes which are ensemble-independent.

Dynamic and stochastic relaxation. The inherent feature of MD dynamical system is the Lyapunov instability. Due to this instability there appears a dynamic memory time t_m which limits the time interval when the Caushy problem is valid for MD numerical integration. For times greater than t_m MD trajectory "forgets" its initial conditions and ceases to correlate with the hypothetical Newtonian trajectory with the same initial conditions [11,12]. We expect that the duration of ensemble-dependent part of the relaxation correlates with t_m.

Number of particles. Correlation lengths do not remain constant during the relaxation. The account of these lengths is of a great importance when the cooperative phenomena are considered, e.g. nucleation or plasma waves. The MD box size should be greater than the maximum correlation length which might appear during the relaxation process.

Instantaneous diagnostics. New procedures are to be used to find characteristics of relaxation at a given moment of time, e.g. parameters which qualitatively show the overall degree of deviation from the Maxwellian velocity distribution and from the Gaussian distribution of crystal particles in space. The distribution form shows the details of the deviations. The diagnostics can be applied separately to different regions in space for non-homogeneous cases.

Time-averaged diagnostics. Averaging over a relatively long period is also needed to obtain e.g. the values of dynamic structure factor. The procedures are suggested [4,13] which imply freezing of the instantaneous non-equilibrium state for the period of measurement. The idea is to introduce the energy exchange

with an external bath keeping constant the non-equilibrium excitation. It is necessary to check that this procedure does not transform the non-equilibrium state during the period of measurement. Then the external source is switched off and the relaxation continues as if there were no period of freezing.

Numerics. Numerical integration with a variable time step is applied if the relaxation includes dramatic increase of velocities of some particles [5]. Pair distribution function is a sensitive tool to detect particles which could approach potentially forbidden small distances due to insufficiently small time step.

Parallel computations. As well known, MD code is hard to parallelise because of many network communications required for the force calculation procedure. An alternative parallelising algorithm is possible in case of simulation of the relaxation. Since the averaging over initial states is required the relaxation from different initial states can be calculated in parallel. The network load is very low in this case. Therefore this algorithm should be used to calculate the relaxation whereas the standard one to obtain an ensemble of initial states (e.g. in subsection 3.1).

Analytics. Relaxation process can be too long to be simulated by MD in particular for multi-scale systems. Models are used in this case. The results obtained are rescaled with the help of analytic expressions which should be found.

3 Validation of Standard

3.1 Relaxation in Two-Component Nonideal Plasmas

We consider a fully-ionized two-component plasma of N electrons and N ions with masses m and M, respectively. The ions are singly ionized. The formation of the bound states should be prohibited in classical MD simulations of the systems of free charges by the appropriate choice of the interaction pseudopotential. We use the so called "corrected Kelbg" pseudopotential [14]. As low bound states are excluded the pseudopotential does not depend on the chemical element. The number of ions in present simulations is $N = 64$–800. The choice of N and another details of simulations are discussed elsewhere [4,11].

The initial nonequilibrium state is characterized by different temperatures of electrons T_e and ions T_i. We used the initial conditions where the velocities of electrons or/and ions are equal to zero. There are also two variants of the initial spacial configuration of the cold ions. The first variant is a crystal structure with cubic lattice which corresponds to the state after ionization in solids. The second one is a quasi-random configuration obtained from the equilibrium isothermal electron-ion plasma with the same number density.

The results of MD runs are averaged over an ensemble of $I = 50 - 200$ initial states. Provided the result is N-independent, the relative error is given by $1/\sqrt{NI}$. The error bars in figures below correspond to the confidence coefficient 0.68. The bars are not indicated if they are smaller than the size of the points. The values of T_e and T_i are obtained as the average kinetic energy of the particles

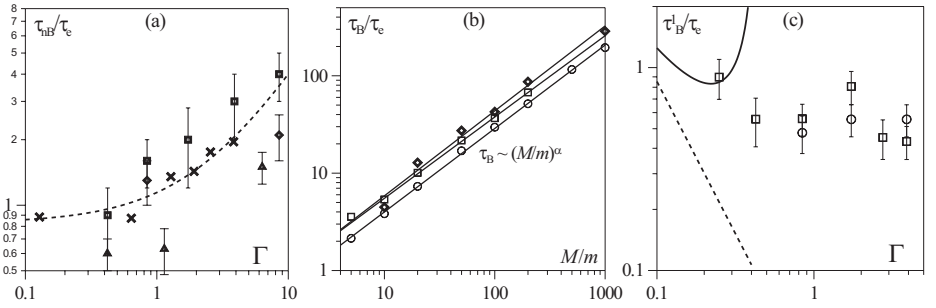


Fig. 1. The dependencies of the duration of the nonexponential relaxation stage τ_{nB} on the nonideality parameter (a) and of the exponential relaxation time on the mass ratio (b) and on the nonideality parameter (c) for different initial conditions: $T_i(0) = 0$ (crystal) — circles, $T_i(0) = 0$ (quasirandom) — squares, $T_e(0) = 0$ — rhombus, $T_e(0) = T_i(0) = 0$ — triangles.
(a) The dynamical memory time t_m — crosses, the dashed line is drawn through the crosses to guide the eye.
(a) Straight lines correspond to the power fits $\tau_B \sim (M/m)^{\alpha}$. $\Gamma = 1.28$.
(c) Solid curve corresponds to the Landau formula (1), dashed curve — the same with the fixed Coulomb logarithm $L_e = 3.2$

$T(t) = \frac{1}{2NI} \sum_{j,k}^{N,I} m v_{jk}^2(t)$. The relaxation is charactarized by the the decrease of the difference of electron and ion temperatures $\Delta T = |T_e - T_i|$ which following the Boltzmann equation has the exponential asymptote $\Delta T \sim \exp\{-t/\tau_B\}$.

Nonexponential relaxation. The distinguishing feature of a nonideal plasma is a nonexponential stage of the relaxation which precedes the exponential one. The duration of this stage τ_{nB} does not depend on the mass ratio for $M/m \geq 100$. It increases with the increase of the nonideality parameter (Fig. 1a), $\tau_e = 2\pi/\omega_p$, $\omega_p = \sqrt{4\pi n_e e^2/m}$ is the period of electron plasma oscillations, n_e is the electron number density. The plasma nonideality is characterized by the parameter $\Gamma = e^2(4\pi n_e/3)^{1/3}/(k_B T)$, where T is the final temperature at the end of relaxation. As seen τ_{nB} depends on the initial conditions but stays comparable with t_m.

The kinetic equation used to calculate relaxation in ideal plasma implies that the collisions are statistically independent and the particle motion is stochastic. However the time of the stochastization or dynamical memory time t_m in nonideal plasma becomes greater than the time between collisions [11].

Dependence of the exponential relaxation time τ_B on plasma properties. The dependence of τ_B on the mass ratio is shown in Fig. 1b. As seen, different initial conditions result in close values of $\tau_B(M)$. The lower relaxation time for the crystal-like initial configuration of ions is caused by an additional heating of ions due to reconfiguration and correlation build up in the ionic subsystem. The mass dependence can be fitted by the power fit $\tau_B/\tau_e \sim (M/m)^{\alpha}$ in all cases in Fig. 1b. The dependence of α on the nonideality parameter can be fitted by the parabolic curve $\alpha(\Gamma) = 1 - 0.15\,\Gamma + 0.035\,\Gamma^2$, $0 < \Gamma < 4$.

The dependencies of τ_B on the ion mass and Γ can be separated as $\tau_B(\Gamma, M) = \tau_B^1(\Gamma)(M/m)^{\alpha(\Gamma)}$. The dependence of $\tau_B^1(\Gamma)$ is presented in Fig. 1c. As seen the initial conditions do not affect τ_B^1 significantly. In the weak nonideality region the MD results are in a good agreement with the Landau theory [15]

$$\tau_B^* = \frac{3}{16} \frac{(mT_i + MT_e)}{\sqrt{2\pi m M}} \frac{1}{e^4 n_e L_e}, \tag{1}$$

where L_e is the Landau logarithm. The deviation becomes crucial with the increase of Γ.

One can estimate the relaxation times in real experimental conditions using given Γ-dependence and mass-dependence. The error of determination of coefficients α_B and α_{nB} is about $\xi_\alpha = 5\%$. The corresponding error of the extrapolation of relaxation times e.g. for aluminium is $\xi_\tau = \log(M_r/M)\xi_\alpha = 40\%$, where M_r/M is the ratio between real and model ion masses. The obtained precision is enough for comparison by an order of magnitude. The results of such extrapolation correlate with the experimental data [16].

3.2 Decay of Metastable Crystals

A system under study is a *fcc* crystal of particles interacting via Lennard-Jones potential $U(r) = 4\epsilon((\sigma/r)^{-12} - (\sigma/r)^{-6})$. Periodic boundary conditions are used.

Decay of crystal subjected to constant rate heating. Superheated solid is a state of matter that can be realized only under particular conditions of high rate energy impacts and/or very low concentration of defects and impurities which could lead to heterogeneous melting (e.g. see [17,18,19]).

Particle structure is prepared for the simulation to be an ideal lattice initially. The system is equilibrated for 10^4 time steps at the temperature $T = 0.04\epsilon$ (extremely low temperature is chosen to preserve the crystal structure completely defect-free). The particular value of temperature is not crucial for the following. The ensemble of initial states can be chosen from the equilibrium run.

Then the model crystal is subject to the isochoric heating at the constant rate $\dot{T} = 6 \times 10^{-4}(\epsilon^3/m\sigma^2)^{1/2}$. The heating is performed by velocity rescaling in the spirit of the Berendsen thermostat technique. As heating is being done the temperature of the model crystal becomes higher than the melting temperature at the given density $T_m = 1.25\epsilon$. When temperature reaches $T = 1.52\epsilon$ crystal structure decays into liquid. Phase transformation leads to the step in the dependence of the averaged potential energy U on time t (see Fig. 2).

Structural changes are well manifested in terms of the Lindemann parameter $\delta_L = \langle \Delta r^2 \rangle / r_{nn}$, where $\langle \Delta r^2 \rangle$ is the displacement of particles from the ideal *fcc* lattice sites averaged over all the particles in the simulation box at the current time step, r_{nn} is the nearest neighbour distance in the lattice. According to the Lindemann criterion of melting $\delta_L = 0.12 - 0.13$ for simple crystals at the melting temperature T_m. This relation holds for the case considered (Fig. 2). However at $T = T_m$ there is no phase transition but the crystal becomes superheated and δ_L achieves the values of 0.4 before decay.

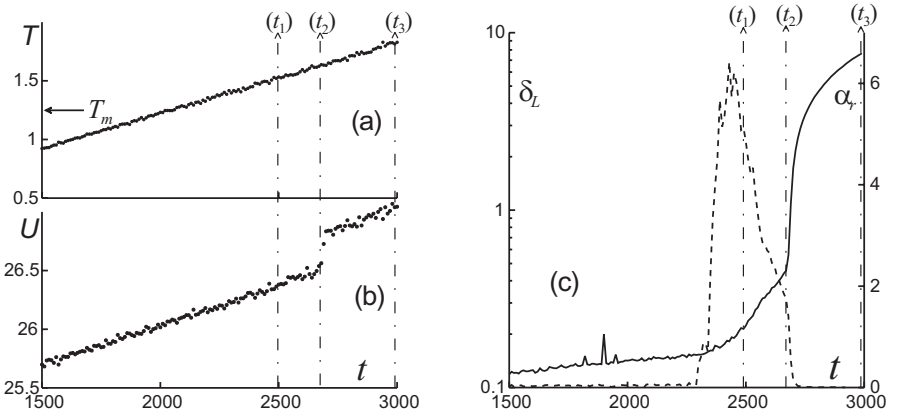


Fig. 2. The time-dependencies: (a) the temperature T, (b) the average potential energy of particles U, (c) the Lindemann parameter δ_L (solid) and the non-Gaussian parameter α_r (dashed). The time is given in units $(m\sigma^2/\epsilon)^{1/2}$. $N = 6912$; $\rho\sigma^3 = 1$

Another parameter $\alpha_r = 0.6(\langle \Delta r^4 \rangle / \langle \Delta r^2 \rangle^2) - 1$ shows the degree of deviation of the distribution of displacements from the Gaussian shape ($\alpha_r = 0$ for the Gaussian distribution). There is a peak on the α_r dependence on time which corresponds to the emergence of collective modes in the motion of particles. This phenomenon can be considered as a precursor of decay under heating. It does not produce any effect on the potential energy time dependence, however Lindemann parameter turns to the faster increase during this peak. After the decay motion of particles becomes irregular, so that α_r vanishes and δ_L gets the diffusion-like dependence on time. The loss of long range order is well seen in the evolution of the radial distribution function (in an animation).

The system evolution can be divided into three parts: 1) heating and superheating up to the appearance of the melting precursor; 2) the emergence of collective modes in the particle motion; 3) the decay to the liquid state. The process of the relaxation in the case of the superheated crystal decay is connected with homogeneous nucleation that allows to apply the formalism of the classical nucleation theory (see [20,6] for details).

Void formation in crystals under negative pressure. A condensed matter can exist at negative pressures. Such states were observed experimentally, e.g. in [21].

Negative pressure states are the metastable ones, so they can exist only a limited period of time. Then the states decay spontaneously and voids are formed. The temperature T and number density ρ are chosen for the simulation to obtain a state near spinodal in order its lifetime can be reached during the MD run. Initial velocities are taken from Maxwellian distribution.

The results are presented for certain examples in Fig. 3-4. Time dependencies for T, pressure P, parameters δ_L, α_r and α_v are given in the Lennard-Jones units. Parameter α_v shows the degree of the deviation from the Maxwellian distribution. The metastable state with approximately constant T and P exists

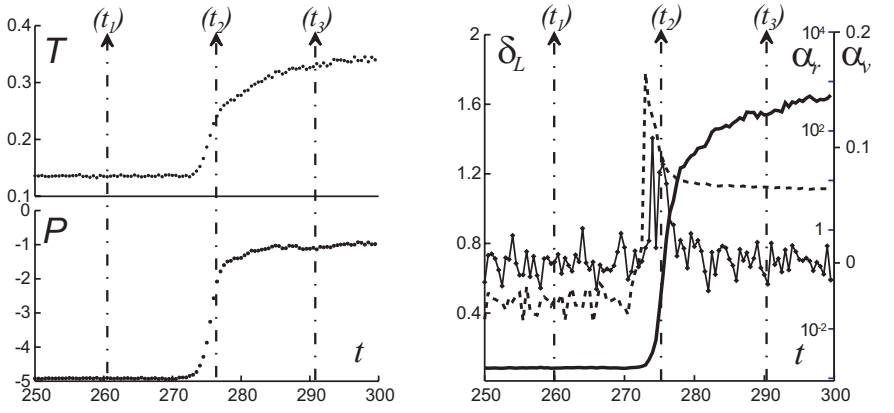


Fig. 3. The time-dependencies of T, P (points), δ_L (bold solid), α_r (dashed) and α_v (solid) along the last interval of the MD run. $N = 4000$

during $\sim 275(m\sigma^2/\epsilon)^{1/2}$ which is much greater than t_m, then spontaneous decay starts. It takes only few time units to form new more or less stationary state. The non-zero value of α_v points to the fact that the Maxwellian distribution is broken during the short decay period. The time t_1 in Fig. 3 corresponds to the long-order structure, while the structures at t_2 and t_3 reveal only short order.

The microscopic picture of the spontaneous decay is presented in Fig. 4. The decay starts with the local disordering of the crystal structure. The voids appear in the melted regions only at the next stage of the decay. Another characteristic feature of the decay is observed: the structure formed is strongly non-uniform. At least three phase states can be distinguished: crystal clusters, disordered regions and voids. In fact the final state of our MD run is not an equilibrium one, since the pressure remains to be negative till the end of our simulation (Fig. 3).

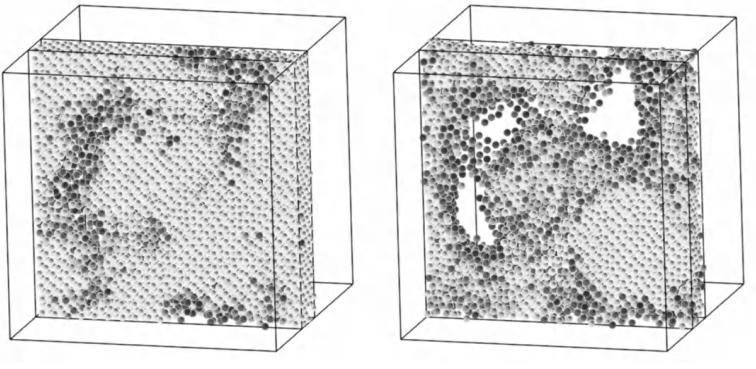


Fig. 4. The microscopic picture of the slab inside the MD cell for two moments of time corresponding to the decay process. The grey scale shows the degree of disordering. $N = 32000$, $\rho = 0.8\ \sigma^{-3}$

4 Conclusions

At attempt is done to formulate standard requirements to MD modeling and simulation of relaxation processes in dense media, which would be more or less universal and independent of the physical systems studied. The main features are considered which are new with respect to the simulations of equilibrium systems: 1) physically proven choice of the ensemble of initial non-equilibrium states for averaging the relaxation MD runs; 2) consideration of both first ensemble-dependent stage of relaxation and subsequent stage which might be ensemble-independent and remind of Boltzmann exponential relaxation; 3) calculation of dynamical memory time t_m and comparison of t_m with the duration of the first stage of relaxation; 4) instantaneous and time-averaged measurement procedures which are specific for studying the relaxation processes; 5) analytical extrapolation approaches to estimate relaxation times which are too long to be simulated by MD, especially for multi-scale systems. Three examples of relaxation are considered: equilibration of electrons and ions in nonideal plasmas, decay of metastable crystals under superheating or stretching.

Acknowledgments. This research is partially supported by grants NWO-RFBS 047.016.007, RFBS 03-11-90272v, "Integracia" U0022 and project "Parallel computations on multiprocessor computer systems" of RAS. AK, IM and VS acknowledge the support from "Dynastia" foundation and the International center of fundamental physics in Moscow. The computations were performed on the cluster granted by DAAD.

References

1. Wang, J., Li, J., Yip, S., Wolf, D., Phillpot, S. Physica A **240** (1997) 396
2. Zwicknagel, G. Contrib. Plasma Phys. **39** (1999) 155
3. Hansen, J.P., McDonald, I.R. Phys. Lett. **97A** (1983) 42
4. Morozov, I.V., Norman, G.E., *et al.* J. Phys. A: Math. Gen. **36** (2003) 8723
5. Kuzmin, S.G., O'Neil, T.M. Phys. Plasmas **9** (2002) 3743
6. Norman, G.E., Stegailov, V.V., Valuev, A.A. Contrib. Plasma Phys. **43** (2003) 384
7. Firsov, D.A., Grigorenko, B.L., *et al.* Chem. Phys. Lett **338** (2001) 317
8. Stella, L., Melchionna, S. J. Chem. Phys., **109** (1998) 10115
9. Snow, C.D., Nguyen, H., Pande, V.S., Gruebele, M. Nature **420** (2002) 102
10. Tanguy, D., Mareschal, M., Lomdahl, P.S., *et al.* Phys. Rev. B **68** (2003) 144111
11. Morozov, I.V., Norman, G.E., Valuev, A.A. Phys. Rev. E **63** (2001) 036405
12. Norman, G.E., Stegailov, V.V. Computer Physics Comm. **147** (2002) 678
13. Gibbon, P., Pfalzner., S. Phys. Rev. E **57** (1998) 4698
14. Ebeling, W., Norman, G.E., *et al.* Contrib. Plasma Phys. **39** (1999) 61
15. Lifshitz, E.M., Pitaevskii, L.P. Physical kinetics (Oxford: Pergamon) (1981)
16. Riley, D., Woolsey, N.C., McSherry, D., *et al.* Phys. Rev. Lett. **84** (2000) 1704
17. Bonnes, D.A., Brown, J.M. Phys. Rev. Lett. **71** (1993) 2931
18. Jin, Z.H., Gumbsch, P., Lu, K., Ma, E. Phys. Rev. Lett. **87** (2001) 055703
19. Sheng-Nian Luo, Ahrens, T.J., Çağin T., *et al.* Phys. Rev. B. **68** (2003) 134206
20. Norman, G.E., Stegailov, V.V. Doklady Physics **47** (2002) 667
21. Kanel, G., Razorenov, S., Baumung, K., Singer, J. J. Appl. Phys. **90** (2001) 136

Implicit and Explicit Higher Order Time Integration Schemes for Fluid-Structure Interaction Computations

Alexander van Zuijlen* and Hester Bijl

Delft University of Technology, Faculty of Aerospace Engineering,
P.O. Box 5058, 2600GB, The Netherlands
A.H.vanZuijlen@lr.tudelft.nl

Abstract. In this paper higher order time integration schemes are applied to fluid-structure interaction (FSI) simulations. For a given accuracy, we investigate the efficiency of higher order time integration schemes compared to lower order methods. In the partitioned FSI simulations on a one-dimensional piston problem, a mixed implicit/explicit (IMEX) time integration scheme is employed: the implicit scheme is used to integrate the fluid and structural dynamics, whereas an explicit Runge-Kutta scheme integrates the coupling terms. The resulting IMEX scheme retains the order of the implicit and explicit schemes. In the IMEX scheme considered, the implicit scheme consists of an explicit first stage, singly diagonally implicit Runge-Kutta (ESDIRK) scheme, which is a multistage, L-stable scheme.

1 Introduction

For many engineering applications, fluid-structure interaction (FSI) phenomena are important for an efficient and safe design. Increased computational power has enabled the simulation of FSI, through coupling of existing flow and structure solvers. However, the simulation of long term dynamic behavior is still very time consuming. Therefore efficiency of the FSI solver is of the utmost importance.

It has already been demonstrated that for flow applications, higher order time integration schemes are computationally more efficient than popular lower order schemes, even for engineering levels of accuracy [2]. This drives the idea to use higher order time integration schemes for fluid-structure interaction simulations as well. So far we only found examples of fluid-structure interaction computations based on at most second order implicit time integration methods [3,5,11]. For the coupled fluid-structure simulations we envisage a partitioned scheme, meaning that an existing flow and structure solver can be used, each solving efficiently their own equations on a separate domain and coupling is obtained through boundary conditions. A partitioned strategy enables the re-use of all the effort put into the development and optimization of such codes, especially the

* Funded by the Netherlands Organisation for Scientific Research (NWO)

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 604–611, 2004.
© Springer-Verlag Berlin Heidelberg 2004

iterative solvers. In the partitioned scheme, the coupling terms between fluid and structure are not solved in a fully coupled system, but are given as a boundary condition. In order to obtain a stronger coupling, a predictor-corrector scheme is sometimes suggested, but we only found applications with first or second order time integration schemes. Also the use of sub-iterations at every time step is sometimes proposed, which comes, however, at an increased computational cost. An other possibility is to integrate one system with an explicit scheme, which can only be done efficiently in the absence of stiffness for that system.

In this paper we consider a mixed implicit/explicit (IMEX) time integration scheme based on higher order implicit and explicit Runge-Kutta methods. The separate fluid and structural dynamics are integrated by an unconditionally stable, implicit scheme and only the coupling component is integrated by the explicit scheme. The IMEX scheme is applied to a linear and a nonlinear, one-dimensional piston, which resembles a simple fluid-structure interaction problem. The order and accuracy of the resulting scheme are investigated and efficiency is compared to lower order methods.

2 Fluid-Structure Interaction Model Problem

The test case considered is the one-dimensional piston problem (Fig. 1), which is often used as an initial test case for fluid-structure interaction simulations [1, 10].

Fig. 1. The one-dimensional piston problem

The flow is modelled as a one-dimensional, isentropic, inviscid flow. Usually, the governing equation for the flow are written in the arbitrary Lagrangian-Eulerian (ALE) to cope with the moving and deforming mesh [4,6]. In this paper, however, we only consider the fluid on a non-moving mesh. The governing equation for the fluid yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_K w_\mathrm{f}\mathrm{d}x + \int_{\partial K} F(w_\mathrm{f})\cdot n\mathrm{d}s = 0, \tag{1}$$

wherein $w_\mathrm{f} = \begin{pmatrix}\rho\\ \rho u\end{pmatrix}$ the fluid state vector, ∂K is the boundary of K, $F(w_\mathrm{f}) = \begin{pmatrix}\rho u\\ \rho u^2+p\end{pmatrix}$ is the flux vector and n is the unit normal vector perpendicular to ∂K pointing outward. Coupling is achieved through an inflow/outflow boundary condition at the interface $u(x=0)=0$, $u(x=L)=\dot{q}$.

The structure is modelled as a simple harmonic oscillator with spring stiffness k, mass m and without any physical damping under an external pressure force due to a difference between the ambient pressure and the pressure in the flow at the interface, yielding

$$m\ddot{q} + kq = p(x = L) - p_0, \tag{2}$$

wherein $\ddot{q}$ denotes the structural acceleration $\frac{\mathrm{d}^2}{\mathrm{d}t^2}q$, $p(x = L)$ is the pressure in the flow at the interface and p_0 is the ambient pressure.

The flow is discretized using a cell-centered Finite Volume (FV) method on a uniform mesh of N cells. A standard second order central scheme and a first order upwind scheme [8] are used. In the fluid domain two ghost cells K_0 and K_{n+1} are introduced to cope with the boundary conditions. The structural dynamics are written as a system of two ordinary differential equations for which the state vector is denoted by w_s. The coupled, nonlinear system is written in semi-discrete form

$$\dot{w}_\mathrm{f} = F_\mathrm{f}(w_\mathrm{f}, w_\mathrm{s}), \qquad \dot{w}_\mathrm{s} = F_\mathrm{s}(w_\mathrm{s}, w_\mathrm{f}), \tag{3}$$

wherein F_f the flux for the fluid under boundary condition w_s and F_s the flux for the structure under boundary condition w_f. The system, linearized around an equilibrium state, yields

$$\dot{w}'_\mathrm{s} = A_\mathrm{s}w'_\mathrm{s} + A_\mathrm{sf}w'_\mathrm{f}, \tag{4}$$

$$\dot{w}'_\mathrm{f} = A_\mathrm{fs}w'_\mathrm{s} + A_\mathrm{f}w'_\mathrm{f}, \tag{5}$$

wherein w'_s and w'_f denote the perturbation in the structure and fluid states respectively and $A_\mathrm{f} = \frac{\partial F_\mathrm{f}}{\partial w_\mathrm{f}}$, $A_\mathrm{fs} = \frac{\partial F_\mathrm{f}}{\partial w_\mathrm{s}}$, $A_\mathrm{sf} = \frac{\partial F_\mathrm{s}}{\partial w_\mathrm{f}}$ and $A_\mathrm{s} = \frac{\partial F_\mathrm{s}}{\partial w_\mathrm{s}}$ all evaluated at the equilibrium. The matrix A_f is a $2N \times 2N$ band-matrix which contains the discretization of the fluid domain and A_s is a 2×2 matrix. The coupling matrices A_fs $(2N \times 2)$ and A_sf $(2 \times 2N)$ will generally only have a relatively small amount of non-zero entries, since the coupling only takes place at the boundary of the domain.

3 Time Integration by Mixed Implicit/Explicit Schemes

We consider any system of the form

$$\dot{w} = F(w, t), \tag{6}$$

which can be any (semi-discrete) system describing e.g. structural and/or fluid dynamics. Since future applications involve engineering problems a large range of eigenvalues will be introduced due to a wide range of scales in the flow (for example in boundary layers [9]), giving rise to the stiffness of the system. Stiffness can cause the time step to be limited by stability rather than accuracy considerations. Hence, we only consider L-stable, implicit time integration methods, which can cope with stiffness in a robust fashion and dissipate the high frequency modes. Some well-known unconditionally stable implicit methods include the first

and second order multi-step Backward Differentiation Formula (BDF) schemes [7] and the trapezoidal rule method. The third and higher order multi-step BDF methods, however, are only $L(\alpha)$-stable, making them impractical in engineering codes. Additionally the multi step methods are not self-starting. Less known and also less applied in engineering codes are implicit Runge-Kutta (IRK) methods [7]. The IRK methods can be designed with arbitrary high order and L-stability.

When a problem with easily separable stiff and nonstiff components is considered, a combination of implicit and explicit Runge-Kutta methods can be used. The implicit method is used to integrate the stiff component in a stable fashion and the nonstiff component of the system is integrated using the explicit scheme. These combined implicit/explicit (IMEX) schemes are already used for convection-diffusion-reaction problems in [9].

The IMEX schemes we consider in this paper consist of an explicit Runge-Kutta (ERK) and a stiffly-accurate explicit, singly diagonally implicit Runge-Kutta (ESDIRK) scheme, for which the solution at t^{n+1} can be made of arbitrary high order by cancellation of the lower order errors. The ESDIRK scheme is an L-stable, implicit scheme with an explicit first stage, which allows the implicit stages to be second order accurate. For every stage k we solve

$$w^{(k)} = w^n + \Delta t \sum_{i=1}^{k} a_{ki} F^{(i)}, \quad k = 1 \dots s, \tag{7}$$

wherein $F^{(i)} = F(w^{(i)})$ is the flux at stage i. After computing s stages, the solution at the next time level is found by

$$w^{n+1} = w^n + \Delta t \sum_{i=1}^{s} b_i F^{(i)}. \tag{8}$$

In this paper we consider third to fifth order IMEX methods. At least 37 different combinations have been found in the literature, but we only use the ones reported in [9], which are optimized for the Navier-Stokes equations. The third, fourth and fifth order schemes consist of a 4, 6 and 8-stage algorithm respectively. In order to distinguish between the implicit scheme and explicit schemes, we denote a_{ki} for the implicit and $\hat{a}_{ki}$ for the explicit schemes. Both schemes use the same b_i coefficients to obtain the higher order solution. An s-stage ESDIRK scheme needs to solve $s-1$ implicit systems within one time step compared to only one for the multi-step methods. So the question is whether the higher computational cost of the ESDIRK methods is compensated by their higher order accuracy.

4 Partitioning Algorithm

When the discretized fluid and structural equations are written as in (6), the monolithic or fully coupled solution is obtained by direct integration of (6) with any time integration scheme. For this academic problem, the monolithic solution

is still manageable, but for real-world applications a partitioned approach is more favorable. In the proposed partitioning algorithm, both the fluid and structure are integrated by the ESDIRK scheme. At every stage a Gauss-Seidel type of partitioning is applied. When the linear systems (4) and (5) are considered, the structure is advanced to stage k by

$$(I - a_{kk}\Delta t A_{\mathrm{s}})w_{\mathrm{s}}^{(k)} = w_{\mathrm{s}}^{n} + \Delta t \sum_{i=1}^{k-1} a_{ki} F_{\mathrm{s}}^{(i)} + \Delta t \sum_{j=1}^{k-1} \hat{a}_{kj} F_{\mathrm{sf}}^{(j)}, \tag{9}$$

showing that the coupling from fluid to structure is integrated by the ERK scheme. The coupling fluxes are treated in a consistent, explicit way in order to retain the higher order of the scheme. Due to the combined implicit/explicit nature of the scheme, we refer to it as IMEX. The same scheme is obtained when we define a flux predictor for $F_{\mathrm{sf}}^{(k)}$ as

$$F_{\mathrm{sf}}^{(*)} = \sum_{i=1}^{k-1} \frac{\hat{a}_{ki} - a_{ki}}{a_{kk}} F_{\mathrm{sf}}^{(i)}. \tag{10}$$

Next the fluid is integrated to stage k by

$$(I - a_{kk}\Delta t A_{\mathrm{f}})w_{\mathrm{f}}^{(k)} = w_{\mathrm{f}}^{n} + \Delta t \sum_{i=1}^{k-1} a_{ki} \left(F_{\mathrm{f}}^{(i)} + F_{\mathrm{fs}}^{(i)} \right) + a_{kk}\Delta t F_{\mathrm{fs}}^{(k)}, \tag{11}$$

wherein the coupling flux $F_{\mathrm{fs}}^{(k)}$ is already known from (9). After solving all s stages of the integration scheme, the state at the next time level is obtained

$$w_{\mathrm{s}}^{n+1} = w_{\mathrm{s}}^{n} + \Delta t \sum_{i=1}^{s} b_i \left(F_{\mathrm{s}}^{(i)} + F_{\mathrm{sf}}^{(i)} \right), \tag{12}$$

$$w_{\mathrm{f}}^{n+1} = w_{\mathrm{f}}^{n} + \Delta t \sum_{i=1}^{s} b_i \left(F_{\mathrm{f}}^{(i)} + F_{\mathrm{fs}}^{(i)} \right), \tag{13}$$

which completes one partitioned integration.

For the nonlinear system (3), the separation of flux contributions to be used with the implicit and explicit schemes needs extra attention. In order to obtain separated flux contributions, a linearization of the flux is made around t_n

$$F_{\mathrm{f}}(w_{\mathrm{f}}^{(k)}, w_{\mathrm{s}}^{(k)}) = F_{\mathrm{f}}(w_{\mathrm{f}}^{n}, w_{\mathrm{s}}^{n}) + \Delta w_{\mathrm{f}}^{(k)} \frac{\partial F_{\mathrm{f}}}{\partial w_{\mathrm{f}}} + \Delta w_{\mathrm{s}}^{(k)} \frac{\partial F_{\mathrm{f}}}{\partial w_{\mathrm{s}}} + \text{h.o.t.}. \tag{14}$$

We define the flux at stage k as

$$F_{\mathrm{f}}^{(k)} = \bar{F}_{\mathrm{f}}^{(k)} + \Delta F_{\mathrm{fs}}^{(k)}, \tag{15}$$

wherein

$$\bar{F}_{\mathrm{f}}^{(k)} = F_{\mathrm{f}}(w_{\mathrm{f}}^{(k)}, w_{\mathrm{s}}^{n}) \quad \text{and} \quad \Delta F_{\mathrm{fs}}^{(k)} \approx \Delta w_{\mathrm{s}}^{(k)} \frac{\partial F_{\mathrm{f}}}{\partial w_{\mathrm{s}}}. \tag{16}$$

The contribution $\bar{F}_{\mathrm{f}}^{(k)}$ can be considered the flux in the flow under a constant boundary condition w_{s}^{n} and $\Delta F_{\mathrm{fs}}^{(k)}$ can be considered the change in flux at the interface caused by a change in structural state within the time step. The integration of the nonlinear system follows the IMEX partitioning algorithm as described in Eqs. (9–13), with $F_{\mathrm{f}} = \bar{F}_{\mathrm{f}}$, $F_{\mathrm{fs}} = \Delta F_{\mathrm{fs}}$ and similarly $F_{\mathrm{s}} = \bar{F}_{\mathrm{s}}$ and $F_{\mathrm{sf}} = \Delta F_{\mathrm{sf}}$.

5 Results and Discussion

The piston problem, for which an exact solution exists in the linear case, has only one single structural node with a mass $m = 2$ and a spring stiffness $k = 1.429$. With these settings the fluid has a strong influence on the structural motion without dominating the structural dynamics. In all computations 64 finite volume cells (FVC) are used to discretize the flow. For the linear simulations the standard central scheme without artificial dissipation is used. However for the nonlinear simulations stabilization is necessary and a first order upwind scheme is employed. The coupled simulations are performed with the IMEX scheme, using third to fifth order ESDIRK schemes for the implicit time integration and third to fifth order ERK schemes for the integration of the coupling terms. In the linear computations a Gaussian elimination is used to solve the implicit systems. In the nonlinear simulations, a Picard iteration is used for the monolithic BDF2 scheme and a Newton iteration is used in the flow solver. Since the iterative solvers have different efficiencies, it is not justified to determine the work by the total amount of CPU time. Therefore the work is defined as the number implicit stages that need to be solved during the simulation.

The linear system has a coupled radial frequency of $\omega_c = 1.01$ and a coupled period of $P = 6.19$. The computational efficiency of the higher order partitioned IMEX schemes is investigated by comparing them to the monolithic (or fully coupled) solution with the second order BDF time integration scheme. This way the efficiency of BDF2 is not diminished by partitioning. The energy error versus work is presented in Fig. 2. Since every time integration scheme has a fixed number of implicit stages to solve within one time step, we can obtain the order of the schemes by measuring the slope of the curves in the asymptotic range, as displayed in Fig. 2. The IMEX schemes show design order in the asymptotic range. In addition, the efficiencies of fourth and fifth order IMEX are much higher than monolithic BDF2. The partitioned third order IMEX performs equally with respect to monolithic BDF2. The effect of the consistent explicit coupling flux treatment can be seen by comparing the result for fifth order IMEX to the partitioned ESDIRK5 scheme, which uses the fluid state at the previous time level as a predictor for the implicit flux calculations. The scheme with the predictor does not show design order and needs about 2.5 times as much work to obtain the same accuracy as third order IMEX.

For the nonlinear problem an exact solution is not obtained. A "temporally exact solution" is obtained with the fifth order IMEX scheme and $\Delta t = 1/1024$. At $t = 0$ the flow is at rest and the piston has an initial displacement $q_0 = 0.5$.

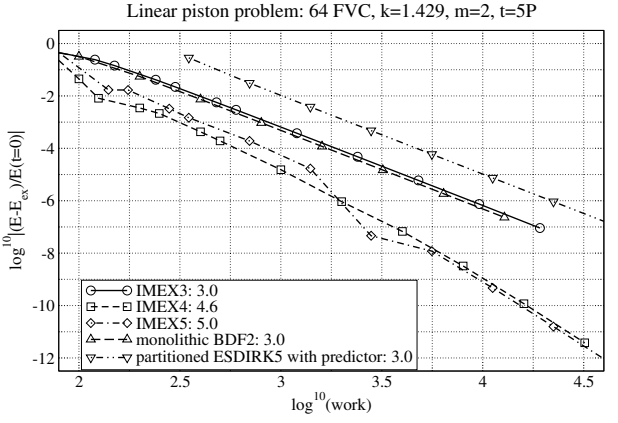


Fig. 2. Energy efficiency of IMEX schemes compared to monolithic BDF2 and partitioned ESDIRK5 without consistent predictor

The piston is released and the simulation is run until $t = 7$ which is a little over one period ($P \approx 6.34$). At $t = 7$ the L_2-norm of the error in the fluid density field is determined. For the structure the error for the displacement and velocity are computed. Simulations are performed with time steps ranging from $\Delta t = 1$ to $\Delta t = 1/512$. In Fig. 3 the L_2-norm of the error in the fluid density field is shown versus work for the third to fifth order IMEX schemes and the monolithic BDF scheme. From the results the order of the scheme are obtained by calculating

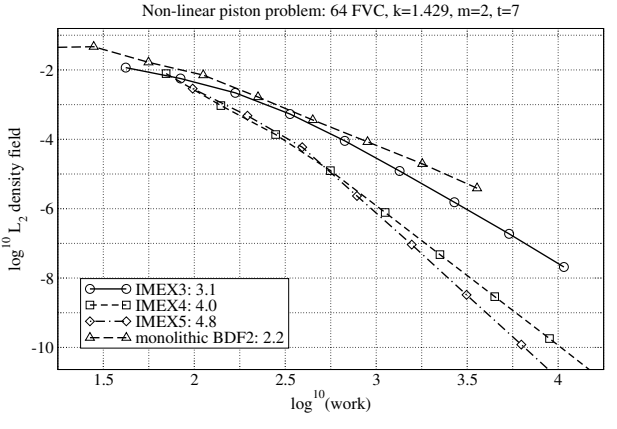


Fig. 3. Fluid density field efficiency for the third to fifth order IMEX schemes compared to monolithic BDF2

the slope of the graphs in the asymptotic range. In the asymptotic range the IMEX scheme have design order. For the larger time steps ($\Delta t = 1, 1/2$), the

order of the scheme is diminished, but the accuracy is still far better than the second order monolithic scheme with the same time step. When the error level is considered between -2 and -4, we find that the monolithic BDF2 scheme needs $\approx$1.3–2.5 times as much work compared to the fourth and fifth order IMEX schemes. For higher accuracy requirements, the efficiency of the higher order schemes increases.

6 Conclusions and Future Work

For the partitioned simulation of fluid-structure interaction third to fifth order IMEX schemes are used. Applied to a linear and nonlinear piston problem, the IMEX schemes are more efficient than the monolithic BDF2 scheme.

Up to this point only a simple one-dimensional problem has been considered. Future research focusses on the application of the IMEX scheme to more realistic, multidimensional problems.

References

1. F.J. Blom, A monolithical fluid-structure interaction algorithm applied to the piston problem, Comp. Meth. Appl. Mech. Engrg. 1998;167:369-91.
2. H. Bijl, M.H. Carpenter, V.N. Vatsa, and C.A. Kennedy, Implicit Time integration schemes for the unsteady compressible Navier-Stokes equations: laminar flow, J. Comput. Phys. 2002;179:1-17.
3. C. Farhat, and M. Lesoinne, Two efficient staggered algorithms for the serial and parallel solution of three-dimensional nonlinear aeroelastic problems, Comp. Meth. Appl. Mech. Engrg. 2000;182:13-22.
4. C. Farhat, P. Geuzaine, and Celine Grandmont, The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids, J. Comput. Phys. 2001;174:669-94.
5. P. Geuzaine, G. Brown, C. Harris, and C. Farhat, Aeroelastic dynamic analysis of a full F-16 configuration for various flight conditions, AIAA Journal 2003;41(3):363-71.
6. H. Guillard, and C. Farhat, On the significance of the geometric conservation law for flow computations on moving meshes, Comp. Meth. Appl. Mech. Engrg. 2000;190:1467-82.
7. E. Hairer, S.P. Norsett, and G. Wanner, Solving ordinary differential equations I, nonstiff problems, Springer Verlag 2000.
8. C. Hirsch, Numerical Computation of internal and external flows. Vol. 1: Fundamentals of numerical discretization, John Wiley & Sons Ltd. 1994.
9. C.A. Kennedy and M.H. Carpenter, Additive Runge-Kutta schemes for convection-diffusion-reaction equations, Appl. Num. Math. 2003;44:139-81.
10. S. Piperno, C. Farhat, and B. Larrouturou, Partitioned procedures for the transient solution of coupled aeroelastic problems – Part I: model problem, theory and two-dimensional application, Comp. Meth. Appl. Mech. Engrg. 1995;124:79-112.
11. S. Piperno, and C. Farhat, Partitioned procedures for the transient solution of coupled aeroelastic problems – Part II: energy transfer analysis and three-dimensional applications, Comp. Meth. Appl. Mech. Engrg. 2001;190:3147-70.

Accounting for Nonlinear Aspects in Multiphysics Problems: Application to Poroelasticity

David Néron[1], Pierre Ladevèze[1], David Dureisseix[2], and Bernard A. Schrefler[3]

[1] LMT-Cachan (ENS Cachan / CNRS / Paris 6 University)
61, avenue du Président Wilson, F-94235 Cachan Cedex, France
{neron,ladeveze}@lmt.ens-cachan.fr
[2] LMGC (Montpellier 2 University / CNRS)
CC 048, place Eugène Bataillon, F-34095 Montpellier CEDEX 5, France
dureisse@lmgc.univ-montp2.fr
[3] Department of Structural and Transportation Engineering (University of Padova)
Via Marzolo 9, I-35131 Padova, Italy
bas@caronte.dic.unipd.it

Abstract. Multiphysics phenomena lead to computationally intensive structural analyses. Recently, a new strategy derived from the LATIN method was described and successfully applied to the consolidation of saturated porous soils.
One of the main achievements was the use of the LATIN method to take into account the different time scales which usually arise from the different physics: a multi-time-scale strategy was proposed.
Here we go one step further and use the LATIN method to deal with some of the classical nonlinearities of poroelasticity problems (such as non-constant stiffness and permeability) and we show that these phenomena do not result in a significant increase of the computation cost.

1 Introduction

For coupled multiphysics problems such as fluid-structure interaction, partitioned procedures and staggered algorithms are often preferred, from the point of view of computational efficiency, to direct analysis (also called the monolithic approach). Moreover, partitioning strategies enable one to use different analyzers for different subsystems, and help keep the software manageable.

Recently, an approach suitable for multiphysics problems was developed based on the LArge Time INcrement method (LATIN) [1] and applied to the consolidation of saturated porous soils, which is a typical example of a highly coupled fluid-structure interaction problem. The term *consolidation* designates the slow deformation of the solid phase accompanied by flow of the pore fluid. One of the consequences of natural consolidation is surface subsidence, i.e. the lowering of the Earth's surface. The consolidation analysis of soils has long been recognized as an important problem in civil engineering design [2].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 612–620, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The principles of the LATIN method and examples of its applicability to such coupled multiphysics problems were given in [3]. This strategy was compared to the Iterated Standard Parallel Procedure (ISPP) [4], which is one of the standard partitioning schemes. In [5], a multi-time-scale strategy was proposed in order to improve the LATIN procedure by taking into account the different time scales. An *ad hoc* radial loading approximation for both kinematic and static quantities was also set up in order to increase the modularity of the approach and to reduce the storage cost. Here, we go one step further and use the LATIN method to deal with some of the classical nonlinearities of poroelasticity problems (such as non-constant stiffness and permeability).

2 The Reference Problem

Let us briefly describe a typical consolidation problem [6]. A structure Ω is made of a saturated porous material undergoing small perturbations and isothermal evolution over the time interval $[0, T]$ being studied.

The loading consists of a prescribed displacement $\underline{U}_d$ on a part $\partial_1\Omega$ of the boundary, a traction force $\underline{F}_d$ on the complementary part $\partial_2\Omega$ of $\partial_1\Omega$, a fluid flux w_d on another part $\partial_3\Omega$ of the boundary and, finally, a prescribed pore pressure p_d on the complementary part $\partial_4\Omega$ of $\partial_3\Omega$. For the sake of simplicity, we assume that there are no body forces.

For solid quantities, strain and stress are denoted $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ respectively; for fluid quantities, the pore pressure gradient is denoted $\underline{Z}$ and the opposite of Darcy's velocity $\underline{W}$; finally, q denotes the rate of fluid mass accumulation in each representative elementary volume.

The state of the structure is given by the set of the fields $\mathbf{s} = (\boldsymbol{\varepsilon}, p, \underline{Z}, \boldsymbol{\sigma}, q, \underline{W})$ defined on the whole structure Ω and over the time interval $[0, T]$ being considered. The problem consists in finding $\mathbf{s}$ in the corresponding space $\mathbf{S}^{[0,T]}$ which verifies at each time step the following equations:

- in the solid, compatibility of strains $\boldsymbol{\varepsilon}$ and equilibrium of stresses $\boldsymbol{\sigma}$:

$$\begin{aligned} \underline{U} \in \mathcal{U}^{[0,T]} \quad \text{and} \quad \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\underline{U}) \quad \text{on } \Omega \\ \underline{\operatorname{div}}\boldsymbol{\sigma} = \underline{0} \quad \text{on } \Omega \quad \text{and} \quad \boldsymbol{\sigma}\underline{n} = \underline{F}_d \quad \text{on } \partial_2\Omega \end{aligned} \tag{1}$$

  $\mathcal{U}^{[0,T]}$ being the set of the finite-energy displacement fields on $\Omega\times[0, T]$ equal to $\underline{U}_d$ on $\partial_1\Omega$;
- in the fluid, flow conservation for Darcy's velocity $-\underline{W}$:

$$\begin{aligned} p \in \mathcal{P}^{[0,T]} \quad \text{and} \quad \underline{Z} = \underline{\operatorname{grad}}\, p \quad \text{on } \Omega \\ q = \operatorname{div} \underline{W} \quad \text{on } \Omega \quad \text{and} \quad \underline{W} \cdot \underline{n} = w_d \quad \text{on } \partial_4\Omega \end{aligned} \tag{2}$$

  $\mathcal{P}^{[0,T]}$ being the set of the finite-energy pressure fields on $\Omega\times[0, T]$ equal to p_d on $\partial_3\Omega$;
- the constitutive relations:

- Hooke's law, which relates the macroscopic stress $\boldsymbol{\sigma}$ to the strain $\boldsymbol{\varepsilon}$ and the pore pressure p so that:

$$\boldsymbol{\sigma} = \mathbf{D}\boldsymbol{\varepsilon} - bp\mathbf{I} \tag{3}$$

- Darcy's law, which relates Darcy's velocity to the pore pressure gradient:

$$\underline{W} = \frac{K}{\mu_w}\underline{Z} \tag{4}$$

- compressibility, which relates the fluid accumulation rate to the pressure rate and couples it with the rate of volume modification:

$$q = \frac{1}{Q}\dot{p} + b\,\mathrm{Tr}\,\dot{\boldsymbol{\varepsilon}} \tag{5}$$

$\mathbf{D}$ is Hooke's tensor of the drained skeleton, b is Biot's coefficient, K is the intrinsic macroscopic permeability and μ_w is the dynamic viscosity of the saturation fluid. Throughout the following sections, the operator $\frac{K}{\mu_w}\mathbf{I}$ will be designated by $\mathbf{H}$. Finally, Q is Biot's modulus.

3 Nonlinear Behavior

Most of the consolidation problems which have been analyzed so far are limited to the assumption of linear elastic constitutive behavior and constant permeability, but in most geotechnical situations the behavior of the soil is nonlinear. Following Kondner and his co-workers [7], the stress-strain curves for both clay and sand in a conventional triaxial compression test (constant σ_3) may be approximated by a hyperbolic equation of the form:

$$\sigma_1 - \sigma_3 = \frac{\varepsilon_1}{A + B\varepsilon_1} \tag{6}$$

which relates the difference between the major principal stress σ_1 and the minor principal stress σ_3 to the major principal strain ε_1. A and B are material constants which can be determined experimentally. Then, Hooke's law is defined by:

$$\boldsymbol{\sigma} = \mathbf{D}(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon} - bp\mathbf{I} \tag{7}$$

However, Kondner's model (6) is available only for one-dimensional analysis. This is the case of the following numerical test. There is also evidence that the intrinsic permeability is not constant, even in the case of full saturation. It seems reasonable [2] to assume a dependency of the permeability on the void ratio (or porosity) as well as on the deformation. We propose to test the LATIN method on a variation of one of the laws given in [8] for the intrinsic permeability:

$$K(\boldsymbol{\varepsilon}) = K_0 \frac{n_0}{1+n_0}\left(1 + \frac{1}{n_0}\left\langle \frac{\mathrm{Tr}\,\boldsymbol{\varepsilon} - \mathrm{Tr}\,\boldsymbol{\varepsilon}_0}{-\mathrm{Tr}\,\boldsymbol{\varepsilon}_0}\right\rangle_+^{\alpha}\right) \tag{8}$$

where $\langle\cdot\rangle_+$ denotes the positive part, K_0 and n_0 the initial intrinsic permeability and porosity, $\boldsymbol{\varepsilon}_0$ the strain below which the intrinsic permeability cannot decrease (typically $\mathrm{Tr}\,\boldsymbol{\varepsilon}_0 = -n_0$), and α a material constant. Darcy's law is then defined by:

$$\underline{W} = \mathbf{H}(\boldsymbol{\varepsilon})\underline{Z} \tag{9}$$

Thus, the consolidation problem which is to be simulated is nonlinear.

4 The LATIN Method for Multiphysics Problems

The LATIN method is a nonincremental iterative approach originally designed for nonlinear time-dependent problems [1]. For coupled multiphysics problems, the method consists in extending the notion of *material interface* (between substructures) [9] to that of *multiphysics interface*. Such an interface must take into account the coupling between the constitutive relations. The development of this strategy was described in [3] and only the main principles will be reviewed here.

At each iteration, the LATIN method produces an approximation of the solution over the whole domain and over the entire time interval being studied. The method is based on three principles:

- **The first principle** consists in separating the difficulties. For coupled field problems, a first set of equations, $\mathbf{A_d}$, containing the so-called admissibility conditions is defined. In order to avoid dealing with both a global and a coupled problem simultaneously, the remaining equations are grouped into a second set of equations, $\boldsymbol{\Gamma}$; these equations, which are local in the space variables, are the constitutive relations.
- **The second principle** of the method consists in using search directions to build approximate solutions of $\mathbf{A_d}$ and $\boldsymbol{\Gamma}$ alternatively until a sufficient convergence level has been reached. Each iteration consists of 2 stages: once an element $\mathbf{s}_n \in \mathbf{A_d}$ is known, the *local stage* of iteration $n+1$ uses an initial search direction E^+ to provide an element $\hat{\mathbf{s}}_{n+1/2} \in \boldsymbol{\Gamma}$:

$$\begin{aligned}
(\hat{\boldsymbol{\sigma}}_{n+1/2} - \boldsymbol{\sigma}_n) + \mathbf{L}_{n-1/2}(\hat{\dot{\boldsymbol{\varepsilon}}}_{n+1/2} - \dot{\boldsymbol{\varepsilon}}_n) &= 0 \\
(\hat{q}_{n+1/2} - q_n) + r(\hat{p}_{n+1/2} - p_n) &= 0 \\
(\hat{\underline{W}}_{n+1/2} - \underline{W}_n) + \mathbf{H}_{n-1/2}(\hat{\underline{Z}}_{n+1/2} - \underline{Z}_n) &= 0
\end{aligned} \tag{10}$$

$\mathbf{L}_{n-1/2}$, $\mathbf{H}_{n-1/2}$ and r are three parameters of the method; they do not influence the solution once convergence has been reached. However, their values modify the convergence rate of the algorithm. In dimensional analysis, r can be chosen in the form $r = \frac{1}{Qt_h}$, where t_h is an arbitrary characteristic time. The choice of $\mathbf{L}_{n-1/2}$ and $\mathbf{H}_{n-1/2}$ will be discussed below.

At each integration point, using the constitutive relations (3,4,5), the local stage leads to the resolution of a small system of ordinary differential equations in the local space variables:

$$\begin{aligned}\mathbf{L}_{n-1/2}\hat{\dot{\boldsymbol{\varepsilon}}}_{n+1/2} + \mathbf{D}(\hat{\boldsymbol{\varepsilon}}_{n+1/2})\hat{\boldsymbol{\varepsilon}}_{n+1/2} - b\hat{p}_{n+1/2}\mathbf{I} &= \mathbf{A}_n \\ \frac{1}{Q}\hat{\dot{p}}_{n+1/2} + r\hat{p}_{n+1/2} + b\,\mathrm{Tr}\,\hat{\dot{\boldsymbol{\varepsilon}}}_{n+1/2} &= \alpha_n \\ (\mathbf{H}_{n-1/2} + \mathbf{H}(\hat{\boldsymbol{\varepsilon}}_{n+1/2}))\hat{\underline{Z}}_{n+1/2} &= \underline{\beta}_n\end{aligned} \tag{11}$$

where $\mathbf{A}_n = \boldsymbol{\sigma}_n + \mathbf{L}_{n-1/2}\dot{\boldsymbol{\varepsilon}}_n$, $\alpha_n = q_n + rp_n$ and $\underline{\beta}_n = \underline{W}_n + \mathbf{H}_{n-1/2}\underline{Z}_n$ are known quantities from local stage $n+1$, and with the initial conditions on the pressure and strain fields. This nonlinear system (11) is solved using a Newton-type scheme.

Once an element $\hat{\mathbf{s}}_{n+1/2} \in \boldsymbol{\Gamma}$ is known, the *linear stage* provides an element $\mathbf{s}_{n+1} \in \mathbf{A}_{\mathbf{d}}$. $\mathbf{s}_{n+1}$, which must satisfy the admissibility relations, is sought along a search direction E^- conjugate of the previous one, so that the mechanical and hydraulic problems remain uncoupled:

$$\begin{aligned}(\boldsymbol{\sigma}_{n+1} - \hat{\boldsymbol{\sigma}}_{n+1/2}) - \mathbf{L}_{n+1/2}(\dot{\boldsymbol{\varepsilon}}_{n+1} - \hat{\dot{\boldsymbol{\varepsilon}}}_{n+1/2}) &= 0 \\ (q_{n+1} - \hat{q}_{n+1/2}) - r(p_{n+1} - \hat{p}_{n+1/2}) &= 0 \\ (\underline{W}_{n+1} - \underline{\hat{W}}_{n+1/2}) - \mathbf{H}_{n+1/2}(\underline{Z}_{n+1} - \underline{\hat{Z}}_{n+1/2}) &= 0\end{aligned} \tag{12}$$

One can note that the search directions in linear stage n and local stage $n+1$ are conjugates if the parameters of these directions are kept constant.

In order to use a finite element approach, the admissibility of $\mathbf{s}_{n+1}$ is expressed using a variational formulation. On the one hand, this admissibility condition consists in $\underline{U} \in \mathcal{U}^{[0,T]}$ and $\boldsymbol{\sigma} \in \mathcal{S}^{[0,T]}$ such that:

$$\forall t \in [0,T], \quad \forall \underline{U}^\star \in \mathcal{U}_0, \quad \int_\Omega \mathrm{Tr}[\boldsymbol{\sigma}\boldsymbol{\varepsilon}(\underline{U}^\star)]d\Omega = \int_{\partial_2\Omega} \underline{F}_d \cdot \underline{U}^\star dS \tag{13}$$

where $\mathcal{U}_0$ is the set of the finite-energy displacement fields on Ω which vanish on $\partial_1\Omega$. On the other hand, the admissibility condition also consists in $p \in \mathcal{P}^{[0,T]}$, $q \in \mathcal{Q}^{[0,T]}$ and $\underline{W} \in \mathcal{W}^{[0,T]}$ such that:

$$\forall t \in [0,T], \quad \forall p^\star \in \mathcal{P}_0, \quad \int_\Omega (qp^\star + \underline{W} \cdot \underline{\mathrm{grad}}\, p^\star)d\Omega = \int_{\partial_4\Omega} w_d p^\star dS \tag{14}$$

where $\mathcal{P}_0$ is the set of the finite-energy pressure fields on Ω which vanish on $\partial_3\Omega$. Equations (13) and (14) define two uncoupled global problems parameterized by time t.

The convergence of this approach is guaranteed provided that $\mathbf{L}_{n+1/2}$, $\mathbf{H}_{n+1/2}$ and r are positive definite operators which remain constant throughout the iterations [1].

– **The third principle** uses the fact that the successive approximations are defined over both the entire domain and the entire time interval to represent the solution on a radial loading basis. This last point was detailed in [1] and developed, for this particular case, in [3,5]. Briefly, this approach enables one to reduce the number of space fields generated and, therefore, the number of global systems to be solved.

A multi-time-scale strategy was also described in [5]. This strategy enables one to use different time steps for the solid and fluid parts of the problem. In particular, in order to perform an iso-quality simulation (i.e. with identical contributions to the global error) the fluid part requires a smaller time step than the solid.

Choice of the search direction $(\mathbf{L}_{n+1/2}, \mathbf{H}_{n+1/2})$ Many choices, all of which ensure the convergence of the LATIN method, are available [1]. The easiest way is to take a constant search direction:

$$\forall n, \quad \mathbf{L}_{n+1/2} = t_m \mathbf{D}(\hat{\boldsymbol{\varepsilon}} = \mathbb{O}) \quad \text{and} \quad \mathbf{H}_{n+1/2} = \mathbf{H}(\hat{\boldsymbol{\varepsilon}} = \mathbb{O}) = \frac{K_0}{\mu_w}\mathbf{I} \tag{15}$$

where t_m is an arbitrary characteristic time. This choice allows one to assemble operators $\mathbf{L}_{n+1/2}$ and $\mathbf{H}_{n+1/2}$ only once at the beginning of the algorithm. In [1], it was shown that optimal convergence of the method is achieved by using a tangent search direction. In the present case of a multiphysics problem, an approximation of such a tangent direction is:

$$\begin{aligned} (\boldsymbol{\sigma}_{n+1} - \hat{\boldsymbol{\sigma}}_{n+1/2}) - \mathbf{D}(\hat{\boldsymbol{\varepsilon}}_{n+1/2})(\dot{\boldsymbol{\varepsilon}}_{n+1} - \hat{\dot{\boldsymbol{\varepsilon}}}_{n+1/2}) &= 0 \\ (q_{n+1} - \hat{q}_{n+1/2}) - r(p_{n+1} - \hat{p}_{n+1/2}) &= 0 \\ (\underline{W}_{n+1} - \underline{\hat{W}}_{n+1/2}) - \mathbf{H}(\hat{\boldsymbol{\varepsilon}}_{n+1/2})(\underline{Z}_{n+1} - \underline{\hat{Z}}_{n+1/2}) &= 0 \end{aligned} \tag{16}$$

which is equivalent to $\mathbf{L}_{n+1/2}(t) = \mathbf{D}(\hat{\boldsymbol{\varepsilon}}_{n+1/2}(t))$ and $\mathbf{H}_{n+1/2}(t) = \mathbf{H}(\hat{\boldsymbol{\varepsilon}}_{n+1/2}(t))$. Such a choice requires the assembly and factorization of the operators not only at each iteration, but also at each time step. A new approximation consists in defining an average of the operators over the time interval $[0, T]$:

$$\mathbf{L}_{n+1/2} = \frac{1}{T}\int_0^T \mathbf{D}(\hat{\boldsymbol{\varepsilon}}_{n+1/2}(t))dt \quad \text{and} \quad \mathbf{H}_{n+1/2} = \frac{1}{T}\int_0^T \mathbf{H}(\hat{\boldsymbol{\varepsilon}}_{n+1/2}(t))dt \tag{17}$$

5 Numerical Results

The proposed test case concerns the consolidation of a Berea sandstone soil. The geometry is shown in Fig. 1 and the material characteristics are given in Table 1. The simulation was performed for the one-dimensional case, since the law (6) is defined only in that case.

Table 1. Water-saturated Berea sandstone poroelastic material

Initial porosity	$n_0 = 0.19$	Initial Young's modulus	$E_0 = 14.4$ GPa
Poisson's coeff.	$\nu = 0.2$	Biot's modulus	$Q = 13.5$ GPa
Biot's coeff.	$b = 0.78$	Initial permeability	$\frac{K_0}{\mu_w} = 2\ 10^{-10}$ m^3.s.kg^{-1}

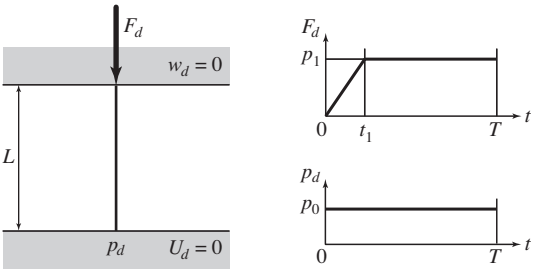


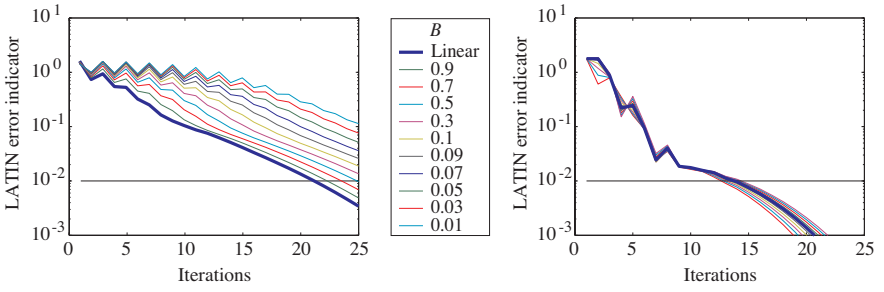
Fig. 1. The force-driven test problem

The time interval was $T = 1$ s with $t_1 = T/2$ and the pressures were $p_1 = 10$ MPa and $p_0 = 0.1$ MPa; the initial condition was $p(t = 0) = p_0$; the height of the structure was $L = 5$ m, discretized into 100 elements (quadratic interpolation for displacements and linear interpolation for pore pressures). The search direction parameters were set to $t_m = 9\ 10^{-3}\ t_c$ and $t_h = 8\ 10^{-3}\ t_c$, where $t_c = 9.3$ s.

Two simulations were performed to illustrate the behavior of the method when nonlinearity increases. The first test was dedicated to the evaluation of the influence of stiffness: in (6), the value of A and σ_3 were set to $A = \frac{1}{E_0}$ and $\sigma_3 = 0$ while the value of B increased from 0 (which corresponds to the linear case) to 1 GPa^{-1}. The second test concerned the evaluation of the influence of permeability: in (8), the value of α was set to $\alpha = 3$ while the initial porosity $n_0 = -\operatorname{Tr} \boldsymbol{\varepsilon}_0$ was no longer that of the Berea sandstone, but was assumed to decrease from 0.9 to 0.01. (The linear case was recovered by taking $n_0 \to +\infty$.)

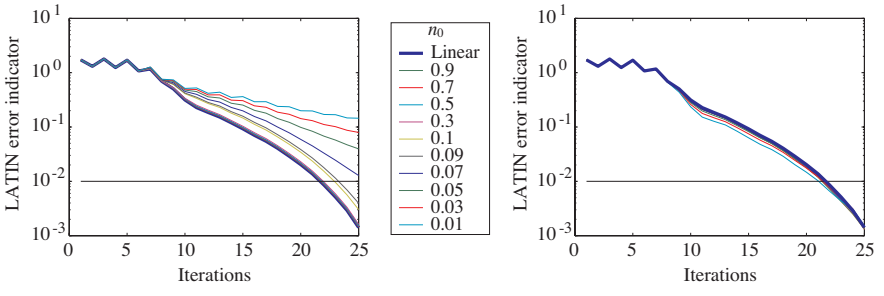
From here on, the error indicator based on the difference between an element $\mathbf{s}$ of $\mathbf{A_d}$ and an element $\hat{\mathbf{s}}$ of $\boldsymbol{\Gamma}$ will be used: $\hat{\eta} = e(\hat{\mathbf{s}} - \mathbf{s})/e(\frac{1}{2}(\hat{\mathbf{s}} + \mathbf{s}))$ with $e^2(\hat{\mathbf{s}} - \mathbf{s}) = \int_0^T (\frac{1}{2}\|\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}\|^2_{\mathbf{D}} + \frac{1}{2}\|\hat{p} - p\|^2_{Q^{-1}})dt$, $\|\boldsymbol{\varepsilon}\|^2_{\mathbf{D}} = \int_\Omega \operatorname{Tr}[\boldsymbol{\varepsilon}\mathbf{D}(\hat{\boldsymbol{\varepsilon}} = \mathbb{O})\boldsymbol{\varepsilon}]d\Omega$ and $\|p\|^2_{Q^{-1}} = \int_\Omega pQ^{-1}pd\Omega$.

Fig. 2**(a)** and Fig. 3**(a)** show that if constant search directions, such as (15), are used (as in [3,5]) the convergence rate is very dependent on the degree of nonlinearity. One can see in Fig. 2**(b)** and Fig. 3**(b)** that if updated average search directions, such as (17), are used at each iteration the convergence rate becomes nearly independent, but in that case, even if the number of iterations is smaller, the strategy could become very expensive because it requires the assembly and factorization of the operators at each iteration. However, one can note that nearly identical results can be obtained by using updated search directions only during the first iterations (usually 4 or 5). This reduces the computational cost significantly. Let us observe that nonlinearities do not increase the number of iterations needed to reach a given error.



(a) Constant search direction **(b)** Variable search direction

Fig. 2. Variable rigidity



(a) Constant search direction **(b)** Variable search direction

Fig. 3. Variable permeability

6 Conclusions

In this paper, we described a partitioned strategy based on the LATIN approach which enables one to take into account some of the classical nonlinearities of consolidation problems. The numerical tests showed that if updated search directions are used during the first iterations, the convergence rate is nearly independent of the level of nonlinearity. Thus, these nonlinear phenomena do not result in a significant increase in the computational costs.

References

1. Ladevèze, P.: Nonlinear Computational Structural Mechanics — New Approaches and Non-Incremental Methods of Calculation. Springer Verlag (1999)
2. Schrefler, B.A., Lewis, R.W.: The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media. 2nd edn. Wiley (1998)
3. Dureisseix, D., Ladevèze, P., Schrefler, B.A.: A computational strategy for multiphysics problems — application to poroelasticity. International Journal for Numerical Methods in Engineering **56** (2003) 1489–1510

4. Matteazzi, R., Schrefler, B., Vitaliani, R. Advances in Computational Structures Technology. In: Comparisons of partitioned solution procedures for transient coupled problems in sequential and parallel processing. Civil-Comp Ltd, Edinburgh, Scotland (1996) 351–357
5. Dureisseix, D., Ladevèze, P., Néron, D., Schrefler, B.A.: A multi-time-scale strategy for multiphysics problems: application to poroelasticity. *To appear in* International Journal for Multiscale Computational Engineering (2004)
6. Coussy, O.: Mechanics of porous continua. John Wiley & Sons (1995)
7. Kondner, R.L.: Hyperbolic stress-strain response: cohesive solids. J. Soil. Mech. Found. Div. ASCE **89 (SM1)** (1963) 115–143
8. Meroi, E.A., Schrefler, B.A.: Biomechanical multiphase approaches in soft biological tissues. In: Proceedings of the 12th International Conference on Mechanics in Medicine and Biology. (2002)
9. Ladevèze, P., Loiseau, O., Dureisseix, D.: A micro-macro and parallel computational strategy for highly heterogeneous structures. International Journal for Numerical Methods in Engineering **52** (2001) 121–138

Computational Modelling of Multi-field Ionic Continuum Systems

John Michopoulos

Special Projects Group, Code 6303, Naval Research Laboratory, U.S.A.
john.michopoulos@nrl.navy.mil

Abstract. The present paper presents a methodology for modelling the behavior of continua under multi-field influence, along with the architecture of its computational implementation. It is being applied for the special case of modelling multi-field ionic conductive material systems. Its computational implementation addresses generation and solution of both the constitutive and the field evolution equations. An application of this methodology for the case of electric multi-hygrothermoelasticity generates a model for materials participating in artificial muscle applications. The corresponding system of nonlinear governing partial differential equations describing the state evolution of large deflection plates made from such materials is derived as an example. Finally, initial numerical solution examples of these electro-hygro-thermally generalized Von-Karman equations are presented.

1 Introduction

The present paper reports on the developmental startup of a computational and algorithmic infrastructure that attempts not to ignore any coupled field and/or transport effects nor it makes any geometric simplifications while it captures the behavioral modelling of associated continuum systems.

The computational implementation and application examples on ionic artificial muscle materials are part of a validation effort at the sub-component level of a more general data-driven environment for multiphysics applications (DDEMA) that has been preliminarily described elsewhere [1,2,3].

Recent progress on processing and development of various multi-field activated materials such as electro-active polymers (EAP) and ionic polymer conductive composites (IPCC) for artificial muscle applications has underlined the general need for a rigorous, complete and preferably automated modelling of their behavior from a continuum coupled multi-field perspective.

There are various generalized approaches for deriving multi-field theories [4, 5,6,7] in the 4D space-time domain. Furthermore, mass transport considerations were combined with continuum theories based on continuous thermodynamics [8,9]. In the continuum mechanics context, the governing partial differential equations (PDEs) can be produced when all constitutive equations have been eliminated through term rewriting of the conservation laws.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 621–628, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Arguably, the goal of developing a general methodology to utilize computational technologies for automating the process of generating and utilizing multi-field theories for demanding applications, has not been reached. The plethora of analytic approaches as well as the pluralism, the quick evolution, and the heterogeneity of the available software and hardware infrastructure was a significant set of reasons that is contributing to this failure.

The novelty of this paper is that it attempts to address these difficulties by presenting both an analytical activity workflow along with an architecture of a computational infrastructure to derive and solve the equational system representing the behavior of multi-field ionic systems.

The paper presents the abstract algorithmic context of the general modelling process in Section 2, from the perspective of continuous multi-field constitutive theories (CFTs). Section 3 presents the general activity workflow of this process along with a computational architecture of a computational infrastructure that implements this workflow. Application of this methodology produced the generalized Von-Karman partial differential equations (PDEs) presented in Section 4 along with a solution example. Conclusions provide the closure of the paper.

2 Multi-field Model Derivation

Every deformable continuum under multi-field generalized loading including reactive substance and charge diffusion can be considered as a system that can be described in space and time by the evolution of its state variables. Some of these variables can be thought of as dependent or output parameters and some as independent or input parameters. A systemic abstraction of such a medium is presented in Fig. (1a) and its corresponding continuum one in Fig. (1b). The systemic abstraction differentiates between the bulk material state behavior and the structural state behavior, while the continuum mechanics traditional approach never makes this differentiation. Constitutive behavior refers to the bulk material state behavior (in the sense of the representative volume element behavior), while systemic behavior is the composition of both the bulk material state behavior with the structural material state behavior.

The bulk behavior of such a system is usually described as a set of relational restrictions $\tilde{F}$ among the state variables selected by the modeler as the ones pertinent to the situation at hand, given by

$$\tilde{F}(\tilde{q}, \tilde{\xi}, \tilde{p}) = 0, \tag{1}$$

where $\tilde{p}, \tilde{\xi}$and $\tilde{q}$ represent the input state variables, the internal state variables and the output state variables respectively. For those cases that these relations can be solved with respect to the output variables they are called constitutive relations of the form

$$\tilde{q} = \tilde{C}(\tilde{\xi},\tilde{p}) \ . \tag{2}$$

Functionals $\tilde{C}$ in equation (2) represent an a-priori definable multi-functional mapping of the form $\mathrm{R}^{\dim(\tilde{p})} \times \mathrm{R}^{\dim(\tilde{\xi})} \xrightarrow{\tilde{C}} \mathrm{R}^{\dim(\tilde{q})}$. In most path-history independent state spaces these functionals can be recovered by differentiation of

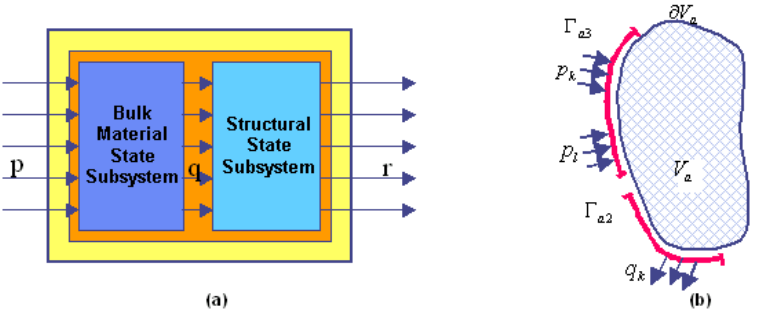


Fig. 1. Systemic (a) and continuum-based (b) descriptions of multi-field systems

an also a-priori definable potential function $\Xi(\tilde{\xi},\tilde{p})$, with respect of the input (independent) state variables. This potential function has to be constructed as a function of the input and the internal state variables and should it be necessary any time derivatives of them. This can be expressed by,

$$\tilde{q} = \nabla_{\tilde{p}}\ \Xi(\tilde{\xi},\tilde{p},\tilde{p}_{,t^n},\tilde{\xi}_{,t^n}) \tag{3}$$

This formalism imposes a conjugation between input and output state variables in a way that allows us to form "cause-effect" pairs $\{q_i, p_i\}$ that have the property that their product has unit dimensionality of energy density per unit of volume or mass. Various researchers have suggested various choices for the potential function required for the constitutive relations. The most practicable and well known among them are those of internal energy, enthalpy, Helmholtz free energy and Gibbs free energy.

In the context of continuous multiphysics, system behavior in terms of state evolution in most continuum systems, is expressed in terms of solutions of partial differential equations (PDEs) that enforce special topological form on the fields described by the spatial and time distributions of the state variables and further restrict the potential values these variables can take, i.e.

$$\aleph_i(\nabla^n\tilde{q},\tilde{q}_{,t^m},...) = 0 \tag{4}$$

The traditional sources of such equations are the so-called "conservation" or "balance" laws of physics. These are the thermomechanical laws of conservation of mass, momentum, moment of momentum, energy, entropy flaw and the electrodynamic laws of conservation of electric displacement (Gauss-Faraday law), magnetic flux, electric charge, rotation of electric intensity (Faraday's law), magnetic intensity (Ampere's law). These are the axioms of continuous physics (ACP) that are not (formally) provable (in their most general form) but rather they are beliefs that we a priori accept to be true. Unfortunately they are not always enough for completely determining the state variable field evolution. For this reason the constitutive functionals are used to provide algebraic closure. Introducing the constitutive equations (2) or (3) into (4) and eliminating the

independent variables or more generally half of the conjugate variables, yields a complete set of PDEs that constitute the so-called field equations of the system. They are of the form,

$$\wp_i(\tilde{p}, \tilde{p}_{,t^m}, ...) = 0 \tag{5}$$

The ACP have been historically expressed in either their global (integral),

$$(\int_V \phi \partial v)_{,t} = \int_{\partial V} \tau\{\phi\} \partial a + \int_V g\{\phi\} \partial v \tag{6}$$

or the local (differential) form of a generalized divergence theorem [11] transformed via the Gauss-Ostrogratsky theorem,

$$\phi_{,t} + \boldsymbol{\nabla} \cdot (\boldsymbol{v}\phi) = \boldsymbol{\nabla} \cdot \boldsymbol{\tau}\{\phi\} + g\{\phi\}, \tag{7}$$

where ϕ is the field in a volume V, $\tau\{\phi\}$ is the influx of ϕ through the surface ∂V, and $g\{\phi\}$is the amount of ϕ produced in the body (source term). Further simplifications on the form of the ACP can be obtained by applying some or all of the additional axiomatic and meta-axiomatic restrictions that are traditionally called the axioms of constitutive theory [7].

3 Computational Workflow and Infrastructure

The general process of simulating the behavior of continuum system usually involves the solution of the PDEs describing its space and time evolution via application of discrimination method over the domain of their applicability and the subsequent solution of a set of ordinary differential and eventually algebraic equations. However, as our ability to produce new material systems has outpaced our ability to model them, the need for automating the process of deriving these PDEs as well as the prerequisite sub-models and technologies has become apparent. In attempting to address this need we have developed a conceptual model of the workflow involved for developing a system of governing multi-field coupled PDEs. Figure 2 shows the activities and their interconnection along with the computational context they are embedded.

The essential features of this diagram capture the outline of the methodology described in the previous section. A crucial aim of the present effort is to extend the symbolic computing context as much to the left as possible and seamlessly integrate all contexts with each-other.

The current state of evolution of a continuously evolving computational infrastructure that implements the workflow of Fig. 2 is shown schematically in Fig. 3.

Lack of space does not allow detailed description of all modules and relation involved. However, the reader can easily follow the logic involved and recognize the programming languages involved. It is essential to underscore that J/Link is the Mathematica [10] to Java and Java to Mathematica application programming interface. MathTensor [11] and Theorema [12] are packages developed to run under Mathematica, addressing the tensor manipulation and theorem proving needs of deriving the constitutive and field equations of the model at hand.

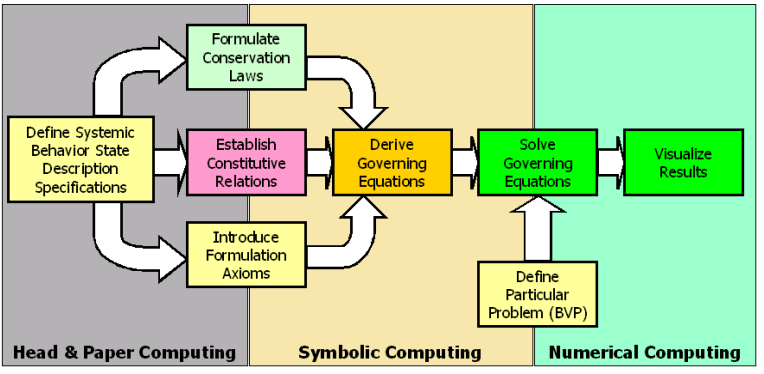


Fig. 2. Workflow diagram of activities and their interconnectivity along with their computational embedding for the process of developing multi-field models of continua

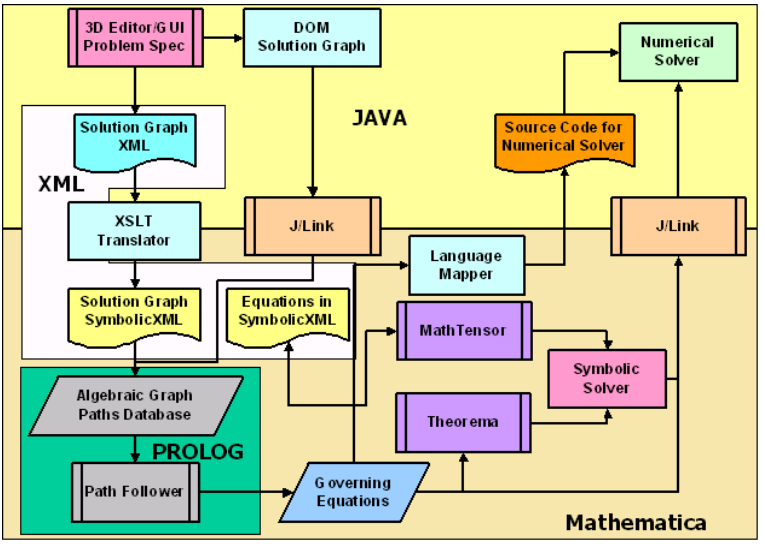


Fig. 3. Abstract architecture of computational infrastructure for implementing the workflow of Fig. 2

4 Multi-component Hygro-Thermo-Electro-Elastic Plates

Ionic material and structural systems that can be used as artificial muscles are falling within the category of multi-field systems. Limited electromechanical modelling of such systems has been already attempted for simple membrane and one dimensional systems and never considered the global continuous multi-field perspective of more than two simultaneously acting fields.

Application of the general process as described in the previous sections for the case of homogeneous mechanically isotropic system generated a set of nine

field governing coupled PDEs and associated constitutive relations that have been originally derived elsewhere [13,14] manually.

Introduction of Lagrange strains to account for large deformations along with insignificant ionic currents and chemical reactivity for multi-component diffusion, electric potential, temperature and mechanical deformation generalized loading yields the following set of modified Von-Karman nonlinear PDEs [16,17]:

$$\nabla^2\nabla^2 w + (1+\nu)\sum_k \nabla^2 X^k = \frac{h}{N}(\frac{q}{h} + F_{,22}w_{,11} - 2F_{,12}w_{,12} + F_{,11}w_{,22}), \quad (8)$$

$$\nabla^2\nabla^2 F + E\sum_k \nabla^2 X^k = E[(w_{,12})^2 - w_{,11}w_{,22}], \quad (9)$$

$$\frac{1-\nu}{2E}(\nabla^2\delta_{ij} - \partial_i\partial_j)F\delta_{ij} + \nabla^2 X^k = 0. \quad (10)$$

Here w, F, X^k are the deflection, the Airy stress function and the additional generalized field state variables and $\delta_{ij}, \nu, E, h, N, q$ are Kronecker's delta, Poisson's ratio, the modulus of elasticity, the plate thickness, the flexural rigidity and the distributed load respectively.

Equation (10) provides the closure of the system of equations (8) and (9) governs the balance of the generalized additional scalar fields X^k [13,14](and appears in its expected divergent theorem form).

Since there is no known closed form solution of the generalized Von-Karman equations an approximate solution based on Finite Element Analysis (FEA) can be utilized. The general purpose code *flexPDE* [15]was used for this task.

Although we solved various cases of boundary conditions due to the space limitations here, we will only present the case of a simply supported plate along all the edges with no lateral mechanical load. The value of $X^k = Temp$ at the boundary varies sinusoidally in time.

Figure 4 shows the distribution of deflection over the entire domain for increment 36 (4a) near the beginning of the cycle, and increment 121 (4b) on the end of the complete cycle. Clearly these two distributions verify the reversible actions of the applied field since the plate evolves from an all-negative to an all-positive deflection field.

Much is already known about how deflection and Airy function relate to each other from the traditional large deflection Von-Karman plate analysis [16, 17]. All other fields and fluxes can be trivially computed by exploitation of the constitutive and flux definition relations given elsewhere [13,14] and are not provided here due to lack of space. The intension of this section is not to detail the derivation and solution methodologies for the corresponding PDEs but rather to provide evidence that is possible.

5 Conclusions

In this effort, we have described an abstract framework for multi-field modelling of material systems with emphasis to ionic materials used for artificial muscle

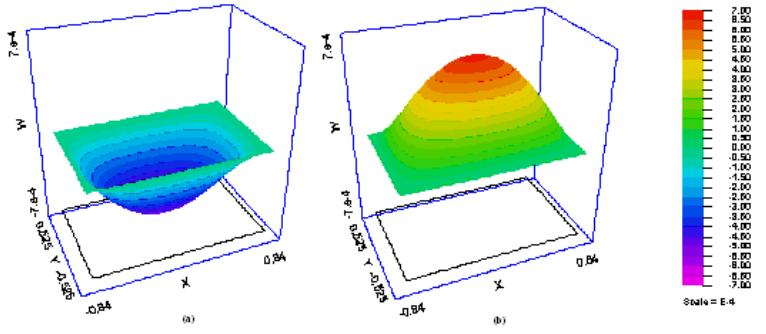


Fig. 4. Two instances of the deflection solution of a non-linear generalized Von-Karman system of PDEs governing a rectangular plate made from an ionic material

applications. Furthermore an abstract workflow of activities has been created to employ this framework. The current status of a computational infrastructure that attempts to instantiate this workflow is also briefly described. Application of this methodology and workflow has been utilized for deriving the generalized Von-Karman equations for multi-field activation of large deflection plates. Finally an example of numerically solving the derived system is presented.

The approach followed still leaves open the issue of determination of the required constants participating in this formulation. Subsequent work in addition to evolving the computational infrastructure, will focus on this task.

Acknowledgement. The author acknowledges the support by the National Science Foundation under grant ITR-0205663 and professor M. Shahinpoor from the University of New Mexico, for his donation of ionic polymer composite material samples.

References

1. Michopoulos, J., Tsompanopoulou, P., Houstis, E., Rice, J., Farhat, C., Lesoinne, M., Lechenault, F., DDEMA: A Data Driven Environment for Multiphysics Applications,in: Proceedings of International Conference of Computational Science - ICCS'03, Sloot, P.M.A., et al. (Eds.) Melbourne Australia, June 2-4, LNCS 2660, Part IV, Springer-Verlag, Haidelberg, (2003) 309-318.
2. Michopoulos, J., Tsompanopoulou, P., Houstis, E., Rice, J., Farhat, C., Lesoinne, M., Lechenault, F., Design Architecture of a Data Driven Environment for Multiphysics Applications, in: Proceedings of DETC'03, ASME DETC2003/CIE Chicago IL, Sept. 2-6 2003, Paper No DETC2003/CIE-48268, (2003).
3. Michopoulos, J., Tsompanopoulou, P., Houstis, E., Farhat, C., Lesoinne, M., Rice, J., Joshi, A., On a Data Driven Environment for Multiphysics Applications, Future Generation Computer Systems, in-print (2004).
4. Truesdell, C., Toupin, R., 1960, "The Classical Field Theories", in Handbuch der Physik (Herausgegeben con S. Flugge) **III/1**, Springer-Verlag, Berlin.

5. Truesdell, C., Noll, W., 1965, "The Non-Linear Field Theories of Mechanics", Handbuch der Physik (Herausgegeben con S. Flugge) Bd**. III/3**, Springer-Verlag, Berlin.
6. Green, A. E., Naghdi, P. M., 1995, "A unified procedure for construction of theories of deformable media. I. Classical continuum physics", Proc. Roy. Soc. London Ser. **A 448 /1934**, pp. 335—356.
7. Eringen, A.C., Maugin., G.A., 1990, "Electrodynamics of Continua I. Foundations and Solid Media." Springer-Verlag, New York.
8. Sih, G.C., Michopoulos, J.G., Chou S.C., 1986, "Hygrothermoelasticity", Martinus Nijhoff Publishers (now Kluwer Academic).
9. Michopoulos, J.G., Sih, G.C., 1984, "Coupled Theory of Temperature Moisture Deformation and Electromagnetic Fields", Institute of Fracture and Solid Mechanics report IFSM-84-123, Lehigh University.
10. Wolfram, S.,1999,"The Mathematica Book", 4th Edition, Wolfram Media, Cahmpaign and Cambridge University Press, Cambridge.
11. Parker, L., Christensen, S. M., 1994, "MathTensor: A System for Doing Tensor Analysis by Computer", Addison-Wesley, Reading.
12. Buchberger B. 1998, "Theorema: Computer-Supported Mathematical Proving", in proc. of Annual Meeting of the Mathematics Department of the University of Wales, Gregynod.
13. Michopoulos, J.G., Shahinpoor, M., 2002, "Towards a Multiphysics Formulation of Electroactive Large Deflection Plates Made from Ionic Polymeric Artificial Muscles", Proceedings of First World Congress on Biomimetics and Artificial Muscles, 9-11 December 2002, Albuquerque NM
14. Michopoulos, J., 2003, "Numerical Solution of the Multiphysics Generalized Von-Karman Equations for Large Deflection Ionic Polymer Plates", in: Proc. of DETC'03, ASME DETC2003/CIE Chicago IL, Sept. 2-6 2003, Paper No DETC2003/CIE-48207.
15. PDE Solutions Inc., "FlexPDE 3 Reference manual", PDE Solutions Inc., (2001).
16. vonKarman, T., 1910, "Festifkeitsprobleme im Maschinebau", Encyk. der Math. Wiss., pp. 349.
17. Levy, S., 1942, "Bending of Rectangular Plates with Large Deflections", NACA Technical Report No. 737.

Formation of Dwarf Galaxies in Reionized Universe with Heterogeneous Multi-computer System

Taisuke Boku[1,5], Hajime Susa[2], Kenji Onuma[3],
Masayuki Umemura[4,5], Mitsuhisa Sato[1,5], and Daisuke Takahashi[1,5]

[1] Institute of Information Sciences and Electronics, University of Tsukuba
[2] Department of Physics, Faculty of Science, Rikkyo University
[3] Doctoral Program of Systems and Information Engineering, Graduate School, University of Tsukuba
[4] Institute of Physics, University of Tsukuba
[5] Center for Computational Physics, University of Tsukuba

Abstract. HMCS (Heterogeneous Multi-Computer System) is a very powerful and ideal computational environment for large scale computational astrophysics simulations including multiple physical phenomena. In this system, general purpose and special purpose parallel processing systems are involved to realize very high performance computation. We have constructed a system with MPP and PC-cluster as general purpose side and GRAPE-6 gravity engine as special purpose side.
We perform 3D radiation smoothed-particle-hydrodynamics (RSPH) simulations on the formation and the photoevaporation of subgalactic objects ($M \sim 10^8 - 10^9 M_\odot$). We confirm the suppression of the formation of small galaxies after the reionization. We also find that the galaxies that undergo violent photoevaporation process, still retain certain amount of stars, which are formed at small scale high density peaks. These cooled components merge with each other when the dark matter halo of the whole system is formed. It is also found these low mass galaxies should have large mass-to-light ratio, and these systems could be the progenitor of dwarf spheroidal galaxies in Local Group.

1 Introduction

Galaxy formation is one of the important issues on computational astrophysics. In this field, we are especially interested into the formation of dwarf galaxies in reionized universe which requires very heavy simulations under multiple physical phenomena. The main part of the simulation consists of two elements: SPH (smoothed particle hydrodynamics) and N-body gravity calculation. We have developed a parallel code for the first half on both massively parallel processor and Linux PC clusters. The latter half is extremely heavy because it requires $O(N^2)$ computation. For this part, we introduced a special purpose processor GRAPE-6[1] which provides 1 TFLOPS peak performance with only a single board. We have developed a combined system with these two parts, named

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 629–636, 2004.
© Springer-Verlag Berlin Heidelberg 2004

HMCS (Heterogeneous Multi-Computer System)[2]. In this paper, we describe the simulation result of the formation of dwarf galaxies on HMCS as well as the brief introduction of HMCS itself.

2 Formation of Low Mass Galaxies and Ultraviolet Background Radiation Field

According to the standard theory of cosmology, first galaxies are small ($M \sim 10^6 - 10^8 M_\odot$[1]), and are formed when the universe is a hundred million years old. These first generation galaxies start to reionize the universe due to the ultraviolet radiation field emitted from the massive stars in them. These emitted ultraviolet photons built up the ultraviolet background radiation field. In fact, recent observations strongly suggest that the universe is highly ionized after the age of the universe ($t_{\rm H}$) is approximately 3×10^8 years [5]. On the other hand, these small galaxies are so tiny that they cannot afford to keep the ionized hot gas in their gravitational potential, because of the gas pressure. Thus, if the background ultraviolet photons penetrate and heat up the galaxy, the gas component escapes from the potential of the host galaxy. This mechanism is always called as *photoevaporation*. In order to evaluate the effects in realistic clumpy forming galaxies, we perform numerical simulations on the formation of small galaxies under the ultraviolet background radiation field. The code includes the effects of radiation transfer of ionizing photons, chemical reaction network, radiative cooling, star formation, gravity, dark matter particles and smoothed particle hydrodynamics (SPH). This type of simulation with radiative transfer on the formation of small galaxies have never been done before, because of the complexity and high cost of radiation transfer. We utilize the newly developed parallel processing platform Heterogeneous Multi-Computer System [2,7] to realize the present numerical simulation which includes various type of physical phenomena.

3 Heterogeneous Multi-computer System

HMCS (Heterogeneous Multi-Computer System) [2] is a paradigm combining heterogeneous parallel processing systems to solve multi-physical or multi-scale problems which cannot be solved ordinary single system architecture such as general purpose MPPs or clusters. In HMCS, basically, two or more high performance parallel processing systems are connected by wide-bandwidth parallel commodity network such as parallel link of Fast- or Gigabit-Ethernet. We have developed a prototype system of HMCS for astrophysics introducing special purpose gravity engine GRAPE-6[1]. GRAPE-6 is developed at University of Tokyo, and we made a cluster with eight boards of GRAPE-6 in a collaborative work with the originators of GRAPE-6[2].

Fig. 1 shows the conceptual overview of our HMCS prototype. As general purpose machines (GPMs hereafter), we are using CP-PACS[4] MPP system

[1] $M_\odot$ denotes the mass of Sun.

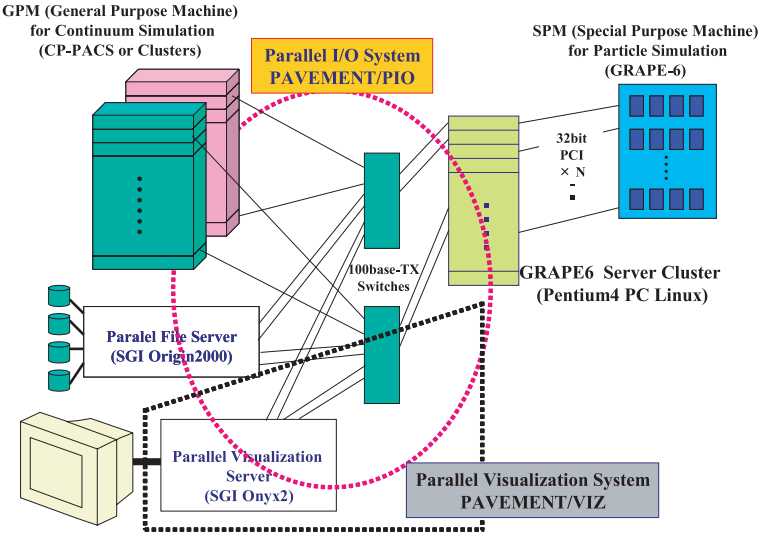


Fig. 1. HMCS overview

with 2048 processors (peak performance = 614 GFLOPS) or commodity-base Pentium Xeon PC-cluster with 37 nodes (74 processors, peak performance = 414 GFLOPS). Eight boards of GRAPE-6 (peak performance = 8TFLOPS) are connected to a small PC-cluster with Pentium4 operated by Linux. Each node of this PC-cluster is connected to a GRAPE-6 board via 32-bit PCI. A parallelized management program written in MPI is provided to control multiple GRAPE-6 boards performing a large scale particle calculation in parallel.

We have designed the system to allow multiple GPMs share a cluster of GRAPE-6 in time-sharing manner. Thus, this GRAPE-6 cluster, named HMCS Server Cluster, works as a server system to provide gravity calculation service, and all GPMs work as client machines. GPMs communicate with HMCS Server Cluster via serial or parallel Ethernet links according to the load of calculation, i.e. the number of particles to be processed. This network with parallel Ethernet links is controlled by user-level middleware named PIO[3] which provides high bandwidth communication with trunk of parallel Ethernet links on TCP/IP level.

Each GRAPE-6 board consists of 32 ASICs for N-body calculation and provides 1 TFLOPS of peak performance on gravity calculation. HMCS Server Cluster operates with MPI-based management program for 1) parallel data exchanging with GPMs through PIO, 2) time-sharing controlling of multiple requests of gravity calculation by multiple GPMs and 3) coordinating parallel gravity calculation on all GRAPE-6 boards. The server program is well designed to maximize the utilization ratio of GRAPE-6 regardless the communication speed between any of GPMs and the server[8].

In HMCS prototype, we can distribute multiple series of simulations with various initial conditions to multiple GPMs surrounding HMCS Server Cluster.

For relatively small scale problems where the number of particles is less than 50,000, one or two GRAPE-6 boards are involved to minimize the overhead for parallel processing among multiple boards. For larger problems, 4 to 8 GRAPE-6 boards are involved. Currently, this distribution control is performed manually with system configuration setting, however we are now developing the automatic load distribution and balancing system to optimize the utilization ratio of eight GRAPE-6 boards.

4 Algorithm and Execution on HMCS

HMCS is an ideal platform for computational astrophysics problems which require both continuum and particle simulations in the target system. In the simulation of dwarf galaxies, there are multiple physical phenomena to be simulated such as hydrodynamics, chemical process and gravity and so on. We simulate all these elements except gravity calculation on GPM while HMCS Server Cluster performs gravity calculation with GRAPE-6.

The basic algorithm of 3D RSPH with gravity on HMCS is as follows.

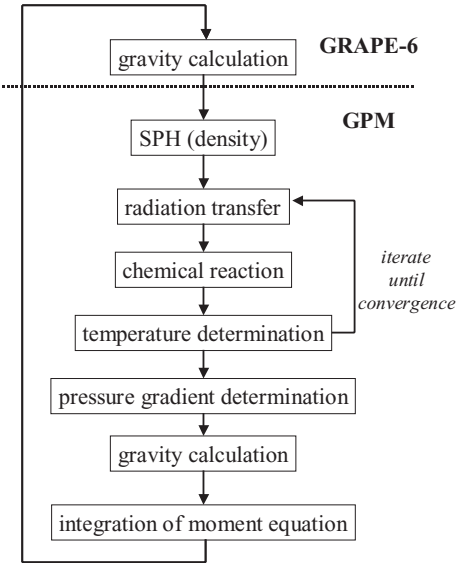


Fig. 2. Basic algorithm and workload of 3D RSPH with gravity

1. Initialize all physical conditions of particles such as mass, location and velocity, and the field such as density, temperature, the amount of chemical seeds, etc.
2. Transfer the particle data from GPM to GRAPE-6, and evaluate the self gravity.
3. Calculate the density, pressure and gradient of pressure with SPH.

4. Determine the time step from the velocity and force.
5. Vary the size of SPH particles.
6. Iterate the calculation of radiation, chemical reaction and energy equation of SPH until they converge, with optional adjustment of time step if necessary.
7. Integral the movement of particles according to SPH dynamic equation.
8. Iterate from Step 2 to 7 for required time length.

In each time step, HMCS Server Cluster (GRAPE-6) and GPM communicate to exchange particle data and acceleration. In this algorithm, most of computation time is spent by GPM. Even if the order of computation for gravity calculation is $O(N^2)$ for N particles, RSPH calculation part is much heavier than gravity part for $N \sim 50,000$. In such cases, GRAPE-6 is mostly idle, therefore we can share HMCS Server Cluster with multiple GPMs in time-sharing manner to perform multiple cases of simulation simultaneously.

5 Simulations and Results

We have performed several runs with two different total masses and various formation epochs of galaxies. The detailed of applied method and algorithm are shown in [7]. In this paper, results from four runs with parameters listed in Table 1 are shown. Projected positions of gas particles, star particles and dark matter particles in model A1 are plotted in Fig. 3. The left three panels represent the epoch prior to the reionization. At this epoch, gas particles are cold ($T < 10^4$K) and almost trace the distribution of dark matter particles. Stellar particles are not formed yet. After the reionization (middle three panels), the low density regions are heated up to $T \lesssim 10^4$K, while the high density peaks are self-shielded to the ultraviolet background radiation field. In the self-shielded dense clumps, stars are formed from cooled clouds. Finally (right three panels), the gas components are blown away due to the photoevaporation. The clumps of the star particles merge with each other, and form a spheroidal system. On the other hand, such as in case B2, the gas is not blown away, and most of the gas and stars settle onto the gravitational potential of dark matter particles, because the gravitational force is strong enough to keep the hot ionized gas with $T \simeq 10^4$K.

Table 1. Model parameters

Model ID	Formation epoch z_c	Total mass M_{tot}
A1	1.5	$2 \times 10^8 M_\odot$
A2	1.5	$2 \times 10^9 M_\odot$
B1	6	$2 \times 10^8 M_\odot$
B2	6	$2 \times 10^9 M_\odot$

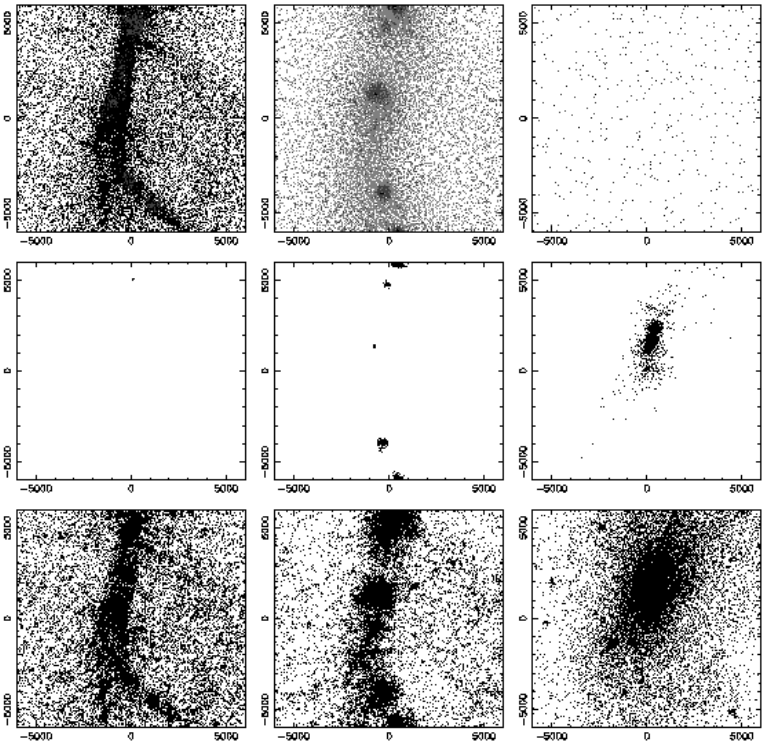


Fig. 3. Projected positions of SPH particles (upper panels), numerically formed star particles (middle) and dark matter particles (bottom panels) are plotted at three different epochs (left:$z = 8.958$, middle:$z = 5.755$ and right:$z = 1.361$) for model A1. Remark that larger redshift represents the earlier epoch. The color of the gas particles represents the gas temperature (red:$T \geq 10^4$K, green:$10^3\text{K} \leq T \leq 10^4$K, blue:$T \leq 10^3$K). The box size is 6 kpc × 6 kpc

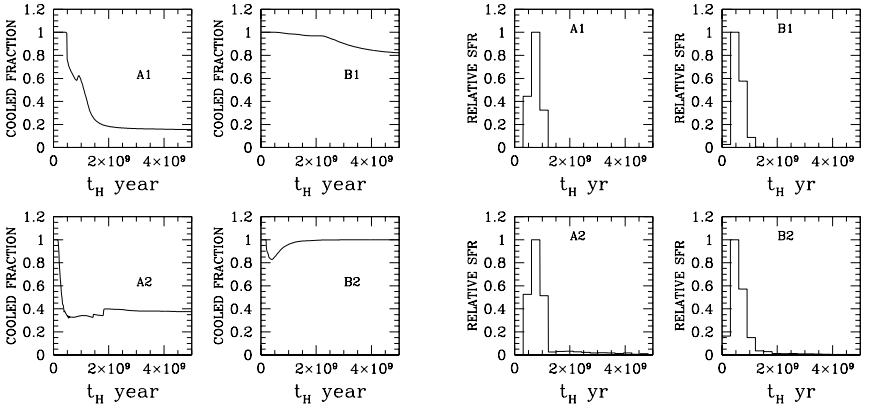
In Fig. 4(a), time evolution of the fraction of stars and gas particles within 5 kpc[2] are plotted for four runs. For models A1 and A2, which correspond the models with later formation epoch (i.e. low z_c), significant amount of gas are lost from galaxy. On the other hand, for models B1 and B2 (earlier formation epoch), most of the gas and stellar components do not escape from the galaxy. Gas particles are converted to star particles more efficiently than the previous case, because the gas clumps are formed prior to the reionization epoch (i.e. when the intensity of UV radiation field is weak), and they are easily self-shielded. Once the gas are converted to star particles, thermal pressure of the ionized gas does not disrupt the system.

These results also infer the large mass-to-light ratio (ratio of luminosity and total mass) for models A1 and A2, because gas are lost by photoevaporation while the dark matter particles are not affected by radiation. The observation of

[2] kpc $= 3.08 \times 10^{21}$ cm

Local Group dwarf spheroidal galaxies[3] tells that those faint galaxies have very large mass-to-light ratio [6], which is similar to our results in models A1 and A2.

In Figure 4(b) relative star formation histories are plotted. We find a clear feature for all models: star formation rates[4] sharply drop at $t_{\rm H} = 1-2$Gyr, which is the direct consequence of the photoevaporation of the gas. It is interesting to point out that this star formation history is again similar to that of dwarf galaxies in Local Group [6].



(a) Cooled fraction (b) Star formation history

Fig. 4. (a) Time evolution of the fraction of cooled baryonic components are plotted for four runs. Horizontal axes denote the cosmological time, and vertical axes represent the fractions of the baryonic components (i.e. SPH particles and star particles) within 5kpc from the center of gravity. Left two panels represent the cases with $z_{\rm c} = 1.5$ (Model A1 and A2). Right two panels denote the cases with $z_{\rm c} = 6$ (Model B1 and B2). (b) Star formation history of four runs are plotted. Vertical axis denotes the star formation rate (mass of the formed stars per unit time) normalized by the peak value. Left two panels represent the cases with $z_{\rm c} = 1.5$ (Model A1 and A2). Right two panels denote the cases with $z_{\rm c} = 6$ (Model B1 and B2)

6 Conclusion

We performed 3D radiation hydrodynamical simulations on the formation of low mass objects with four set of parameters on HMCS prototype system with general purpose PC-clusters and special purpose cluster with GRAPE-6 gravity engine. Through these product runs, we confirmed that HMCS is an ideal and powerful tool for real applications of computational astrophysics. The suppression of the formation of low mass objects at later epoch is confirmed. It is also

[3] Small galaxies in our neighbourhood.
[4] Mass of the formed stars per unit time.

found that the low mass galaxies formed at low redshift should have very large mass-to-light ratio and characteristic star formation history. Observational counter part of these systems might be the spheroidal components of Local Group dwarf galaxies.

Acknowledgment. The authors truly thank Prof. Jun'ichiro Makino at University of Tokyo for his valuable suggestions and comments to our work. We also thank to members of Center for Computational Physics, University of Tsukuba for their support.

References

1. Makino, J., et.al., "A 1.349 Tflops simulation of black hole in a galactic center on GRAPE-6", Proc. of SC2000 (CD-ROM), IEEE, Los Alamitos, 2000.
2. Boku, T., Makino, J., Susa, H., Umemura, M., Fukushige, T., Ukawa, A., "Heterogeneous Multi-Computer System: A New Platform for Multi-Paradigm Scientific Simulation", Proc. of ICS'02, pp.26–34, New York, Jul. 2002.
3. Boku, T., Matsubara, M., Itakura, K., "PIO: Parallel I/O System for Massively Parallel Processors", Proc. of European High Performance Computing and Network Conference 2001 (LNCS-2110), pp.383-392, Amsterdam, Jun. 2001.
4. Boku, T., Itakura, K., Nakamura, H., Nakazawa, K., "CP-PACS: A massively parallel processor for large scale scientific calculations", Proc. of ACM ICS'97, pp.108-115, Vienna, Jul. 1997.
5. Kogut, A. et al., "First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Temperature-Polarization Correlation", Astrophysical Journal Supplement, Volume 148, 161, 2003.
6. Mateo, M., "Dwarf Galaxies of the Local Group", Annual Review of Astronomy & Astrophysics, Volume 36, 435, 1998.
7. Susa, H., Umemura, M., "Formation of Dwarf Galaxies during the Cosmic Reionization", Astrophysical Journal, Volume 600, Issue 1, pp. 1–16, 2004.
8. Onuma, K., Boku, T., Sato, M., Takahashi, D., Susa, H., Umemura, M., "Heterogeneous Remote Computing System for Computational Astrophysics with OmniRPC", Proc. of Workshop on High Performance Grid Computing and Networking, Proc. of SAINT2004, pp.623–629, Tokyo, Jan. 2004.

A Multi-scale Numerical Study of the Flow, Heat, and Mass Transfer in Protective Clothing

Michal P. Sobera[1,2], Chris R. Kleijn[1], Paul Brasser[2], and Harry E.A. Van den Akker[1]

[1] Kramers Laboratorium voor Fysische Technologie, Delft University of Technology, Prins Bernhardlaan 6, 2628BW, Delft, The Netherlands
[2] TNO Prins Maurits Laboratory, P.O. Box 45, 2280AA, Rijswijk, The Netherlands

Abstract. A multi-scale study of the performance of protective clothing has been performed by coupling various types of numerical simulation of flow, heat and mass transfer. From meso-scale Direct Numerical Simulation, it was found that the flow underneath the clothing is laminar and periodic, with a magnitude much smaller than the free stream velocity. Micro-scale Direct Numerical Simulation revealed a simple relation between textile porosity and permeability. A good agreement was found between flow and heat transfer predictions of Direct Numerical Simulation and Reynolds Averaged simulation. From the latter, an engineering correlation for heat and mass transfer was deduced.

1 Introduction

The demands on NBC (Nuclear-Biological-Chemical) protective clothing are strict and mostly contradictory, requiring the combination of optimal protection and optimal comfort. This makes the design and development of effective NBC protective clothing complex. It is believed, that computational models of flow, heat and mass transfer can simplify this process. Flow, heat and mass transfer in NBC protective clothing depend on phenomena at very different scales, ranging from the sub-millimeter scales of textile fibers and turbulent flow structures, via the centimeter to decimeter scale of the flow around body parts, to the meter scale of a full clothed person. Therefore, a coupled multi-scale modeling approach is necessary. In this paper we describe the development of such an approach. Using Computational Fluid Dynamics (CFD), we study NBC protective clothing on different scales, focusing on different aspects of flow, heat and mass transfer. Information from small scales is used to extract lumped models for larger scales. Large-scale simulations are used to determine the conditions at which small scale simulations are performed.

2 Multi Scale Modeling Strategy

We aim to develop predictive models for the performance of protective clothing at full body scale. At this (macro-) scale it is not feasible to account for phenomena at the scale of the textile structure, nor is it possible to account for

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 637–644, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the detailed properties of the turbulent flow. Therefore, models are based on large-scale, time-averaged Computational Fluid Dynamics descriptions of the flow, using the so called Reynolds Averaged Navier Stokes (RANS) approach. The textile material is modeled as a porous material with particular macroscopic properties. This approach was used in our previous study [1], were we investigated the influence of flow and clothing properties on heat and mass transfer. Two aspects of this kind of simulation require more detailed studies at smaller scales: Firstly, the performance of RANS models in laminarizing flow through a porous material is unknown. Therefore, in this paper we extend and compare our previous RANS study with DNS at meso-scale. We perform a detailed study of the flow around a single clothed limb modeled as a circular cylinder sheathed by a porous layer (see Fig. 1) using DNS. From these simulations we evaluate the performance of RANS turbulence models for the flows of interest. Secondly, engineering models for the lumped porous properties of the clothing material can be deduced from micro-scale DNS analyzes of the flow through the actual geometry of the textile structure. Input from both the meso-scale DNS and the micro-scale textile studies are used to improve and validate RANS simulations.

2.1 Meso-scale DNS

The flow around a cylinder at subcritical Re has been studied by several authors [2,3,4,5,6,7], both numerically and experimentally. In the range of Reynolds number $\mathrm{Re} = 10^3$ to 10^7 (based on cylinder diameter and free stream velocity), the flow is periodic and transitional in character. In the present work, the focus was on a subcritical flow regime ($10^3 \leq \mathrm{Re} \leq 10^5$, corresponding to air velocities of 0.1-10m/s around a typically sized limb) [2], in which, due to the vortex shedding the flow is strongly unsteady. The boundary layer remains fully laminar up to the separation point and transition to turbulence occurs in the wake.

The flow around a cylinder sheathed by a second, larger, porous cylinder (see Fig. 1), has received little attention. The general characteristics of the flow around a such an obstacle are expected to be similar to those of the flow around a single solid cylinder. However, not much is known about the interaction between the unsteady, turbulent flow outside the porous cylinder and flow inside the gap between the porous and the solid cylinder. For our DNS we used the commercial CFD solver Fluent 6, which is based on an unstructured finite volume formulation. Unstructured solvers have relatively high memory requirements (in our case $\approx$ 1 GB of RAM per 1M of computational cells). In order to reduce the number of cells, local grid refinement was applied, with a fine mesh in the vicinity of the cylinder and in the near wake region. The grid spacing in this region was approximately $0.008D$ similar to that reported by Tremblay [7], who studied the flow around a cylinder by means of DNS with a structured CFD code. In the far field we used a slightly lower grid resolution. As a result, the total number of cells in our simulations was $\sim 6 \cdot 10^6$, as compared to $\sim 40 \cdot 10^6$ in the simulations of Tremblay.

The porous material was treated as a fluid zone, where the pressure drop was imposed as a sink in the momentum equation according to Darcy's law. For its

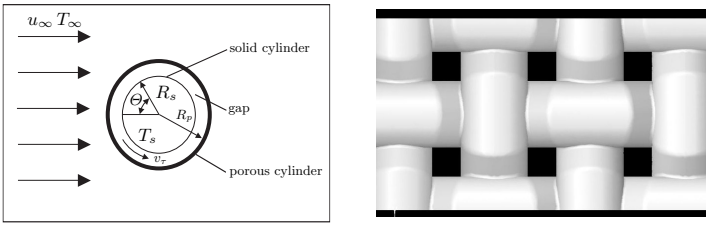


Fig. 1. Schematic problem description for meso-scale approach (left) and computational representation of the net of fibers (right)

resistance, values were set that were obtained experimentally from a sample of real protective clothing. The cylinder length was πD, with periodic boundaries at the end walls. Thus, in fact we simulated an infinitely long cylinder. The simulation was run with a constant time step $\Delta t = 0.001\frac{D}{u_\infty}$ and was carried out for 9 vortex shedding cycles after the flow had developed, which was 33000 time steps. Before applying our computational DNS approach to a cylinder covered by a porous layer, we have first tested it extensively for the well studied case of flow around a solid cylinder at subcritical Re. For this case, our results (e.g. time averaged velocity profiles and turbulence spectra) were in very good agreement with results published by Ma et al. [6] and Tremblay [7].

2.2 Micro-scale DNS

For the micro-scale simulations of the flow around the textile fibers we again used Fluent. We created a so-called virtual textile, schematically presented in the right-hand part of Fig. 1. Since NBC clothing consists of woven textile, the geometric structure is ordered with only minor irregularities. The shape, diameter and distribution of the textile fibers were based on average values observed in a microscopic image of an actual NBC protective textile. A 3-dimensional flow domain was simulated, with the virtual textile plane halfway and parallel to the inflow and outflow boundary of the domain. The inflow boundary conditions were obtained from the meso-scale DNS simulations, from which it was found that the flow has laminarized in the vicinity of the textile. Periodic boundary conditions were applied at the edges of the virtual textile. The total size of the computational domain was about 200k cells. Due to the low fiber diameter and low velocity ($10^{-2} \leq \mathrm{Re} \leq 10^{1}$, with Re based on the fiber diameter), the flow remains laminar and steady. A comprehensive study of the influence of different fiber volume fractions and different inflow conditions on the pressure drop and permeability has been performed. As outcome of these studies we obtained realistic values for the range of the textile resistances and the range of the superficial velocities through the porous layer. These values have subsequently been used as input for the meso-scale DNS and meso-scale RANS simulations.

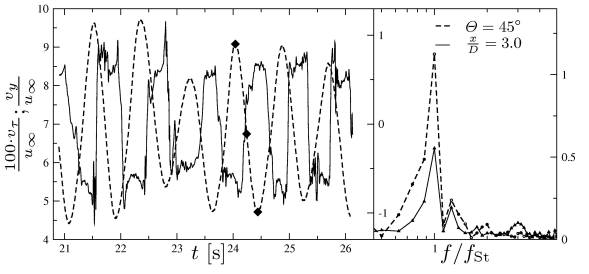


Fig. 2. Dimensionless velocity v_y in the wake at $x/D = 3$ and tangential velocity v_τ inside the air gap at $\Theta = 45°$, both in time (left) and Fourier (right) space

2.3 Meso-scale RANS

Two-dimensional engineering simulations of flow, heat and mass transfer at meso-scale were performed earlier in a comprehensive parameter study, details of which have been reported in [1]. Based on this parameter study, a correlation has been formulated, predicting heat and mass flux through the protective clothing as a function of clothing properties and flow conditions. Here we compare these earlier results to results from present DNS.

3 Results and Discussion

3.1 Meso-scale DNS

As expected, the flow inside the air gap (i.e. in the space between the outer porous cylinder and the inner solid cylinder) is transient. Its dynamics is due to the vortex shedding in the wake of the outer flow. In Fig. 2 the velocity inside the air gap is compared to the velocity in the wake. The inner flow velocity is orders of magnitude smaller than the outer flow, but exhibits the same periodic behavior, as can be seen from the figure's left-hand part. It can also be seen that only the low shedding frequency is pronouncedly present in the inner flow, whereas higher frequencies are filtered out. The right-hand part of the graph confirms that the frequency of the flow inside the air gap is locked to the frequency of vortex shedding in the outer flow.

Fig. 3 (left) shows the tangential velocity component along the air gap centerline. The velocity is presented as time averaged quantity, together with three instantaneous time realizations. The latter were taken at instances corresponding to the maximal, intermediate and minimal value of the velocity at $\Theta = 45°$, respectively. The inset figure is a simplified representation of the periodic inner flow behavior from Fig. 2, with the three time realizations indicated by black dots. The velocity distribution in the air gap at different time instances confirms its periodic nature. The time averaged velocity profile is in good agreement with our experimental data obtained by Laser Doppler Anemometry [1].

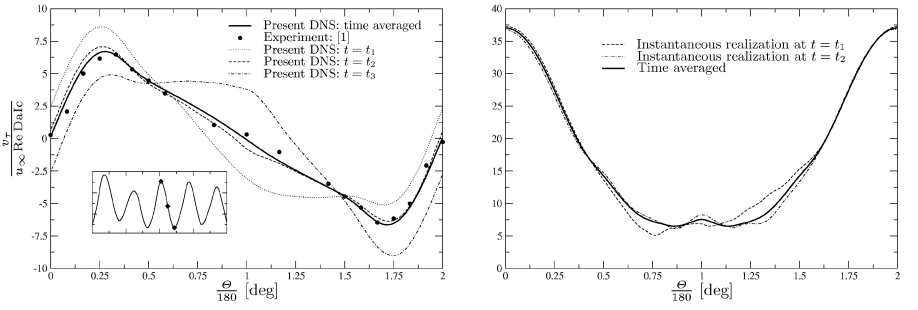


Fig. 3. Dimensionless, tangential velocity component at the air gap centerline (left) and local Nu number (right) both as a function of angle

Fig. 3 (right-hand part) presents the Nusselt number, which is a dimensionless measure of heat transfer from the solid cylinder surface to the air, as a function of angle. The time averaged quantity is compared to the instantaneous ones gathered at two different time instances. The Nusselt number distribution shows that the highest heat flux to the solid cylinder occurs in the vicinity of the front stagnation point, as expected. The heat transfer is strongly correlated to the radial velocity component, with high heat transfer at locations were there is a high inward air velocity.

3.2 Micro-scale DNS

For the virtual textile, simulations for varied $\mathrm{Re} = \frac{\varrho \langle v \rangle 2 r_\mathrm{f}}{\mu}$ were conducted, where r_f is fiber radius. Note that here $\langle v \rangle$ is the velocity in the vicinity of the textile, which is orders of magnitude smaller than the free stream velocity. Furthermore, at fixed Re, simulations were conducted for a range of textile porosities ϵ, defined as the ratio of the open frontal area to the total frontal area of the textile. When the textile is considered to be a 2-D square array of cylinders of radius r_f, at a mutual half distance δ, then ϵ is given by

$$\epsilon = \frac{\delta^2}{(r_\mathrm{f} + \delta)^2} \qquad \text{or} \qquad \frac{\delta}{r_\mathrm{f}} = \frac{\sqrt{\epsilon}}{1 - \sqrt{\epsilon}} \tag{1}$$

In the left-hand part of Fig. 4, the pressure gradient ∇P through one specific textile as a function of Re is presented. Numerical results are compared to the unpublished experimental data obtained at the Prins Maurits Laboratory of the Netherlands Organization for Applied Scientific Research (TNO). It is clearly seen that for two decades of Re the pressure gradient through the textile depends linearly on Re, in agreement with Darcy's law, which assumes that the pressure drop is due to viscous effects only. Inertia effects become important for larger values of Re, and a super-linear dependence of the pressure gradient on Re is observed.

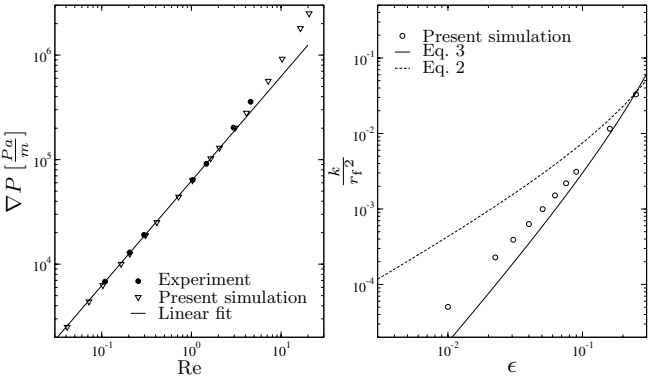


Fig. 4. Computed pressure gradient versus Re (left) and dimensionless permeability as a function of textile porosity (right)

The right-hand part of Fig. 4 presents the hydraulic permeability k, which is a measure of the fluid conductance in a porous layer, as a function of porosity. It is related to the pressure gradient by Darcy's law $\nabla P = \mu \frac{\langle v \rangle}{k}$. Based on a simple analysis combining Stokes flow and mass continuity, Clague et al. [8] proposed the following scaling estimate for the hydraulic permeability of an infinite square array of cylinders:

$$k = const \cdot \delta^2 \qquad \text{or} \qquad \frac{k}{{r_\mathrm{f}}^2} = const \cdot \frac{\epsilon}{\left(1 - \sqrt{\epsilon}\right)^2} \tag{2}$$

This result was obtained using $\langle v \rangle$ as the characteristic scale for the velocity through the array. This is reasonable for large porosities, but for low porosities it is more appropriate to use $\frac{\langle v \rangle}{\epsilon}$ as velocity scale. This leads to:

$$k = const \cdot \delta^2 \epsilon \qquad \text{or} \qquad \frac{k}{{r_\mathrm{f}}^2} = const \cdot \frac{\epsilon^2}{\left(1 - \sqrt{\epsilon}\right)^2} \tag{3}$$

By fitting the constant in Eq. 3, we found good agreement to our numerical data, as can be seen in the right-hand side of Fig. 4. The small deviation for $\epsilon \to 0$ is to be expected, since friction is becoming important here. It is clear that Eq. 3, which resembles to the Kozeny-Carmen equation for flow through a packed bed of particles, represents a better scaling estimate than Eq. 2.

3.3 Meso-scale RANS

Ultimately, we wish to develop an engineering model of NBC protective clothing at the full body macro scale [9], based on lumped parameter descriptions of the textile and on RANS modeling of the air flow. Earlier [1] we performed such simulations at the meso-scale (see Fig. 1), focusing on the flow in the air gap

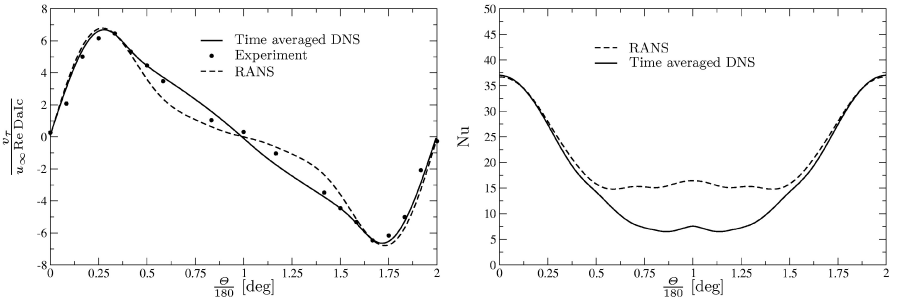


Fig. 5. Tangential velocity component at the air gap centerline (left) and local Nu distribution (right)

and on heat and mass transfer from the flowing air to the inner cylinder surface. In the present paper we validate these results against novel meso-scale DNS data. As shown in [1], for the studied cases there is a one-to-one correspondence between the Nusselt number for heat transfer and the Sherwood number Sh for mass transfer. Because of this similarity, only results for heat transfer will be presented here.

In the left hand side of Fig. 5 the dimensionless tangential velocity component in the air gap is presented as a function of angle. The velocity distribution obtained from RANS is compared to velocity profile obtained from DNS after time averaging (shown earlier in Fig. 3) and to our experimental LDA data. The overall performance of the RANS model is satisfactory. The maximum velocity is well predicted, a slight under prediction occurs for $90° < \Theta < 270°$.

The right-hand side of Fig. 5 presents a comparison between time averaged Nu numbers obtained from DNS and RANS computations. In the vicinity of the front stagnation point ($-90° < \Theta < 90°$), where heat transfer is largest and most critical, the agreement is very good. In the downstream region, $90° < \Theta < 270°$, where the flow in the gap is unsteady due to the vortex shedding in the outer flow, the steady-state simplification of RANS leads to significant over-predictions of the heat transfer. Nevertheless, averaging the Nu number over all angles, gives a satisfactory agreement between DNS ($\langle \mathrm{Nu} \rangle = 18$) and RANS ($\langle \mathrm{Nu} \rangle = 20$) results.

From the above comparisons, we conclude that the RANS model performs sufficiently accurate for engineering purposes, particularly when we focus on global (non-local) analyses of heat and mass transfer. We used this model to study the influence of the free-stream air velocity, the hydraulic permeability and thickness of the porous layer, and the distance between the inner and outer cylinder on the average Nu number at the inner cylinder surface. Simulations were carried out for typical values of the porous layer permeability, expressed in dimensionless form as $\mathrm{DaIc} = \frac{k}{4R_p \cdot r_f}$, and the dimensionless air gap thickness Ig. Based on the results from 2-D RANS simulations, an empirical correlation, that can be used as an engineering tool, was proposed and reported in [1].

4 Conclusions

The large differences of scale between the phenomena of interest necessitate a multi-scale approach for modeling the performance of NBC protective clothing. In our attempts to ultimately develop a computational model for the design and evaluation of NBC protective clothing at the full body macro scale, we performed various studies of flow, heat and mass transfer at micro and meso scales. Direct Numerical Simulations at meso-scale of the turbulent flow around a sheathed cylinder (mimicking a clothed limb) showed that, for the free flow velocities of interest, the flow underneath the clothing is laminar and periodic, with a frequency locked to that of the vortex shedding. For the range of clothing permeabilities of interest, the velocity of the flow through the clothing was found to be orders of magnitude smaller than the free stream velocity. Direct Numerical Simulations at micro-scale of the laminar flow around the textile fibers showed that, for the range of air velocities and textile porosities of interest, the pressure drop over the textile can be accurately described by Darcy's law. A simple scaling rule was proposed for the hydraulic permeability as a function of porosity and fiber thickness. The above findings from micro and meso scale DNS validate the simplifications that were made in a RANS type engineering model. From a comparison between meso scale DNS and RANS simulations, it is concluded that the RANS models has shortcomings in accurately predicting local Nusselt and Sherwood numbers, but performs satisfactorily accurate in predicting global heat and mass transfer. The model is now ready to be used for macro-scale full body studies.

References

1. Sobera, M., Kleijn, C., Brasser, P., van den Akker, H.: Convective heat and mass transfer to a cylinder sheathed by a porous layer. AIChE Journal **49** (2003)
2. Zdravkovich, M.: Flow around circular cylinders. Oxford University Press, (1997)
3. Norberg, C.: Efects of Reynolds number and low-intensity freestream turbulence on the fow around a circular cylinder. TR 87/2, Department of Applied Thermodynamics and Fluid Mechanics, Chalmer University of Technology, Sweden (1987)
4. Ong, L., Wallace, J.: The velocity field of the turbulent very near wake of a circular cylinder. Exp. Fluids **20** (1996) 441–453
5. Breuer, M.: Large eddy simulation of the subcritical flow past a circular cylinder: numerical and modeling aspects. Int. J. Numer. Meth. Fluids **28** (1998) 1281–1302
6. Ma, X., Karmanos, G.S., Karniadakis, G.: Dynamics and low-dimensionality of a turbulent near wake. J. Fluid Mech. **410** (2000) 29–65
7. Tremblay, F.: Direct and large-eddy simulation of flow around a circular cylinder at subcritical Reynolds numbers. PhD thesis, Munich University of Technology, Germany (2001)
8. Clague, D., Kandhai, B., Zhang, R., Sloot, P.: Hydraulic permeability of (un)bounded media using lattice boltzman method. Physical Review E **61** (2000) 616–625
9. Barry, J., Hill, R., Brasser, P., Sobera, M., Kleijn, C., Gibson, P.: Computational fluid dynamics modeling of fabric systems for intelligent garment design. MRS Bulletin **28** (2003) 568–573

Thermomechanical Waves in SMA Patches under Small Mechanical Loadings

Linxiang Wang and Roderick V.N. Melnik

MCI, Faculty of Science and Engineering,
University of Southern Denmark,
Sonderborg, DK-6400, Denmark

Abstract. 2D thermo-mechanical waves in SMA (Shape Memory Alloys) patches are simulated with a model derived for a special case of material transformations. The mathematical model includes the coupling effect between thermal and mechanical fields. It is shown that the classical 1D Falk dynamical model of SMA is a special case of the formulated 2D model. The differential algebraic approach is adopted for the numerical analysis. Computational experiments are carried out with small distributed mechanical loadings to analyze thermo-mechanical waves and coupling effects. Numerical results from 2D structures are compared with its 1D analog which is already been verified.

Keywords: Nonlinear waves, thermo-mechanics coupling, differential-algebraic solver.

1 Introduction

A better understanding of the dynamics of SMA is an important task in many areas of applications. However, even for the one dimensional case, the analysis of this dynamics is quite involved due to martensitic phase transformation and a strongly nonlinear pattern of interaction between mechanical and thermal fields ([13,2] and references therein). It is not trivial either to analyze numerically the thermo-mechanical interaction even when the phase transformation is not included, because the strong nonlinearity and thermo-mechanical coupling are still there. For a number of practical applications, the understanding of the dynamics of SMA structures with dimensions higher than one is required, which makes the nonlinear thermo-mechanical waves in higher dimensions essential for the applications.

Many instructive investigations have been carried out to understand dynamics during the process of martensitic phase transition. They provided a firm background for the application development, in particular inthe one-dimensional cases where the model for shape memory alloysis usually based on the Landau-Ginzburg free energy function ([12,3,5] and references therein). Although various approximations to the free energy function have been proposed in both one dimensional and three dimensional cases ([5,16] and references therein), results of two or three-dimensional cases are rarely available in the literature in the context

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 645–652, 2004.
© Springer-Verlag Berlin Heidelberg 2004

of modelling the thermo-mechanical waves and phase transformations in SMA. Recently,for the simulation of 2D microstructures in ferro-elastic materials, several free energy functions were proposed [6,7]. Although the authors proposed an interesting approach, no thermo-mechanical coupling was discussed and only static simulation of microstructures with a fixed temperature were presented in [6,7].

In this paper, we propose a two-dimensional dynamical model for modelling the thermo-mechanical coupling behavior. The model has been originally formulated to describe square-to-rectangular phase transformations in materials with shape memory effects. However, the phase transformations will not be discussed in this contribution. Our major emphasis here will be given to the first step in validation of the model: we will investigate the thermomechanical behavior of 2D SMA structures in the case of small mechanical loadings (not large enough to induce phase transformations). In this model, a similar free energy function as that used in [10,1,9] is employed to construct the 2D dynamical models.

2 The 2D Model for SMA Dynamics

Based on conservation laws for linear momentum and energy, the system describing coupled thermo-mechanical wave interactions for the first order phase transitions in a two dimensional SMA structure can be written as follows [16,14]

$$\begin{aligned} \rho\frac{\partial^2 u_i}{\partial t^2} &= \nabla_x \cdot \boldsymbol{\sigma} + f_i, \quad i,j = 1,2, \\ \rho\frac{\partial e}{\partial t} - \boldsymbol{\sigma}^T : (\nabla \mathbf{v}) + \nabla \cdot \mathbf{q} &= g, \end{aligned} \tag{1}$$

where ρ is the density of the material, $\mathbf{u} = \{u_i\}|_{i=1,2}$ is the displacement vector, $\mathbf{v} = \partial\mathbf{u}/\partial t$ is the velocity vector, $\boldsymbol{\sigma} = \{\sigma_{ij}\}$ is the stress tensor, $\mathbf{q}$ is the heat flux, e is the internal energy, $\mathbf{f} = (f_1, f_2)^T$ and g are mechanical and thermal loadings, respectively. Let ϕ be the free energy function of a thermo-mechanical system described by (1). Then, the stress and the internal energy are connected with the free energy function ϕ by the following relationships:

$$\boldsymbol{\sigma} = \rho\frac{\partial \phi}{\partial \boldsymbol{\eta}}, \quad e = \phi - \theta\frac{\partial \phi}{\partial \theta}, \tag{2}$$

where θ is the temperature, and the Cauchy-Lagrangian strain tensor $\boldsymbol{\eta}$ is given by its components as follows(with the repeated-index convention used):

$$\eta_{ij}(\mathbf{x}, t) = \left(\frac{\partial u_i(\mathbf{x}, t)}{\partial x_j} + \frac{\partial u_j(\mathbf{x}, t)}{\partial x_i}\right)/2, \tag{3}$$

where $\boldsymbol{x}$ is the coordinates of a material point in the domain of interest.

In the two-dimensional case, the square-to-rectangular transformations could be regarded as a 2D analog of the cubic-to-tetragonal and tetragonal-to-orthorhombic transformations observed in general three-dimensional cases [6,

7]. It was established earlier for this kind of transformations that the free energy function ϕ can be constructed by taking advantage of a Landau free energy function F_L [6,7]. Here, a simple free energy function is chosen for our purpose following the proposal in ([10,6,7,17,8] and references therein):

$$\phi = -c_v\theta\ln\theta + \frac{1}{2}a_1e_1^2 + \frac{1}{2}a_3e_3^2 + F_L, \quad F_L = \frac{1}{2}a_2\left(\theta-\theta_0\right)e_2^2 - \frac{1}{4}a_4e_2^4 + \frac{1}{6}a_6e_2^6, \quad (4)$$

where c_v is the specific heat constant, θ_0 is the reference temperature for transformation, a_i, $i = 1, 2, 3, 4, 6$ are the material-specific coefficients, and e_1, e_2, e_3 are dilatational, deviatoric, and shear components of strain, respectively. The later are defined as follows:

$$e_1 = \left(\eta_{11} + \eta_{22}\right)/\sqrt{2}, \;\; e_2 = \left(\eta_{11} - \eta_{22}\right)/\sqrt{2}, \;\; e_3 = \left(\eta_{12} + \eta_{21}\right)/2. \quad (5)$$

By substituting the free energy function defined by Eq.(4) into the model (1)–(3), the following coupled system of equations are obtained. For the numerical analysis, the displacement are replaced by the strains in the governing equations:

$$\begin{aligned}
\frac{\partial^2u_1}{\partial t^2} &= \frac{\sqrt{2}}{2}\frac{\partial}{\partial x}(a_1e_1 + a_2(\theta-\theta_0)e_2 - a_4e_2^3 + a_6e_2^5) + \frac{1}{2}\frac{\partial}{\partial y}(a_3e_3) + f_1,\\
\frac{\partial^2u_2}{\partial t^2} &= \frac{1}{2}\frac{\partial}{\partial x}(a_3e_3) + \frac{\sqrt{2}}{2}\frac{\partial}{\partial x}(a_1e_1 - a_2(\theta-\theta_0)e_2 + a_4e_2^3 - a_6e_2^5) + f_2,\\
c_v\frac{\partial\theta}{\partial t} &= k(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2}) + \frac{\sqrt{2}}{2}a_2\theta e_2\frac{\partial e_2}{\partial t} + g.
\end{aligned} \quad (6)$$

The above mathematical model is expected be able to capture the thermomechanical interactions and phase transformations in the 2D SMA structures. However, as we have mentioned before, the phase transformations are excluded from our analysis here. We aim only at initial validation of the model with relatively simple model examples. For this purpose, only a small constant mechanical loading, which is not strong enough to induce phase transformations, is applied to the above model. Then, we compare our results to the results obtained with the corresponding 1D model. This 1D mathematical model is the well known Falk model [4,14]. As we pointed out in [18], this model can be obtained directly by a reduction of the above 2D model:

$$\begin{aligned}
\rho\frac{\partial^2u}{\partial t^2} &= \frac{\partial}{\partial x}\left(k_1\left(\theta-\theta_1\right)\frac{\partial u}{\partial x} - k_2(\frac{\partial u}{\partial x})^3 + k_3(\frac{\partial u}{\partial x})^5\right) + F,\\
c_v\frac{\partial\theta}{\partial t} &= k\frac{\partial^2\theta}{\partial x^2} + k_1\theta\frac{\partial u}{\partial x}\frac{\partial v}{\partial t} + G,
\end{aligned} \quad (7)$$

where k_1, k_2, k_3, c_v and k are re-normalized material-specific constants, θ_1 is the reference temperature for 1D martensitic transformations, and F and G are distributed mechanical and thermal loadings.

3 Numerical Methodology

For the convenience of numerical simulation, the 2D model is re-written in the following form in terms of the dilatational and deviatoric strains introduced above:

$$
\begin{aligned}
&\frac{\partial e_1}{\partial t} = (\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y})/\sqrt{2}, \quad \frac{\partial e_2}{\partial t} = (\frac{\partial v_1}{\partial x} - \frac{\partial v_2}{\partial y})/\sqrt{2},\\
&\frac{\partial v_1}{\partial t} = \frac{\partial \sigma_{11}}{\partial x} + \frac{\partial \sigma_{12}}{\partial y} + f_1, \quad \frac{\partial v_2}{\partial t} = \frac{\partial \sigma_{12}}{\partial x} + \frac{\partial \sigma_{22}}{\partial y} + f_2,\\
&c_v \frac{\partial \theta}{\partial t} = k(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2}) + \frac{\sqrt{2}}{2} a_2 \theta e_2 \frac{\partial e_2}{\partial t} + g,\\
&\sigma_{11} = \frac{\sqrt{2}}{2}(a_1 e_1 + a_2(\theta - \theta_0)e_2 - a_4 e_2^3 + a_6 e_2^5),\\
&\sigma_{12} = \sigma_{21} = \frac{1}{2}(a_3 e_3),\\
&\sigma_{22} = \frac{\sqrt{2}}{2}(a_1 e_1 - a_2(\theta - \theta_0)e_2 + a_4 e_2^3 - a_6 e_2^5).
\end{aligned}
\tag{8}
$$

The above formulated model should be simulated together with the compatibility relations in terms of the strains [7]:

$$
\frac{\partial^2 e_1}{\partial x_1^2} + \frac{\partial^2 e_1}{\partial x_2^2} - \sqrt{8}\frac{\partial^2 e_3}{\partial x_1 \partial x_2} - \frac{\partial^2 e_2}{\partial x_1^2} + \frac{\partial^2 e_2}{\partial x_2^2} = 0. \tag{9}
$$

The 2D model given by Eq.(8) is a differential-algebraic system, it is obtained by keeping the constitutive relations as algebraic equations while the stress components are kept as independent variables to be solved for. The idea of simulating the thermo-mechanical waves by the differential-algebraic approach is stimulated by the idea of [12] where the same approach was adopted for the simulation of phase transformations in SMA. We note that to deal with a strong (cubic and quintic) nonlinearities, a smoothing procedure similar to that proposed in [15] has been employed. In particular, we have used the following expansions:

$$
\epsilon^3 = \frac{1}{4}\sum_{i=0}^{3} \epsilon_n^i \epsilon_{n-1}^{3-i}; \quad \epsilon^5 = \frac{1}{6}\sum_{i=0}^{5} \epsilon_n^i \epsilon_{n-1}^{6-i}, \tag{10}
$$

where ϵ_n is the strain on the current time layer n while ϵ_{n-1} is the strain on the previous time layer $n-1$ (For 1D case, $\epsilon = \partial u/\partial x$, while for 2D problem, $\epsilon = e_2$). We summarise this smoothing procedure as follows: Nonlinear terms are averaged here in the Steklov sense, so that for nonlinear function $f(\epsilon)$ (in particular, ϵ^3 and ϵ^5), averaged in the interval $[\epsilon_{n-1}, \epsilon_n]$, we have

$$
g(\epsilon_{n-1}, \epsilon_n) = \frac{1}{\epsilon_n - \epsilon_{n-1}} \int_{\epsilon_{n-1}}^{\epsilon_n} f(\eta) d\eta, \quad \epsilon_{n-1} = \epsilon(t_{n-1}), \quad \epsilon_n = \epsilon(t_n). \tag{11}
$$

System (8) is discretized on a staggered grid. There are totally eight variables need to be solved for the two dimensional problems. The variables e_1, e_2, θ, σ_{11}, σ_{12}, σ_{22} are discretized at the same nodes, while velocity components v_1 and v_2 are computed at the flux points. For the time integration, the backward differentiation formula methodology is applied to obtain the numerical solution of the problem.

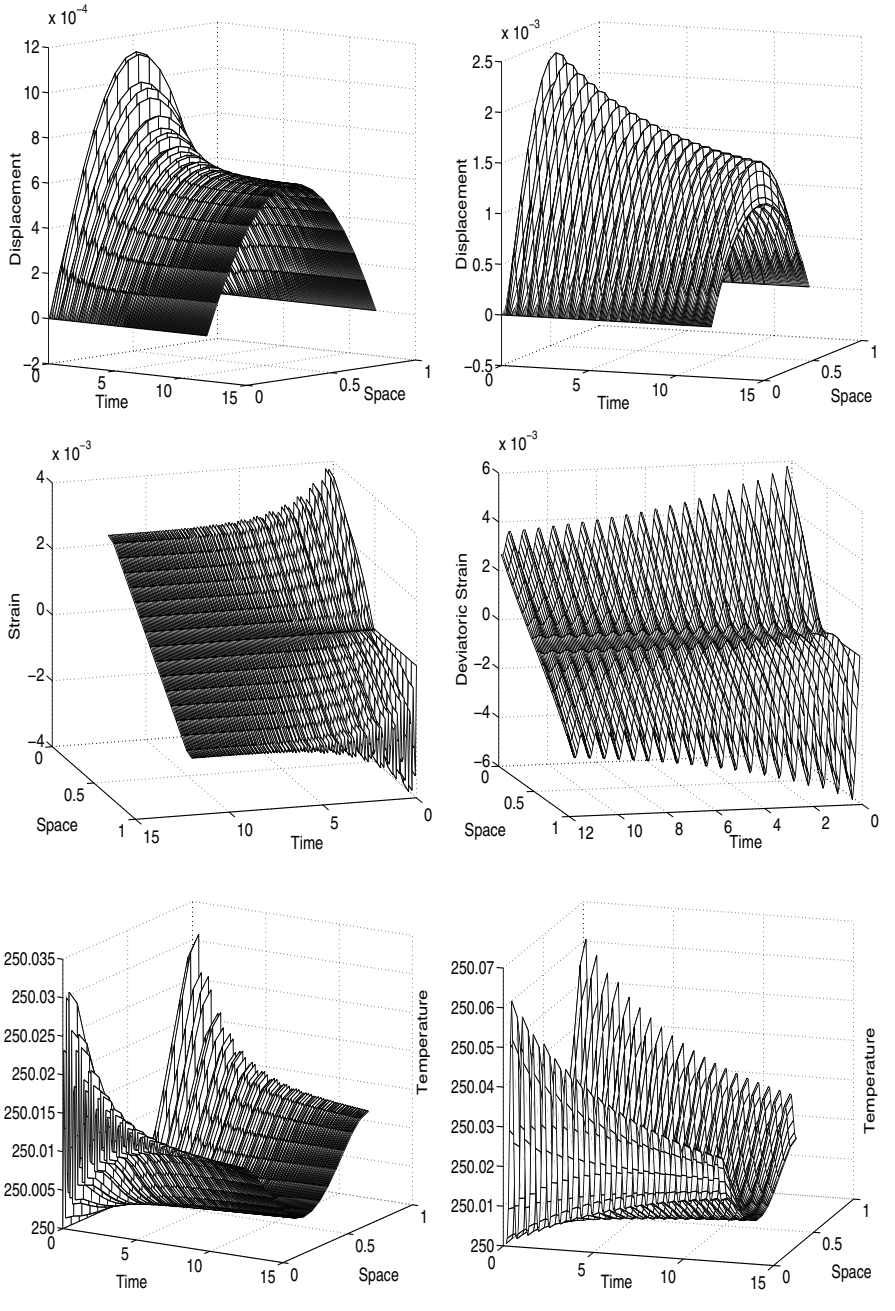


Fig. 1. Thermo-mechanical waves in SMA caused by small mechanical loadings. Left and right columns present the results obtained with the 1D and 2D models, respectively.

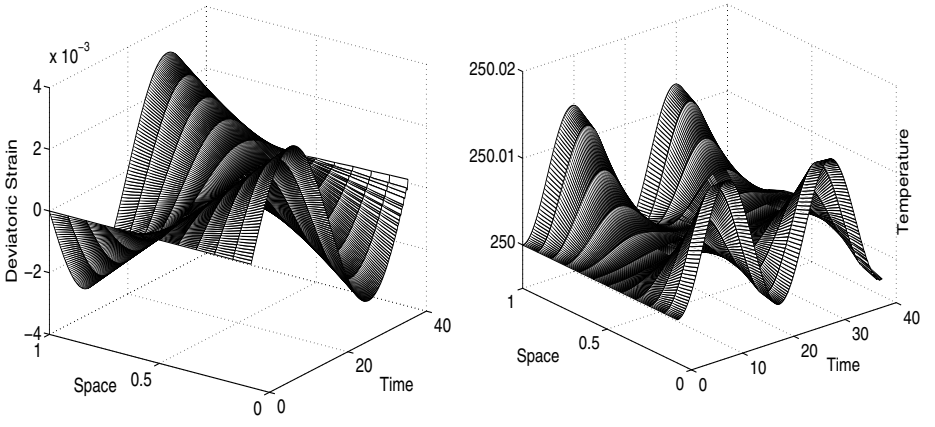


Fig. 2. Thermo-mechanical waves in SMA patch caused by varying mechanical loadings.

4 Numerical Results

Here, we demonstrate first the similarity of the response of 1D and 2D SMA structures under a small constant mechanical loadings, based on the numerical simulation. All simulations are carried out here on the same material, which is $Au_{23}Cu_{30}Zn_{47}$. For this specific material, all necessary physical parameters are taken the same as in [12].

The first simulation is performed in a long thin $Au_{23}Cu_{30}Zn_{47}$ strip of length $L = 1$cm based on the 1D model given by Eq.(7). The initial conditions for this simulation are taken as: $\theta^0 = 250^o K$, $\epsilon^0 = v^0 = s^0 = 0$ where $\epsilon = \partial u/\partial x$, $s = k_1(\theta - \theta_1)\epsilon - k_2\epsilon^3 + k_3\epsilon^5, v = \partial u/\partial t$. Boundary conditions are taken as pinned-end mechanically and insulated thermally. The distributed mechanical loading is assumed constant as $F = 100\text{g}/(\text{ms}^3\text{cm})$, and there are no distributed thermal loadings. There are 17 nodes used for ϵ, s and θ in the computational domain and 16 nodes are used for v approximated at flux points. Time span of the simulation is $[0, 12]$ and time stepsize is set to 0.00025. Displacement, strain, and temperature distributions in the SMA strip are presented in Fig.1, from top to bottom on the left side. The MATLAB program for this case is already validated in [18,12], where the phase transformations are also investigated using the same model and algorithm.

The second simulation is performed in a $Au_{23}Cu_{30}Zn_{47}$ patch with an area of $1 \times 0.4\text{cm}^2$. For this simulation, the coefficients in the 2D model are taken as follows: $a_2 = k_1, a_3 = k_2$, $a_4 = k_3, a_1 = k_1$, $a_3 = 2k_1$, and all other coefficients are taken the same value as those in the 1D case. The initial temperature is also

250^o while all other variables are initially zero. Boundary conditions are taken as follows:

$$\begin{aligned} &\frac{\partial \theta}{\partial x} = 0, \quad \frac{\partial u_2}{\partial x} = 0, \quad u_1 = 0, \quad \text{on left and right boundaries,} \\ &\frac{\partial \theta}{\partial y} = 0, \quad \frac{\partial u_1}{\partial y} = 0, \quad u_2 = 0, \quad \text{on top and bottom boundaries.} \end{aligned} \tag{12}$$

Because the displacement components are already replaced by the strains in Eq.(8), the above boundary conditions are enforced in terms of velocities when the mechanical boundary conditions are concerned.

There are 17 nodes used in the x-direction and 7 in the y-direction in the computational domain. As before, v_1 and v_2 are computed at the flux points. The same time span and stepsize are used as those in the 1D case. The distributed mechanical loading is enforced only in the x-direction as: $f_1 = 100\text{g}/(\text{ms}^3\text{cm})$. Displacement, and strain, and temperature distributions on the line $y = 0.2\text{cm}$ (the central horizontal line) in the SMA patch are presented in Fig.1, from top to bottom on the right side. It is shown clearly from the comparison of strain and temperature evolutions that, when the oscillations in the other direction can be neglected, the 2D SMA structure respond to the mechanical loadings in the same way as its 1D analog. It is also seen from the computation that the oscillations in 1D case are dissipated much faster than those in the 2D case. This implies that there might be a numerical instability problem in 2D simulations when the phase transformations take place due to the slow dissipation of oscillations. This comparison can be used as an initial validation of the 2D model.

The third numerical experiment is to investigate the dynamical thermomechanical response of the SMA patch under varying distributed mechanical loadings, but still small enough not to induce phase transformations. The SMA patch is now covers an area of $1 \times 1\text{cm}^2$, and there are 12 nodes in each direction. The loadings employed for this simulation are: $f_1 = 200 \sin(20\pi t/40)\text{g}/(\text{ms}^3\text{cm})$ in the x-direction, and $f_2 = f_1$ in the y-direction. The time span for this simulation is $[0, 40]$ (one period of loading) and the time stepsize is set at 4×10^{-5}. All other simulation conditions and parameters are the same as those in the second case. The deviatoric strain and temperature distributions are presented in Fig.2. The simulation shows clearly that both temperature and strain field are driven periodically by the distributed mechanical loading because of the thermomechanical coupling. Under such a small loading, the SMA patch behaves just like a classical thermoelastic material.

Based on these three numerical experiments, we conclude that the formulated 2D model is able to capture the thermo-mechanical coupling in SMA. It gives the same prediction about the dynamical behavior of the SMA structure as that of the 1D model, when oscillations in one of the directions are negligible.

References

1. Ball, R., James, M.: Fine phase mixtures as minimizers of energy. Archive. Rat. Mech. Anal. **100** (1988) 13-52.

2. Birman,V.: Review of mechanics of shape memory alloys structures, Appl.Mech.Rev. **50** (1997) 629-645.
3. Bunber, N.: Landau-Ginzburg model for a deformation-driven experiment on shape memory alloys. Continuum Mech. Thermodyn. **8** (1996) 293-308.
4. Falk, F.: Model free energy, mechanics, and thermomechanics of shape memory alloys. Acta Metallurgica, **28**, (1980) 1773-1780.
5. Falk, F., Konopka, p.: Three-dimensional Landau theory describing the martensitic phase transformation of shape memory alloys. J.Phys.:Condens.Matter **2** (1990) 61-77.
6. Ichitsubo, T., Tanaka, K., Koiva, M.,Yamazaki, Y.: Kinetics of cubic to tetragonal transformation under external field by the time-dependent Ginzburg-Landau approach. Phys. Rev. B **62** (9) (2000) 5435-5441.
7. Jacobs, A.: Landau theory of structures in tetragonal-orthorhombic ferro-elastics. Phys. Rev. B **61** (10) (2000) 6587-6595.
8. Kartha, S.: Disorder-driven pretransitional tweed pattern in martensitic transformations. Phys. Rev. B **52** (2) (1995) 803-823.
9. Luskin, M.: On the computational of crystalline microstructure. Acta Numerica **5** (1996) 191-256.
10. Lookman, T., Shenoy, S., Rasmusseh, D., Saxena, A., Bishop, A.: Ferro-elastic dynamics and strain compatibility. Phys. Rev. B **67**(2003) 024114.
11. Matus, P., Melnik, R., Wang, L., Rybak, I.: Application of fully conservative schemes in nonlinear thermoelasticity: Modelling shape memory materials. Mathematics and Computers in Simulation (to appear)
12. Melnik, R., Roberts, A., Thomas, K.: Phase transitions in shape memory alloys with hyperbolic heat conduction and differential algebraic models. Computational Mechanics **29** (1) (2002) 16-26.
13. Melnik, R., Roberts, A., Thomas, K.: Coupled thermomechanical dynamics of phase transitions in shape memory alloys and related hysteresis phenomena. Mechanics Research Communications **28** 6 (2001) 637-651.
14. Melnik, R., Robert, A., Thomas, K.: Computing dynamics of Copper-based SMA via center manifold reduction models. Computational Material Science **18**(2000) 255-268.
15. Niezgodka, M., Sprekels, J.: Convergent numerical approximations of the thermomechanical phase transitions in shape memory alloys. Numerische Mathematik **58**(1991) 759-778.
16. Pawlow, I.: Three dimensional model of thermomechanical evolution of shape memory materials. Control and Cybernetics **29** (2000) 341-365.
17. Saxena, A., Bishop, A., Shenoy, S., Lookman, T.: Computer simulation of martensitic textures. Computational Materials Science **10** (1998) 16-21.
18. Wang, L., Melnik, R.: Nonlinear coupled thermomechanical waves modelling shear type phase transformation in shape memory alloys. in Mathematical and Numerical Aspects of Wave Propagation, Eds.G.C.Cohen, et al,Springer,723-728 (2003).

Direct and Homogeneous Numerical Approaches to Multiphase Flows and Applications

Roman Samulyak[1], Tianshi Lu[2], and Yarema Prykarpatskyy[1]

[1] Center for Data Intensive Computing,
Brookhaven National Laboratory, Upton, NY 11973, USA
{rosamu, yarpry}@bnl.gov
[2] Department of Applied Mathematics and Statistics,
SUNY at Stony Brook, Stony Brook, NY 11794, USA
tlu@sunysb.edu

Abstract. We have studied two approaches to the modeling of bubbly and cavitating fluids. The first approach is based on the direct numerical simulation of gas bubbles using the interface tracking technique. The second one uses a homogeneous description of bubbly fluid properties. Two techniques are complementary and can be applied to resolve different spatial scales in simulations. Numerical simulations of the dynamics of linear and shock waves in bubbly fluids have been performed and compared with experiments and theoretical predictions. Two techniques are being applied to study hydrodynamic processes in liquid mercury targets for new generation accelerators.

1 Introduction

An accurate description of cavitation and wave propagation in cavitating and bubbly fluids is a key problem in modeling and simulation of hydrodynamic processes in a variety of applications ranging from marine engineering to high energy physics. The modeling of free surface flows imposes an additional complication on this multiscale problem.

The wave propagation in bubbly fluids have been studied using a variety of methods. Significant progress has been achieved using various homogeneous descriptions of multiphase systems (see for example [1,2,13,15] and references therein). The Rayleigh-Plesset equation for the evolution of the average bubble size distribution has often been used as a dynamic closure for fluid dynamics equations. This allows to implicitly include many important physics effects in bubbly systems such as the drug, viscosity, and surface tension. Numerical simulations of such systems require relatively simple and computationally inexpensive numerical algorithms. Nevertheless, homogeneous models cannot capture all features of complex flow regimes and exhibit sometimes large discrepancies with experiments [13] even for systems of non-dissolvable gas bubbles. Homogeneous models are also not suitable for modeling phase transitions in bubbly fluids such as boiling and cavitation.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 653–660, 2004.
© Springer-Verlag Berlin Heidelberg 2004

A powerfull method for multiphase problems, direct numerical simulation, is based on techniques developed for free surface flows. Examples of numerical simulations of a single vapor bubble undergoing a phase transition on its surface are given in [9,14]. Systems of bubbles in fluids were modeled in [7] using the incompressible flow approximation for both fluid and vapor and a simplified version of the interface tracking. In this paper, we describe a direct numerical simulation method for systems of compressible bubbles in fluids using the method of front tracking. Our FronTier code is capable of tracking and resolving topological changes of a large number of fluid interfaces in 2D and 3D spaces. We present the simulation results of the wave dynamics of linear and shock waves in bubbly systems and compare them with classical experiments.

The direct numerical simulations of wave dynamics in bubbly fluids in large 3D domains remain, however, prohibitively expensive even on supercomputers. Homogeneous models can effectively be used for such systems, especially if the resolving of spatial scales smaller then the distance between bubbles is not necessary. To model cavitating and bubbly fluids within the homogeneous approximation, we have recently developed and implemented in the FronTier code a two-phase equation of state (EOS) model based on the isentropic approximation. Therefore both direct and homogeneous approaches have advantages and disadvantages and can be used to resolve different temporal and spatial scales in numerical simulations.

Two numerical approaches are being used to study hydrodynamic processes involving cavitation and bubble dynamics in liquid mercury targets for new generation accelerators such as the Spallation Neutron Source (SNS) and the Muon Collider/Neutrino Factory (MC). Hydrodynamic insbabilities and cavitation in the MC mercury jet target will create complications for the machine operation. The collapse of cavitation bubbles in the SNS mercury target, resulting in the pitting of steel walls, has been the most critical problem reducing the target lifetime. The injection of layers of gas bubbles in mercury has been proposed as a possible pressure mitigation technique. These processes must be studied by means of large-scale numerical simulations.

The paper is organized as follows. In Section 2, we describe briefly the direct and homogeneous methods for the modeling of bubbly flows. Section 3 contains results of the numerical simulation of linear and shock waves in bubbly liquids using the direct and homogeneous techniques. We discuss classical shock tube experiments and applications to liquid mercury targets. Finally, we conclude the paper with a summary of our results and perspectives for future work.

2 Modeling of Multiphase Flows

2.1 Direct Method

In the direct method, we model a liquid – vapor or liquid – non-dissolvable gas mixture as a system of one phase domains separated by free interfaces. The FronTier code represents interfaces as lower dimensional meshes moving through a

volume filling grid [4,5]. The traditional volume filling finite difference grid supports smooth solutions located in the region between interfaces. The dynamics of the interface comes from the mathematical theory of Riemann solutions, which are idealized solutions of single jump discontinuities for a conservation law. The FronTier code is capable of tracking simultaneously a large number of interfaces and resolving their topological changes (the breakup and merger of droplets) in two- and three- dimensional spaces. Away from interfaces, FronTier uses high resolution hyperbolic techniques. Different equation of state models are used for gas/vapor bubbles and the ambient liquid. Though computationally intensive, such an approach is very accurate in treating important effects in bubbly flows including phase transitions in boiling and cavitating fluids. The method makes it possible to resolve spatial scales smaller than the typical distance between bubbles and to model some non-equilibrium thermodynamics features such as finite critical tension in cavitating liquids.

2.2 Homogeneous Method

The advantage of the homogeneous method compared to the direct one is its computational simplicity and, as a result, low computational requirements for the simulation of large systems. The homogeneous flow approximation provides a simple technique for analyzing two-phase (or multiple phase) flows. It is accurate enough to handle a variety of practically important processes. Suitable averaging is performed over the length scale which is large compared to the distance between bubbles and the mixture is treated as a pseudofluid that obeys an equation of state (EOS) of a single component flow [15].

We have recently developed [11] a simple isentropic homogeneous equation of state for two-phase liquids and implemented the corresponding software library in the FronTier code. The isentropic approximation reduces by one the number on independent variables defining the thermodynamic state. As a result, all thermodynamic states in our EOS are functions of only density. The proposed EOS consists of three branches. The pure vapor and liquid branches are described by the polytropic and stiffened polytropic [10] EOS models, respectively, reduced to a single isentrope. The two branches are connected by a model for the liquid-vapor mixture

$$P = P_{sat,l} + P_{vl}\log\left[\frac{\rho_{sat,v}{a_{sat,v}}^2(\rho_{sat,l} + \alpha(\rho_{sat,v} - \rho_{sat,l}))}{\rho_{sat,l}(\rho_{sat,v}{a_{sat,v}}^2 - \alpha(\rho_{sat,v}{a_{sat,v}}^2 - \rho_{sat,l}{a_{sat,l}}^2))}\right],$$

where $\rho_{sat,v}$, $\rho_{sat,l}$, $a_{sat,v}$, $a_{sat,l}$ are the density and the speed of sound of vapor and liquid in saturation points, respectively, $P_{sat,l}$ is the liquid pressure in the saturation point, α is the void fraction

$$\alpha = \frac{\rho - \rho_{sat,l}}{\rho_{sat,v} - \rho_{sat,l}},$$

and the parameter P_{vl} is

$$P_{vl} = \frac{\rho_{sat,v}{a_{sat,v}}^2\rho_{sat,l}{a_{sat,l}}^2(\rho_{sat,v} - \rho_{sat,l})}{\rho_{sat,v}{}^2{a_{sat,v}}^2 - \rho_{sat,l}{}^2{a_{sat,l}}^2}.$$

These expressions were derived by integrating an experimentally validated model for the sound speed in bubbly mixture [15]. A set of the EOS input parameters most of which are measurable quantities allows to fit the two-phase EOS to thermodynamics data for real fluids. Details on the EOS model are presented in [11].

The FronTier code with two-phase EOS modes has been applied to study the interaction of mercury with an intensive proton pulse in the geometry typical for mercury thimble neutrino factory experiments. The use of two-phase EOS has led to improvement over single phase EOS simulations [12] of the mercury surface evolution.

3 Numerical Results

3.1 Direct Numerical Simulations

In this section, we present results of the direct numerical simulation of the dynamics of linear and nonlinear waves in bubbly liquids. The schematic of the numerical experiment setup is depicted in Figure 1. The liquid contains non-dissolvable gas bubbles at normal conditions. The bubble radius is 0.12 mm and the void fraction is $1.55 \cdot 10^{-4}$. The region around a long column of bubbles was chosen as the computational domain. As the first order approximation, we can assume that the pressure waves are axisymmetric. The influence of neighboring bubbles can be effectively approximated by the Neumann boundary condition on the domain walls. Therefore the wave propagation in bubbly flows was reduced to an axisymmetric two-dimensional problem.

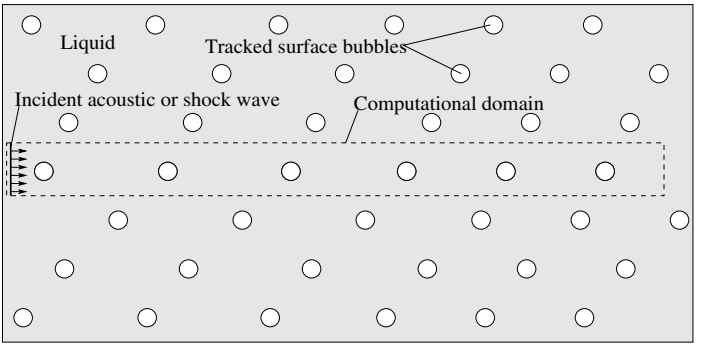


Fig. 1. Schematic of the numerical experiment. The computational domain contained 100 bubbles

Our first numerical experiments were performed with small amplitude linear waves in bubbly fluids. Measuring the dispersion relation and the attenuation rates from simulations, we found that results are in good agreement with Fox,

Curley and Larson's experiments [3] as well as theoretical predictions [15]. The numerical and theoretical results are depictured in Figure 2.

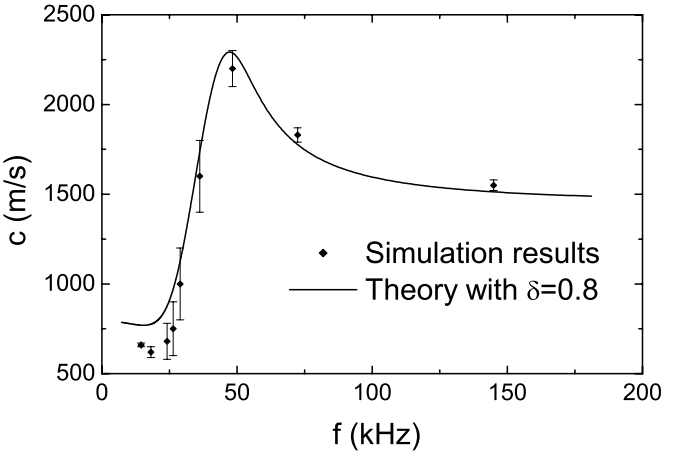


Fig. 2. Dispersion relation in bubbly flows. The bubble radius is 0.12mm, void fraction is $1.55 \cdot 10^{-4}$, and the pressure is 1.0 bar. The amplitude of the incident pressure wave is 0.1 bar. Grid size is 90×10800, δ is the damping coefficient

Figure 3 depicts results of experiments and numerical simulations of the interaction of shock waves with bubbly layers. We measured shock velocities and plotted shock profiles for bubbles consisting of different kinds of gases. The shock speeds agreed with the theoretical prediction of [13] very well (with the difference less than 3%). The shock profiles agreed with experiments of Beylich and Gulhan [1] qualitatively and partly quantitatively. Some discrepancy in the amplitude of pressure oscillations can be explained by grid related numerical errors. To improve the accuracy and performance of the direct method, we have been working on the adaptive mesh refinement method for the FronTier code. Current simulations performed on 90×10800 grids required several days of CPU time on a parallel cluster of Pentium processors. Both simulations and experiments showed that the amplitude of pressure ocsillations in the bubbly layer after the passage of the shock front is smaller for the gas with larger polytropic index.

We have also performed preliminary numerical simulations of the interaction of bubbly mercury with a strong proton pulse in the SNS target. The use of layers of gas bubbles has been proposed as a pressure mitigation technique which may reduce the cavitation induced erosion and extend the target lifetime. Preliminary results confirmed the usefulness of this mitigation method.

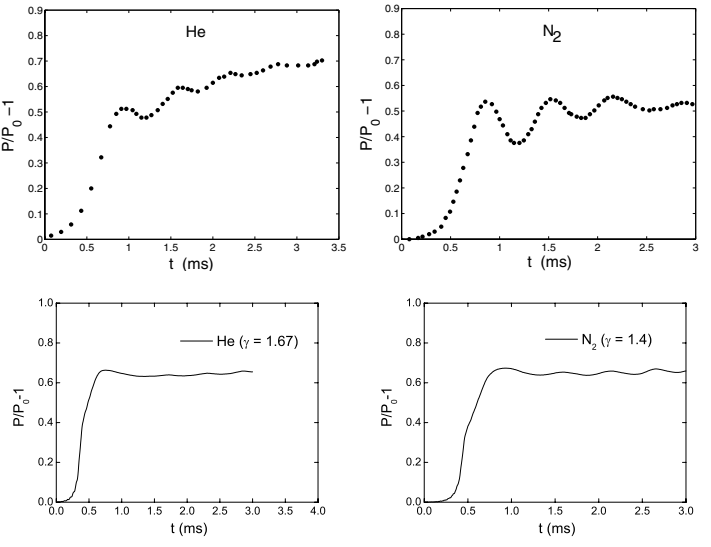


Fig. 3. Experiments of Beylich and Gulhan [1] (top) and FronTier numerical simulations (bottom) of shock waves in bubbly liquids. The bubble radius is 1.18mm, the void fraction is $2.5 \cdot 10^{-3}$, the fluid pressure is 1.1 bar, and the initial shock pressure is 1.727 bar. Grid size is 35×7000. Shock profiles were measured 40 cm from the interface

3.2 Numerical Simulation of Interaction of Mercury with Protons Using the Homogeneous Model

In this section, we present numerical simulation results of the interaction of free surface mercury with strong proton pulses using FronTier code with the homogeneous two-phase EOS model. To study the influence of proton pulse induced thermal shocks on mercury targets, a series of experiments were conducted at the Alternating Gradient Synchrotron (AGS) at BNL and On-Line Isotope Mass Separator facility (ISOLDE) at CERN [6,8]. We will discuss here some experimental and numerical simulation results of the mercury thimble studies.

The volume of the thimble excavated in a stainless steel bar is 1.3 cm^3. It consists from bottom to top of a half sphere (r = 6 mm), and a vertical cylinder (r = h = 6 mm). The mercury has a free surface in up-direction, where it can expand to. The mercury interacts with a proton pulse of approximately Gaussian distribution. The pulse intensity range is $0.6 - 17 \cdot 10^{12}$ protons at energy 24 GeV.

We have performed numerical simulations of the mercury splash evolution in the thimble using the FronTier code with the two-phase EOS (Figure 5). We have studied the evolution of the mercury splash in the thimble at different values of the proton intensity and the r.m.s. spot size of the beam, and obtained a good agreement with experimental data. However, numerical simulation do not capture some experimentally observable fine effects in the splash evolution such as the reduction of the splash velocity during first 2 microseconds after

the arrival of the proton pulse [6]. This discrepancy can be explained by the incomplete thermodynamics of our EOS. In the future, these simulations will also be performed using the direct method with full thermodynamics and the phase transition support.

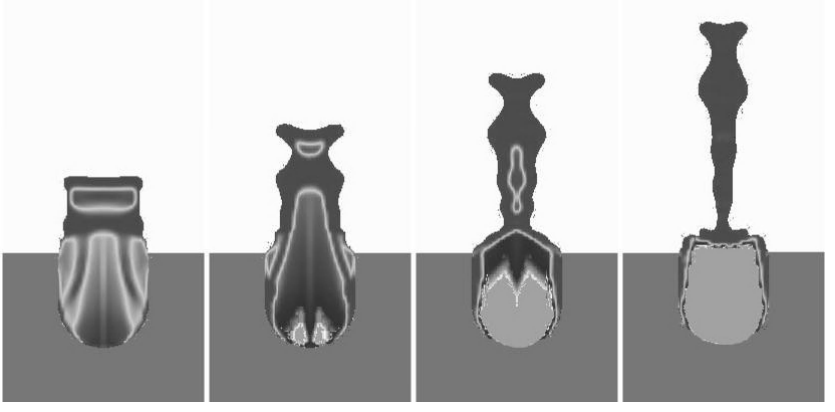


Fig. 4. Numerical simulation of the mercury splash in the thimble

4 Conclusions

We have developed the direct and homogeneous approaches within the FronTier code for the simulation of bubbly fluids. Through the comparison of numerical simulations with experiments and theoretical predictions on the propagation of linear and shock waves in bubbly fluids, the direct approach which uses the method of front tracking and the FronTier code has been validated. The direct method has a variety of current and potential applications. It is being applied for numerical simulations of the interaction of bubbly mercury with strong proton pulses in the SNS target. The use of layers of gas bubbles has been proposed as a pressure mitigation technique which may reduce the cavitation induced erosion and extend the target lifetime. Preliminary simulations have demonstrated an effective reduction of the peak pressure. The mass transfer across bubble surfaces due to the phase transition is being implemented in the code. This will allow to simulate systems of vapor bubbles in boiling and cavitating liquids. To improve the numerical resolution and performance of the code, an adaptive mesh refinement method will be used in future simulations.

To model cavitating and bubbly fluids within the homogeneous approximation, we have developed and implemented in the FronTier code a two-phase equation of state (EOS) model based on the isentropic approximation. This approach is especially suitable if spatial scales smaller than the distance between bubbles can be neglected for global studies of the wave dynamics. It requires

coarser grids compared to the direct numerical simulation method and often significantly smaller CPU time. The code has been applied to study the interaction of mercury with an intensive proton pulse in the geometry typical for mercury thimble neutrino factory experiments. The simulations are in good quantitative agreement with experiments. The direct and homogeneous approaches are complementary and can be used to resolve different temporal and spatial scales in numerical simulations.

Acknowledgments. The authors are grateful to James Glimm, Xiaolin Li, and Harold Kirk for fruitful discussions. Financial support has been provided by the USA Department of Energy, under contract number DE-AC02-98CH10886.

References

1. Beylich, A.E., Gülhan, A.: On the structure of nonlinear waves in liquids with gas bubbles. Phys. Fluids A **2** (1990) 1412 – 1428
2. Caflisch, R.E., Miksis, M.J., Papanicolaou, G.C., Ting, L.: Effective equations for wave propagation in bubbly liquids. J. Fluid Mech. **153** (1985) 259 – 273
3. Fox, F.E., Curley, S.R., Larson,G.S.: J. Acoust. Soc. Am. **27** (1955) 534
4. Glimm, J., Grove, J.W., Li, X.L., Shyue, K.M., Zhang, Q., Zeng, Y.: Three dimensional front tracking. SIAM J. Sci. Comp. **19** (1998) 703 – 727
5. Glimm J., Grove J., Li X.-L, and Tan D.C.: Robust computational algorithms for dynamic interface tracking in three dimensions. SIAM J. Sci. Comp. **21** (2000) 2240 – 2256
6. Fabich, A., Lettry J.: Experimental Observation of Proton-Induced Shocks and Magneto-Fluid-Dynamics in Liquid Metal. In Proceedings NuFact 01, NIM A, (2001)
7. Juric, D., Tryggvason, G.: Computation of boiling flows. Int. J. Multiphase Flow **24** (1998) 387 – 410
8. Kirk, H., et al.: Target studies with BNL E951 at the AGS. Particles and Accelerators 2001, June 18-22 (2001) Chicago IL
9. Matsumoto, Y., Takemura, F.: Influence of internal phenomena on gas bubble motion. JSME Int. J. **37** (1994) 288 – 296
10. Menikoff, R., Plohr, B.: The Riemann problem for fluid flow of real materials. Rev. Mod. Phys. **61** (1989) 75 – 130
11. Samulyak, R., Prykarpatskyy, Y.: Richtmyer-Meshkov instability in liquid metal flows: influence of cavitation and magnetic fields. Mathematics and Computers in Simulations (2003). In press
12. Samulyak, R.: Numerical simulation of hydro- and magnetohydrodynamic processes in the Muon Collider target. Lecture Notes in Comp. Sci., Vol. 2331. Springer-Verlag, Berlin Heidelberg New York (2002) 391 – 400
13. Watanabe, M., Prosperetti, A.: Shock waves in dilute bubbly liquids. J. Fluid Mech. 274 (1994) 349 – 381
14. Welch, S.W.: Local simulation of two-phase flows including interface tracking with mass transfer. J. Comp. Phys. **121** (1995) 142 – 154
15. Wijngaarden, L. Van: One-dimensional flow of liquids containing small gas bubbles. Ann. Rev. Fluid Mech. 4 (1972) 369-396

Molecular Dynamics and Monte Carlo Simulations for Heat Transfer in Micro and Nano-channels

A.J.H. Frijns[1], S.V. Nedea[1], A.J. Markvoort[2], A.A. van Steenhoven[1], and P.A.J. Hilbers[2]

[1] Eindhoven University of Technology, Department of Mechanical Engineering, P.O.Box 513, 5600MB Eindhoven, the Netherlands
[2] Eindhoven University of Technology, Department of Biomedical Engineering, P.O.Box 513, 5600MB Eindhoven, the Netherlands

Abstract. There is a tendency to cool mechanical and electrical components by microchannels. When the channel size decreases, the continuum approach starts to fail and particle based methods should be used. In this paper, a dense gas in micro and nano-channels is modelled by molecular dynamics and Monte Carlo simulations. It is shown that in the limit situation both methods yield the same solution. Molecular dynamics is an accurate but computational expensive method. The Monte Carlo method is more efficient, but is less accurate near the boundaries. Therefore a new coupling algorithm for molecular dynamics and Monte Carlo is introduced in which the advantages of both methods are used.

1 Introduction

There is a tendency that mechanical and electrical components become smaller and smaller. Since most components produce heat when operating, it is essential to cool them in order to perform well and to ensure the life span of such components. For example in computer chips, the power increases with a factor of 10 every 6 years [9]. The standard cooling techniques start to fail. Therefore, more efficient cooling techniques, like microchannel cooling with phase transition, are necessary.

However when the channel size decreases, the continuum approach starts to fail. At Knudsen numbers $\gtrsim 0.1$, particle based methods should be used. From a physical point of view, molecular dynamics is a suitable method. From a computational point of view, this method is too expensive with respect to time to use for microchannels. To handle larger time frames, a Monte Carlo approximation is employed, but at the expense of less accuracy near the boundaries.

Our goal is to combine molecular dynamics with a Monte Carlo method such that we can use the advantages of both methods: molecular dynamics near the boundaries, because of the accuracy, and Monte Carlo in the bulk, because of the lower computational costs.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 661–666, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper, we consider both methods of modelling heat transfer in micro and nano-channels and introduce a new algorithm in which molecular dynamics and Monte Carlo codes with different time stepping are coupled.

2 Molecular Dynamics and Monte Carlo Methods

Molecular dynamics (MD) is a computer simulation technique where the time evolution of a set of interacting particles is followed. This is done by numerically solving the equations of motion (Newton's Second law) of classical multi-body systems. Given the positions, masses and velocities of all particles in the system and the forces on the particles, the motion of all (individual) particles can be followed in time by calculating the (deterministic) single particle trajectories. However, to calculate these trajectories is in general still very computationally intensive, because in principal it is possible that every particle interacts with every other particle in the system, resulting in N(N-1)/2 pair interactions for a set of N particles that need to be calculated.

The interactions between the particles are given in the form of potentials. In order to model a gas by almost hard sphere interactions we use a truncated and shifted Lennard-Jones potential with a cut-off radius $R_c = 2^{1/6}\sigma$. Where σ is the collision diameter. For the simulations, we use the PumMa code developed at the Eindhoven University of Technology [4].

The Monte Carlo (MC) method is based on the Direct Simulation Monte Carlo method (DSMC) developed by Bird [3]. This DSMC method does not calculate the collisions exactly as in molecular dynamics, but generates collisions stochastically with scattering rates and post-collision velocity distributions determined from the kinetic theory of a dilute gas. Several authors [7,8,10] used this method successfully to study flow and heat transfer in microchannels for a dilute gas. However, for cooling purposes (high pressure or phase transition) we have also to model a dense gas in a microchannel. Therefore the DSMC method is modified by using the Enskog equation instead of the Boltzmann equation [5, 6]. In this MC method the collisions take place between particles inside a cell and particles of neighbouring cells. The Enskog equation preserves the momentum and energy of the system.

In order to be able to couple molecular dynamics with Monte Carlo, both methods should give the same results for the limit situation where one artificial particle in the Monte Carlo simulation represents one molecule in the molecular dynamics simulation. As validation the particle density distribution and the temperature in a nanochannel are computed with both models. A nanochannel is chosen because of the computational costs. However, the method holds for microchannels too. We approximate the nanochannel by two infinite parallel plates kept at two constant temperatures $T_c = 120$K and $T_h = 240$K. Gas molecules are confined between these walls. The interaction between the gas molecules and the walls is by thermal boundary conditions. A dense Argon gas is considered, with a particle diameter $R = 0.191$ nm, the mean free path length $\lambda_R = 0.346$ nm, and

the initial uniform particle density $n_0 = 3.43$ particles nm^{-3}. The total number of particles is $N = 20000$.

The results for the MD and the MC codes are similar (fig. 1) and are in good agreement with Frezzotti's results [6]. The heat is transferred from the cold to the hot wall by the kinetic energy of the particles.

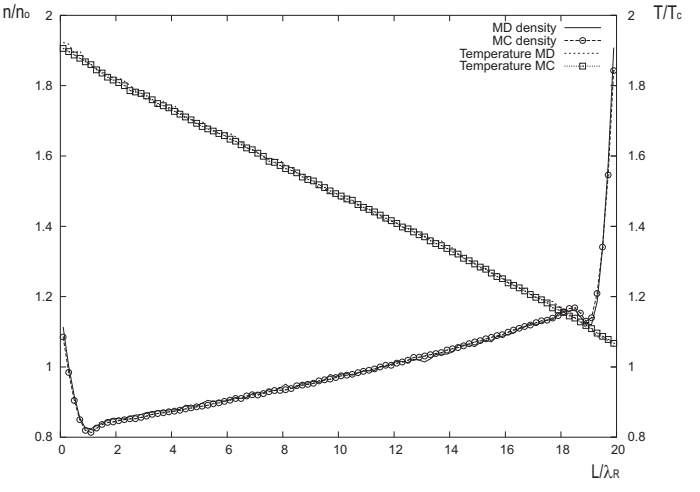


Fig. 1. Particle density profile n/n_0 and temperature profile T/T_c in a nanochannel. Left: hot wall; right: cold wall

3 Coupling of MD with MC

As shown, both methods can be used to model the nanochannels. The MD method is accurate but computational expensive, while the MC method, in which an artificial particle represents several molecules, is more efficient. However, the boundary effects are computed less accurate, because the artificial particles are too large. The positions of the maxima and minima depend on the particle size. Therefore the MC method deviates from the MD method when several molecules are put in one MC particle. In order to get an accurate solution near the walls, a detailed model is needed in which the particle diameter is equal to the molecular diameter. MD is suited for that. In the bulk of the channel, the particle diameter is not so critical. In this region the computations can be sped up by MC with several molecules inside one artificial particle and by using larger time steps. So, a coupling of MD with MC results in a fast and accurate solution.

In our coupling algorithm, the domain is divided into MD and MC subdomains and an interface between these subdomains is defined (figure 2). The coupling algorithm is as follows.

1. *Assign particles to MD or MC part.*
   Particles are assigned to the MD or to the MC part depending on their position. All particles in regions I and II (see figure 2A) are assigned to the MD part and all particles in regions III and IV to the MC part. In the figure the particles that have been assigned to molecular dynamics are white whereas the MC particles are grey.
2. *Perform N Molecular Dynamics time steps.*
   Depending on the respective time step sizes in the MD and MC part of the simulation a certain number (N) of MD steps corresponds with one single MC step. These N MD steps are performed while keeping the positions of the MC particles fixed, as shown in figure 2B. The particles in region III are necessary because, although they are fixed in position, they can interact with the MD particles.
3. *Send particles in range II and III from MD to MC.*
   The molecular dynamics time steps should be followed by a time step of the MC part. Before this can be done the particles in the interface regions II and III should be sent to the MC. As shown in figure 2C, the particles are removed from the MD simulation and added to the MC simulation.
4. *Perform one MC time step.*
   One MC time step consists of two parts: an advection step and a collision step. The advection step, i.e. moving the particles, is done only for the MC (solid) particles, as shown in figure 2D. During the collision step collision pairs are randomly selected with the restriction that at least one of the particles should be a MC particle.
5. *Send particles in range II and III from MC to MD.*
   After the MC step the particles that are now in ranges II and III are sent to the MD simulation. Extra care should be taken here, because a MC particle can overlap with another particle, whereas this is impossible for particles in the molecular dynamics simulation. A solution to this problem is to check whether a MC particle in range II is too close to any MD particle or other MC particle in range II, and if so adding it a new random y and z position while keeping the x position fixed such that it no longer overlaps with any other particles.
6. *Start over with step 1.*
   By converting the MC particles that ended up in range II to MD particles and vice versa MD particles that ended up in range III to MC particles (step 1) the next cycle of the simulation can be started.

Note that the MD and the MC steps are computed by different codes. They are coupled by an interface written in Python.

4 Results and Discussion

The algorithm is validated for the case in which the particles for the MC part have the same size as in the MD part. However, the time steps are different in both domains: one MC time step equals five MD time steps. We consider the

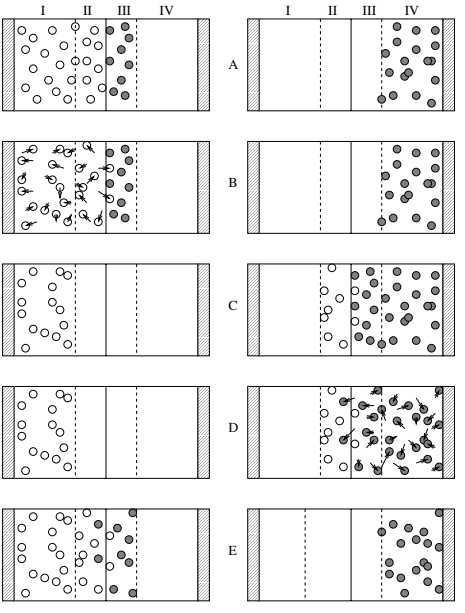


Fig. 2. Schematic view of the coupling algorithm. Left: MD steps; right: MC steps

nanochannel with the constant wall temperatures $T_c = T_h = 120$K. The other system parameters are chosen the same as in the first example. The results are plotted in figure 3.

It is shown that the coupling of MD with MC by this interface is functioning. Some care must be taken at the artificial boundary between the MC and the MD domains, because the MC particles do not move while the MD steps are computed. In this way an artificial highly porous wall can be created. For dilute gases, this effect is negligible. However for a dense gas a small local disturbance in the particle density occurs.

5 Future Research

In this paper, it is shown that the algorithm for the coupling of MD with MC is functioning, but for a dense gas a small local disturbance in the particle density will occur at the interface. For the future, we will adapt the algorithm such that the local disturbance in the particle density will be minimised. To that end the interaction between the MD and MC particles in region III (step 2) will be improved by allowing small displacements of the MC particles. The next step will be that the particles in the MC domain consist of several MD particles. So, an extra step for combining several MD particles into one MC particle and disintegrating of one MC particle into several MD particles will be added to the algorithm. Finally, the efficiency of the algorithm will be studied in more detail.

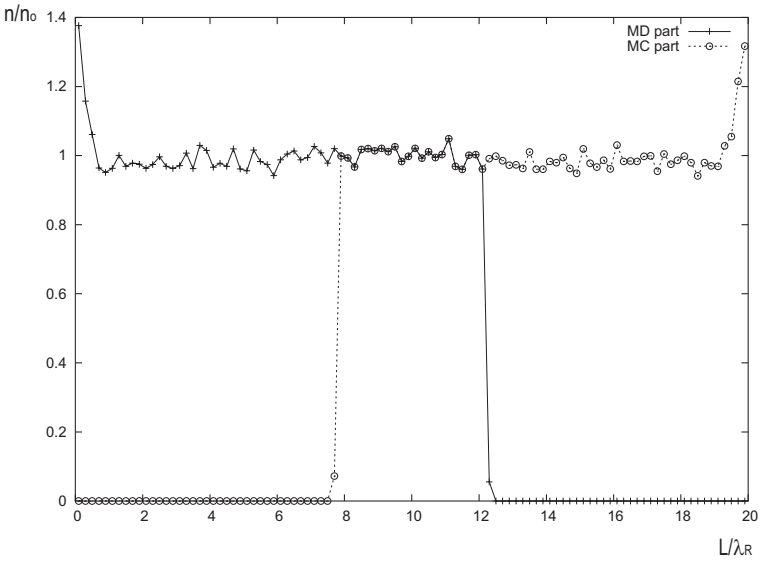


Fig. 3. Dimensionless particle density n in a nanochannel with the coupling of the molecular dynamics method with the Monte Carlo method

References

1. Alexander, F.J., and Garcia, A.L., The Direct Simulation Monte Carlo method, Computers in Physics, **11** (1997), 588–593.
2. M.P. Allen and D.J. Tildesley, Computer Simulation of Liquids, Oxford University Press,1987.
3. Bird, G.A., Molecular gas dynamics and the direct simulation of gas flows, Clarendon Press, Oxford, 1994.
4. K. Esselink, B. Smit, and P.A.J. Hilbers, Efficient Parallel Implementation of Molecular Dynamics on a Toroidal Network. Part I. Parallelizing Strategy, Journal of computational physics, **106** (1993), pp. 101-107.
5. Frezzotti, A., A particle scheme for the numerical solution of the Enskog equation, Phys. Fluids, **9** (1997), 1329–1335.
6. Frezzotti, A., Monte Carlo simulation of the heat flow in a dense sphere gas, European journal of mechanics,**18**, (1999), 103–119.
7. Hadjiconstantinou, N.G. and Simek, O., Constant-Wall-Temperature Nusselt Number in Micro and Nano-Channels, Journal of Heat Transfer, **124** (2002), 356–364.
8. Mavriplis, C., and Ahn, J.C., and Goulard, R., Heat Transfer and Flowfileds in Short Microchannels Using Direct Simulation Monte Carlo, Journal of Thermophysics and Heat Transfer, **11** (1997), 489–496.
9. Schmidt, R.R., and Notohardjono, B.D., High end server low temperature cooling, IBM Journal of Research and Development, **46** (2002),739–751.
10. Yan, F., and Farouk, B., Computations of Low Pressure Fluid Flow and Heat Transfer in Ducts Using the Direct Simulation Monte Carlo Method, Journal of Heat Transfer, **124** (2002), 609–616.

Improved Semi-Lagrangian Stabilizing Correction Scheme for Shallow Water Equations

Andrei Bourchtein and Ludmila Bourchtein

Pelotas State University, Department of Mathematics
Rua Anchieta 4715 bloco K, ap.304, Pelotas 96020-250, Brazil
burstein@terra.com.br

Abstract. Improved splitting method based on stabilizing correction scheme is used for solving shallow water equations. Recently introduced by Douglas et al. technique is applied to reduce splitting errors. Accuracy and stability analysis showed that the developed scheme allows to chose extended time steps and is more accurate as compared with standard splitting method. The results of numerical experiments confirmed that presented scheme has almost the same computational efficiency as primitive stabilizing correction method and it continues to be accurate and stable for extended time steps up to 1 hour when primitive method fails.

1 Introduction

For a given spatial resolution, the maximum allowable time step for systems with multiple time scales is primarily determined by the fastest processes treated explicitly in numerical scheme. In the large scale atmospheric models the main part of flow energy is in advective motions, which are much slower than the gravitational wave speeds present in the system. Since overall accuracy of numerical solution is mainly defined by approximation of slow modes, the use of explicit schemes in numerical weather prediction models is not justifiable: in such schemes the time step needed for stable integration of the high-frequency motions is more than 10 times as small as the value required by accuracy and stability conditions of the principal low-frequency modes. As a mere fact, explicit time integration has no been used in the last 20 years in the majority of atmospheric centers.

Semi-implicit time differencing is an attractive alternative for atmospheric modeling because the terms responsible for the fastest waves appear in linear form in the primitive (hydrostatic or shallow water) equations. In fact, the semi-implicit Eulerian and semi-Lagrangian methods are the most popular techniques used in numerical weather prediction and simulation of the atmospheric processes of different space and time scales: from general circulation and climate models to regional and mesoscale modeling and numerical weather prediction [9, 17]. In the last two decades a semi-Lagrangian treatment of advection was shown to be more efficient because it allows to overcome the Courant-Friedrichs-Lewy (CFL) condition with respect to advection at the low cost of solving a set of trajectory equations. Semi-Lagrangian semi-implicit methods are used in the majority of actual research and operational atmospheric models [6, 9, 11, 14, 15, 16, 17].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 667–672, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The implicit terms of unsplit semi-implicit models require solution of 2D or 3D linear elliptic equations that is a computationally expensive task. To circumvent this problem, some time splitting techniques have been applied to atmospheric models [2, 3, 5, 10, 12, 18]. Splitting methods allow to decouple multi-dimensional elliptic equations into a set of 1D problems, which are solved very efficiently by direct Gelfand-Thomas algorithm. Unfortunately, to the best of our knowledge, all reports on application of splitting techniques in atmospheric models indicate the fast grow of splitting truncation error when time step exceeds CFL advection criterion [2, 3, 4, 13, 18, 19]. To reduce the splitting errors, Douglas et al. recently have proposed small modifications to splitting construction in the case of parabolic equations [7, 8]. In this study we apply the last technique to semi-Lagrangian stabilizing correction (SLSC) scheme for shallow water equations, analyze stability and accuracy properties of the obtained scheme and present the results of numerical tests.

2 SC Scheme for Linearized Shallow Water Equations

Shallow water equations linearized about a state of rest can be written in the form:

$$\partial_t u = fv - \partial_x \Phi \ , \ \partial_t v = -fu - \partial_y \Phi \ , \ \partial_t \Phi = -c^2 \left(\partial_x u + \partial_y v \right) . \tag{1}$$

The equations are presented in rotating reference system using Cartesian spatial coordinates x, y, time coordinate t and common denotations for unknown functions and parameters: u and v are velocity components, $\Phi = gz$ is the geopotential, g is the gravitational acceleration, z is the height of pressure surface, $c^2 = \Phi_0$ (Φ_0 is a mean value of the geopotential), $f = 2\Omega \sin\varphi$ is the Coriolis parameter, Ω is the modulus of angular velocity of earth's rotation. In system (1) we assume $f = const$.

The semi-discrete form of a single time step of considered SC scheme applied to (1) can be written as follows:

$$\frac{\hat{u} - u}{\tau} = fv - \partial_x \Phi \ , \ \frac{\hat{v} - v}{\tau} = -fu - \partial_y \Phi \ , \ \frac{\hat{\Phi} - \Phi}{\tau} = -c^2 \left(\partial_x u + \partial_y v \right) \ ; \tag{2}$$

$$\frac{\tilde{u} - \hat{u}}{\tau} = f \frac{\tilde{v} - v}{2} - \partial_x \frac{\tilde{\Phi} - \Phi}{2} \ , \ \frac{\tilde{v} - \hat{v}}{\tau} = 0 \ , \ \frac{\tilde{\Phi} - \hat{\Phi}}{\tau} = -c^2 \partial_x \frac{\tilde{u} - u}{2} \ ; \tag{3}$$

$$\frac{u^\tau - \tilde{u}}{\tau} = 0 \ , \ \frac{v^\tau - \tilde{v}}{\tau} = -f \frac{u^\tau - u}{2} - \partial_y \frac{\Phi^\tau - \Phi}{2} \ , \ \frac{\Phi^\tau - \tilde{\Phi}}{\tau} = -c^2 \partial_y \frac{v^\tau - v}{2} \ . \tag{4}$$

Here τ is the time step, n is the step number, φ ($\varphi = u, v, \Phi$) denotes the values at the "current" time level $n\tau$, $\hat{\varphi}, \tilde{\varphi}, \varphi^\tau$ denote the values at the "new" time level $(n+1)\tau$. By eliminating the intermediate functions, one can reduce system (2)-(4) to

$$\frac{u^\tau - u}{\tau} = f\frac{v^\tau + v}{2} - \partial_x \frac{\Phi^\tau + \Phi}{2} + \frac{\tau f^2}{4}\left(u^\tau - u\right) + \frac{\tau f}{4}\partial_y\left(\Phi^\tau - \Phi\right) - \frac{\tau c^2}{4}\partial_{xy}\left(v^\tau - v\right), \tag{5}$$

$$\frac{v^\tau - v}{\tau} = -f\frac{u^\tau + u}{2} - \partial_y \frac{\Phi^\tau + \Phi}{2}, \quad \frac{\Phi^\tau - \Phi}{\tau} = -c^2\partial_x \frac{u^\tau + u}{2} - c^2\partial_y \frac{v^\tau + v}{2}. \tag{6}$$

This implies the second order of accuracy and the following characteristic equation:

$$(\lambda - 1)\left[\left(1 - \frac{\tau^2 f^2}{4}\right)(\lambda - 1)^2 + \left(\frac{\tau^2 c^2}{4}\left(d_x^2 + d_y^2\right) + \frac{\tau^4 c^4}{16} d_x^2 d_y^2 + \frac{\tau^2 f^2}{4}\right)(\lambda + 1)^2\right] = 0 \ . \tag{7}$$

Here d_x, d_y are the traces of partial derivative operators and λ is the amplification factor. Obviously, all roots of equation (7) are on the unit circle iff

$$\tau^2 f^2 \le 4 \ , \tag{8}$$

which is the linear stability criterion for scheme (2)-(4) and its spatial discretizations. This is very lenient restriction because $f_{\max} \approx 1.4 \cdot 10^{-4}\ \sec^{-1}$.

3 Improved SC Scheme for Full Shallow Water Equations

The full shallow water equations have a form

$$d_t u = fv - \partial_x \Phi \ , \ d_t v = -fu - \partial_y \Phi \ , \ d_t \Phi = -\Phi\left(\partial_x u + \partial_y v\right) \ , \tag{9}$$

where d_t is a symbol of the substantive derivative operator.

One of the principal difficulties in applying the splitting methods to atmospheric models is the fast growth of truncation errors due to operator splitting when time step increases. Reports on computational experiments with different splitting techniques show that the overall solution errors become to be unacceptable when time step exceeds 30-40 min, although the stability criterion allows using time steps up to 1.5-2 hours or the scheme is even absolutely stable [2, 3, 4, 13, 18, 19]. One of the recent approaches to reduce these splitting errors was proposed by Douglas et al [7, 8] for parabolic equations. Applying a similar idea to approximation of equations (9) along with efficient semi-Lagrangian treatment of advection, we obtain the following modified SLSC scheme:

$$\frac{\hat{u} - u}{\tau} = fv - \partial_x \Phi - \left[\frac{\tau f^2}{4}\left(u - u^{-\tau}\right) + \frac{\tau f}{4}\partial_y\left(\Phi - \Phi^{-\tau}\right) - \frac{\tau c^2}{4}\partial_{xy}\left(v - v^{-\tau}\right)\right],$$

$$\frac{\hat{v} - v}{\tau} = -fu - \partial_y \Phi \ , \ \frac{\hat{\Phi} - \Phi}{\tau} = -\Phi\left(\partial_x u + \partial_y v\right) \ ; \tag{10}$$

$$\frac{\tilde{u} - \hat{u}}{\tau} = f\frac{\tilde{v} - v}{2} - \partial_x \frac{\tilde{\Phi} - \Phi}{2} \ , \ \frac{\tilde{v} - \hat{v}}{\tau} = 0 \ , \ \frac{\tilde{\Phi} - \hat{\Phi}}{\tau} = -c^2\partial_x \frac{\tilde{u} - u}{2} \ ; \tag{11}$$

$$\frac{u^{\tau}-\tilde{u}}{\tau}=0\ ,\ \frac{v^{\tau}-\tilde{v}}{\tau}=-f\frac{u^{\tau}-u}{2}-\partial_y\frac{\Phi^{\tau}-\Phi}{2}\ ,\ \frac{\Phi^{\tau}-\tilde{\Phi}}{\tau}=-c^2\partial_y\frac{v^{\tau}-v}{2}\ . \qquad (12)$$

In (11), (12) we use constant divergence coefficient, which simplifies solving the implicit equations and has no effect on practical stability and accuracy if geopotential deviations from the mean value are sufficiently small. Hereinafter all the values are located at the points of the particle trajectories calculated at each time step by the formulas $d_t x = u$, $d_t y = v$. Namely, φ denotes the values at the departure points of the 2D trajectories at "current" time level $n\tau$, $\varphi^{-\tau}$ denotes the values at "double departure" points at the "old" time level $(n-1)\tau$, $\hat{\varphi}, \tilde{\varphi}, \varphi^{\tau}$ denote the values at the arrival points of the trajectories at "new" time level $(n+1)\tau$. Arrival points are chosen to be the nodes of a uniform spatial grid and the departure points are calculated by fixed point iterations (which is usually applied algorithm) with a sufficient condition of convergence given in the form [17]:

$$\tau \le 1/V_d\ ,\ V_d = \max\left(|u_x|,|u_y|,|v_x|,|v_y|\right)\ . \qquad (13)$$

For spatial grid with meshsize $h \approx 50km$ we have $V_d \approx 1.6\cdot 10^{-4}\,\mathrm{s}^{-1}$ and the maximum allowable time step is $\tau \approx 100\,\mathrm{min}$. Analysis of the linear stability of scheme (10)-(12) reveals that besides (13) there is a counterpart of (8) to be satisfied: $\tau^2 f^2 \le 2$. However, the last inequality admits time step up to $\tau \approx 170\,\mathrm{min}$, which is a less restrictive than (13), and therefore it does not cause any inconvenience.

Note that if bracketed expression in (10) is omitted we obtain standard SC scheme, which is analogous to (2)-(4), but with semi-Lagrangian approximation of advective terms. This scheme has to satisfy the same condition (13).

4 Numerical Experiments

In Table 1 we present the results of comparison of 24-h geopotential height forecasts produced by standard (SLSC) and modified (MSLSC) schemes with "exact" forecasts obtained by using leap-frog scheme. The spatial domain Q of 5000x5000 km^2 centered at Porto Alegre city ($30^0 S$, $52^0 W$) was covered by uniform spatial grid. To reveal an influence of a spatial resolution, two grids with meshsizes of 100 km and 50 km were used. The initial and boundary value conditions for geopotential and wind components on the 500hPa pressure surface were obtained from objective analysis and global forecasts of National Centers for Environmental Prediction (NCEP). To define the splitting errors of SLSC and MSLSC schemes as functions of time step, numerical integrations were performed for $\tau = 10, 20, 40, 60\,\mathrm{min}$. Note that computations using leap-frog scheme were carried out with time step $\tau = 1\,\mathrm{min}$ (maximum allowable time step for meshsize $h = 50\,\mathrm{km}$). Evidently, some initial error is inherent to this comparison, that is, even if τ approaches 0, the differences

between two forecasts do not vanish because of different spatial truncation errors for semi-Lagrangian and leap-frog methods. However, this error is quite small (less than 2 meters) and has no influence on comparison of the schemes. The results in Table 1 clearly show an advantage of modified scheme for time steps exceeding 30 min and comparison between 100-km and 50-km results confirms that principal part of truncation error consists of additional time splitting error for extended time steps.

Table 1. Root-mean-square differences between presented and reference (leap-frog) schemes. τ is the time step used for integrations (in minutes); h is the meshsize of spatial grid (in kilometers); the shown numbers are root-mean-square height differences between 24-h forecasts of chosen scheme and leap-frog scheme (in meters).

scheme	h	$\tau = 10$	$\tau = 20$	$\tau = 40$	$\tau = 60$
SLSC	100	2.1	2.7	4.9	27.8
MSLSC	100	2.1	2.6	3.5	4.8
SLSC	50	2.0	2.6	4.7	27.3
MSLSC	50	2.0	2.6	3.4	4.6

In the following series of experiments, ten 24-h forecasts were calculated using leap-frog, SLSC and MSLSC schemes with 1-min, 40-min and 60-min time steps, respectively. The same 50-km spatial grid and the same initial and boundary conditions were used. To evaluate a quality of geopotential forecastings, two common used mean scores were calculated: the root-mean-square differences between 24-h forecasts and NCEP analysis and the correlation coefficient between observed and forecast changes [1]. The results of these estimations presented in Table 2 show good level of accuracy and efficiency of MSLSC scheme. Overall values of these measures of forecast skill are quite characteristic for shallow water model [1].

Table 2. Mean objective scores of the 24-h geopotential forecasts. ε and r are the root-mean-square differences (in meters) and correlation coefficients (nondimensional) between analysis and 24-h forecasting fields at the height 500hPa; T_{CPU} is the CPU time for one forecast (in seconds).

scheme	τ	ε	r	T_{CPU}
leap-frog	1	48	0.83	41.4
SLSC	20	46	0.84	11.1
SLSC	40	50	0.82	5.6
MSLSC	40	47	0.84	6.0
MSLSC	60	48	0.83	4.1

Finally, extended in time 120-h forecasts were performed with SLSC and MSLSC models using 40-min and 60-min time steps, respectively. Both integrations were stable, but forecast accuracy was lost after 72 hours of calculations. All computations were carried out on a DEC 3000 computer.

This research was supported by brazilian science foundations CNPq and FAPERGS under grants 302738/2003-7 and 02/0588.7.

References

1. Antes R.A.: Regional models of the atmosphere in middle latitudes. Mon. Wea. Rev., 111 (1983) 1306-1335.
2. Bates J.R.: An efficient semi-Lagrangian and alternating direction implicit method for integrating the shallow water equations. Mon. Wea. Rev., 112 (1984) 2033-2047.
3. Bourchtein A.: Semi-Lagrangian semi-implicit space splitting regional baroclinic atmospheric model. Appl. Numer. Math. 41 (2002) 307-326.
4. Browning G.L., Kreiss H.-O.: Splitting methods for problems with different timescales. Mon.Wea.Rev. 122 (1994) 2614-2622.
5. Cohn S.E., Dee D., Isaacson E., Marchesin D., Zwas G.: A fully implicit scheme for the barotropic primitive equations. Mon. Wea. Rev., 113 (1985), 436-448
6. Côté J., Gravel S., Methot A., Patoine A., Roch M., Staniforth A.: The operational CMC-MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation. Mon. Wea. Rev. 126 (1998) 1373-1395.
7. Douglas J., Kim S.: Improved accuracy for locally one-dimensional methods for parabolic equations. Mathematical Models and Methods in Applied Science, 11 (2001) 1563-1579.
8. Douglas J., Kim S. Lim H.: An improved alternating-direction method for a viscous wave equation, In “Current Trends in Scientific Computing", Z. Chen, R. Glowinski, and Kaitai Li, eds., Contemporary Mathematics, 329 (2003) 99-104.
9. Durran D.: Numerical Methods for Wave Equations in Geophysical Fluid Dynamics. Springer, New York (1999).
10. Kar S.K., Turco R.P., Mechoso C.R., Arakawa A.: A locally one-dimensional semi-implicit scheme for global gridpoint shallow-water models. Mon. Wea. Rev. 122 (1994) 205-222.
11. Kiehl J.T., Hack J.J., Bonan G.B., Boville B.A., Williamson D.L., Rasch P.J.: The National Center for Atmospheric Research Community Climate Model: CCM3. J. Climate 11 (1998) 1131-1149.
12. Mcdonald A.: A semi-Lagrangian and semi-implicit two time level integration scheme. Mon. Wea. Rev. 114 (1986) 824-830.
13. McDonald A., Haugen J.: A two-time-level, three-dimensional semi-Lagrangian, semi-implicit, limited-area gridpoint model of the primitive equations. Mon. Wea. Rev. 120 (1992) 2603-2621.
14. Moorthi S.: NWP Experiments with a gridpoint semi-Lagrangian semi-implicit global model at NCEP. Mon. Wea. Rev. 125 (1997) 74-98.
15. Qian J.H., Semazzi F.H.M., Scroggs J.S.: A global nonhydrostatic semi-Lagrangian atmospheric model with orography. Mon. Wea. Rev. 126 (1998) 747-771.
16. Ritchie H., Temperton C., Simmons A., Hortal M., Davies T., Dent D., Hamrud M.: Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model. Mon. Wea. Rev. 123 (1995) 489-514.
17. Staniforth A., Côté J.: Semi-Lagrangian integration schemes for atmospheric models - A review. Mon. Wea. Rev. 119 (1991) 2206-2223.
18. Tanguay M., Robert A.: Elimination of the Helmholtz equation associated with the semi-implicit scheme in a grid point model of the shallow water equations. Mon. Wea. Rev. 114 (1986) 2154-2162.
19. Yakimiw E., Robert A.: Accuracy and stability analysis of a fully implicit scheme for the shallow water equations. Mon. Wea. Rev. 114 (1986) 240-244.

Bose-Einstein Condensation Studied by the Real-Time Monte Carlo Simulation in the Frame of Java Applet

Monika Gall, Ryszard Kutner, Andrzej Majerowski, and Dariusz Żebrowski

Institute of Experimental Physics, Department of Physics, Warsaw University, Smyczkowa 5/7, Pl-02678 Warsaw, Poland

Abstract. A Monte Carlo algorithm was constructed combined with a Java applet for the simulation of statistical physics quantities characterizing noninteracting bosons within micro- and macroscales. By this approach the Bose-Einstein condensate was considered within a three-dimensional isotropic harmonic oscillator in real-time. The algorithm can be used to study both the static properties of ideal bosons within other trapping potentials and the relaxation of the system to the condensate. The algorithm can be extended to cover also collisions between bosons. Conluding, our approach can be used for studying and visualizing both educational and professional problems regarding quantum statistical physics of bosonic systems.

1 Introduction

In educational and professional computational physics lattice gas models have attracted much attention since they can be successfully used in Monte Carlo simulations [1], particularly in real-time computer experiments. The principal goal of this paper is to show the main possibilities of our software to study, within micro- and macroscales, the noninteracting bosons. We show that it can mimic the quantum statistical physics features of particles by using standard MC simulations and a lattice gas model provided with a peculiar requirement (cf. Sec.2). We trace the consequences of this mimic, mainly by simulation of the Bose-Einstein condensate (BEC), which can be treated as a macroscopic system. Since below the transition temperature the ground level is occupied by a marcoscopic amount of bosons, the situation can be considered within the macroscale in distinction from the case where the temperature exceeds the transition temperature. Here, a microscopic amount of bosons occupy each level which defines the case that should be considered within the microscale.

By macroscopic occupancy of the ground energy level we understand that the ratio $\langle N_0\rangle/N$ is larger than zero even at the limit where the total amount of bosons N in the system (which possesses infinitely many energy levels) increases to infinity; of course, the normalization $\sum_{h=0}^{\infty}\langle N_h\rangle/N = 1$ is always obeyed, where $\langle N_h\rangle$, $h = 0, 1, 2, \ldots$, is the average number of bosons occupying the energy level h. We can say that the occupancy of any energy level is microscopic

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 673–678, 2004.
© Springer-Verlag Berlin Heidelberg 2004

if $\lim_{N\to\infty}\langle N_h\rangle/N = 0$. Of course, in our numerical experiments performed within finite systems we can observe only the finger print of macroscopic and microscopic energy level occupancies.

We hope that our MC algorithm offers a complementary possibility for studying statistical and thermal physics of Bose-Einstein condensate (i.e. its static and dynamic properties) both on educational and professional levels. Since 1995, when this exotic state of matter was experimentally observed in dilute atomic gases for the first time [2,3], this state is the subject of intense analytical, numerical, experimental and educational studies [4]. The present work is a direct continuation of our previous one [5] extended by including the interactive Java applet, which is a modern, platform-independent programming technology useful both for educational and professional purposes (our applet is working under the Java 2 version 1.4.1.1 or higher).

2 Model and Simulation Procedure

We define the model together with the algorithm for Monte Carlo simulation. The algorithm allows us to calculate relevant quantities within different statistical ensembles [6,7]. This algorithm makes possible to study both static and dynamic statistical physics quantities such as ground- and excited state occupancies of energy levels by bosons and the corresponding fluctuations, internal energy of the system, its entropy and specific heat. Thus, it provides an approach to consider thermal and statistical physics of quantum (and classical) gases in equilibrium and non-equilibrium states both on educational and professional levels. It should be noted that none of the known numerical methods has so wide possibilities. For example, by using partition functions one can calculate only equilibrium quantities [6,8]; an alternative MC simulation of BEC for ideal bosonic gas but enclosed in a rigid box was performed in [9].

In the case of a classic lattice gas identical (point) particles are considered to be distinguishable and any number of particles in the same site can be understood as a single-particle quantum-mechanical state. However, the classical particles require in some sense an even more rafined treatment than the quantum ones since they are genuinely indistinguishable when occupying the same single-particle quantum-mechanical state but otherwise they are entirely distinguishable and can then be treated as entirely independent particles. Such particles obey the Boltzmann statistics which is quantum-mechanically incorrect but can serve, e.g., as a reference high-temperature case.

The basic feature of our lattice gas model consists in that it mimics quantum indistinguishability, where the net result of a simple interchange of two identical particles is that no new state of the whole lattice gas is obtained. In distinction from the classical lattice gas, it does not matter which particle is in a given single-particle state, but only how many particles are in this state. The simplest example considered below should make this general idea clearer and make possible the construction of an algorithm.

2.1 Strategy of the Algorithm

Let us consider, for example, a lattice gas consisting of only two identical particles and call them 1 and 2, which is an auxiliary numeration which helps us to describe the idea of our algorithm, (we also consider in higher dimensions an equivalent algorithm where no auxiliary numbering of particles is necessary). We denote two lattice sites, i.e. different single–particle states as $h = 0$ and $h = 1$. According to quantum indistinguishability, the configuration where particle 1 is at state $h = 0$ and particle 2 is at state $h = 1$ is no longer counted as distinct from the configuration where these two particles are interchanged. Thus, we see the **basic step** of our algorithm since we can decide that particle number 1 never has a higher energy than particle number 2 (the opposite situation is disregarded). In other words, the sequence of particles is dynamically preserved during the whole simulation which defines the conservation principle which is a **generalization of the Pauli exclusion principle** valid for fermions or fermionic lattice gas. As it is seen, the space of states for this system is shrunk and the correlation between particles exist when the total number of particles in the system is fixed (e.g., for canonical or microcanonical ensembles). The above introduced conservation principle has already been used by us to study one-dimensional bosonic lattice gas (BLG) [10].

2.2 Local Dynamics

We confine our attention to spinless particles; for simplicity we assume that there is no mutual interaction between particles apart from the above introduced conservation principle which, however, can introduce correlations between them.

The initial particle configuration was randomly selected but other initial configurations are also accepted since the final equilibrium result does not depend on the choice of the initial condition.

We deal with a three-dimensional isotropic harmonic oscillator. Then each (energy) level, $h(= 0, 1, 2, \ldots)$, of the energetic ladder is degenerated and consists of $g(h) = \frac{1}{2}h(h + 3) + 1$ single-particle states (sites); the energy level spacing (which is fixed) is denoted as $\Delta\varepsilon$. Every state is defined here by the state-vector $\underline{h} = (h_1, h_2, h_3)$, where $h_1,\ h_2,\ h_3 \geq 0$, while $h = h_1 + h_2 + h_3$.

Note that two states can be connected here by a direct jump of a boson only if the chosen components of the initial and final states differ by ± 1. This configuration of states together with the above rule for jumps determine, in fact, the bosonic lattice gas defined by the three-dimensional lattice consisting of site-vectors having non-negative coordinates. As, for simplicity, particle jumps are assumed to occur only between the nearest-neighbor sites, there are no direct jumps between states belonging to the same energy level (since they would only slow down the relaxation).

From the strategy developed in Sec.2.1 it follows that one and only one particle from all $N_{\underline{h}}$ ones currently occupying the given single-particle state $\underline{h}$ performs an upward jump and it is unimportant which one, since the particles are indistinguishable in the quantum-mechanical sense. We have a similar situation

for any excited state when a downward jump is performed. The above defines the **strategy of an alternative algorithm**, particularly effective when energy levels are degenerated (as is the case of the three-dimensional isotropic harmonic oscillator). Namely, instead of using the above defined numbering of particles it is sufficient to choose at random a particle from an occupied (earlier chosen) state with probability $1/N_{\underline{h}}$ and then determine the direction of the jump. It should be noted that we choose an occupied state simply by drawing a particle; hence, the probability that a state is chosen is proportional to the current number of particles occupying that state. The above procedure defines the most effective algorithm (since no Monte Carlo step is lost) which defines the relaxation process to the bosonic condensate.

The particles perform thermally activated jumps between the lattice sites since the lattice gas is coupled to a heat bath (canonical ensemble). To define the local dynamics the jump (transition) rates $\Gamma_\uparrow$ and $\Gamma_\downarrow$ are assumed which obey the so-called detailed balance condition for the jump rates: $\Gamma_\downarrow/\Gamma_\uparrow = \exp(\Delta\varepsilon/k_BT)$. In fact, this condition is necessary to prove that the classical, non–interacting particles obey the Boltzmann statistics. Since this condition does not uniquely define the jump rates we can assume their simplest form $\Gamma_{\downarrow\uparrow} \sim \exp(\pm\Delta\varepsilon/2k_BT)$ as energy levels are equally distant; this leads to the following jump (transition) probabilities: $p_{\downarrow\uparrow} = p_0 \exp(\pm\Delta\varepsilon/2k_BT)$, where $p_0 = \frac{1}{2}\frac{1}{\cosh(\Delta\varepsilon/2k_BT)}$ is the normalization factor. Of course, other choices of jump rates (obeying a detailed balance condition) have no influence on the equilibrium quantities; they could play a role when one would study dynamic properties, e.g. relaxation in the system.

We assume the blocking boundary condition in the vertical direction although the length of the energetic ladder, L, and the ratio $\Delta\varepsilon/k_BT$ are limited so that in statistical equilibrium the macroscopic part of jumping particles, in principle, does not reach the top of the ladder, i.e. $(p_\uparrow)^L \ll p_\downarrow$.

3 Results

We observed that already the system including several hundreds of particles reproduces quite well the Bose-Eintein condensation or λ-transition. Hence, it is possible to obtain the condensation and the corresponding phase diagram in real time by the Monte Carlo simulation within the interactive Java applet, cf. Fig.1, where temporal distribution of bosons among different energy levels is also shown (e.g., for $k_BT/\Delta\varepsilon = 4.8$) to directly show the condensation phenomenon (here below the relative transition temperature $k_BT_C/\Delta\varepsilon \approx 6.9$).

By using the applet we are able to calculate and show the characteristic thermodynamic quantities such as boson occupancies and fluctuations, internal energy, entropy and the specific heat of the bosonic system as a function of temperature. We are able to show (with good approximation) that internal energy and entropy are continuous but non-differentiable functions of the relative temperature $k_BT/\Delta\varepsilon$ at the transition temperature. Hence, we show that the specific heat has characteristic discontinuity at this temperature which is a

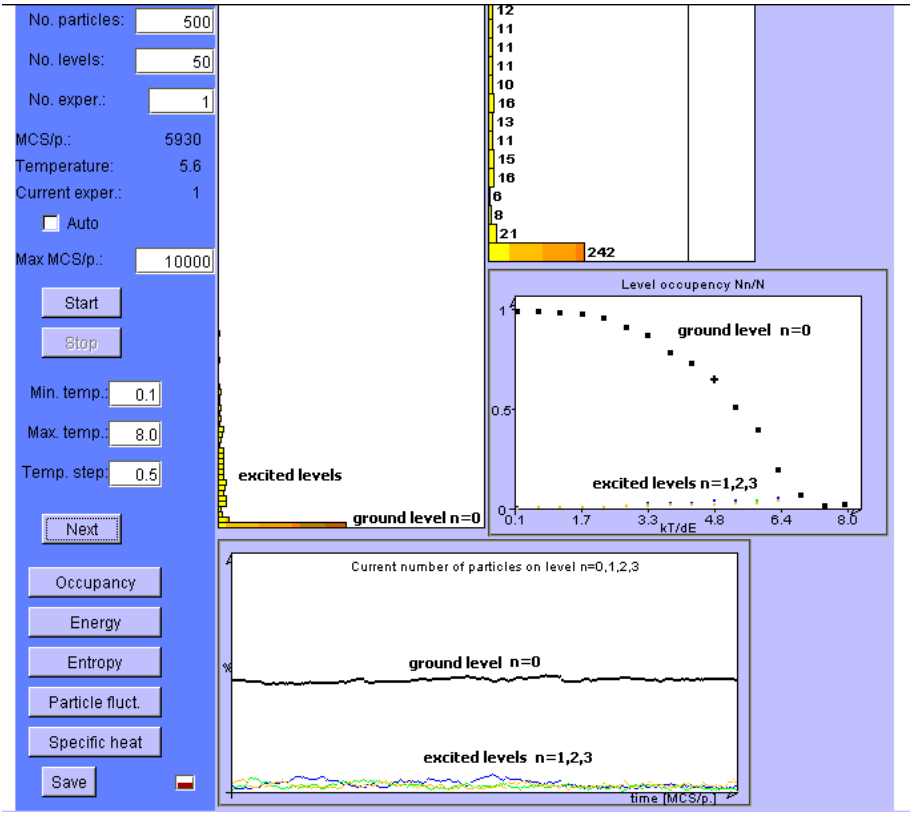


Fig. 1. Screen's picture given by our applet: (i) the phase diagram of the Bose-Einstein condensate (denoted by large black squares shown in the plot placed in the middle right window) i.e. the equilibrium ground-level occupancy N_0/N vs. relative temperature, $k_BT/\Delta\varepsilon$, for example, for $N = 500$ bosons in the system obtained in real time by our interactive Java applet (within a single experiment but with sufficiently good accuracy). (ii) The equilibrium distribution of bosons among energy levels obtained after a sufficiently long time (counted by Monte Carlo steps/particle) is shown (by horizontal indicators) in the diagram placed in the left middle window; its main part, including occupancies of the first 15 energy levels, is placed in the upper window. (iii) The lower window presents occupancies of the ground-level (the highest plotted curve) and three consecutive excited levels versus time (curves at the bottom); in this simple way we make it clear when the system reaches the statistical equilibrium, i.e. when there are no trends in the time-dependences. Temporal distributions considered in (ii) and (iii) relate to the condensate, for example, in the macrostate denoted in the phase diagram by plus (+).

clear finger print of λ-type phase transition. Moreover, there is no latent heat of the phase transition and therefore we can conclude that it is not of the first order. Note the pronounced λ-shape of the curve reminds the specific heat curve of He^4 near the transition to superfluidity.

4 Concluding Remarks

An algorithm is presented and its possibilities are discussed for studying the condensation phenomenon occurring within the bosonic lattice gas. Below the transintion temperature we observe the physical phenomenon characteristic for the macroscale as the macroscopic amount of bosons occupies the ground level in distinction from the situation where the temperature exceeds the transition one; then, a microscopic amount of bosons occupy each energy level. Applying our Monte Carlo simulations we are able to study statistical physics and thermodynamics of the Bose-Einstein condensate which could also have meaning for studying similar aspects of superfluidity. The approach well illustrates the properties of the bosonic lattice gas and can help in understanding the open problems concerning, e.g., the dynamical behavior of bosonic systems. Our approach is promising also for education since already small systems consisting of a few hundreds of lattice bosons, reproduces quite well the characteristic properties of macroscopic systems.

References

1. Landau D.P., Binder K., A Guide to Monte Carlo Simulations in Statistical Physics. Springer-Verlag, Berlin (2000)
2. Anderson M.H., Ensher J.R., Matthews M.R., Wieman C.E., Cornell E.A.: Observation of Bose-Einstein Condensation in a Dilute Atomic Vapor. Science **269** (1995) 198-201.
3. Townsend Ch., Ketterle W., Stringari S.: Bose-Einstein condensation. Physics World **10** No 3 (1997) 29-34.
4. Bardou F., Bouchaud J.-P., Aspect A., Cohen-Tannaoudji C.: Lévy Statistics and Laser Cooling. How Rare Events Bring Atoms to Rest. Cambridge Univ. Press, Cambridge (2002)
5. Kutner R., Regulski M.: Bose-Einstein condensation shown by Monte Carlo simulation. Comp. Physics Comm. **121-122** (1999) 586-590
6. Wilkens M., Weiss Ch.: Particle number fluctuations in an ideal Bose gas. J. Mod. Optics **44** (1997) 1801-1814.
7. Navez P., Bitouk D., Gajda M., Idziaszek Z., Rzążewski K.: Fourth Statistical Ensemble for the Bose-Einstein Condensate. Phys. Rev. Lett. **79** (1997) 1789-1792.
8. Reif F.: Fundamentals of Statistical and Thermal Physics. McGraw–Hill, New York (1965)
9. Gould H., Spornick L., J. Tobochnik J.: Thermal and Statistical Physics Simulations. CULPS. J. Wiley & Sons, New York (1995)
10. Kutner R., Kehr K.W., Renz W., Przeniosło R.: Diffusion in a one-dimensional bosonic lattice gas. J. Phys. A: Math. and General **28** (1995) 923-940.

Life History Traits and Genome Structure: Aerobiosis and G+C Content in Bacteria

Jean R. Lobry

Université Claude Bernard - Lyon I
Laboratoire de Biométrie, Biologie Évolutive
CNRS UMR 5558 - INRIA Helix project
43 Bd 11/11/1918
F-69622 VILLEURBANNE CEDEX, FRANCE
look@my.home.page.invalid,
http://pbil.univ-lyon1.fr/members/lobry/

Abstract. Evolution is a tinkerer not an engineer: the term exaptation was coined to signify that old structures, that could be not significant in terms of fitness, get re-used when environmental conditions changed. Here I show that the average protein composition of G+C rich bacteria were exapted to the switch from anaerobic to aerobic conditions. Because the proteome composition is under the strong control of directional mutation pressure, this is an example of exaptation at the molecular level for which the underlying mechanism is documented.

1 Introduction

During the last 20 years, genomic sequence data have been produced in a exponential way, with a doubling time close to 18 months, reminiscent of Moore's law in computer sciences (Fig. 1). We don't know whether this is just an anecdotical coincidence or evidence that the limiting factor for the production of genomic sequence data is related to computer performances, although the latter interpretation is my favorite given the perpetual struggle for disk space we are facing in my laboratory just to store primary data.

Whatever the underlying reason for this doubling time, we have a huge amount of data available and the problem is how to make sense from this. This paper is an example, admitly modest, of what is called data mining, or post-mortem data analysis, in which I have used previously published results to interpret them my own way. This paper is basically an attempt to make a connection between two previously published results that are summarized thereafter to provide background material.

1.1 Some Biological Terms

– Bacteria: this is a subset of living organisms on Earth. It is used here in its broad sense (*i.e.* Archae + Eubacteria) to designate small unicellular organisms without complex subcellular structures such as a nucleus.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 679–686, 2004.
© Springer-Verlag Berlin Heidelberg 2004

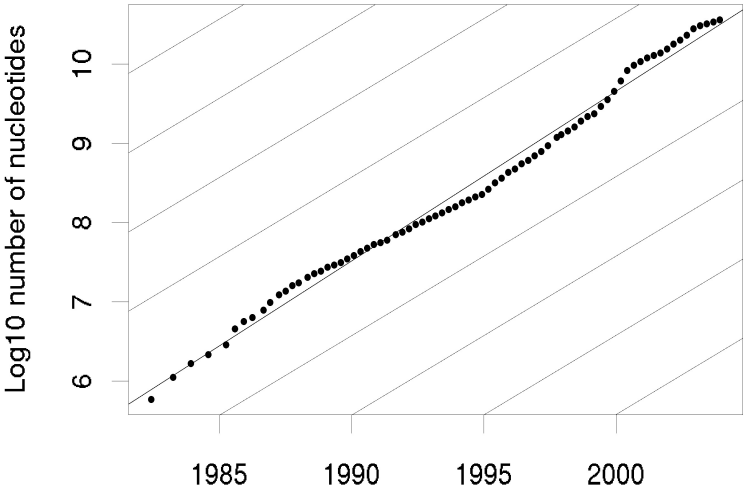


Fig. 1. The exponential growth of genomic sequence data mimics Moore's law. The source of data is the december 2003 release note (realnote.txt) from the EMBL database available at *http://www.ebi.ac.uk/*. External lines correspond to what would be expected with a doubling time of 18 months. The central line through points is the best least square fit, corresponding to a doubling time of 16.9 months.

- Aerobic bacteria: is used here to designate bacteria that can live only in presence of oxygen in their environment.
- Anaerobic bacteria: is used here to designate bacteria that can live only in absence of oxygen in their environment.
- Exaptation. Modern evolutionary theories are all based on Markov processes in which the future is influenced by the past only through the present state. To avoid using the term preadaptation, that would be misleading in this context as it may suggest some kind of knowledge of the future, Gould and Vrba have introduced the term exaptation [1]. This term is used to designate features of organisms that are non-adapted, but available for useful cooptation in descendants.

1.2 Genomic G+C Content and Aerobiosis

The G+C content is an example of global genomic structure that was used early in bacterial taxonomy, *i.e.* before the genomic era, because it was possible to estimate its value experimentally without knowing the sequence of a genome. The G+C content of bacterial chromosomes is the molar ratio of bases G and C over all bases, so that we could express this by the following lines of R implementation [2] of the S language:

```
> urn <- c("A", "C", "G", "T") # the bases in DNA
> dna <- sample(urn, size = 1000, replace = TRUE) # simulated DNA
```

```
> gc.content <- function(dna) {
    length(dna[dna == "G" | dna == "C"])/length(dna)
}
> gc.content(dna) # should be close to 0.5 in this case
[1] 0.467
```

Observed values range from 0.25 to 0.75 in bacteria and this was interpreted early as the result of differences in mutation rates between AT and GC pairs (see [3] and references therein). In the last 40 years, all the attempts to find an adaptative value for the G+C content of bacterial chromosome have failed, for instance there is no connection with the optimum growth temperature of bacteria [4] despite one may have expected from the extra hydrogen bond in GC pairs as compared to AT pairs.

Recently, Naya *et. al.* showed [5] that the G+C content is undoubtedly higher in aerobic bacteria than in anaerobic bacteria, linking for the first time a genome structure and a life history trait, and then raising the exciting possibility of a non-zero impact of the genomic G+C content on the cell fitness in bacteria. This kind of relationship between a genome structure and a life history trait is typically what makes sense for biologists because they are always looking for evidences of adaptation. Strickly speaking aerobiosis is not a life history trait *per se* but through its consequences because growth in aerobic conditions is much more efficient than in anaerobic conditions, allowing for smaller generation times.

1.3 Protein Metabolic Cost in Aerobic Conditions

Recently, Akashi and Gojobori have shown [7] that proteins produced in high amounts (*e.g.* ribosomal proteins) tend to avoid amino-acids that are expensive in terms of metabolic cost in aerobic conditions. This is an evidence that amino-acid composition of proteins is under the control of natural selection to enhance metabolic efficiency. On an other hand, the influence of the G+C content on the average amino-acid composition of proteins has been documented for a long time (see [6] and references therein): in G+C rich genomes, the encoded proteins tend to use amino-acid that are coded by G+C rich codons. Three groups of amino-acids can be defined to reflect their expected dependence on the G+C content [6]. A visual representation of Akashi and Gojobori data [7] taking into account these three groups of amino-acids is given in Fig. 2 that was generated with the following S code:

```
cost <- list(Ile = 32.3, Phe = 52.0, Lys = 30.3, Tyr = 50.0,
Asn = 14.7, Leu = 27.3, Met = 34.3, Asp = 12.7, Glu = 15.3,
Ser = 11.7,  Val = 23.3, Thr = 18.7, His = 38.3, Gln = 16.3,
Cys = 24.7,  Trp = 74.3, Arg = 27.3, Ala = 11.7, Pro = 20.3,
Gly = 11.7)gc.groups <- factor(rep(c(1, 2, 3), c(6, 10, 4)),
ordered = TRUE, label = c("low G+C", "mid G+C", "high G+C"))
stripchart(unlist(cost)~gc.groups, pch = 19, ylim = c(0.5,3.5),
xlim = c(0, max(unlist(cost))) )bx <- boxplot(unlist(cost)
~gc.groups, horizontal = TRUE, add = TRUE)
```

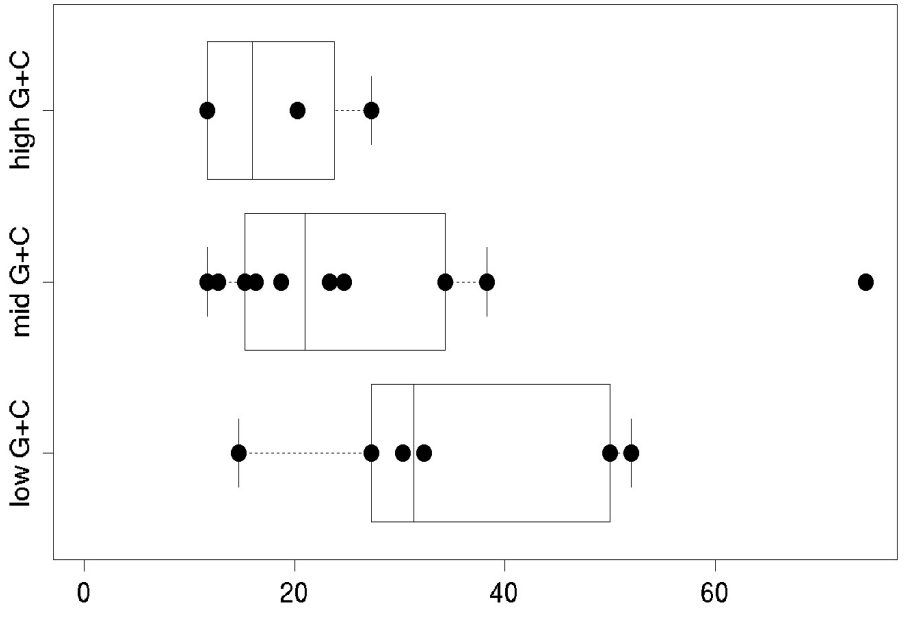


Fig. 2. Metabolic cost of the 20 amino-acids expressed in high-energy phosphate bond equivalent, ∼P, per amino-acid. Data are from table 1 in [7]. The box-and-whisker plot is a simple summary of data: the box represents the first quartile, the median and the last quartile ; the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

Fig. 2 shows that there is a trend for amino-acids that are favoured in high G+C genome to be less expensive in term of metabolic cost than those favoured low G+C genomes. It is therefore tempting to connect this with the result from Naya *et. al.* showing [5] that the G+C content is higher in aerobic bacteria. This is however not sufficient to conclude because we have to take into account the frequencies of amino-acid in proteins. For instance, the rigth outlier in the middle G+C group corresponds to Trp (*i.e.* tryptophan) which is known to be one of the rarest amino-acid in proteins (*cf* for instance table 1 in [8]).

2 Material and Methods

2.1 Source of Data

The G+C content in 225 anaerobic and 326 aerobic bacteria is from [5] and was download from `http://oeg.fcien.edu.uy/` and copied at: `http://pbil.univ-lyon1.fr/R/donnees/gc02.txt`. The amino-acids metabolic costs are from table 1 in [7]. The amino-acid frequencies in the proteins of 293 bacteria are from [9] and are available at: `ftp://pbil.univ-lyon1.fr/pub/datasets/JAG2003/`. This dataset, based on GenBank [10] release 130 including daily updates on the date of 13-JUL-2002, contains 97,095,873 codon counts.

2.2 Data Analyses

All analyses were done under R [2]. R is an Open Source implementation of the S language and similar to the commercial implementation S-Plus. S is both a general programming language and an extensible interactive environment for data analysis and graphics. See `http://www.r-project.org` for information on the project and CRAN (the Comprehensive R Archive Network) `http://cran.r-project.org` for available software and packages.

The model that predicts the frequency of a given amino-acid, aa, as function of the G+C content, θ, in absence of selective constraints is defined by:

$$P(\theta, aa) = \frac{f(\theta, aa)}{8 - (1-\theta)^2(1+\theta)}$$

with $\theta \in [0, 1]$ and

$$f(\theta, aa) = \begin{cases} (1-\theta)^2(2-\theta) & \text{if } aa \in \{\text{Ile}\} \\ (1-\theta)^2 & \text{if } aa \in \{\text{Phe, Lys, Tyr, Asn}\} \\ 1-\theta^2 & \text{if } aa \in \{\text{Leu}\} \\ (1-\theta)^2\theta & \text{if } aa \in \{\text{Met}\} \\ (1-\theta)\theta & \text{if } aa \in \{\text{Asp, Glu, His, Gln, Cys}\} \\ 2(1-\theta)\theta & \text{if } aa \in \{\text{Val, Thr}\} \\ 3(1-\theta)\theta & \text{if } aa \in \{\text{Ser}\} \\ (1-\theta)\theta^2 & \text{if } aa \in \{\text{Trp}\} \\ \theta(\theta+1) & \text{if } aa \in \{\text{Arg}\} \\ 2\theta^2 & \text{if } aa \in \{\text{Gly, Pro, Ala}\} \end{cases}$$

This is a simple probabilistic model in which coding sequences are generated by random sampling from a DNA urn with a given G+C content. The numerator reflects the structure of the genetic code and the denominator is a correcting factor due to stop codons (see [6] for details).

To allow for the reproducibility of the results presented here, the R source code that was used to produce Fig. 3 is available at the URL: `http://pbil.univ- lyon1.fr/members/lobry/exapt/fig3.R`. If you don't have R at hand you can copy and paste this script in our RWeb interface at the following URL: `http://pbil.univ-lyon1.fr/Rweb/Rweb.general.html`.

3 Results and Discussion

3.1 Results

Results are summarized by Fig. 3 which is divided into two panels sharing as common x-axis the genomic G+C content, ranging from 0.25 to 0.75, as expected in bacteria.

The bottom panel of Fig.3 recalls Naya *et. al.* recent breakthrough [5]: the G+C content is higher in aerobic bacteria (on the right) than in anaerobic bacteria (on the left). A direct representation of data, with a small amount of noise

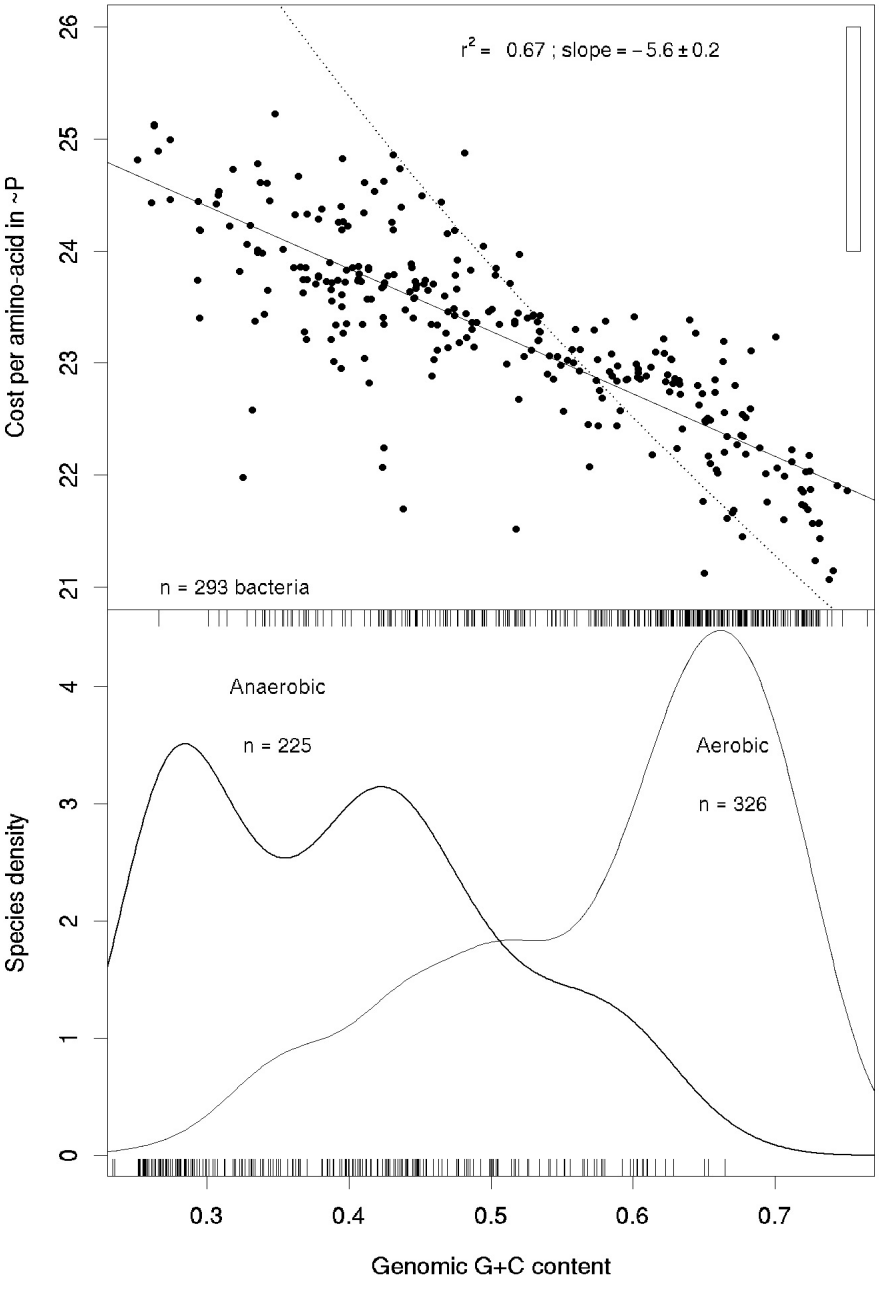


Fig. 3. Decrease of the average protein aerobic cost and distribution of anaerobic and aerobic bacteria with G+C content. See text for explanations.

added to break ties, is given on the top for aerobic species and at the bottom for anaerobic species. Although the two distributions are overlapping, there is clearly a trend for aerobic species to be G+C rich.

The top panel of Fig. 3 has a common y-scale expressed in aerobic metabolic cost (in high-energy phosphate bond equivalent, ∼P, per amino-acid) for the four following items:

1. The bar on the top right gives the range of observed selective effects [7] between proteins with a $\sim 10^5$-fold difference in terms of intracellular concentration. This bar is important to show a biologically relevant scale: a 2 ∼P per amino-acid difference in aerobic cost is enough to be selected in highly expressed genes. This bar gives also an idea of the within-species variability for the average protein aerobic cost.
2. The doted line represents what would be the average protein aerobic cost if protein composition was under the sole control of directional mutation pressure [6]. This model shows that under neutral conditions, if there were no selective constraints on the average amino-acid composition of protein, there would be an interest of being G+C rich under aerobic conditions because the cost decreases significantly (as compared to the the reference bar) when the G+C content increases. Note that the observed trend has a lower intensity, as expected, because the average protein composition is not completely free of selective constraints, so that the model is not realist.
3. The points represent the average (uniform protein weighting) aerobic cost for 293 bacteria, whose protein composition was deduced from a previously described dataset [9]. The actual average cost for a cell is expected to be lower because the uniform protein weighting is not realist. We should weight individual protein composition to take into account their intra-cellular concentrations in the cell, but this information is not available.
4. The line is the best least-squares fit: there is a significant decrease from low G+C to high G+C bacteria: from 24.7 to 21.9 ∼P per amino-acid. This 2.8 ∼P per amino-acid variation compares well the within-bacteria variation between highly and poorly expressed proteins [7] depicted by the reference bar.

3.2 Discussion

As noted by an anonymous referee of this paper, and I would like to take this opportunity to thanks him/her for valuable suggestions, I have assumed in the following discussion thatdirectional mutational pressure (responsible for generation of G+Ccontent) is free from selection. However, we can not exclude the possibility that themutational pressure is subjected in some way to selection (*e.g.* repairsystems may be selected to prefer some mutational defects than other leading tocomposition bias). I think this is unlikely, but even if this was true we would still have an example of exaptation at the molecular level. Features coopted as exaptations have two possible previous statuses. They may have been

adaptations for another function, or they may have been non-adaptative features (*cf* section VI C in [1]).

It would be tempting to connect the top and the bottom of Fig. 3, assuming that no confounding factor is present, by a simple regular selective scenario: aerobic low cost amino acid are encoded by G+C rich codons so that the selection for low cost amino-acids at the proteome level has induced a G+C enrichment in coding sequences. This is, however, not defendable because in G+C rich bacteria the whole genome, including non-coding regions and synonymous positions, are also enriched in G+C content (*cf* [3] and references therein). The selective advantage results from the long-term effects of a directional mutation pressure. This is an example of exaptation at the molecular level: having a high G+C content is interesting under aerobic conditions, but this was unforeseeable before the oxygen concentration was enough on Earth.

References

1. Gould, S.J., Vrba, E.S.: Exaptation-A missing term in the science of form. Paleobiology **8** (1982) 4–15
2. Ihaka, R., Gentleman, R.: R: A Language for Data Analysis and Graphics. J. Comp. Graph. Stat. **3** (1996) 299–314
3. Lobry, J.R., Sueoka, N.: Asymmetric directional mutation pressures in bacteria. Genome Biology **3** (2002) 58.1–58.14
4. Galtier, N., Lobry, J.R.: Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. **44** (1997) 632–636
5. Naya, H., Romero, H., Zavala, A., Alvarez, B., Musto, H.: Aerobiosis increases the genomic guanine plus cytosine content (GC %) in prokaryotes. J. Mol. Evol. **55** (2002) 260–264
6. Lobry, J.R.: Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene **205** (1997) 309–316
7. Akashi, H., Gojobori, T.: Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **99** (2002) 3695–3700
8. Lobry, J.R., Gautier, C.: Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucl. Acids Res. **22** (1994) 3174–3180
9. Lobry, J.R., Chessel, D.: Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J. Appl. Genet. **44** (2003) 235–261
10. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: GenBank. Nucl. Acids Res. **30** (2002) 17–20

Differential Gene Survival under Asymmetric Directional Mutational Pressure

Paweł Mackiewicz[1], Małgorzata Dudkiewicz[1], Maria Kowalczuk[1], Dorota Mackiewicz[1], Joanna Banaszak[1], Natalia Polak[1], Kamila Smolarczyk[1], Aleksandra Nowicka[1], Mirosław R. Dudek[2], and Stanisław Cebrat[1⋆]

[1] Department of Genetics, Institute of Microbiology, University of Wroclaw, ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
{malgosia, pamac, nowicka, kowal, dorota, polak, smolar, cebrat}@microb.uni.wroc.pl
http://smORFland.microb.uni.wroc.pl

[2] Institute of Physics, University of Zielona Góra, ul. A. Szafrana 4a, PL-65516 Zielona Góra, Poland
mdudek@proton.if.uz.zgora.pl

Abstract. We have simulated, using Monte Carlo methods, the survival of prokaryotic genes under directional mutational pressure. We have found that the whole pool of genes located on the leading DNA strand differs from that located on the lagging DNA strand and from the subclass of genes coding for ribosomal proteins. The best strategy for most of the non-ribosomal genes is to change the direction of the mutational pressure from time to time or to stay at their recent position. Genes coding for ribosomal proteins do not profit to such an extent from switching the directional pressure which seems to explain their extremely conserved positions on the prokaryotic chromosomes.

1 Introduction

Most of the natural DNA sequences are asymmetric. There are two main mechanisms introducing DNA asymmetry: the replication-associated directional mutational pressure and the selection for protein coding sequences (see for review: [1], [2]). The replication-associated mutational pressure generates some kind of a global asymmetry between the two strands called the leading and the lagging DNA strands. On the other hand, the selection for coding sequences generates a local asymmetry between sense (coding) and anti-sense (complementary to the sense) strands of genes. This asymmetry results from the coding function requirement of genes. Thus, as in the case of two chiral molecules, the two possible ways of superposition of a coding sequence on the asymmetric bacterial chromosome are not equivalent. For example, if the sense strand of a gene located on the leading strand has more G than C, and C is more often substituted by other

⋆ To whom all correspondence should be sent.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 687–693, 2004.
© Springer-Verlag Berlin Heidelberg 2004

nucleotides than G on the leading strand, then inversion of this sequence, which transfers the C-rich anti-sense strand of the gene to the leading strand, would increase the mutation rate of the gene. Thus, a gene sequence remaining for a long time on one DNA strand tends to acquire some asymmetry characteristic for the mutational pressure while sequences occasionally inverted oscillate between the two compositional stages and their composition depends on the time which they spend on each strand and on how frequent they are translocated. In this paper we have simulated the effect of changing the mutational pressure on the gene survival.

2 Methods

Simulations have been performed on 564 leading strand genes and 286 lagging strand genes from the *Borrelia burgdorferi* genome [3], whose sequence and annotations were downloaded from GenBank. The replication-associated mutational pressure (RAMP) describing the nucleotide substitution frequencies has been parameterized as described by Kowalczuk et al. [4]. The matrix describing RAMP of the lagging strand is the mirror reflection of the RAMP for the leading DNA strand. In one Monte Carlo Step (MCS) each nucleotide of the gene sequence (its sense strand) was drawn with a probability $p_{mut} = 0.01$, then substituted by another nucleotide with the probability described by the corresponding parameter in the substitution matrix. After each round of mutations, we translated the nucleotide sequences into the amino acid sequences and compared the resulting composition of the proteins with the original. For each gene we calculated the selection parameter (T) for the amino acid composition which is the sum of absolute values of differences between fractions of amino acids as follows:

$$T = \sum_{i=1}^{20} |f_i(0) - f_i(t)|, \tag{1}$$

where: $f_i(0)$ is a fraction of a given amino-acid in the original sequence (before mutations) and $f_i(t)$ is a fraction of a given amino acid in the sequence after mutations in t MCS. If T was below the assumed threshold, a gene stayed mutated and went to the next round of mutations (the next MC step). If T trespassed the threshold - the gene was "killed" and replaced by its allele from the second genomic sequence, originally identical, simulated parallely. As a value of the threshold we have assumed the average value T between pairs of orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum* which equals 0.3. All simulations were performed for 1000 Monte Carlo steps, repeated 100 times and averaged. For comparison, the numbers of killed genes from different sets were normalized by the number of genes in the given set. In the simulations we have applied both stable and changing replication associated mutational pressure (RAMP). Stable RAMP means that during the whole simulation genes were subjected only to one pressure characteristic for the leading or the lagging strand. In the simulations with changing RAMP genes were

alternately under the RAMP characteristic for the leading or the lagging DNA strand, changing with different frequencies. These simulations were carried out in different conditions described by the two parameters: F - the fraction of MC steps during the whole simulation in which the genes were subjected to mutational pressure characteristic for the strand on which they are normally located in the genome, N - Number of switches of the RAMP from leading to lagging one or *vice versa*. In sum, we have analyzed 87 different conditions of RAMP changing (different combinations of values F and N).

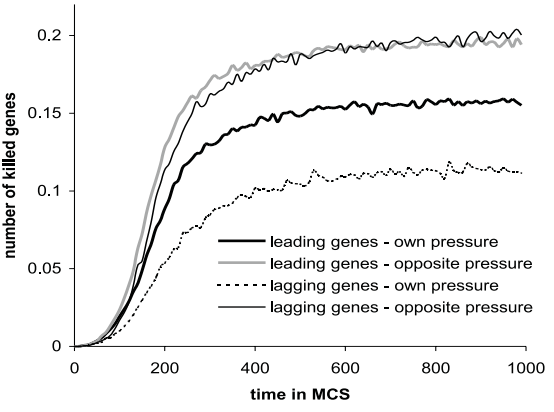


Fig. 1. The normalized number of killed genes from the leading and lagging strands of the *B. burgdorferi* genome. The genes were subjected to mutational pressure characteristic for them (their own pressure) and the mutational pressure characteristic for the complementary DNA strand (the opposite pressure)

3 Results and Discussion

After simulations of genes subjected to stable mutational pressure we found that (Fig. 1): (i) The effect of killing grew in time and approximated to a relatively high level. (ii) The killing effect for the genes staying under their own pressure is higher for the leading strand genes than for the lagging strand genes. (iii) Both sets of genes are better adapted to the mutational pressure characteristic for their recent positions in the genome than to the pressure from the opposite strand. Furthermore, the killing effect under the opposite RAMP is equally deleterious for both sets of genes.

In the earlier studies we have found that frequent changes of RAMP could be the best general strategy for gene survival [5]. In the present studies we are showing the relationship between the frequency of gene transpositions (inversions)

between differently replicating DNA strands and their survival. The diagram in Fig. 2 shows which percent of a given set of genes has the highest survival chance under one of the 87 combinations of tested parameters (F and N) after 1000 MCS of simulation. Generally, genes prefer to stay longer under the RAMP to which they are actually subjected, but there are no preferred positions for the ribosomal genes located in the *B. burgdorferi* genome on the leading strand.

In Fig. 3 we have presented how the number of killed genes depends on N for different F values. These analyses show that too frequent switching the direction of mutational pressure does not enhance significantly the gene survival. Usually switching every several hundreds of steps is close to the optimal gene survival. Relationship between the number of killed genes and F has a distinct minimum (Fig. 4). Ribosomal genes do not profit as much from switching their positions (data not shown).

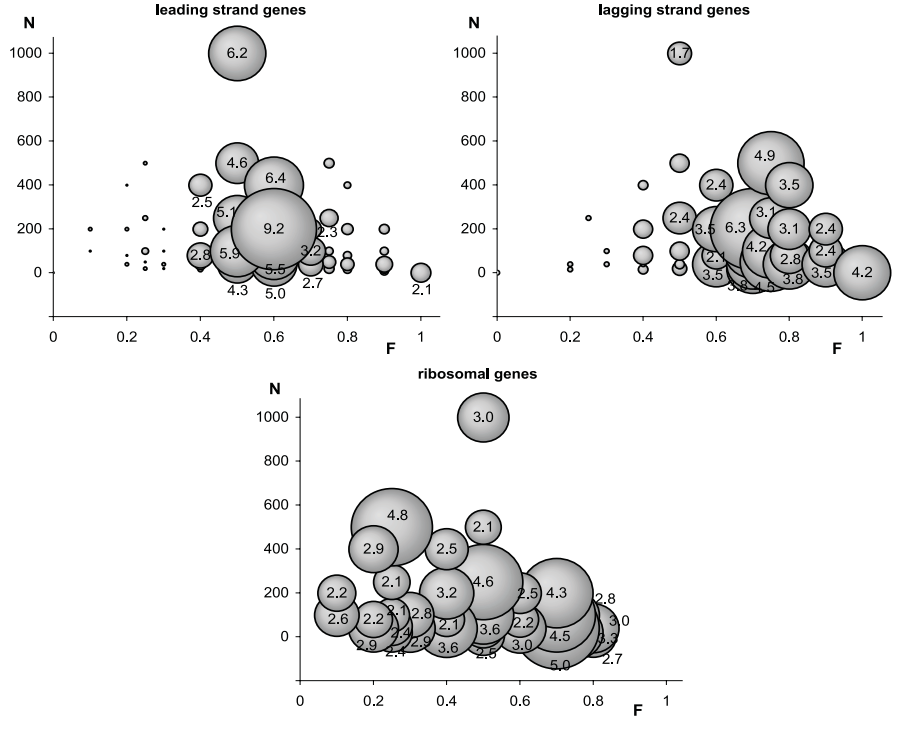


Fig. 2. Diagram presenting the best survival strategy for three sets of genes. This diagram shows which percent of a given set of genes has the highest survival chance under one of the 87 combinations of tested parameters (F and N) of changing mutational pressure after 1000 MCS of simulation

As it can be seen in Fig. 5 the number of accepted amino acid substitutions in coded proteins per site (substitutions which did not eliminate the gene function)

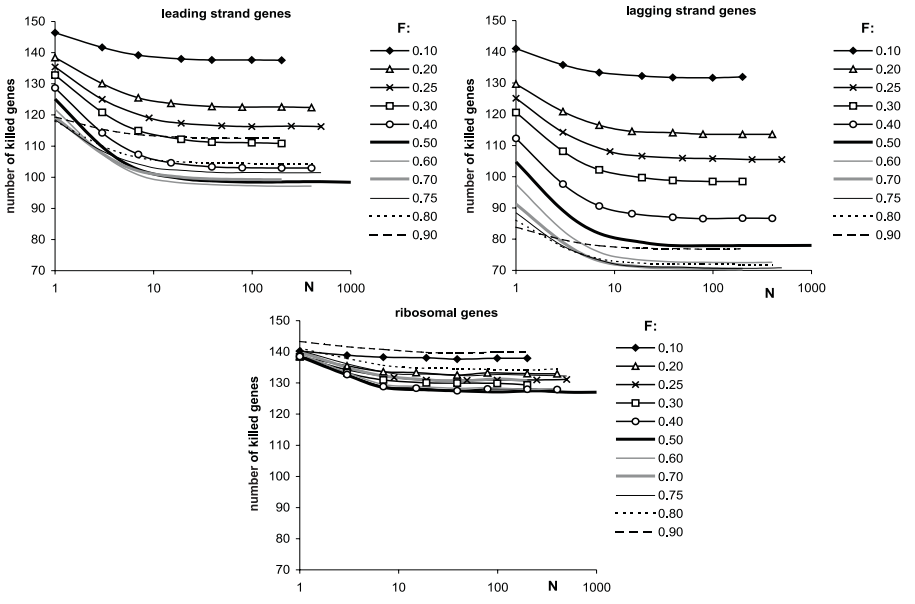


Fig. 3. Relationship between the number of killed genes and N for different F values for three sets of genes after 1000 MCS of simulation

is also higher. That means that the observed divergence of genes which recently changed their positions on chromosome should be higher, which was actually observed in numerous genomic analyses ([6] – [8]). In Fig. 5 it is also clear that the number of accepted substitutions is the lowest for the ribosomal proteins which are actually extremely conserved.

The last observations, these from simulations as well as from genome analyses lead to the conclusion that switching the direction of the mutational pressure does not diminish the total frequency of mutations but rather introduces intragenic suppression mutations which complement the former mutations in the same gene. Such intragenic suppression should be much more effective for longer genes (see accompanying paper). The ribosomal genes, in all the genomes analyzed thus far are usually located on the leading strand [9]. Our simulations have shown that they do not profit very much from transpositions (switching the mutational pressure) and the deleterious effect of the prolonged opposite mutational pressure is the same for the leading and lagging DNA strands. Since these genes are very intensively transcribed it is important for them to concert the direction of replication fork movement and the direction of transcription. This eliminates the possible deleterious effect of head on collisions of replication and transcription complexes ([10], [11]). The location of sense strands of these genes on the leading strand eliminates this effect.

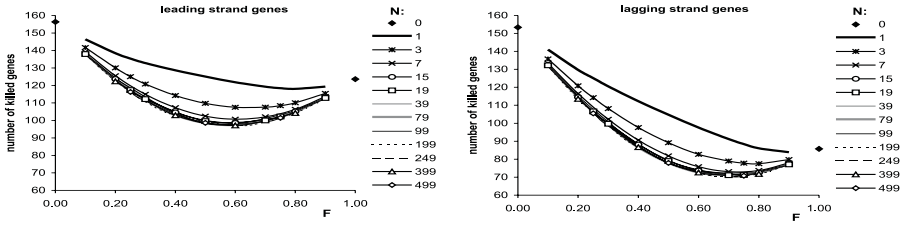


Fig. 4. Relationship between the number of killed genes and F for different N values for two sets of genes after 1000 MCS of simulation

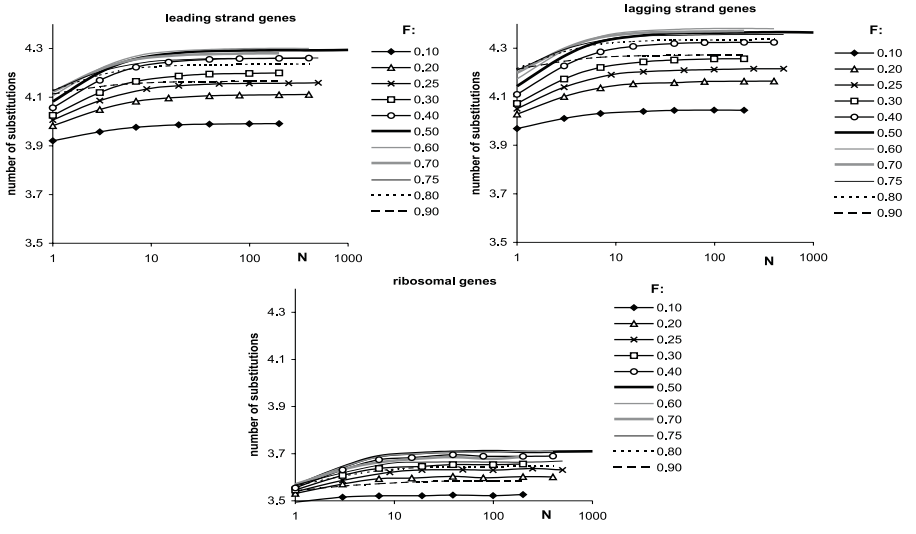


Fig. 5. Relationship between the number of accepted amino acid substitutions in coded proteins per site and N for different F values for three sets of genes after 1000 MCS of simulation

Acknowledgements. The work was supported by the grant number 1016/S/IMi/03 and is done in the frame of COST Action P10 program. M.K. was supported by the Foundation for Polish Science.

References

1. Frank, A.C., Lobry, J.R.: Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238** (1999) 65–77
2. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, A., Dudek, M.R., Cebrat, S.: DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. **42** (2001) 553–577

3. Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K. et al.: Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390** (1997) 580–586
4. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. **1** (2001) (1):13
5. Dudkiewicz, M., Mackiewicz, P., Nowicka, A., Kowalczuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Dudek, M.R., Cebrat, S.: Properties of Genetic Code under Directional, Asymmetric Mutational Pressure. Lect. Notes Comput. Sc. **2657** (2003) 343–350
6. Tillier, E.R., Collins, R.A.: Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. **51** (2000) 459–463
7. Szczepanik, D., Mackiewicz, P., Kowalczuk, M., Gierlik, A., Nowicka, A., Dudek, M.R., Cebrat, S.: Evolution rates of genes on leading and lagging DNA strands. J. Mol. Evol. **52** (2001) 426–433
8. Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. J. Appl. Genet. **44** (2003) 561– 584
9. McLean, M.J., Wolfe, K.H., Devine, K.M.: Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47** (1998) 691–696
10. Brewer, B.J.: When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell **53** (1988) 679–686
11. French, S.: Consequences of replication fork movement through transcription units *in vivo*. Science **258** (1992) 1362–1365

How Gene Survival Depends on Their Length

Natalia Polak[1], Joanna Banaszak[1], Paweł Mackiewicz[1], Małgorzata Dudkiewicz[1], Maria Kowalczuk[1], Dorota Mackiewicz[1], Kamila Smolarczyk[1], Aleksandra Nowicka[1], Mirosław R. Dudek[2], and Stanisław Cebrat[1⋆]

[1] Department of Genomics, Institute of Genetics and Microbiology, University of Wroclaw,
ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
{malgosia, pamac, nowicka, kowal, dorota, polak, smolar, cebrat}@microb.uni.wroc.pl
http://smORFland.microb.uni.wroc.pl

[2] Institute of Physics, University of Zielona Góra, ul. A. Szafrana 4a, PL-65516 Zielona Góra, Poland
mdudek@proton.if.uz.zgora.pl

Abstract. Gene survival depends on the mutational pressure acting on the gene sequences and selection pressure for the function of the gene products. While the probability of the occurrence of mutations inside genes depends roughly linearly on their length, the probability of elimination of their function does not grow linearly with the length because of the intragenic suppression effect. Furthermore, the probability of redefinition of the stop and start codons is independent of the gene length while shortening of gene sequences by generating stop codons inside gene sequences depends on gene length.

1 Introduction

One of many different mechanisms introducing mutations into genomes are single nucleotide substitutions which happen during DNA replication. There are four different kinds of nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) and substitution of one of them by any of the three others are random but highly biased. Some nucleotides are more often substituted than others and the substituting nucleotides are also unevenly "chosen" [1]. Thus, for each of the twelve possible kinds of nucleotide substitutions a specific probability of the event can be experimentally estimated and put into the "matrix of substitutions" (Tab. 1) [2].

The most stable genes should be built of the most stable nucleotides. On the other hand, the selection for gene function demands rather specific composition of the gene products which restricts not only the nucleotide composition of genes but, which is more important, the proper length of the coding sequence. A substitution inside the coding sequence can exert very different effects on the amino

⋆ To whom all correspondence should be sent.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 694–699, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Table 1. Frequencies of substitutions in the leading strand of the *B. burgdorferi* genome (nucleotide in a column is substituted by a nucleotide in a row). All substitution frequencies between different nucleotides sum up to 1

	A	T	G	C
A	0.300	0.103	0.067	0.023
T	0.066	0.480	0.035	0.035
G	0.164	0.116	0.149	0.015
C	0.070	0.261	0.047	0.073

acid sequence of its product. There are silent mutations which do not change the sense of the coding sequence, some substitutions change one amino acid in the gene product for another one though very similar, but some substitutions can change the properties of the coded amino acid significantly and such substitutions are potentially dangerous - they can lead to the nonfunctional gene product. The other dangerous point mutations are substitutions which eliminate the start or stop codons. These codons are responsible for initiation and termination of the protein synthesis, respectively (there is one start codon in the universal genetic code - ATG, and three stop codons: TAA, TAG, TGA). While elimination of the stop causes additional elongation of the coding sequence, elimination of the start could shorten it. Furthermore, start and stop codons could be generated inside the coding sequence. In that case, generation of stops is dangerous because it shortens the length of the gene product. The effect of generation of a start codon can be considered as another amino acid substitution, since the start codon inside the gene codes for methionine (ATG in the universal code). In this paper we analyze, using the Monte Carlo methods, the stability of the real genes of different length found in the *Borrelia burgdorferi* genome under the mutational pressure experimentally described for this genome.

2 Material and Methods

Simulations have been performed on 850 genes taken from the *B. burgdorferi* genome [3], whose sequence and annotations were downloaded from GenBank (*ftp://ftp.ncbi.nih.gov*). The gene sequences were subjected to the replication-associated mutational pressure (RAMP) described by the matrix of nucleotide substitution frequencies - Table 1 [2]. Since in this genome the RAMP is significantly different for the two differently replicating DNA strands: leading and lagging [4], we have applied two different matrices respectively for the genes located on these strands. The matrix describing RAMP of the lagging strand is the mirror reflection of the RAMP for the leading DNA strand. In one Monte Carlo Step (MCS) each nucleotide of the gene sequence was drawn with a probability $p_{mut} = 0.001$ then substituted by another nucleotide with the probability described by the corresponding parameter in the substitution matrix. We have applied two kinds of selection for gene survival: selection for the amino acid composition and selection for start and stop codons. After each round of mutations, we

translated the nucleotide sequences into the amino acid sequences and compared the resulting composition of the proteins with the original one. For each gene we calculated the selection parameter T for the amino acid composition as follows:

$$T = \sum_{i=1}^{20} |f_i(0) - f_i(t)|, \tag{1}$$

where: $f_i(0)$ is a fraction of a given amino-acid in the original sequence (before mutations) and $f_i(t)$ is a fraction of a given amino acid in the sequence after mutations in t MCS. If T was below the assumed threshold, the gene stayed mutated and went to the next round of mutations (the next MC step). If T trespassed the threshold - the gene was "killed" and replaced by its allele from the second genomic sequence, originally identical, simulated parallely.

We have applied three variants of selection for start and stop codons. Gene was killed when: (i) its start codon was substituted by a non-start codon, (ii) its stop codon was substituted by a non-stop codon, (iii) a stop codon was generated inside the gene sequence. In *B. burgdorferi* genes ATG, TTG, GTG are used as start codons and TAA, TAG, TGA - as stop codons. We have assumed that substitutions between these start codons and between these stop codons are neutral. After each MCS the number of gene replacements (the number of killed genes) was counted. All simulations were performed for 1000 Monte Carlo steps, repeated 10 times and averaged.

3 Results and Discussion

In the first simulations we have assumed that genes are "killed" only because of changes in the amino acid composition of their products. The results of the simulations done for the whole set of genes from the *B. burgdorferi* genome are shown in Fig. 1.

It is obvious that the number of mutations which hit the genes roughly linearly depends on their length. If the assumed tolerance is very low - even one substituted amino acid can eliminate the gene, the intragenic suppressions (complementation of mutation by consecutive substitutions) are very rare and are slightly more probable for longer genes. That is why in such conditions the probability of elimination of genes depends almost linearly on their length only slightly decreasing with length (Fig. 1a). When increasing the tolerance, the probability of intragenic suppression grows (that could be compared with the buffer capacity) and the probability of killing the gene decreases (Fig. 1b-d). The longer genes can deal with the mutational pressure more successfully than shorter ones. However, one can suspect that the effect of higher sensitivity of shorter genes is caused by the biased nucleotide composition of the short and long genes. To eliminate this effect, we have constructed artificial genes composed of different numbers of repetitions of the same unit being the coding sequence of one short *B. burgdorferi* gene. The observed effect was similar to that for the real set of genes though much more regular because the effect of nucleotide composition of different genes has been eliminated (Fig. 2).

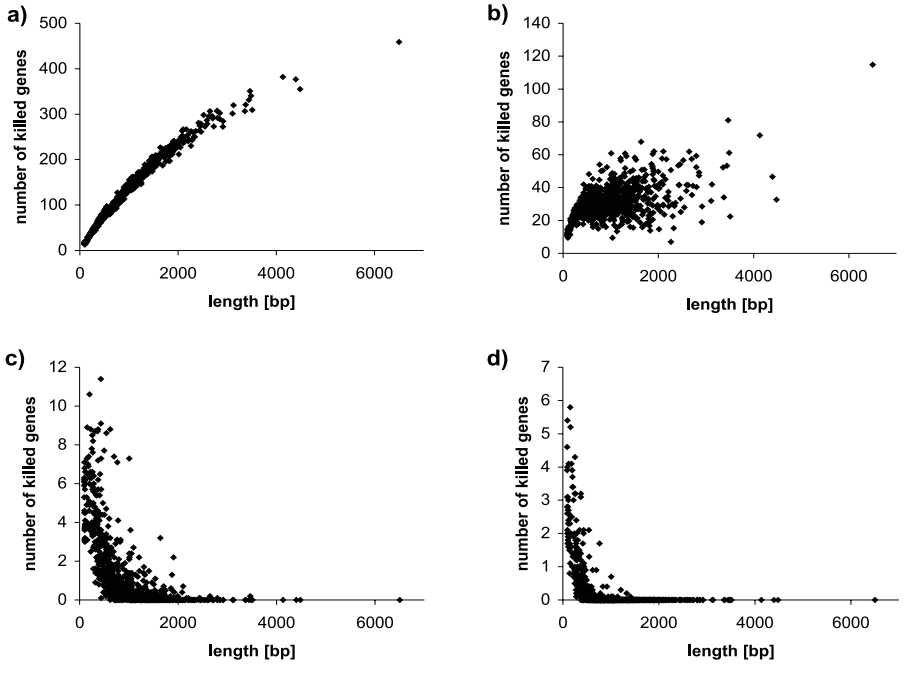


Fig. 1. Elimination of genes from *B. burgdorferi* genome by the selection pressure on their global amino acid composition. In all four cases the genes were under their specific mutational pressure but under different strength of selection pressure - different threshold T: a) $T = 0.01$; b) $T = 0.1$; c) $T = 0.25$; d) $T = 0.33$

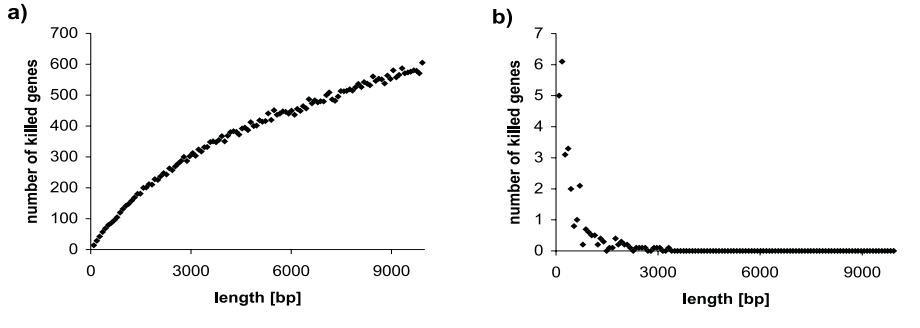


Fig. 2. Elimination of virtual genes of different length by the selection pressure on their global amino acid composition. All sequences - virtual genes - were produced by repeating different number of times one coding sequence of 31 codons long (deprived start and stop codons). These sequences were under their specific mutational but under different strength of selection pressure - different threshold T: a) $T = 0.01$; b) $T = 0.25$

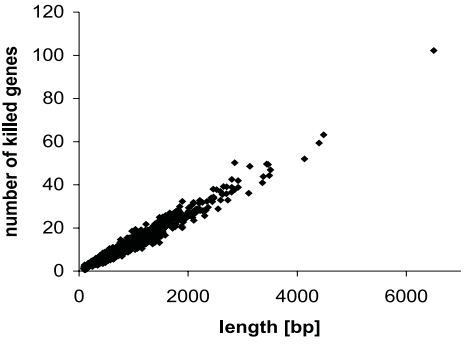


Fig. 3. Killing effect of the stop codon generation inside the coding sequences. Any other mechanisms of selections were switched off

In the next simulations we have eliminated genes when their start codon was substituted by a non-start codon. Since each gene has one start codon, it is obvious that the rate of gene elimination did not depend on the gene size (results not shown). A similar effect was observed for the elimination of stops. Elimination of the stop does not necessarily lead to the gene elimination because these mutations elongate the gene products. We have found that the frequency of stop codons usage in the *B. burgdorferi* genes corresponds almost exactly to the usage counted from the nucleotide composition of DNA in the equilibrium with the mutational pressure. This suggests that it is the mutational pressure which structures the nucleotide composition of stop codons and there is no selection pressure on specific stops usage. More dangerous for the gene function could be the shortening of the coding sequences by generation of stop codons inside the genes. Simulations of this effect have shown that the killing effect depended strongly on the length of the genes (Fig.3). Analytical calculations of the probability of generation of the stop codons inside genes (normalized per length unit) have shown that the generation of stops is not correlated with the gene length (Fig. 4). These results suggest that the selection pressure for the longer genes has not resulted in the decreasing the probability of generating stops - or the longer genes do not avoid codons which could be mutated to the stop codons with higher probability.

4 Conclusions

Simulations of the relationships between the genes' length and their survival have shown that while the short genes and the long ones are equally sensitive for killing by elimination the stop or start codons, the killing effect by amino acid substitutions seems to be relatively stronger for shorter genes, because of the lower probability of intragenic suppression. This effect can be compensated by

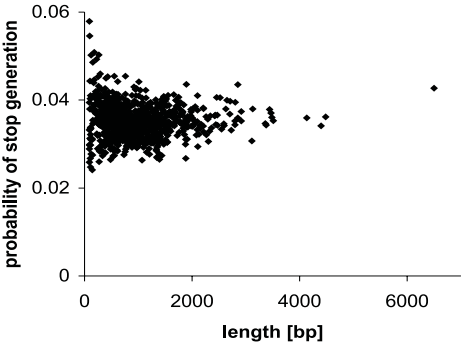


Fig. 4. Relationship between probability of the stop codon generation in genes of *B. burgdorferi* and their length. For each codon the probability of transformation into the stop codon were counted according to the matrix of nucleotide substitution. Then, for each gene the probabilities were weighted by the fractions of codons and summed

the effect of stops generation, which can not be suppressed by other intragenic mutations and the longer genes are more susceptible to such mutations. Since the effect of the mutations at the borders of the coding sequences concerns also mutations at the border of introns and exons in the eukaryotic genes, it seems that the price which these genes have to pay for increasing the probability of defects has to be compensated by other profits from using this risky and complicated coding strategy.

Acknowledgements. The work was done in the program COST Action P10 and supported by the grant number 1016/S/IMi/03. M.K. was supported by the Foundation for Polish Science.

References

1. Frank, A.C., Lobry, J.R., Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238** (1999) 65–77
2. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. **1** (2001) (1):13
3. Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K. et al.: Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390** (1997) 580–586
4. Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Łaszkiewicz, A., Dudek, M.R., Cebrat, S.: Replication associated mutational pressure generating long-range correlation in DNA. Physica A **314** (2002) 646–654

Super-tree Approach for Studying the Phylogeny of Prokaryotes: New Results on Completely Sequenced Genomes

Alexandra Calteau[1], Vincent Daubin[2], and Guy Perrieère[1]

[1] Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558,
Université Claude Bernard - Lyon 1, 43, bd. Du 11 Novembre 1918,
69622 Villeurbanne Cedex, France
{calteau, perriere}@biomserv.univ-lyon1.fr
http://pbil.univ-lyon1.fr
[2] Department of Biochemistry and Molecular Biophysics,
229 Life Sciences South, University of Arizona
Tucson, Arizona 85721, USA
daubin@email.arizona.edu

Abstract. In this paper we describe a super-tree approach that is able to use the information brought by hundreds of individual gene trees in order to establish a reliable phylogeny of prokaryotes. Indeed, it has been often shown that the use of single genes is not sufficient to solve such a complex case. This is mainly due to problems like reconstruction artifacts, existence of hidden paralogies or the presence of numerous lateral transfers. Here, our super-tree approach allowed us to build two trees containing 86 prokaryotic organisms. All the major bacterial and archeal taxonomic groups are monophyletic in these trees, which present some striking differences with the classical view of prokaryotic phylogeny represented by the ribosomal RNA trees.

1 Introduction

Many attempts have been made since the precursor work of Woese [1] in order to establish a reliable phylogeny of prokaryotic species (archaea and bacteria). For a long time, the different phylogenies produced using the Small Subunit of ribosomal RNA (SSU rRNA) were considered as the best reference to reveal the evolutionary history of the prokaryotic world [1]–[3]. On these trees, the only well-supported positions were the taxonomic groups located near the base, these groups being represented by hyperthermophilic bacteria, like those from the *Thermotoga* or *Aquifex* genera. The lack of resolution between the other divisions was interpreted as the proof of a rapid radiation of the organisms considered. This phenomenon has also been observed in many phylogenies based on protein genes [4]. But improvements realized in methodology led to the conclusion that some parts of Woese's trees were in fact reconstruction artifacts, this especially for the position of hyperthermophiles [5]–[7]. Moreover, it seems that even the supposed very stable rRNA genes could be horizontally transferred

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 700–708, 2004.
© Springer-Verlag Berlin Heidelberg 2004

between species, as some biological experiments suggest [8],[9]. Also, if the hypotheses about the massive occurrence of horizontal transfers between bacterial species are true (see [10]–[14] for many examples), then it seems difficult to use other gene markers to reconstruct the phylogeny of these organisms. Another problem is the fact that a lot of hidden paralogies exist among bacterial protein genes, and therefore risks of obtaining erroneous phylogenies are high when using some markers. Since the availability of an increasing number of complete genomes, several methods have been proposed to build genome-based trees, or to test whether this concept makes sense for bacterial species. The most common approaches are genes concatenation [15]–[17] and orthologous [18] or gene families content [19]–[21] comparisons among species. For concatenation methods, the problem is that the number of genes that are both shared by the different species studied and alignable is very limited (only 23 in the case of [16]). And for the comparison methods, it seems that they are only able to find the relationships between closely related organisms. In that context, we introduced a super-tree method based on the Matrix Representation using Parcimony (MRP) coding scheme [22],[23]. This method already allowed us to produce a tree of life containing 45 species for wich complete genomes were available [24]. At that moment some important bacterial divisions were not represented in our data set (mainly Fusobacteria and Chlorobiales), and incertitudes remained for some parts of the tree. In this paper, we present a more complete study, realized on 86 prokaryotic genomes. The results of our analysis are partially in agreement with the rRNA reference. However, some important differences bring raises questions about bacterial phylogeny.

2 Material and Methods

2.1 Gene Families

A special release of the HOBACGEN database [25] called HOBACGEN-CG was made, gathering all protein sequences into families of homologous genes from the completely sequenced genomes of 76 bacteria and 10 archaea. We retained as orthologous gene families only those containing one gene per species. Though this approach may miss some hidden paralogy, especially in the case of archaea for which only a few organisms are completely sequenced, this definition of orthology has been shown to be much more accurate than a reciprocal BLAST hit based one [26]. Protein sequences from hyperthermophilic bacteria having orthologues only in archaea were removed from the family they belong to since these genes are suspected to have been acquired by lateral transfers [27]–[29]. At last, only families containing at least 30 species were considered for further analysis.

2.2 Alignments and Gene Trees Construction

The protein sequences of each family were aligned using CLUSTAL W [30], with all default parameters. To select the parts of the alignments for which homology

between sites can be assumed with good confidence, we used the GBLOCKS program [31]. It has been shown to give alignments that are almost independent to the different options of CLUSTALÂW. We retained for tree construction only the alignments having conserved at least twice more sites than species. For each family, we computed two trees: one using PHYML [32], a Maximum Likelihood (ML) method, and one using BIONJ [33] an implementation of the Neighbour-Joining (NJ) algorithm. For both methods, we used the Jones-Taylor-Thorton (JTT) model of amino acid substitution [34]. In the case of BIONJ trees, the distances were computed using PROTDIST, from the PHYLIP package [35]. Heterogeneities between sites were estimated under a gamma law based model of substitution, and the computation of the alpha parameter was carried out by PUZZLE [36]. All trees were bootstrapped (1000 replicates) using programs SEQBOOT and CONSENSE from PHYLIP.

2.3 Final Selection of Families

In order to reduce the impact of inter-domain lateral transfers, we applied the same criteria as in [16], i.e. we screened the trees where bacteria were not monophyletic and we removed these families from the data set or corrected them by removing the transferred sequences from the alignment when it was evident. At last, we ended with 122 trees suitable for super-tree construction.

2.4 Super-tree Computation

Trees chosen for the super-tree computation were coded into a binary matrix using the MRP coding scheme [22],[23]. Before we applied this method to complete genomes, it has been used to infer a phylogeny of Eutheria [37]. Each tree obtained for a set of species from a single gene family is coded into a binary matrix of informative sites with respect to bootstrap values as shown in Fig. 1. The matrices obtained are concatenated into a super-matrix in which species absent from a gene family are encoded as unknown state. The super-tree is calculated on the super-matrix using program PAUP* [38] with all default options. 1000 bootstrap replicates are performed on the supermatrix with SEQBOOT and CONSENSE.

3 Results

The two super-trees we obtain are quite similar regarding to their topology (Fig. 2 and 3). In both of them, the great bacterial divisions and subdivisions are found to be monophyletic, as well as the Euryarcheota. The lack of resolution of the archaeal part of the tree is due to the low number of genes available for these species. As expected, the bacterial part presents higher bootstrap values and appears thus more resolved, especially with the BIONJ-based trees. The monophyly of Proteobacteria and all their subdivisions, high G+C Gram

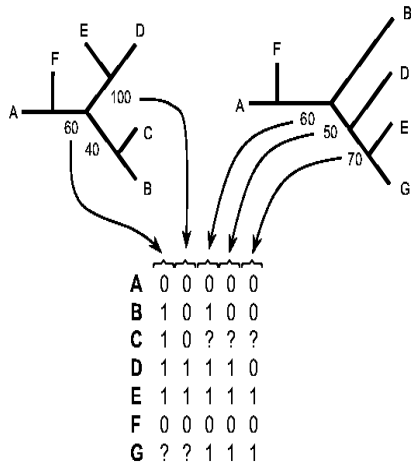


Fig. 1. Construction of super-trees by MRP with bootstrap weighting. Each tree obtained for a set of species from a single orthologous gene family is coded into a binary matrix of informative sites. Only branches having a bootstrap value over 50% are coded. The matrices obtained are concatenated into a super-matrix in which species absent from a gene family are encoded as unknown state (?). The supertree is computed on the super-matrix with PAUP*.

positives and Spirochetes are strongly supported (bootstrap $\geq 80\%$). The monophyly of low G+C Gram positives is also strongly supported in the BIONJ-based super-tree, but has a weaker support in the PHYML-based super-tree. As in our previous study [24], we find that D.Âradiodurans is grouped with the high G+C Gram positives. Also, *C.Âtepidum* is grouped with the Chlamidiales. Like in the rRNA trees, hyperthermophilic bacteria are located at a basal position, this with a strong support. This is an important difference from our previous results, where this position was occupied by Spirochetes [24]. Here, the Spirochetes are located much deeper in the two super-trees, and this positioning is not supported The remaining parts of the super-trees, and particularly the deep nodes, is not supported. This difficulty of resolving deep branches may be related to the increasing probability of horizontal transfers, hidden paralogies and long branch artifacts with separation time in the individual gene trees.

4 Discussion – Conclusion

Although the deep nodes have low support, the level of resolution of the super-tree is in strong disagreement with the "genome space" vision of the prokaryotic world predicting a "star phylogeny" [13]. One could argue that grouping of species in the super-tree would only reflect the frequency of gene exchanges between these species. This interpretation can be excluded since the super-tree method would then not be expected to give a tree topology radically different from gene-

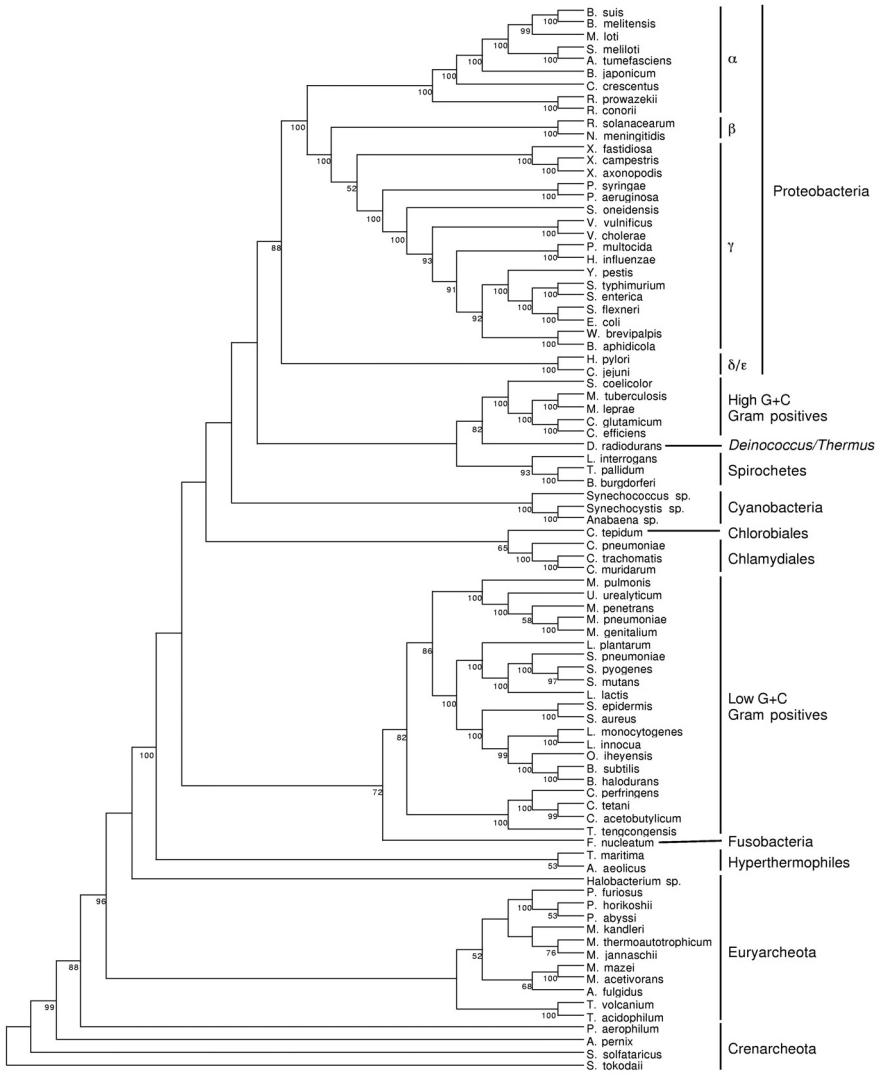


Fig. 2. Super-tree based on 122 gene trees computed with BIONJ. The tree has been arbitrary rooted with *S. tokodai* and only bootstrap values over 50% are shown.

content based trees [18],[19],[21] which are predicted to be very sensitive to this problem. It is worth to insist on the fact that a particularly stringent selection of protein families was exercised for building the supertree. In particular, a phylogenetic definition of orthology rather than a definition based on reciprocal best

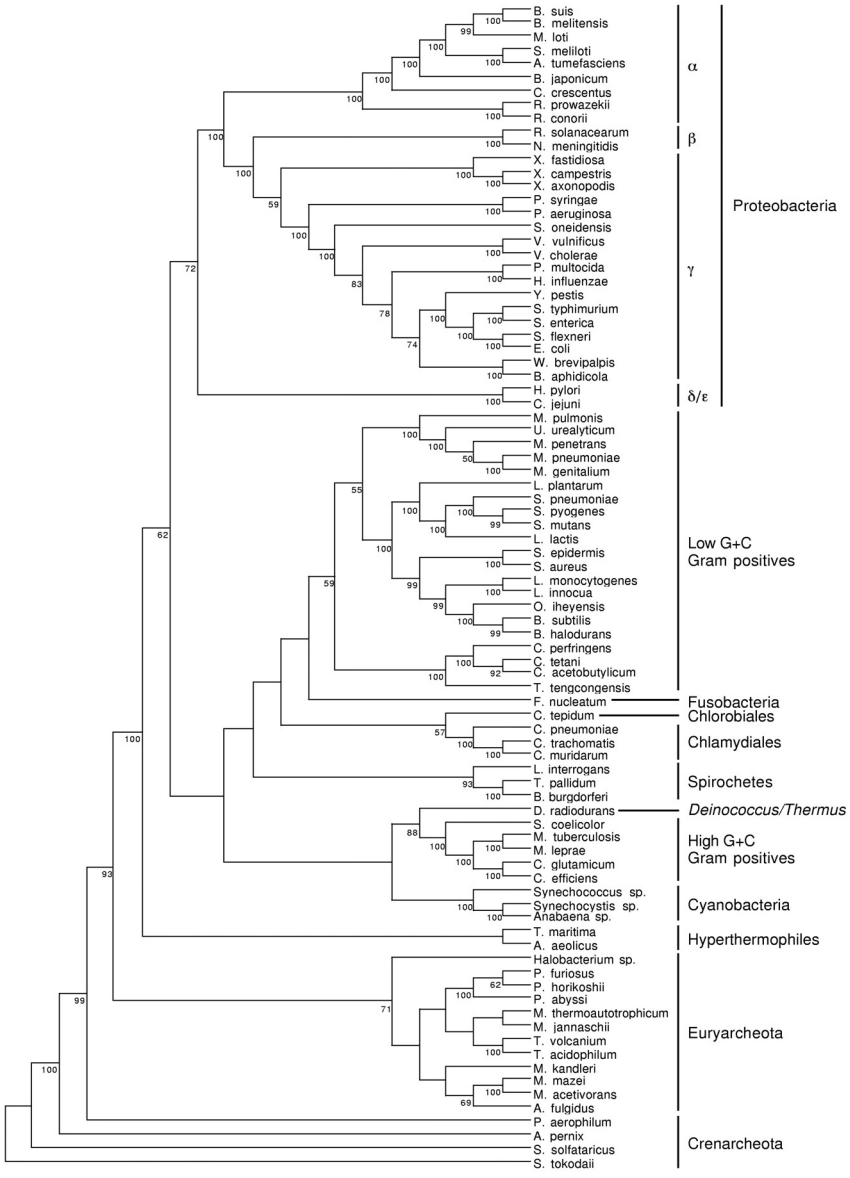


Fig. 3. Super-tree based on 122 gene trees computed with PHYML. The tree has been arbitrary rooted with *S. tokodai* and only bootstrap values over 50% are shown.

BLAST hits - as is often the case for practical reasons - was used. Thus, all gene trees where a species was represented more than once were excluded from analysis. This selection allowed to make absolutely no *a priori* assumption on the topology of the trees, except for the monophyly of bacteria, and to reduce the probability of taking hidden paralogies into account. The phylogeny of Proteobacteria appears is well resolved at this level and is in agreement with the rRNA phylogeny and protein-based works (see[39] for review). Their monophyly (including *H.pylori* and *C.jejunii*) is well supported and this last result is particularly valuable because it has rarely been found with genome-tree methods [15],[19],[21]. Equally interesting is the low (or lack of) support for clustering the two hyperthermophilic bacteria, *A.aeolicus* and *T.maritima*. Indeed, since proteins of hyperthermophilic bacteria and archaea have been shown to possess a very peculiar amino acid composition [40], it is possible that the grouping of *A.aeolicus* and *T.maritima* is linked to a systematic artifact present in the majority of our gene trees. Also, the positioning of these organisms in a basal part of the tree brings us back to a more traditional view of the bacterial world. An explanation for that change could be the presence of horizontally transferred genes in our original data set. Here, due to the fact we had to increase the number of species represented in each family, we mechanically decreased the number of families represented in each species. Therefore, we may have removed horizontally transferred genes responsible for the basal position of Spirochetes, as it was done in [16]. The monophyly of low G+C Gram-positives (including *Bacillus* and *Mycoplasma*) on one side, and of high G+C Gram-positives on the other side is supported in both super-trees. On the other hand, Gram-positive bacteria are polyphyletic, suggesting that two independent losses of the external membrane occurred in the high- and low G+C phyla. The significant support for the position of *D.radiodurans* is very striking since this organism is usually considered to have a much more basal position among bacteria [1]. This bacterium was already placed there in our previous study [24] and [16] also gives strong support to this position. On the other hand, it remains possible that this position is due to the high G+C content of the genome of *Deinococcus*. Indeed, *D.radiodurans* is a close relative of *Thermus aquaticus*, which is a Gram-negative thermophilic bacterium. Though D.radiodurans is positive to the Gram coloration, it has been shown to possess an external membrane unlike Gram-positives. Thus, although this position seems supported, it still needs to be confirmed, in particular by the addition of *Thermus* in the super-tree.

Acknowledgements. This work has been supported by CNRS and MENRT. A.C. is a recipient of a fellowship from the MENRT.

References

1. Woese, C.: Bacterial evolution. Microbiol. Rev. 51 (1987) 221-271
2. Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R.: Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA 93 (1996) 9188-9193

3. Brown, J.R. et Doolittle, W.F.: Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. 61 (1997) 456-502
4. Koonin, E.V., Makarova, K.S., Aravind, L.: Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55 (2001) 709-742
5. Forterre, P.: A hot topic: the origin of hyperthermophiles. Cell 85 (1996) 789-792
6. Philippe, H., Laurent, J.: How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8 (1998) 616-623
7. Galtier, N., Tourasse, N., Gouy, M.: A nonhyperthermophilic common ancestor to extant life forms. Science 283 (1999) 220-221
8. Asai, T., Zaporojets, D., Squires, C., Squires, C.L.: An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc. Natl. Acad. Sci. USA 96 (1999) 1971-1976
9. Yap, W.H., Zhang, Z., Wang, Y.: Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. 181 (1999) 5201-5209
10. Karlin, S., Campbell, A.M., Mrzek, J.: Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. 32 (1998) 182-225
11. Doolittle, W.F.: Lateral genomics. Trends Cell Biol. 9 (1999) M5-8
12. Doolittle, W.F.: Phylogenetic classification and the universal tree. Science 284 (1999) 2124-2129
13. Bellgard, M.I., Itoh, T., Watanabe, H., Imanishi, T., Gojobori, T.: Dynamic evolution of genomes and the concept of genome space. Ann. NY Acad. Sci. 870 (1999) 293-300
14. Garcia-Vallv, S., Romeu, A., Palau, J.: Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10 (2000) 1719-1725
15. Teichmann, S.A., Mitchison G.: Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. 49 (1999) 98-107
16. Brown, J.R., Douady., C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J.: Universal trees based on large combined protein sequence data sets. Nature Genet. 28 (2001) 281-285
17. Brochier, C., Bapteste, E., Moreira, D., Philippe, H.: Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18 (2002) 1-5
18. Snel, B., Bork, P., Huynen, M.A.: Genome phylogeny based on gene content. Nature Genet. 21 (1999) 108-110
19. Tekaia, F., Lazcano, A., Dujon, B.: The genomic tree as revealed from whole proteome comparisons. Genome Res. 9 (1999) 550-557
20. Eisen, J.A.: Assessing evolutionary relationships among microbes from whole-genome analysis. Curr. Opin. Microbiol. 3 (2000) 475-480
21. Lin, J., Gerstein, M.: Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10 (2000) 808-818
22. Baum, B.R.: Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41 (1992) 3-10
23. Ragan, M.A.: Phylogenetic inference based on matrix representation of trees. Mol. Phyl. Evol. 1 (1992) 53-58
24. Daubin, V., Gouy, M., Perrire, G.: A phylogenomic approach to bacterial phylogeny: evidence for a core of genes sharing common history. Genome Res. 12 (2002) 1080-1090
25. Perrière, G., Duret, L., Gouy, M.: HOBACGEN: database system for comparative genomics in bacteria. Genome Res. 10 (2000) 379-385

26. Koski, L.B., Golding, G.B.: The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52 (2001) 540-542
27. Nelson, K.E., Clayton, R.A. Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A. et al.: Evidence of lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399 (1999) 323-329
28. Logsdon J.M., Faguy, D.M.: Thermotoga heats up lateral gene transfer. Curr. Biol. 9 (1999) R747-R751
29. Nesbo, C.L., L?Haridon, S. Stetter, K.O., Doolittle, W.F.: Phylogenetic analyses of two ?archaeal? genes in Thermotoga maritima reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. 18 (2001) 362-375
30. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (1994) 4673-4680
31. Castresana, J.: Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17 (2000) 540-552
32. Guindon, S., Gascuel, O.: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52 (2003) 696-704
33. Gascuel, O.: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14 (1997) 685-695
34. Jones, D.T., Taylor, W.R., Thornton, J.M.: The rapid generation of mutation data matrices from protein sequences. Comput. Applic. Biosci. 8 (1992) 275-282
35. Felsenstein, J.: PHYLIP ? Phylogeny inference package (Version 3.2). Cladistics 5 (1989) 164-166
36. Strimmer, K., von Haeseler, A.: Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13 (1996) 964-969
37. Liu, F.G., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Young, T.S., Gugel, K.F.:. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291 (2001) 1786-1789
38. Swofford, D. L.: PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts (2003)
39. Gupta, R.S.: The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol. Rev. 24 (2000) 367-402
40. Kreil, D.P., Ouzounis, C.A.: Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 29 (2001) 1608-1615

Genetic Paralog Analysis and Simulations

Stanisław Cebrat[1], Jan P. Radomski[2], and Dietrich Stauffer[3]

[1] Institute of Genetics and Microbiology, University of Wroclaw,
ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
cebrat@microb.uni.wroc.pl
[2] Interdisciplinary Center for Computational and Mathematical Modeling,
Warsaw University, PL-02-106 Warsaw, Poland
[3] Institute for Theoretical Physics, Cologne University, D-50923 Köln, Euroland

Abstract. Using Monte Carlo methods, we simulated the effects of bias in generation and elimination of paralogs on the size distribution of paralog groups. It was found that the function describing the decay of the number of paralog groups with their size depends on the ratio between the probability of duplications of genes and their deletions, which corresponds to different selection pressures on the genome size. Slightly different slopes of curves describing the decay of the number of paralog groups with their size were also observed when the threshold of homology between paralogous sequences was changed.

1 Introduction

It is widely accepted that evolution is driven by two random processes - mutations and recombinations and a directional process - selection. Recombination not only re-shuffles genes inside genomes or between genomes but it is also responsible for amplification or elimination of sequences. Duplication of complete coding sequences produces additional copies of genes called paralogs. Thus, paralogous genes are homologous sequences arisen through gene duplication and parallel evolution in one genome. Paralogs can also appear by duplication of large fragments of chromosomes or even by fusion of different genomes (allopolyploidization). Before the fusion, corresponding sequences in the two genomes which had a common ancestor in the past are called orthologs [1]. Since it would be very difficult to reproduce their real history, when they appear in the genome of one organism they are recognized as paralogs. Paralogs are a source of simple redundancy of information, making the genome more stable and resistant to mutational effect by complementing the function of one copy when the other is damaged by mutation [2] or by reinforcing the function of the amplified gene. Most importantly, gene duplication generates a sequence with a defined function but released from the selection pressure. Redefinition of the duplicated gene function may ameliorate the biological potential of the individual. Taking under consideration all the profits brought by paralogs one can ask why the number of paralogs seems to be limited. First of all, a higher number of gene copies, frequently causing a higher level of products does not mean a more concerted

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 709–717, 2004.
© Springer-Verlag Berlin Heidelberg 2004

expression of the gene function. The best example - the Down syndrome - is caused by redundant information. Second, limitation comes also from the cost of replication and translation of information, which leads to selection pressure on the genome size. The genome size is the result of compromise between the trends for accumulating information and keeping the costs of replication in the reasonable limit. Nevertheless, the level of redundancy in genetic information is high, for example in a uni-cellular eukaryote organism - Saccharomyces cerevisiae (baker's yeast) - probably no more than 20 % of genes fulfill essential functions and stay in unique copies. The function of the rest of genes can be complemented, probably mostly by paralogous sequences [3], [4].

According to the definition, all the genes in the genome which have a common ancestor belong to one paralog family or group. However, the genome analysis does not give us direct information about the descent of sequences from the common ancestor because we can only conclude about the common progenitor on the basis of homology between compared currently "living" sequences. The level of homology could additionally indicate the time when the two sequences have diverged. Approximately, the number of mutations which have occurred in the diverging sequences grows with time linearly, though it may depend on the topological character of the duplication itself (i.e. duplication with or without inversion)[7]. Furthermore, the fraction of positions in which the two sequences differ does not grow linearly because of multiple substitutions (substitutions which have occurred in the same position several times) and reversions whose probability grows in time. Thus, the level of homology is not an exact measure of divergence time (branching time). At large time distances the homology between two paralogs could be too low to recognise properly whether the observed homology is accidental or the compared sequences actually descend from one progenitor sequence. That is why a threshold of homology is assumed - if the homology level is below the threshold, the compared sequences are considered as independently evolved. Since the threshold is arbitrary, and differs in different analyses, it is important to find whether the size distribution of paralog families depends on the cutoff level of homology.

In all analyzed genomes the distribution of paralog families follows a specific rule. Some authors claim an exponential function [5], others a power law ruling the frequency of the occurrence of the folds or protein families [6], [8], [9]. The latter authors assumed a limited number of the initial sequences evolving into the full genome of the contemporary organism. In our simulations we have assumed that the evolution of the contemporary genomes has started with all the genes indispensable for survival of the individuals and these initial genes were independent progenitors of all paralog families. The organisation of these genes in higher hierarchy (families or folds) was neglected. We have analysed how the size distribution of paralog families depends on the selection pressure, on genome size and on the arbitrarily accepted threshold of homology deciding about the grouping of the sequences into paralog families. The selection pressure is an objective force influencing the genome evolution while the paralog identification errors are connected with our ignorance, rather. In our simulations we

used two different ways for measuring the distance between paralogs: the first one was somewhat absolute because it measured the real time of duplication and the second one corresponded to the homology analysis - the Hamming distance between two sequences (bit-strings) was measured.

2 Experimental Distributions

Analysis of the first completely sequenced genome demonstrated that distributions of sizes for paralog families indicate a high level of gene duplication [10]. Initial comparison, of bacterial, archeal and eukaryotic genomes has shown that the number of sequences in protein families vs. corresponding family sizes displays power law distributions [8,11].

In contrast, Slonimski et al. [5] in an one page note, reported that for protein families of $N = 2$ to $5 - 6$ members, the clusters of $N + 1$ contain half the number of proteins observed in clusters of N, independently of the microbial genome size. Their methodology [12], [13] used Smith-Waterman scores $SW > 22$, the Z-significance values, and connective-clusters in which a given sequence had similarity of $Z_{value} \geq 8$ with one or more other sequences. The analysis have been performed on yeast and 4 microbial genomes.

Yanai et al. [9] have compared paralog distributions for 20 genomes, using BLAST and E-significance values ranging from $E = 10^{-10}$ to as large as $E = 10^{-3}$. They report linear fits on log-log scale for all genomes, with somewhat noisy behaviour for larger families. Qian et al. [14] have linked the power law distribution of gene families in genomes, with the distribution of structural motifs and protein folds, all three displaying identical slope on log-log plots. Their analyses involved again 20 microbial genomes, and also inter-genome comparisons within analogous functional and structural families.

Unger et al. [15] compared orthologous gene distributions in three large curated databases: COG, ProtoMap, and Predom (28031, 81286 and 278584 sequences respectively), and also performed partial analysis of a human genome. They again observed a power law behaviour relating the number of sequences in structural and functional families $F(N)$ of a given size N, by $F(N) \propto N^{-b}$, where b - the slope of linear fits on log-log plots. Additionally they have linked the slopes for small families, and those for large families by $b_{50} = 1 + 1/b_{500}$, where b_{50} and b_{500} stand for the 50 smallest, and the 500 largest families, after ranking them by size.

Nimwegen [16] has observed power laws, comparing the number of genes in functional categories vs. total number of genes in a genome, with exponents varying both between bacterial, archeal and eukaryotic genomes, and especially between functional categories: from 0.13 for the protein synthesis in bacteria, to as high as 3.36 for the defense response in eukaryotes.

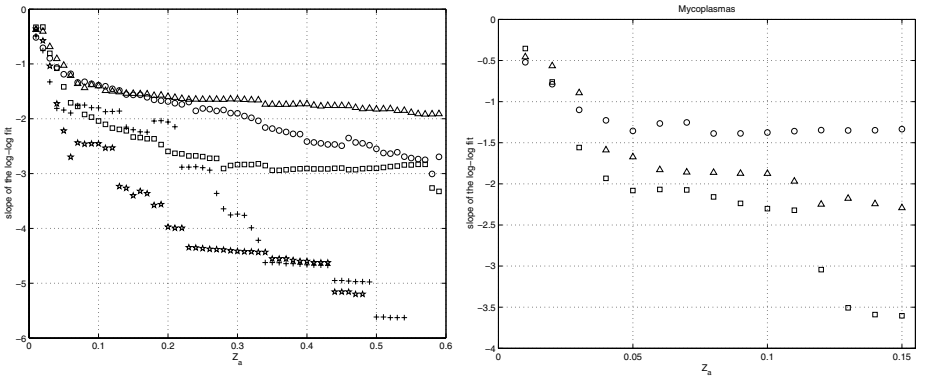


Fig. 1. Part a: Slopes of the log-log fittings as a function of the Z_a cut-off values. *Borrelia burgdorferi* - pentagrams (850 genes); *Haemophilus influenzae* - crosses (1712 genes); *Metanococcus jannaschii* - squares (1721 genes), *Sulfolobus solfataricus* - triangles (2939 genes), *Arabidopsis thaliana* - circles (26462 genes). Part b: Dependence of the power law exponents on genome size for three *Mycoplasma* bacteria: *M.genitalium* (486 genes) - squares, *M.pneumoniae* (687 genes) - triangles, and *M.pulmonis* (778 genes) - circles

3 Current Work

The Z_{value} [12, 13] data of all intragenomic pairwise alignments for 61 complete genomes [18] have been used. In no case an exponential decay for a distribution of paralogous family sizes was found, independently of the cut-off threshold of the Z_{value} used as a similarity measure. As the Z_{value} depends much on the length of compared sequences [12, 13], here we use an amended similarity measure between sequences A and B, $Z_a = Z_{value}(A,B)/\max[Z_{value}(A), Z_{value}(B)]$. For identical sequences $Z_a = 1$, and it tends to zero with increasing dissimilarity.

Figure 1a presents the slopes of the log-log fittings as a function of the Z_a cut-off values between 0.01 and 0.6 used, for several genomes. For $Z_a < 0.04 - 0.05$, for all genomes there are only one or two huge super-clusters, and small fractions of singletons and doublets (sometimes also triplets). Clearly such a small cut-off is too low to distinguish anything of interest. For high values of Z_a, but obviously depending much on the genome size, most sequences are similar only to themselves, and there are mostly singletons, with few still remaining doublets/triplets. At the less stringent similarity cut-off there are regions of gradual change, interspersed by sharp changes in behaviour - corresponding obviously to the splitting events, when clusters are broken, and a possible relationship between homology and function within family/cluster is disrupted. Somewhere in between these two extremes there is a small region of usefulness, when the slope of the log-log fits seems to depend more or less linearly on the cut-off Z_a value. Tentatively it might be attributed to a Z_a range of 0.04 - 0.1, as for most genomes analysed, we can see a relative plateau of the log-log slope changes with increasing Z_a.

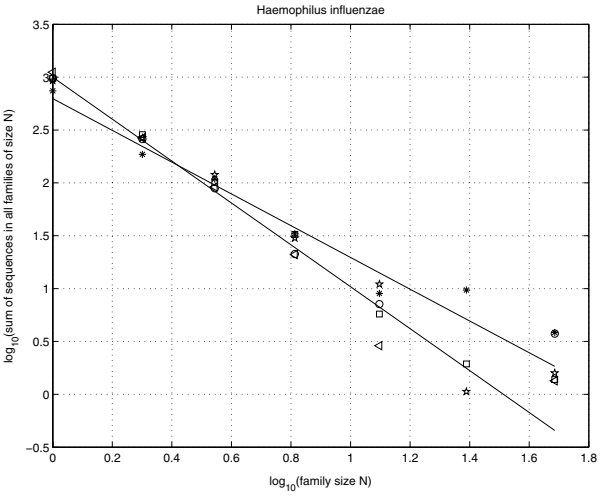


Fig. 2. Comparison of the log-log plots for *Haemophilus influenzae* between the data from Brenner et al. [10] - circles; the current work: $Z_a = 0.04$ - stars, $Z_a = 0.5$ - squares, and $Z_a = 0.06$ - triangles; and when using [12],[13] $Z_{value} \geq 8$ - pentagrams. The steeper solid line, for fitting all but the last point of Brenner's data, has the slope –2, the more shallow one the slope of –1.5 uses all points. All three methods are based on the use of significance for the Smith-Waterman local alignments.

Moreover, as can be seen in Fig. 1a, any comparisons between genomes must depend to a high degree on the cut-off value of the similarity measure actually used. For example, the data of Brenner et al. [10] for $Haemophilus\ influenzae$ would suggest the slope of the log-log plot equal –1.50, which would imply, if compared to the Fig. 1a, the Z_a in between of 0.02 and 0.04, clearly in a twilight zone before the supposedly useful region of linear dependence of the slope on Z_a. However, the last point (Fig. 2, circles) changes the slope of the fit significantly, the slope after its exclusion equals –1.98. The corresponding analysis using Z_a reveals (Fig. 2, stars $Z_a = 0.04$, squares $Z_a = 0.05$, triangles $Z_a = 0.06$) that best agreement between ref. 10 and the current work is at $Z_a = 0.05$, and that in both cases power law approximation underestimates big-sized families (rightmost points, Fig. 2), especially at higher Z_a. Finally, the results of cut-off $Z_{value} = 8$, used by Slonimski et al. [5], [12] (Fig. 2, pentagrams), again agree with both Brenner's and current results.

The often emphassized dependence of the fitted log-log slopes on the genome size can be observed only as a general trend, with many exceptions. *Metanococcus janaschii* and *Haemophilus influenzae* are of almost identical size of about 1700 sequences, but their behaviour is strikingly different, with *H. influenzae* displaying the quickest change of slopes with increasing Z_a of all genomes analysed. Also, *H. influenzae* large clusters are breaking down to singletons much faster (e.g. the rightmost crosses of *Haemophilus* in Fig. 2, correspond to the bi-

partite composition of vast majority of singletons, and a very small remainder of what was before one or two big families). *Sulfolobus solfataricus* - at approximately one tenth the genome size of *Arabidopsis thaliana* - shows the most shallow dependence of slopes on Z_a of all genomes under study, comparable to that of *Arabidopsis*. Even for the smallest genomes (Fig. 1b) of *Mycoplasma genitalium* (486 genes), *Mycoplasma pneumoniae* (687 genes), and *Mycoplasma pulmonis* (778 genes), Fig. 1b (squares, triangles, and circles respectively), which because of their taxonomical proximity can be compared directly relatively easy, the size dependence of the power law exponent is rather perturbed.

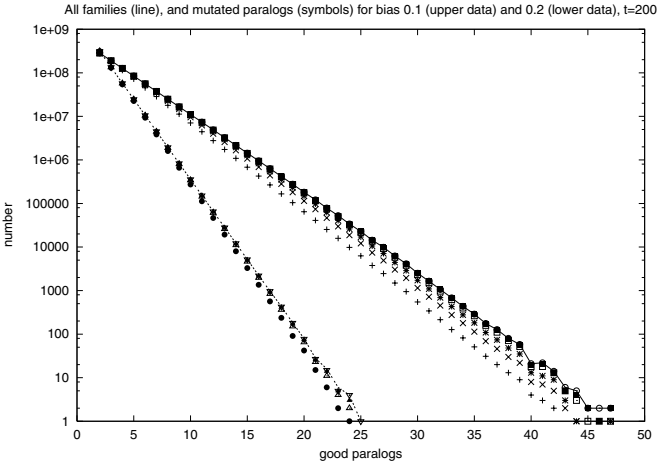


Fig. 3. Line shows the number n_k of families with k paralogs each, independent of the bit-string status. The symbols give, for x = 0,1,2,4,,,, from bottom to top, the normalized number of paralog pairs within such families of size k. $p_{mut} = 0.01$, $b = 0.1$ (upper data) and 0.2 (lower data).

4 Simulations

The results of earlier modeling efforts can be found in Refs. 2,9,11,14-17,19. In our simulations we return to the problem emphasized in the Introduction, the number of paralogs for one given function or gene. Thus, in contrast to what was described in preceding sections, we assume to know for every part of the genome its function. In a simulation that is easy, since we can follow the whole evolution since the beginning; for real genomes, such knowledge in general still lies in the future. Our model is a simplification of our earlier one [2], which was shown [19] to give reasonable ageing results.

The simulations start with N bit-strings of length L each, which are zero everywhere. Then at each iteration each bit-string with mutation probability

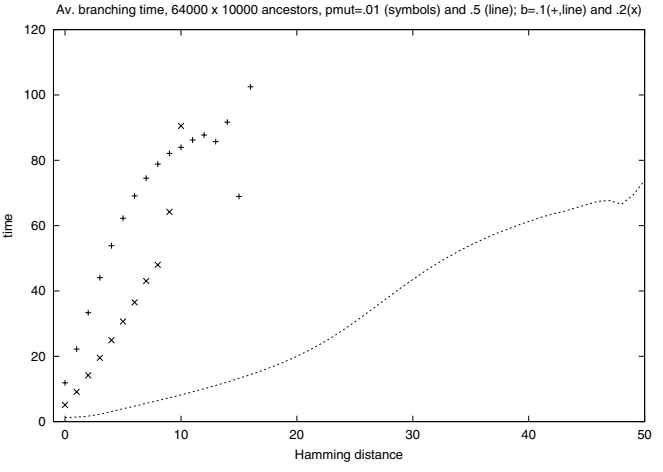


Fig. 4. Average branching time, defined as the number of iterations since the last creation of the paralog, versus Hamming distance, from 64,000 samples of 10,000 ancestors each, with 200 iterations. The fluctuations in this time are about as large as the average. Plus signs: $p_{mut} = 0.01,\ b = 0.1$; x crosses: $p_{mut} = 0.01,\ b = 0.2$; line: $p_{mut} = 0.5,\ b = 0.1$.

p_{mut} selects randomly one of its bits and flips it. Before that, also at every time step, for each family (offspring of one ancestor) either the last bit-string is deleted (with probability $1/2+b$) or a randomly selected bit-string is duplicated (with probability $1/2-b$) and then becomes the last; the positive bias b keeps the number k of copies ("paralogs") for each of the N original bit-strings limited. Also, the number k is not allowed to become negative. Thus at any time we have for each of the N ancestors a family consisting of the first bit-string and possibly $k-1$ additional copies or paralogs, amounting to k bit-strings in total for one ancestor (=gene = function).

The Hamming distance (= number of bits different in a bit-by-bit comparison of two bit-strings) was calculated for each paralog with all other bit-strings in the same family at the same time, giving $(k-1)k/2$ Hamming distances.

The simulations mostly used $L = 64,\ N = 10000,\ b = 0.1,\ p_{mut} = 0.01$ for $t = 200$ iterations and averaged over 64000 samples. Simulations for $L = 8,\ 16,\ 32$ barely differed in the results when a comparison was possible. The average number k of paralogs was nearly 3, i.e. we had nearly two additional bit-strings (plus the first one) for each ancestor. Semilogarithmic plots, Fig. 3, of the number of paralog pairs for one ancestor with Hamming distance not exceeding x different bits typically gave straight lines with slopes only slightly depending on x. x was taken as 0,1,2,4,8,16,32, and 64. For large x the curves nearly overlap. For clarity we divided for our figure the number of pairs by the normalizing factor $[k(k-1)/2]$ and thus for $L = 64$ get the total number of families.

The overall distributions n_k, lines in Fig. 3, decay exponentially, proportional to $[(0.5-b)/(0.5+b)]^k = 1/1.5^k$, in the stationary state achieved after dozens of iterations for n_k, even when the Hamming distances still grow. This formula follows from a detailed balance condition that as many families move on average from size k to size $k+1$ as move in the opposite direction from size $k+1$ to size k. Thus the fraction of families with only a single bit-string is $1-(0.5-b)/(0.5+b) = 4b/(1+2b)$ in this geometric series.

We define the creation of a new paralog as a branching event and store this time. At the end of the simulation we determine for each pair within each family the last event they branched away from each other; the time between this last event and the last iteration of the simulation is the branching time. Within each family the branching times fluctuate strongly but their average value for one given Hamming distance increases roughly linearly with that Hamming distance, until for large Hamming distances the statistics becomes poor, Fig. 4. For longer times (500 and 1000 iterations) the linearity improves.

The above model follows ref.2 except that no selection of the fittest and similar complications are included now. Each of the ancestors is interpreted as one function (or gene) in the whole organism. The bit-string for this ancestor then records important mutations at different places within this gene. The paralogs formed in the simulation from this ancestor all refer to this one function. The first bit-string undergoes mutations just as its paralogs and has the same properties except that it can never be removed. It makes no sense to compare bit-strings for different functions; 00101001 means something entirely different for the function "brain" than for the function "hair". The L bits of each bit-string correspond to 2^L possible alleles for one function, not to L base pairs.

The N initial ancestors can also be interpreted as N different samples simulated for the same function; more generally, they could be M different genomes simulated for a genome of N/M functions.

5 Summary

We presented here two different sets of plots: In the experimental section we found power-law decay for the number of paralogs found by looking through the whole genome. In the simulation section we found exponential decays for the number of paralogs belonging to one known function. The latter exponential decay agrees nicely with simple arguments based on detailed balance; the slopes in these semilogarithmic plots (Fig.3) are determined by our bias in favour of removal instead of addition of a paralog, and the slopes barely depend on the cut-off parameter x for the Hamming distance. This agreement of theory with simulation also makes clear that our results would be quite different if the bias would not be the same for all functions.

Acknowledgements. DS thanks the Julich supercomputer center for time on their Cray-T3E and M. Dudek for help with LNCS formats. JPR

was partially supported from the 115/E-343/S/ICM/853/2003 and 115/E-343/BW/ICM/1624/2003 grants.

References

1. Fitch WM: Distinguishing homologous from analogous proteins, Syst. Zool. **19** (1970) 99-113
2. Cebrat S, Stauffer D: Monte Carlo simulation of genome viability with paralog replacement. J.Appl.Genet. **43** (2002) 391-395
3. MIPS 2002 Database, http://mips.gsf.de/proj/yeast/.
4. Mackiewicz P, Kowalczuk M, Mackiewicz D, Nowicka A, Dudkiewicz M, Laszkiewicz A, Dudek MR, Cebrat S: How many protein-coding genes are there in the Saccharomyces cerevisiae genome? Yeast **19** (2002) 619-629
5. Slonimski PP, Mosse MO, Golik P, Henaut A, Risler JL, Comet JP, Aude JC, Wozniak A, Glemet E, Codani JJ: The first laws of genomics. Microb. Comp. Genomics **3** (1998) 46.
6. Koonin EV, Galperin MY: Sequence - Evolution - Function, Computational approaches in Comparative Genomics (2003), Kluwer Academic Publishers
7. Mackiewicz P, Mackiewicz D, Kowalczuk M, Dudkiewicz M, Dudek MR, Cebrat S: High divergence rate of sequences located on different DNA strands in closely related bacterial genomes. J. Appl. Genet. **44** (2003) 561-584
8. Gerstein M, A structural census of genomes: Comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure: J.Mol.Biol., **274** (1997) 562-574
9. Yanai I, Camacho CJ, DeLisi C: Predictions of gene family distributions in microbial genomes: Evolution by gene duplication and modification. Phys. Rev. Lett., **85** (2000) 2641-2644
10. Brenner SE, Hubbard T, Murzin A, Chotia C: Gene duplications in *Haemophilus influenzae*. Nature, **378** (1995) 140
11. Huynen MA, van Nimwegen E: The frequency distribution of gene family sizes in complete genomes. Mol.Biol.Evol., **15** (1998) 583-589
12. Codani JJ, Comet JP, Aude JC, Glemet E, Wozniak A, Risler JL, Henaut A, Slonimski PP: Automatic analysis of large-scale pairwise alignments of protein sequences. Methods Microbiol., **28** (1999) 229-244
13. Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP, Codani JJ: Significance of Z-value statistics of Smith-Waterman scores for protein alignments. Comput.Chem., **23** (1999) 317-331
14. Qian J, Luscombe NM, Gerstein M: Protein family and fold occurrence in genomes: Power-law behaviour and evolutionary model. J.Mol.Biol. **313** (2001) 673-681
15. Unger R, Uliel S, Havlin S: Scaling law in sizes of protein sequence families: From super-families to orphan genes. Proteins **51** (2003) 569-576
16. van Nimwegen E: Scaling laws in the functional content of genomes. Trends Genet. **19** (2003) 479-484
17. Koonin EV, Wolf Yi, Karev GP: The structure of the protein universe and genome evolution, Nature, **420** (2002) 218-223
18. TERAPROT project (CEA, Gene-It, Infobiogen), June 2002
19. Alle P: Simulation of gene duplication in the Penna bit-string model of biological ageing. Master's thesis, Cologne University 2003.

Evolutionary Perspectives on Protein Thermodynamics

Richard A. Goldstein

Division of Mathematical Biology, National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK
richard.goldstein@nimr.mrc.ac.uk

Abstract. While modern evolutionary theory has emphasized the role of neutral evolution, protein biochemistry and biophysics has interpreted the properties of proteins as largely resulting from adaptive evolution. We demonstrate that a number of these properties can be seen as emerging from neutral evolution acting on sequence entropy, that is, the fact that larger numbers of viable sequences have these properties. In this paper, we use a computational model of populations of evolving lattice proteins to describe how the observed marginal stability of proteins as well as their robustness to mutations can result from neutral evolution.

1 Introduction

Imagine a trained physicist from another world, who had never seen a computer before, was given one to analyze. She could disassemble it, making notes regarding the copper wires, the small chips of silicon with particularly-placed impurities, the ferro-magnetic material on a plastic substrate, etc. At some level, however, we would feel that she had somehow missed the essence of the computer, that understanding such an instrument would require a *functional* explanation including the role and purpose of the memory, cpu, and storage devices. This functional explanation is possible only because the computer has gone through a design process which has determined an appropriate form for its intended function. Conversely, an understanding of the functioning of the computer requires knowledge of the properties of doped semiconductors and ferro-magnetic materials, as these properties determined the constraints and opportunities given to the computer designer.

Similarly, an understanding of organisms requires combining a mechanistic description (the physico-chemical properties of organs, cells, proteins, DNA, membranes) with a functional description (the role and "purpose" of the heart, nucleus, histone, enzyme). As in the computer example, a purely mechanistic description would miss the essence of the biological subsystems by neglecting their functional roles, while the functional aspects cannot be understood independently of the constraints and opportunities presented by the mechanistic properties. This duality is again based on history, in this case the process of biological evolution. As in the computer example, evolution has been able to take

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 718–727, 2004.
© Springer-Verlag Berlin Heidelberg 2004

advantage of the physico-chemical properties of the evolving systems, investing the resulting components with functional roles and purposes. In this instance, however, "design" would not be the appropriate term. Rather we must consider the evolutionary context; by analyzing biological systems in this context, we can work to unify these separate perspectives and understand how evolution utilizes, adopts, and changes the properties of the evolving elements while being constrained by these properties. This particular perspective leads us directly into the heart of what makes biological systems different from non-biological systems, why we can talk about the "purpose" of a lung in a way we cannot talk about the purpose of the electrons in an atom of carbon.

In this paper I focus on the properties of proteins and how they can be understood in an evolutionary context. One of the more interesting aspects of evolution is the separation between genotype and phenotype, between the molecules that are evolving (generally DNA) and the resulting traits that are acted on by evolutionary selection. Proteins can be seen as representing one of the lowest, simplest levels of phenotype, providing an important model for the evolution of higher organizational forms. Proteins are also interesting and important on their own. Various processes that proteins perform, such as folding, are of great theoretical interest – it is difficult to construct theoretical models of protein folding that can explain how such an enormous search problem is solved so quickly. Proteins are also involved in almost all functions in a living system, including respiration, signalling, replication, locomotion, transportation, etc., and are the basis of understanding these processes at a mechanistic and atomistic level. Proteins are the most common targets of pharmaceutical intervention and are thus intrinsically important for biomedical research. In addition, much work is proceeding trying to understand how to engineer proteins with modified or new properties and functions.

Evolution is a complicated procedure, particularly since exceptions to any general principle can lead to an attractive evolutionary niche. One of the more important axes for various models involves the distinction between adaptation and neutral evolution. It is clear that adaptation has had an important role in evolution, in making humans and other biological systems the way that they are. There are, in particular, episodes of adaptive evolution where change is clearly favoured. Some systems, such as pathogens avoiding a host immune response, are likely often undergoing adaptionist evolution. It is also clear that much, if not most, of molecular evolution is neutral in nature, that the vast majority of mutations that occur are either negative, slightly deleterious, or neutral, and that the chance acceptance of neutral or slightly deleterious mutations may often greatly exceed the smaller number of positive changes that might occur. Since the rise of the neutral theory in the late 1960s ([1,2]) much of evolutionary theory has been based on the emphasis of neutral evolution. In contrast, when confronted by the almost miraculous molecular properties of living systems, biochemists have generally thought in terms of adaptation and seen their characteristic traits as having arisen from the "survival of the fittest". Much less work has been performed analyzing specific proteins from a neutralist perspective.

Such a neutralist perspective is, however, important in understanding proteins for a number of reasons. Often it is possible to show that neutralism is sufficient to explain the observed properties. Because neutral evolution is always occurring, neutrality represents, when possible, the most parsimonious explanation. For this reason, adaptation is a reasonable explanation only when neutral evolution can be eliminated. Conversely, attempts to explain features based on neutral evolution can highlight when neutrality is *not* an adequate explanation, where an adaptive mechanism might be required. In addition, neutral evolution protects us against the so-called "Panglossian Paradigm" [3] where the current role of a feature is used to explain how and why the feature emerged in the way that Pangloss in Voltaire's Candide explains how we ended up with noses in order to support spectacles and legs in order to fit into trousers. Features that evolved based on one dynamic may end up being used for a completely different function. Finally, neutralist perspectives make us remember that evolution is decidedly *not* an optimization process. There are well-defined stochastic equations regarding the origin, fixation, and elimination of genetic variation, involving the fixation of deleterious mutations as well as the elimination of favourable mutations. Simplistic images that have been imposed on evolution, such as the afore-mentioned "survival of the fittest", may possibly represent more the projection of our psychological need to imagine ourselves at *some* peak of perfection rather than an inherent characteristic of the evolutionary process.

Another important aspect of the evolutionary process is the fact that evolution occurs in finite populations. The stochastic nature of the process results directly from this aspect. It also is important in elimination and fixation, as subpopulations that fall from one to zero can never recover. The stochastic nature of the evolutionary process does not mean that we cannot come up with general principles. Just as the thermodynamic notions of pressure, temperature, and heat represent the random motions of a large ensemble of particles, so we can generate principles based on the dynamics of populations. One important distinction, however, is the size of the populations. There are no populations close to Avogadro's number. For this reason the stochastic element can never be completely averaged away, and we are often left with tendencies and probabilities rather than fixed rules and laws.

The complicated nature of biomolecular evolution involving specifics of protein structure, function, thermodynamics, as well as population and evolutionary dynamics, makes this area attractive for theoretical simulations. The simulations, however, have to make numerous approximations and simplifications. In a number of publications we have described a simplified, lattice protein model which, combined with simulations of population evolution, have provided some interesting perspectives on why proteins are the way that they are. In this paper, we summarize and advance these perspectives, focusing on the thermodynamic and mutational properties of observed proteins including making connection with relevant experiments.

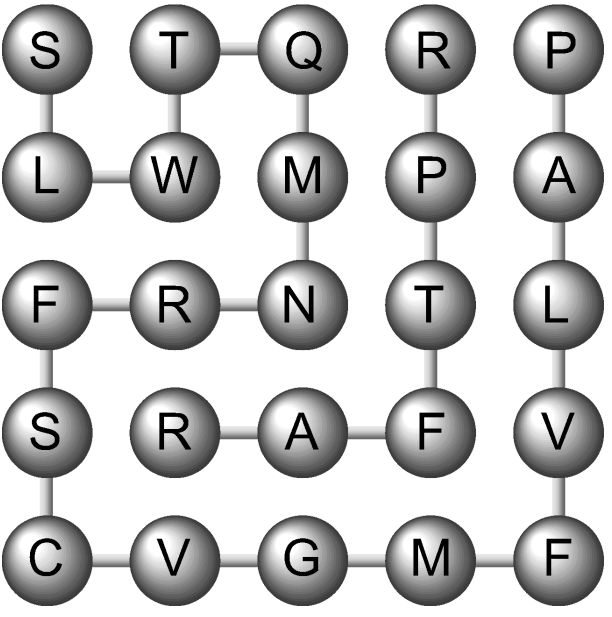


Fig. 1. Model of lattice protein

2 Models

2.1 Modelling the Evolving Proteins

The model of proteins is shown in Fig. 1. Proteins are represented as 16-monomer polypeptides forming a self-avoiding walk on a 5×5 square lattice. Each amino acid occupies exactly one lattice point. There are exactly 1081 possible conformations neglecting structures related by rotation, reflection, or inversion. While a two-dimensional lattice is highly inappropriate for folding simulations – the space of possible conformations is non-ergodic [4] – such a model allows us to have a reasonable ratio of buried to exposed residues with a moderately-sized protein.

The energy $E(S, C)$ of a given sequence S in any particular conformation C is a pairwise contact energy equal to

$$E(S, C) = \sum_{<i,j>} \gamma(A_i, A_j) u_{i,j} \tag{1}$$

where $\gamma(A_i, A_j)$ is the interaction energy between the amino acid at locations i and j in the sequence (such as between the serine and threonine in the upper-left corner of the protein in Fig. 1) and $u_{i,j}$ is equal to 1 if residues i and j are in contact (that is, are not covalently connected but are on neighbouring lattice points) and zero otherwise. The values of $\gamma(A_i, A_j)$ are taken from the contact energies derived by Miyazawa and Jernigan based on a statistical analysis

of the available protein database [5]. Because of the nature of the derivation, interactions with solvent, including entropic terms, are explicitly included in the contact energies.

We can assume that the conformation of lowest energy is the native-state C_{ns}, and compute the probability that a protein at equilibrium would be in this state

$$P_{\mathrm{ns}} = \frac{\exp(E(S, C_{\mathrm{ns}}/kT)}{\sum_C \exp(E(S, C)/kT)} \tag{2}$$

which allows us to compute $\Delta G_{\mathrm{folding}}$

$$\Delta G_{\mathrm{folding}} = -kT \log\left(\frac{P_{\mathrm{ns}}}{1 - P_{\mathrm{ns}}}\right) \tag{3}$$

2.2 Modelling Population Evolution

As mentioned before, the population dynamics are essential to any form of evolutionary simulation. In general, we start with a given population of protein sequences, initially identical. For each sequence we calculate the various thermodynamic quantities described above. We then choose a given number of sequence locations, chosen from an appropriate Poisson distribution, to change to another random amino acid. The thermodynamic quantities for the resulting sequences are again calculated. We then apply truncation selection where we decide on which sequences are to be considered viable based on whatever criteria we choose. Non-viable sequences are eliminated from the population, and the remaining sequences are chosen at random, with replacement, to form the next generation with the same population size.

3 Results

3.1 Protein Thermostability

It has been long observed that proteins are marginally stable with typical stabilities (-$\Delta G_{\mathrm{folding}}$) of approximately 10 kcal/mol, equal to a few hydrogen bonds. Two different classes of theories have been advanced for why this occurs, both adaptionist in nature. The first theory is that there is a fitness advantage to marginal stability. This might be for a number of different reasons. Protein functionality might require flexibility, which might be more common in marginally-stable proteins [6,7]. Marginal stability would weaken binding with ligands by requiring an entropy loss upon binding. This might make it easier to modulate binding affinities through mutation or post-translational modification [8,9,10]. Finally, there may be advantages to marginal stability in ensuring sufficiently rapid protein degradation. The second class of explanation revolves around the need for the protein to fulfil multiple selective criteria including functionality, stability, rigidity, etc. There would naturally be trade-offs in these criteria, so

that proteins can optimize stability given the constraints imposed by the other types of selective pressure.

Given the above discussion, it becomes important to investigate whether the observed marginal stability in proteins can result from neutral evolution. We modelled the population dynamics of 3000 sequences, allowing the dynamics to equilibrate for 30,000 generations and gathering data for an additional 30,000 generations [11]. At each generation, 0.2% of the sequences were mutated to an alternative residue. Proteins were considered viable if they were "adequately" stable, that is, with $\Delta G_{\text{folding}}$ less than some "critical" ΔG_{crit}. The result of these simulations were populations of proteins that were marginally stable, with $\Delta G_{\text{folding}} \approx \Delta G_{\text{crit}}$.

These results can be made intuitive if we consider the space of all possible sequences. This space is high-dimensional (as many dimensions as the length of the sequence) but extremely sparse (only 20 points along each dimension). The vast bulk of this space consists of proteins that would not fold nor be stable – unviable sequences in our model. There are regions in this space, characterized as hyperspheres, which contains sequences that are viable, that is, will fold into a stable, functional protein. It is a characteristic of high-dimensional spaces that the volume of objects in that space will be dominated by the periphery of that object. (99.95% of the volume of a 150-dimensional sphere is in the outermost 5%.) If the volume of sequence space consists of foldable, stable sequences, while the exterior of the volume consists of unfoldable, unstable sequences, the vast majority of the sequences in this volume will be marginally foldable and marginally stable, purely as a result of the high dimensionality. This will result whenever a) the objects in the space are roughly convex, and b) the fitness criteria are smoothly-varying in the space. If the vast majority of protein sequences are marginally stable, there is no problem explaining the observed marginal stability without resorting to adaptionist arguments, either selective pressure for marginal stability or optimization given constraints. Neutral evolution will be strongly affected by "sequence entropy", the number of sequences consistent with a given property. Sequence entropy will strongly drive protein sequences towards marginal stability.

What of the observation that modifying proteins to increase their thermostability sometimes results in decreased function [12,13,14,15]? Firstly, this is not necessarily always the case [16]. Secondly, if proteins evolve functionality in the context of a natural tendency towards marginal stability, it is not surprising to find mechanisms for functionality that are dependent upon, or at least consistent with, marginal stability. Taverna and Goldstein modelled this behaviour by considering competitive dynamics between three sets of lattice protein models, each with a different mechanism of action [11]. While the exact mechanism was unspecified, one set was modelled as requiring marginal stability, another set requiring moderate stability, while the third required high stability. A member of the first set in a marginally-stable protein had exactly the same fitness as a member of the second set in a moderately-stable form, which had exactly the same fitness as a member of the third set that was highly stable. Conversely,

any member of the first set that was not marginally stable, any member of the second set that was not moderately stable, or any member of the third set with other than high stability, was considered non-viable and was eliminated during the truncation selection. After separate equilibration, the three populations were allowed to compete against each other. In 24 out of the 25 runs, the mechanism consistent with marginal stability became the only form in the population, with the two other mechanisms being completely eliminated. (One run resulted in the domination of the mechanism requiring moderate stability, emphasizing the stochastic nature of the evolutionary process.) With the absolute equivalent fitness of these three populations, the entropic forces again combined with neutral evolution to result in proteins that required marginal stability. In the non-intuitive causality of evolution, marginal stability became required for proteins because they were marginally stable!

This is maybe less non-intuitive than it appears. Globular proteins generally require aqueous environments in order to fold and be stable and functional. One could argue that cells are generally aqueous because this environment is required for the globular proteins. This would correspond to the idea that proteins are marginally stable because this is required for protein function. In reality, the aqueous environment came first, and proteins evolved to function in this milieu. Proteins require aqueous environments to function because they evolved in aqueous environments and adapted themselves to this context. The concept that proteins adapted mechanism consistent to their context, a marginal stability induced by neutral evolution acting through sequence entropy, is not any more surprising.

This is not to say that marginal stability does not result from adaptation, only that the assumption of adaptation is not required to explain the observation. Thus neutral evolution is the most parsimonious explanation, and the observation of marginal stability does not provide any evidence for any selective pressure for marginal stability.

3.2 Proteins and Evolutionary Robustness

Another way to consider the role of sequence entropy is to consider the role of robustness in evolution. Fitness can be defined as the average expected number of viable offspring produced. Generally these offspring will be mutated forms of the parent, and so the fitness of an individual depends upon the fitness of the neighbouring genomes (sequences) in the genome (sequence) space. If mutations are more likely to lead to non-viable offspring, this reduces the fitness of the parent. If a protein depends upon being one of the few sequences with high stability, many more mutants will have reduced stability and thus would be non-viable. Conversely, if a protein has a mechanism consistent with marginal stability, the probability that a mutant would have marginal stability is much higher, ergo a higher fitness.

This robustness to mutations can be seen directly, in what has been described as "the survival of the flattest" [17]. It can be observed in the simulation

of the population evolution of lattice proteins, as described above. We can consider the results of the population evolution, and observe the consequence of random mutations. Lattice proteins that evolve based on truncation selection ($\Delta G_{\text{folding}} < \Delta G_{\text{crit}}$) are surprisingly robust to mutations, so that, on average, about half the protein sequences do not have reduced stability upon a random mutation – even when multiple residues are changed [18]. Sequences chosen at random, also subject to the same truncation selection, have almost no probability that a mutation would not be destabilizing. We can have two sequences with the same initial stability, same structure, same observed properties, one derived from population evolution, the other from being selected at random, yet the robustness to mutation is extremely different. In fact, the higher the stability requirement (corresponding to a more-negative ΔG_{crit}), the more likely a mutation will have negligible effect on the stability. Of all of the viable protein sequences, population evolution selects those networks of sequences that have the fewest non-viable neighbours. This perceived robustness of proteins to substitutions has been observed experimentally. For instance, Reddy et al. catalogued a wide range of mutations, observing that approximately 25% actually increased thermal stability [19].

4 Conclusion

Biochemists and molecular biologists have tended to imagine evolution as a constant march to higher and higher fitness levels, while modern evolutionary theory has increasingly emphasized randomness and neutral evolution. When confronted by a characteristic property of living systems, the response of biochemists has been to see the "blind watchmaker" at work, and to inquire how this property increases the fitness of the organisms. In reality, many of these properties can be explained by the process of neutral evolution acting on sequence entropy, taking advantage of the fact that many more sequences have some properties rather than others. In previous work, we have demonstrated how this can result in some structures much more common than others [20,21] (also see [22,23,24]), in proteins that fold into the state of lowest free energy [25], and in work described here, proteins that are marginally stable [11,26] and naturally robust to site mutations [18].

There are direct consequences of this perspective. For instance, the observed robustness of proteins is interpreted as explaining that the mapping of sequence to structure is rather robust and plastic. The alternative viewpoint provided here is that proteins have evolved to be robust to the particular experiments that are being performed. In general, the widespread attempts to understand the relationship between a protein's sequence and its resultant properties through site mutagenesis must take into the fact that proteins have evolved so that these mutations are less likely to change fundamental and important properties. Proteins have, to some extent, prepared for these changes. Finally, robustness is given as evidence why certain properties are not important for the protein. If these properties were under strong selective pressure, it is argued, they should be

"optimized" and therefore highly susceptible to mutational change [27]. Neutral evolution would suggest the opposite conclusion – highly important properties would be "buffered" against change during mutations.

Other conclusions relate to the possibilities and opportunities of protein engineering. If protein sequences are robust to change, this suggests that there are many options to modifying naturally-occurring sequences, in that the sequence can be highly altered while important properties such as stability and foldability will be maintained. Conversely, the evolutionary selection of highly robust sequences suggest that nature finds the flatter peaks in the fitness landscape, even if higher (but narrower) peaks exist elsewhere. This means that it might be possible to design proteins de novo with properties that exceed those found by natural evolution.

Acknowledgements. The work described here was largely performed by Sridhar Govindarajan, Darin Taverna, and Paul Williams. Computer support was provided by Kurt Hillig, Todd Raeker, and Michael Kitson. Funding was provided by NIH grant by NIH grant LM05770.

References

1. Kimura, M., Evolutionary rate at the molecular level. Nature (London) **217** (1968) 624–626
2. King, J.L., Jukes, T.H., Non-Darwinian evolution. Science **164** (1969) 788-798
3. Gould, S.J., Lewontin, R.C., The spandrels of San Marco and the Panglossian Paradigm: A critique of the adaptationist programme. Proc. Royal Soc. London, Series B **205** (1979) 581-598
4. Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I., Impact of local and non-local interactions on thermodynamics and kinectics of protein folding. J. Mol. Biol. **252** (1995) 460-471
5. Miyazawa, S., Jernigan, R.L., Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromol. **18** (1985) 534-552
6. Wagner, G., Wuthrich, K., Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the basic pancreatic trypsin inhibitor. J. Mol. Biol. **130** (1979) 31–37
7. Tang, K.E.S., Dill, K.A., Native protein fluctuations: The conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. J. Biomol. Struct. Dyn. **16** (1998) 397–411
8. Dunker, A.K., et al., Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. Pacific Symp. Biocomputing **3** (1998) 473–484
9. Wright, P.E., Dyson, H.J., Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol. **293** (1999) 321–331
10. Dunker, A.K., Obradovic, Z., The protein trinity–linking function and disorder. Nat. Biotechnol. **19** (2001) 805–806
11. Taverna, D.M., Goldstein, R.A., Why are proteins marginally stable? Proteins: Struct., Funct., Genet. **46** (2002) 105–109

12. Alber, T., Wozniak, J.A., A genetic screen for mutations that increase the thermal-stability of phage-T4 lysozyme. Proc. Natl. Acad. Sci. USA **82** (1985) 747–750
13. Bryan, P.N., et al., Proteases of enhanced stability: Characterization of a thermostable variant of subtilisin. Proteins: Struct. Funct. Genet. **1** (1986) 326–334
14. Liao, H., McKenzie, T., Hageman, R., Isolation of a thermostable enzyme variant by cloning and selection in a thermophile. Proc. Natl. Acad. Sci. USA **83** (1986) 576–580
15. Shoichet, B.K., et al., A relationship between protein stability and protein function. Proc. Nat. Acad. Sci. USA **92** (1995) 452-456
16. Giver, L., et al., Directed evolution of a thermostable esterase. Proc. Nat. Acad. Sci. USA **95** (1998) 12809-12813
17. Nimwegen, E.v., Crutchfield, J.P., Huynes, M., Neutral evolution of mutational robustness. Proc. Nat. Acad. Sci. USA **96** (1999) 9716-9720
18. Taverna, D.M., Goldstein, R.A., Why are proteins so robust to site mutations? J. Mol. Biol. **315** (2002) 479-484
19. Reddy, B.V.B., Datta, S., Tiwari, S., Use of propensities of amino acids to the local structure environment to understand effect of substitution mutations on protein stability. Protein Eng'g **11** (1998) 1137-1145
20. Govindarajan, S., Goldstein, R.A., Searching for foldable protein structures using optimized energy functions. Biopolymers **36** (1995) 43–51
21. Govindarajan, S., Goldstein, R.A., Why are some protein structures so common? Proc. Natl. Acad. Sci. USA **93** (1996) 3341–3345
22. Li, H., et al., Emergence of preferred structures in a simple model of protein folding. Science **273** (1996) 666-669
23. Shakhnovich, E.I., Protein design: a perspective from simple tractable models. Folding & Design **3** (1998) R45-R58
24. Finkelstein, A.V., Ptitsyn, O.B., Why do globular proteins fit the limited set of folding patterns. Prog. Biophys. Mol. Biol. **50** (1987) 171–190
25. Govindarajan, S., Goldstein, R.A., On the thermodynamic hypothesis of protein folding. Proc. Natl. Acad. Sci. USA **95** (1998) 5545–5549
26. Williams, P.D., Pollock, D.D., Goldstein, R.A., Evolution of functionality in lattice proteins. J. Mol. Graphics Modell. **19** (2001) 150–156
27. Kim, D.E., Gu, H., Baker, D., The sequences of small proteins are not extensively optimized For rapid folding by natural selection. Proc. Natl. Acad. Sci. USA **95** (1998) 4982-4986

The Partition Function Variant of Sankoff's Algorithm

Ivo L. Hofacker[1] and Peter F. Stadler[1,2]

[1] Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria
http://www.tbi.univie.ac.at/∼ivo
[2] Bioinformatics, Department of Computer Science, University of Leipzig, Kreuzstrasse 7b, D-04103 Leipzig, Germany
http://www.bioinf.uni-leipzig.de/∼studla

Abstract. Many classes of functional RNA molecules are characterized by highly conserved secondary structures but little detectable sequence similarity. Reliable multiple alignments can therefore be constructed only when the shared structural features are taken into account. Sankoff's algorithm can be used to construct such structure-based alignments of RNA sequences in polynomial time. Here we extend the approach to a probabilistic one by explicitly computing the partition function of all pairwise aligned sequences with a common set of base pairs. Stochastic backtracking can then be used to compute e.g. the probability that a prescribed sequence-structure pattern is conserved between two RNA sequences. The reliability of the alignment itself can be assessed in terms of the probabilities of each possible match.

1 Introduction

Sankoff's algorithm [1] simultaneously predicts a consensus structure for two (or, in its general version, more) RNA secondary structures and at the same time constructs their alignment. It is quite expensive in both CPU and memory requirements, $\mathcal{O}(N^6)$ and $\mathcal{O}(N^4)$, respectively. A further complication is that it requires the implementation of the full loop-based RNA energy model [2]. Currently available software packages such as `foldalign` [3] and `dynalign` [4] therefore implement only restricted versions. A complementary approach is taken in the `pmmatch` program [5]. Instead of attempting to solve the alignment and the structure prediction problem simultaneously, `pmmatch` utilizes the base pairing probability matrices predicted by means of McCaskill's algorithm [6] (implemented in the `RNAfold` program of `Vienna RNA Package` [7,8]). The problem then becomes the alignment of the base pairing probability matrices. This appears to be an even harder threading problem, which in general is known to be NP-complete [9]. In the RNA case, the threading problem remains tractable as long as we score the alignment based on the notion of a common secondary structure. In fact, it reduces to a variant of the Sankoff algorithm in which the

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 728–735, 2004.
© Springer-Verlag Berlin Heidelberg 2004

energy model for structure prediction part is reduced to base weights on the base pairs.

Suppose we are given two sequences A and B (of length $n = |A|$ and $m = |B|$) together with their pair probability matrices P^A and P^B, resp. A natural way of determining the similarities of P^A and P^B is to search for the secondary structure of maximal "weight" that P^A and P^B have in common. Let $S_{i,j;k,l}$ be the score of the best matching of the subsequences $A[i..j]$ and $B[k..l]$. Furthermore, let $S^M_{i,j;k,l}$ be the best match subject to the constraint that (i, j) and (k, l) are matched base pairs. With this definition one obtains dynamic programming recursions

$$\begin{aligned} S_{i,j;k,l} &= \max \begin{cases} S_{i+1,j;k,l} + \gamma, \\ S_{i,j;k+1,l} + \gamma, \\ S_{i+1,j,k+1,l} + \alpha(A_i, B_k), \\ \max_{h \le j, q \le l} \left(S^M_{i,h;k,q} + S_{h+1,j;q+1,l} \right) \end{cases} \\ S^M_{i,j;k,l} &= S_{i+1,j+1,k+1,l+1} + \tau(P^A_{ij}, A_i, A_j; P^B_{kl}, B_k, B_l) \end{aligned} \tag{1}$$

with the initialization $S_{i,j;k,l} = |(j - i) - (l - k)|\gamma$ for $j - i \le M + 1$ or $l - k \le M + 1$, where M is the minimum size of a hairpin loop, usually $M = 3$. The constant $\gamma < 0$ is a gap penalty. The scores $\alpha_{ik} = \alpha(A_i, B_k)$ and $\tau_{ij,kl} = \tau(P^A_{ij}, A_i, A_j; P^B_{ij}, B_k, B_l)$ describe the substitution of unpaired bases and base pairs, respectively. The latter term may depend on both the structures and the underlying sequences. Backtracking can be used to retrieve both the common secondary structure and the associated sequence alignment [5]. For both RNA folding and sequence alignment it is possible to compute partitions functions instead of optimal scores with essentially the same resources. In a second step probabilistic versions of the optimal structure of alignment can be constructed; see [6] for RNA folding and [10,11,12,13,14]. In this contribution we describe a "partition function version" of the Sankoff algorithm that computes the probabilities of matches in the structure-based alignments of two RNA molecules, thereby providing an intrinsic measure of the *local* quality of the structure-based alignments. In the thermodynamic interpretation of the simultaneous folding and alignment problem a state θ is a pair $\theta = (\mathfrak{S}, \mathcal{A})$ of secondary structure $\mathfrak{S}$ consisting of all matched base pairs $(ij; kl)$, where (A_i, A_j) is a base pair in structure A and (B_k, B_l) is a base pair in structure B, and an alignment $\mathcal{A}$ of the underlying sequences A and B such that A_iB_k and A_jB_l are matches. Note that the alignment $\mathcal{A}$ in general contains further matches corresponding to unpaired nucleotides. The probability of a particular state is then

$$\mathrm{Prob}[\theta] = Z^{-1} \exp(-\sigma(\theta)) \tag{2}$$

where the score is given explicitly in the form

$$\sigma(\theta) = \sum_{(ij;kl) \in \mathfrak{S}} \tau_{ij,kl} + \sum_{i \in A, k \in B \ \notin \mathfrak{S}} \alpha_{ik} + \gamma\big(m + n - 2|\mathcal{A}|\big) \,. \tag{3}$$

In the last term, $N_{\text{gap}} = n + m - 2|\mathcal{A}|$ is the number of gaps in the alignment. The normalization constant

$$Z = \sum_{\theta} \exp(-\sigma(\theta)) \tag{4}$$

is the partition function of the model. The probability of a feature Ω can now be computed as the sum of the probabilities of all states $\theta \in \Omega$. In particular, we are interested in $\Omega^{(p,q)}$, the set of all states in which A_pB_q is a match in the alignment.

2 Recursions

We first observe that equ.(1) can easily be transformed into a recursion for the partition function $Z_{ij;kl}$ of the model restricted to the subsequences $A[i..j]$ and $B[k..l]$. Explicitly, we obtain

$$\begin{aligned} Z_{ij;kl} = {} & Z_{i+1,j;kl}e^{\gamma} + Z_{ij;k+1,l}e^{\gamma} + Z_{i+1,j;k+1,l}e^{\alpha_{ik}} \\ & + \sum_{\substack{(k,q) \text{ paired in } B \\ (i,p) \text{ paired in } A}} Z_{i+1,p-1;k+1,q-1}Z_{p+1,j;q+1,l}e^{\tau_{ij,pq}} \end{aligned} \tag{5}$$

Let us now consider all states that contain the match A_xB_y. We have to distinguish four cases: (i) there is no matched base pair in $\mathfrak{S}$, (ii) $(i,x;j,y) \in \mathfrak{S}$, (iii) $(x,k;y,l) \in \mathfrak{S}$, and (iv) A_xB_y is "immediately interior" to a matched pair $(i,j;k,l) \in \mathfrak{S}$ in the sense that $i < x < j$, $k < y < l$ and there no other pair $(i',j',k',l') \in \mathfrak{S}$ such that $i < i' < x < j' < j$ and $k < k' < y < l' < l$. Fig. 1 gives a graphical description: Clearly, these four cases are pairwise disjoint and cover all possibilities. We can therefore write the partition function Q^{xy} of all states that contain the match A_xB_y in the alignment as follows:

$$\begin{aligned} Q^{xy} = {} & Z_{1,x-1;1,y-1}e^{\alpha_{ij}}Z_{x+1,n;y+1,n} + \\ & \sum_{i<x,\, k<y} Z_{i+1,x-1;k+1,y-1}e^{\tau_{i,x;k,y}}\widehat{Z}_{i,x;k,y} + \\ & \sum_{j>x,\, l>y} Z_{x+1,j-1;y+1,l-1}e^{\tau_{x,j;y,l}}\widehat{Z}_{x,j;y,l} + \\ & \sum_{i<x,\, j>x\; k<y,\, l>y} e^{\tau_{i,j;k,l}}Z_{i+1,x-1;k+1,y-1}e^{\alpha_{xy}}Z_{x+1,j-1;y+1,l-1}\widehat{Z}_{i,j;k,l} \end{aligned} \tag{6}$$

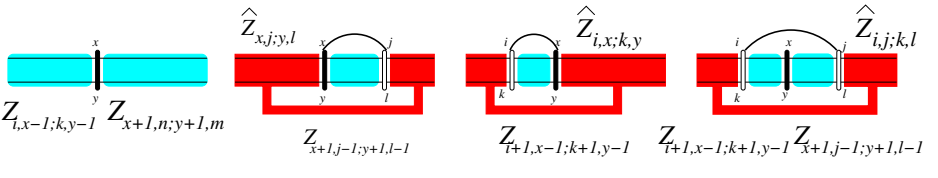


Fig. 1. Decomposition of the restricted partition function Q_{xy} into unconstrained partition functions $Z_{\ldots}$ and $\widehat{Z}_{\ldots}$ of sub-problems. For details see text

where $\widehat{Z}_{i,j;k,l}$ denotes the partition function over all partial states *outside* the aligned interval $[i,j][k,l]$, i.e., excluding the positions i, j, k, and l. This corresponds to the states of the sub-problem with $A' = A[1..i-1]A[j+1..n]$ and $B' = B[1..k-1]B[l+1,m]$. We can easily find recursions for computing $\widehat{Z}_{i,j;k,l}$ from shorter subproblems (i.e., those with a larger "missing" interval) and the values of $Z_{i,j;k,l}$:

$$\begin{aligned}\widehat{Z}_{i,j;k,l} =& \widehat{Z}_{i,j+1;k,l}e^{\gamma} + \widehat{Z}_{i,j;k,l+1}e^{\gamma} + \widehat{Z}_{i,j+1;k,l+1}e^{\alpha_{j+1,l+1}} + \\ & \sum_{p<i,\, q<k} \widehat{Z}_{p,j+1;q,l+1}e^{\tau_{p,j+1;q,l+1}} Z_{p+1,i-1;q+1,k-1} + \\ & \sum_{p>j+1,\, q>l+1} \widehat{Z}_{i,p;k,q}e^{\tau_{j+1,p;l+1,q}} Z_{j+2,p-1;l+2,q-1}\end{aligned} \tag{7}$$

The probability for the match $A_x B_y$ given the input data and scoring scheme is simply

$$P^{xy} = Q^{xy}/Z \tag{8}$$

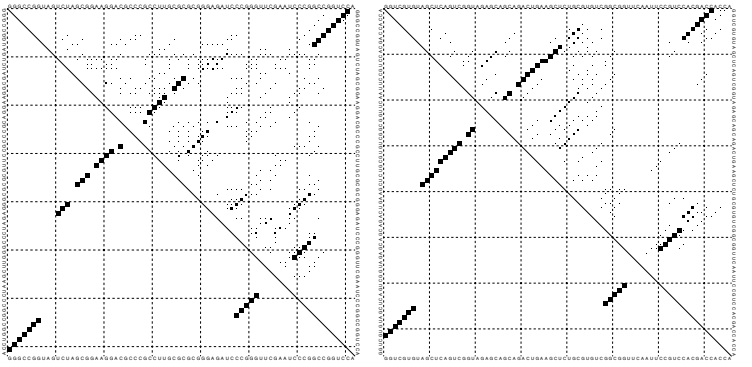
Tabulating the $\mathcal{O}(n^4)$ entries of the partition functions $Z_{ij;kl}$ requires $\mathcal{O}(n^6)$ operation, just as the solution of the optimization problem. Then $\widehat{Z}_{ij;kl}$ can be computed also in $\mathcal{O}(n^6)$ operations. Given these two tables, recursion 6 can be evaluated in $\mathcal{O}(n^4)$ steps for each value x and y. The matrix of matching probabilities can therefore be computed in $\mathcal{O}(n^4)$ memory and $\mathcal{O}(n^6)$ CPU. Just as in the case of sequence alignments and secondary prediction, the partition function version is therefore not more expensive than the associated optimization problem.

3 Stochastic Backtracking

As described in [5], backtracking in the recursions (1) can be performed in $\mathcal{O}(n^3)$ to obtain a score-optimal alignment. When the partition functions $Z_{i,j;k,l}$ for the sub-problems are known, it is possible sample from the distribution of the alignments by means of "stochastic backtracking". This approach has recently been implemented for pairwise sequence alignment [14] and for RNA structure prediction in the latest release of the Vienna RNA Package [8,16], see also [17, 18], and [19], where the idea was used to generate random RNA structures with uniform distribution. This method generalizes in a straightforward way to the Sankoff algorithm: From equ.(5) we obtain immediately that the sub-alignment of $x[i..j]$ with $y[k..l]$ can be of one of the four types listed below with their corresponding probabilities p:

Deletion of x_i	$p = Z_{i+1,j;kl}e^{\gamma}/Z_{ij;kl}$
Deletion of y_k	$p = Z_{ij;k+1,l}e^{\gamma}/Z_{ij;kl}$
Unpaired Match of x_i and y_j	$p = Z_{i+1,j;k+1,l}e^{\alpha_{ik}}/Z_{ij;kl}$
Matched pair (x_i, x_p), $(y_k - y_q)$	$p = Z_{i+1,p-1;k+1,q-1}Z_{p+1,j;q+1,l}e^{\tau_{ij,pq}}/Z_{ij;kl}$

Choosing in each step of the backtracking procedure one of these alternatives with the correct probability results again in an algorithm that produces an



DA0980 DF1140

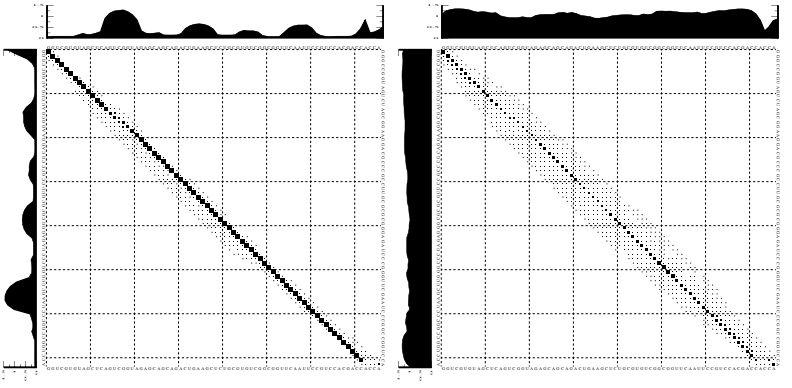
```
GGGCCGGUAGUCUAGC-GGAAGGACGCCCGCCUUGCGCGCGGGAGAUCCCGGGUUCGAAU-CCCGGCCGGUCC-A--
(((((((..(((...-....)))..(((((.......)))))..)..(((((.......-)))))))))))-.--
GGUCGUGUAGCUCAGUCGGUAGAGCAGCAGACUGAAGCUCUGCGUGUCGGCGGUUC-AAUUCCGUCCACGACCACCA
(((((((..(((........)))..(((((.......)))))..)..(((((...-....)))))))))))....
recalculated score:  27.29029  6
GGGCCGGUAGUCUAGCGGAA-GGACGCCCGCCUUGCGCGCGGGAGAUCCCGGGUUCGAAU-CCCGGCCGGUCC-A--
(.(((((..(((.......-)))).(((((.......)))))).....((.((.(...).-)).)))))).)-.--
GGUCGUGUAGCUCAGUCGGUAGAGCAGCAGACUGAAGCUCUGCGUGUCGGCGG-UUCAAUUCCGUCCACGACCACCA
(.(((((..(((........)))).(((((.......))))))).....((.((-(...)..)).)))))).)....
recalculated score:  24.59171  6
GGGCCGGUAGUCUA-GCGGAAGGACGCCCGCCUUGCGCGCGGGAGAUCCCGGGUUCGAAUCCCGGCCGGUC---CA
(((.((((.((((.-......)))).(((((.......)))))..)..((((((...)..)))))))).))---).
GGUCGUGUAGCUCAGUCGGUAGAGCAGCAGACUGAAGCUCUGCGUGUCGGCGGUUCAAUUCCGUCCACGACCACCA
(((.((((.(((........)))).(((((.......)))))..)..((((((...)..)))))))).))...).
recalculated score:  29.47619  4
```

low temperature: $T = 0.5$, $\gamma = -1.5$

```
GGGCCGGUAGUCUAGCGGA-A-GGACGCCCGCCUUGCGCGCGGGAGAUCCCGGGUUCGAAUCCCGGCCGGUCCA---
(((((((.(((.......-.-.))).(((((.......)))))..).)(.(((((...)..)))).).)))))).---
GGUCGUGUAGCUC-AGUCGGUAGAGCAGCAGACUGAAGCUCUGCGUGUCGGCGGUUCAAUUCCGUCCACGACCACCA
(((((((.(((.-.........))).(((((.......)))))..).)(.((((...)..)))).).)))))....
recalculated score:  21.14102  6
GGGCCGGUAGUC-UAGCGGAAGGACGCCCGCCUUGCG--CGCGGGA-GAUCCCGGGUUC-GAAUCCCGGCCGGUCC--A-
(((.(((...((-.........)...(((((.......--..)))).-.)..(((((...-...).)))))).)))).-.
GGUCGUGUAGCUCAGUCGGUAGAGCAGCAGAC-U-GAAGCUCUGC-GUGUCGGCGGUUCAA-UUCCGUCCACGACCACCA
(((.(((...((.........)...((((..-.-.-......))))-..)..(((((.....-.).)))))))).)))....
recalculated score:   5.04139  12
GGGCCGGUAGUCUA-GC--GGAAGGACGCCCGCCUUGCGCGCGGGAGAUCCCGGGUUCGAAUCC-CGGCCGGUCC-A--
.((((((...(...-..--.).....(..((.((.......)).)))...).(.((((....).))-).)))))).-.--
GGUCGUGUAGCU--CAGUCGGUAGAGCAGCAGACUGAAGCUCUGCGUGUCGGCGGU-UCAAUUCCGUCCACGACCACCA
.((((((...(.--.....)....(..((.((.......)).)))...).(.((((-...).)).).)))))).....
recalculated score:   6.33874  10
```

High temperature: $T = 1$, $\gamma = -2$

Fig. 2. Left: Two base pairing probability matrices of tRNAs taken from M. Sprinzl's tRNA database [15]: DA0980 (TGC from *Thermoproteus tenax* and DF1140 (GAA from *Mycoplasma capricolum*). Right: examples of pairwise alignments generated with two different parameter sets. The number of gaps (second column of numbers below the alignment) increases with temperature T even though $-\gamma/T$ decreases



$T = 0.5$, $\gamma = -1.5$ $T = 1$, $\gamma = -2$

Fig. 3. Match probabilities for the pairwise alignments of the two tRNAs DA0980 and DF1140 from Fig. 2. The area of the squares at position x, y is proportional to P^{xy}. The small panels along the axes show the position-wise entropies relative to each sequence

alignment in $\mathcal{O}(n^3)$ steps, such that the probability of an alignment with score σ is $p = \exp(-\sigma)$, Fig. 2. The advantage of this procedure is that an ensemble of on the order of n^3 sample alignments can be computed economically (since we need $\mathcal{O}(n^6)$ time for the forward recursion and only $\mathcal{O}(n^3)$ for backtracking a single alignment). These samples can then be used to estimate the probabilities of features such as particular multiloops or non-local sequence-structure combinations.

4 An Example

As an example we consider here the alignment of two rather disparate tRNA sequences, Figs. 2 and 3. We use here

$$\tau(P^A_{ij}, A_i, A_j; P^B_{kl}, B_k, B_l) = 2 \ln n + \ln P^A_{ij} + \ln P^B_{kl} \tag{9}$$

for the pair score and neglect sequence similarity altogether, i.e., $\alpha_{ik} = 0$. Note that for both sequences the predicted optimal secondary structures is not the clover-leaf, as shown in the l.h.s. of Fig. 2. Nevertheless, most stochastic backtrackings retrieve the clover-leaf as consensus structure of the two molecules. Since sequence similarity was not used in the scoring, the exact position of gaps within loop regions is arbitrary. For low temperatures (upper right panel in Fig. 2) alignments differ almost exclusively in the D-loop and at the 3' end of the tRNAs. The local reliability of the alignment can be measured by the entropy of the match probabilities

$$S(x) = -\sum_y P^{xy} \ln P^{xy} - p^0(x) \ln p^0(x) \tag{10}$$

where $p^0(x) = 1 - \sum_y P^{xy}$ is the probability the position x is unmatched (i.e., opposite to a gap in the alignment). As can be seen in Fig.3, the alignment is typically much more well-defined in paired regions. For large values of the temperature T this difference disappears, however.

5 Concluding Remarks

We have introduced here a partition function version of the Sankoff algorithm. The algorithm is quite expensive both in memory and CPU time; the resource requirement is, however, essentially the same as for the "classical" version that computes the optimal alignment only. From the partition functions we can, in addition to the optimal alignment, also discriminate reliable from unreliable parts of a structure-based alignment of RNA molecules. Stochastic pairwise alignments are useful in many different contexts: Numerous tools in bioinformatics require pairwise sequence alignments as input data. The present approach thus provides a tool that can be used to produce alignments with realistically distributed errors and varying overall quality (by choosing the temperature parameter T). These can be used to investigate the sensitivity of the method with respect to realistic variations of the input alignments. In particular, used as an input of a multiple alignment methods such as `t-coffee` [20] it can be used to produce multiple alignments together with estimates of local alignment quality. While the Sankoff algorithm is too slow to scan large portions of a genome for conserved RNAs, it is still useful to post-process candidates for structurally conserved RNA detected by other methods, e.g. `qrna` [21]. The current implementation uses simple linear gap costs. A generalization to affine gap costs is straightforward along the lines of Gotoh's algorithm [22] for sequence alignments and should improve the placement of scattered gaps.

Acknowledgments. This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893 and the DFG Bioinformatics Initiative BIZ-6/1-2.

References

1. Sankoff, D.: Simultaneous solution of the RNA folding, alignment, and protosequence problems. SIAM J. Appl. Math. **45** (1985) 810–825
2. Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J. Mol. Biol. **288** (1999) 911–940
3. Gorodkin, J., Heyer, L.J., Stormo, G.D.: Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucl. Acids Res. **25** (1997) 3724–3732
4. Mathews, D.H., Turner, D.H.: Dynalign: An algorithm for finding secondary structures common to two rna sequences. J. Mol. Biol. **317** (2002) 191–203

5. Hofacker, I.L., Bernhart, S., Stadler, P.F.: Alignment of rna base pairing probability matrices. Bioinformatics (2003) submitted.
6. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29** (1990) 1105–1119
7. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P.: Fast folding and comparison of RNA secondary structures. Monatsh. Chemie **125** (1994) 167–188
8. Hofacker, I.L.: Vienna RNA secondary structure server. Nucl. Acids Res. **31** (2003) 3429–3431
9. Lathrop, R.H.: The protein threading problem with sequence amino acid interaction preferences is np-complete. Protein Eng. **7** (1994) 1059–1068
10. Bucher, P., Hoffmann, K.: A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States, D.J., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R.F., eds.: Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB '96), Menlo Park, CA, AAAI Press (1996) 44–50
11. Kschischo, M., Lassig, M.: Finite-temperature sequence alignment. Pacific Symposium Biocomputing **1** (2000) 624–35
12. Miyazawa, S.: A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng. **8** (1994) 999–1009
13. Yu, Y.K., Hwa, T.: Statistical significance of probabilistic sequence alignment and related local hidden markov models. J. Comp. Biol. **8** (2001) 249–282
14. Mückstein, U., Hofacker, I.L., Stadler, P.F.: Stochastic pairwise alignments. Bioinformatics **S153-S160** (2002) 18 ECCB 2002.
15. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S.: Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. **26** (1998) 148–153
16. Flamm, C., Hofacker, I.L., Stadler, P.F.: Computational chemistry with RNA secondary structures. Kemija u industriji (2004) Proceedings CECM-2, Varaždin, June 19-21, 2003.
17. Ding, Y., Lawrence, C.E.: Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res. **29** (2001) 1034–1046
18. Ding, Y., Lawrence, C.E.: A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. **31** (2003) 7280–7301
19. Tacker, M., Stadler, P.F., Bornberg-Bauer, E.G., Hofacker, I.L., Schuster, P.: Algorithm independent properties of RNA structure prediction. Eur. Biophy. J. **25** (1996) 115–130
20. Notredame, C., Higgins, D., Heringa, J.: T-coffee: A novel method for multiple sequence alignments. J. Mol. Biol. **302** (2000) 205–217
21. Rivas, E., Eddy, S.R.: Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics **2** (2001) 19 pages
22. Gotoh, O.: An improved algorithm for matching biological sequences. J. Mol. Biol. **162** (1982) 705–708

Simulation of Molecular Evolution Using Population Dynamics Modelling

Sergei V. Semovski[1]

Limnological Institute SB RAS, Ulanbatorskaya 3, Irkutsk 664033 Russia
semovsky@lin.irk.ru

Abstract. Neutral evolution of nucleotide sequences is simulated in the framework of adaptive population dynamics. Simulated speciation events, changes in demographic history, and migration are traced using resulting molecular data. Various scenarios are considered including one- and two-dimensional populations, and competition lottery system.

1 Introduction

It has been shown recently that DNA sequences retain much more information on the past events than simply the degree of relation between species. The main successes in the direction of interpreting this information have been achieved in the framework of the coalescence theory which has allowed scientists to obtain the demographic history of human beings. We use mathematical modelling to simulate evolutionary and demographic processes in populations consisting of objects with different ecological parameters such as mobility, number of progeny etc., and include into each object tracer mutable "DNA sequences", which are extracted after each simulation and analysed with various molecular phylogenetic approaches. This will enable us to match the inferences to the known history.

Population dynamics is a flexible tool, giving opportunities to investigate various practically important ecological situations. In the Lab of molecular systematics, at the Limnological Institute we concentrate mainly on studies of benthic invertebrates (see [1])). Several models of molecular evolution valid for low-mobile species are presented here including a one-dimensional population, which is the model of littoral of an ancient lake, and different ecological and evolutionary patterns in a two-dimensional community.

2 Model

Let us consider a population existing in the environment with limited resource, according to a generalisation [2] of the well-known logistic equation:

$$\frac{dS(x,t)}{dt} = r \cdot S(x,t) \cdot \left[1 - \int \frac{C(x-y)S(y,t)}{K(x)} dy\right]. \quad (1)$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 736–741, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Here $S(x,t)$ denote the number of individuals with value of ecological character (polygen) x in time t. Resource distribution and competition between individuals are considered to be gaussian functions:

$$K(x) = K_0 \exp\left[-\frac{(x-x_0)^2}{2\sigma_K^2}\right] \tag{2}$$

and

$$C(z) = \exp\left[-\frac{z^2}{2\sigma_C^2}\right]. \tag{3}$$

The probability of progeny survival depends on the genetic distance between the parents. The local minimum in zero point is due to inbreeding, and the local maximum is due to heterosis. Hybrid depression at genetic distances larger then K_0 causes reproductive isolation of genetically different organisms.

Rare mutations occur with the probability equal to 0.01 and their deviation from parent genotype corresponds to gaussian distribution

$$\Psi(x) = (2\pi\sigma_\Psi^2)^{-1} \exp\left[-x^2/\sigma_\Psi^2\right], \tag{4}$$

where σ_Ψ^2 is dispersion of the genotype of the rare mutant. The individual-oriented model has been used for all calculations.

For spatially distributed individuals the generalization of the model [2] is proposed in [3]. The general idea of application of (1) is in the introduction of the Gaussian competition between individuals in the space. Then integral operator on the right side of (1) becomes multidimensional and contains an additional kernel $C_{sp}(z) = \exp\left[-\frac{z^2}{2\sigma_{sp}^2}\right]$ responsible for competition in the space. Here z is the distance between individuals in the space and σ_{sp} defines the intensity of spatial competition. Note that for the sake of calculation time it is possible to use rectangular kernel instead of Gaussian. Such formulation will correspond to the limited capacity of the spatial cell. It is possible to use the same attempt to describe the dynamics of interacting populations. If competition for the same resource takes place in the space, individuals from different populations interact according to the integral operator with kernel C_{sp}, however, in the "genetic" space of the quantative trait competition is intra-specific.

In order to bring the neutrally evolving marker into the individually oriented model we consider that each individual bears a neutrally evolving nucleotide sequence represented here by integer vector with elements having values from 1 to 4 We consider all mutations in this sequence neutral that have no impact on the adaptation of individuals. The progeny inherits parental sequence with mutations occurring with the probability of .001 per generation. In the case of a bisexual population this sequence would be considered to be inherited from one of the parents only (maternal inheritance), which corresponds to mitochondrial DNA. The set of neutral sequences resulting from a simulation was used for phylogenetic inferences, which were performed with the program fitch from the package PHYLIP [5]. The phylogenetic trees were visualized with the program njplot [4]. Program system IDL (Reseach Systems Inc.) has been used intensively for computations and results presentation.

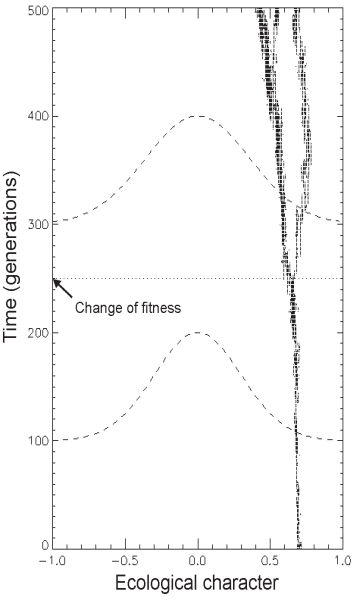


Fig. 1. Simulation of evolutionary history using model (1)

2.1 Molecular Evolution

Incorporating the molecular evolution in the population dynamics model (1) produces a suitable tool to validate those methods of bioinformatics which have been developed to reconstruct details of evolutionary history based on the present samples of those DNA sequences which are selectively neutral. One of important questions is the correlation of demographic changes in population history and speciation events with past changes in the environment. In [6] some details are presented of the application of the model (1) to reconstruct the evolutionary history for different scenarios of environmental change impact on speciation intensity. In Fig. 1 one can see formation of genetically isolated subpopulations after change in fitness. Fig. 2 demonstrates application of molecular phylogeny to resulting molecular data.

2.2 Spatial Patterns of Speciation

Bottom communities of ancient lakes demonstrate various spatial patterns in speciation. In the shallow littoral zone along the shore the benthic community presents approximately a one-dimensional population. In [3] using a generalization of the model (1) different situations of speciation in space have been demonstrated depending on mobility of organisms. For low-mobile species formation of different, genetically isolated colonies is possible, which is close to allopatric speciation but without geographical barriers. For moderate mobility parapatric

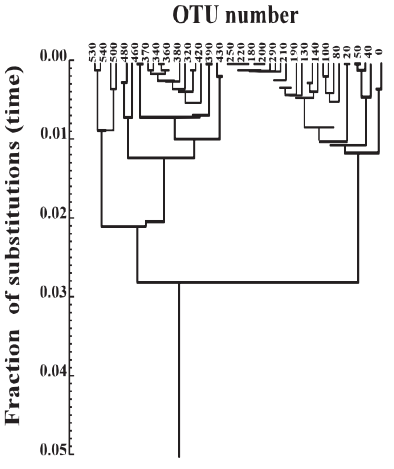


Fig. 2. Reconstruction of evolutionary history of the case presented in Fig. 1 using resulting molecular data

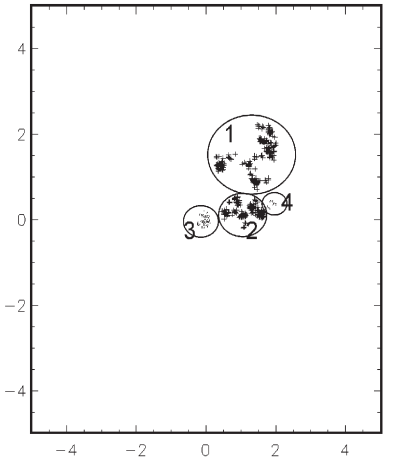


Fig. 3. Speciation patterns for two competitive populations on a plane. 1 and 2 are spatial positions of two genetically different morphs of population A, 3 and 4 are spatial positions of two genetically different morphs of population B

speciation takes place, for which zones of genetic contact are typical, both permanent and temporary, with migration or without it. For highly mobile organisms situation becomes close to zero-dimensional case (1) and we have sympatric speciation with full mixing of newly formed subpopulations.

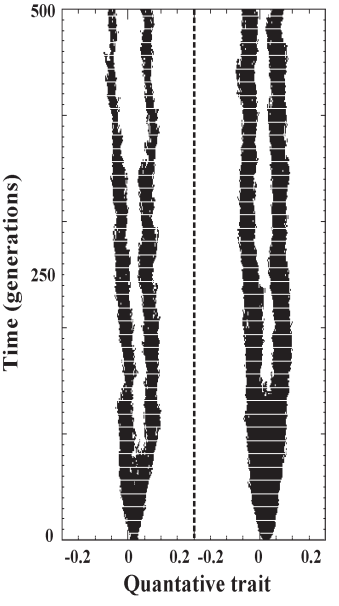


Fig. 4. Co-evolution patterns of two competing populations on a plane

2.3 Co-evolution in Competing Populations

Co-evolution and co-speciation is known for many forms of interaction between species in nature. It is possible for host and parasite pairs, preys and predators. We have a plan to investigate co-evolutionary patterns for different forms of ecological interaction. Results presented here are derived using a model of co-evolution of spatially distributed organisms which compete for a limited resource.

For many natural communities the well-known principle of competitive exclusion is not valid. One of the examples one can find in benthic ecosystems. One of verbal models is based on the idea of "competition lottery", when newborn offspring have equal probability to occupy areas which are free due to organisms mortality. The numerical realization of this idea is based on a three-dimensional generalization of the model (1), where one dimension is a quantative trait (polygenus) and two domensions are spatial. Fig. 3. demonstrates spatial patterns of new parapatric (being in contact) genetically diverging morphs for two competitive species on a plane. To neglect boundary effects in calculations we use torus. Co-evolution patterns in the "genetic" space are presented in Fig. 4.

Changes in the environment during evolutionary history can have different consequences for competing populations, correlative and anti-correlative. Fig. 5 demonstrates the analysis of evolutionary history using resulting molecular data and frequency diagram of pair-wise genetic distances of organisms in the population. Peaks in this diagram correspond to moments of population growth[7], or to moments of speciation [6], minima - to population decline or to decrease in

speciation rate. In Fig. 5 two populations demonstrate opposite reactions to the changes in the model environment. Note that a similar anti-correlation has been noted recently for sand-dwelling and rock-dwelling cichlids in lake Tanganyika [8] using a statistical analysis based on a big array of molecular data.

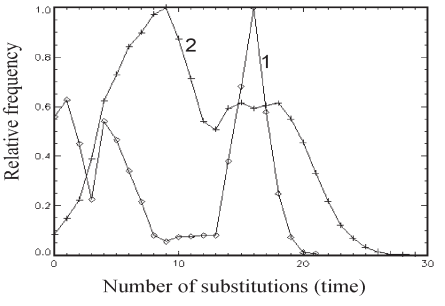


Fig. 5. Frequency diagram for genetic distance (number of substitutions) in two competing populations on a plane which react in different ways to changes in the model environment

References

1. Sherbakov D.Yu.: Molecular phylogenetic studies on the origin of biodiversity in Lake Baikal. Trends in Ecol.and Evolution **14** (1999) 92–95.
2. Dieckmann U., Doebeli M.: On the origin of species by sympatric speciation, Nature, **400** (1999) 354–357.
3. Semovski S.V., Bukin Yu.S., Sherbakov D.Yu.,: Speciation in one-dimensional population. Int.J.of Modern Phys. **14** (2004), (to appear).
4. Gouy, M. : NJPLOT(m.gouy@biomserv.univ-lyon1.fr)
5. Felsenstein, J.: PHYLIP, Phylogeny Inference Package *http://evolution.genetics.washington.edu/phylip.html*
6. Semovski S.V., Verheyen E., Sherbakov D.Yu.,: Simulating the evolution of neutrally evolving sequences in a population under environmental changes. Ecol.Modelling, (to appear)
7. Rogers, A.R., Harpending, H. : Population growth makes waves in the distribution of pairwise genetic differences, Mol.Biol.Evol., **9** (1992) 552–569.
8. Nevado B., Sherbakov D.Yu., Verheyen E., : private communications.

Lotka-Volterra Model of Macro-Evolution on Dynamical Networks

François Coppex[1], Michel Droz[1], and Adam Lipowski[1,2]

[1] Institut de Physique Théorique, Université de Genève,
quai E. Ansermet 24, 1211 Genève 4, Switzerland,
[2] Faculty of Physics, A. Mickiewicz University, 61-614 Poznań, Poland

Abstract. We study a model of a multi-species ecosystem described by Lotka-Volterra-like equations. Interactions among species form a network whose evolution is determined by the dynamics of the model. Numerical simulations show power-law distribution of intervals between extinctions, but only for ecosystems with sufficient variability of species and with networks of connectivity above certain threshold that is very close to the percolation threshold of the network. Effects of slow environmental changes on extinction dynamics are also examined.

1 Introduction

Lotka-Volterra models of interacting species have a well established role in population ecology [1]. Being inspired by an oscillatory behavior in some prey-predator systems, they are typically used to model populations on time scale shorter than lifetime of describing species. It means that long-term properties of ecosystems (macro-evolution) are usually not captured within such an approach. On the other hand, models used to describe macro-evolution very often use a dynamics that operates at the level of species rather than individuals. Such coarse-grained models usually refers to the notion of fitness of a species that is not commonly accepted [2]. Recently, there has been some attempts to study macro-evolution using models equipped with dynamics that operates at the level of individuals [3,4,5]. Taking into account that Lotka-Volterra models are relatively successful in describing many aspects of population dynamics it would be desirable to apply such an approach also to macro-evolution. Some time ago Abramson introduced a discrete version of Lotka-Volterrra ecosystem [6] and studied certain characteristics of extinctions. His model is an example of a one-dimensional food chain with $N(\sim 100)$ trophic levels and a single species occupying a given trophic level. Since in realistic food webs $N \sim 4-6$ with typically many species belonging to a given trophic level [7,8], these are highly nonrealistic assumptions. Nevertheless, extinction dynamics in Abramson's model shows some features that are characteristic to Earth biosystem.

In the present paper we introduce a Lotka-Volterra model that describes a simplified ecosystem of N species of predators and one species of preys. Our model can be thus considered as a simple food web model with only two trophic levels. Competition between predator species is described by a certain random

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 742–749, 2004.
© Springer-Verlag Berlin Heidelberg 2004

network [9] of interactions whose evolution is coupled with dynamics of the model. Namely, when a certain species becomes extinct (i.e., its density falls below a certain threshold) it is replaced by new species with a newly created set of interactions with some of existing species. Despite obvious simplifications the model exhibits some properties that are typical to more complicated ecosystems, as for example power-law distributions of intervals between extinctions. Within our model we can also examine how robust this power-law distribution is. We find that under certain conditions, as for example very sparse interactions between species, or too strong dominance of a small group of species, these power-law characteristics disappear and the model is driven into a regime where extinctions have exponential distributions or where there are no extinctions and the ecosystem enters a steady state. In our opinion, such regimes might be relevant when a restricted (either in space or time) evolution of an ecosystem or its part is studied. Interestingly, a threshold value of connectivity that separates power-law extinctions and steady state is very close to the percolation threshold of the random network of inter-species interactions.

One of the important results coming from applying statistical physics models to biological evolution is that avalanches of extinctions do not require external factors to trigger them, but might be a natural consequence of the dynamics of an ecosystem. As a result, these external factors, as e.g., climate changes, solar activity or impact of a big meteorite, are very often neglected in such studies [10]. But such factors certainly affect the ecosystem and there is a good evidence of it [11]. One possibility to take external factor(s) into account in our model is to modify a growth rate of prey. modify a growth rate of prey. Since dynamics of the model is nonlinear, such a change might have more dramatic consequences than merely a change of densities of species. And indeed we noticed that dynamics of extinctions is strongly dependent on the growth rate. It turns out, that in our model abundance of preys leads to a larger frequency of extinctions, and in periods of hunger there are less extinctions. This is clearly due to nonlinearity of the dynamics. Larger growth rate increases the density of preys that in turn increases densities of predators. With increased densities, dynamics becomes more competitive and extinctions become more frequent. Such a periodically modulated growth rate leaves some traces also in the probability distribution of extinctions. It might be intersting to notice that paleontological data also show some traces of periodic events, but their proper understanding is still missing [11, 12]

2 The Model and Numerical Calculations

We study a Lotka-Volterra ecosystem that consists of N species of predators with densities ρ_i $(i = 1, 2, \ldots, N)$ who are all feeding on one species of preys with density ρ_0. We assume that each predator species i is characterized by a parameter k_i $(0 < k_i < 1)$ that enters evolution equations of the model through death and growth terms

$$\dot{\rho_0} = g(t)\rho_0(1 - \rho_0) - \frac{\rho_0}{N}\sum_{i=1}^{N} f(k_i)\rho_i \tag{1}$$

$$\dot{\rho_i} = -\rho_i(1-\rho_0) + f(k_i)\rho_i\rho_0\left(1 - \frac{k_i\rho_i + \sum_j' k_j\rho_j}{k_i + \sum_j' k_j}\right), \tag{2}$$

where $i = 1, 2, \ldots, N$. In our model we assume that species interact mainly through environmental capacity terms (the last term in Eq. (2)). Namely, the growth rate of a given species i is reduced not only due to its density but also due to weighted (with the factor k) densities of a group of randomly selected neighbor species. In Eq. (2) summation over these neighbouring species is denoted by $(\sum')$. Approximately, we might interpret the coefficient k_i as the size of organisms of i-th species – the bigger they are the stronger their role in the environmental capacity term. We also assume that the growth rate of preys is corrected by the environmental capacity term and due to external factors might be a slowly varying function of time. In principle, external factors might affect also other terms of model (1)-(2), but for simplicity we restrict its influence only to the growth rate of preys. Functions $g(t)$ and $f(k)$ will be specified later.

Differential equations (1)-(2) are solved using Runge-Kutta fourth-order method. Each time a density of a certain species drops below a threshold value which we fix as $\varepsilon = 10^{-7}$ we consider such a species as extinct [13]. Such species is then replaced by a new species with a randomly assigned density (from the interval (0,1)), the coefficient k ($0 < k < 1$) that is randomly drawn from the distribution $p(k)$, and a new set of neighbors (all links of the 'old' species are removed). With such rules the model rather describes N niches, and we assume that a time to create a species that will occupy a niche is relatively short comparing to the typical lifetime of species [14].

We assume that a newly created species makes z links with randomly selected neighbors. Links are not directional so a newly created species will also enter the evolution equation of species it is neighbouring. If the extinct species would be chosen randomly the network of interactions would have been a random graph. However, it is the dynamics (1)-(2) that determines which species are extinct. Thus, extinct species are not selected randomly and the resulting network is in general not a random graph.

3 Results

In the following we describe numerical results obtained for some particular cases of model (1)-(2).

3.1 Intervals between Extinctions

Various paleontological data suggest that dynamics of extinctions has some power-law distributions of sizes or durations [11]. In our model we measured time intervals t between successive extinctions. In this calculations we used a constant growth term of preys $g(t) \equiv 1$. We examined two cases: (i) model I: $f(k_i) \equiv 1$ and (ii) model II: $f(k_i) = k_i$. Unless specified otherwise we select k_i randomly with a homogeneous distribution on the interval (0,1) ($p(k) = 1$). Our results are shown in Fig. 1. In the simplest case, model I with $z = 2$ and $k_i \equiv 1$

(i.e., all species during the evolution have identical k_i) we obtain exponentially decaying distribution of intervals between extinctions $P(t)$. Such a decay is also seen for model I (z=4) with linear distribution of k_i namely $p(k) = 2k$. We expect that such a behavior appears when a distribution of k_i in the ecosystem is relatively narrow and shifted toward unity. Such an effect might be due to the small width of distribution $p(k)$ (i.e., a distribution from which we draw k_i) or might be dynamically generated as in model II. In this case even though k_i are chosen from a homogeneous distribution, the dynamics favours large k_i species (due to their larger growth rate) and they dominate the ecosystem. When the distribution of k_i in the ecosystem is more uniform (model I with $p(k) = 1$) our simulations suggest that $P(t)$ decays as a power law. Let us notice, however, that a power-law behavior is seen only on approximately one decade and we cannot exclude that on a larger time scale a different (perhaps exponential) behavior appears as was already observed in some other macroevolutionary models [3]. Let us also notice that for model I with $p(k) = \frac{1}{2}k^{-1/2}$ the power-law distribution $P(t)$ seems to decay as t^{-2}, i.e., with the exponent consistent with some paleontological data [11] as well as with predictions of some other models [4].

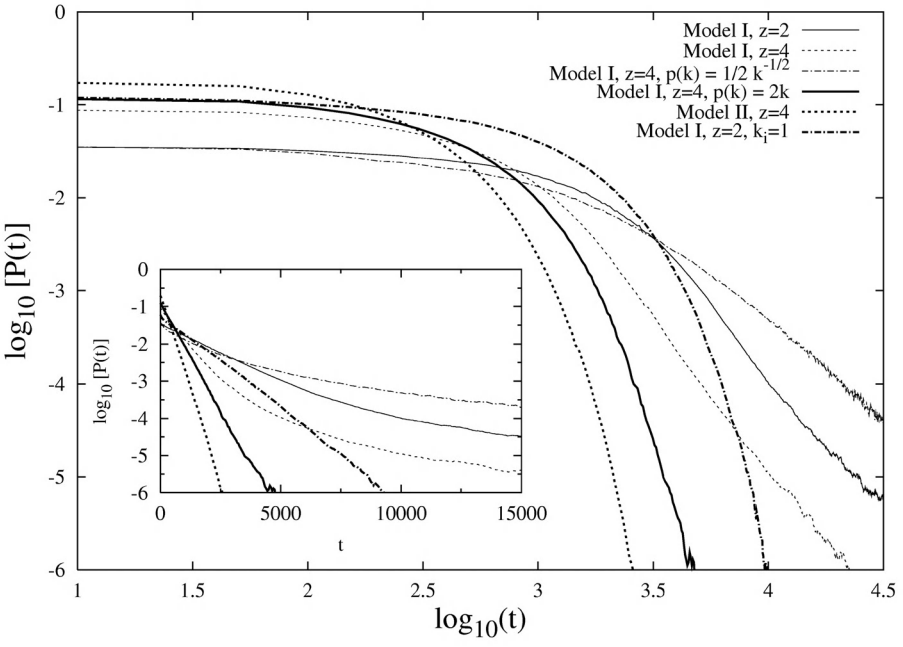


Fig. 1. Probability distribution of intervals between successive extinctions $P(t)$ calculated for some particular cases of model (1)-(2) for $N = 100$. Inset shows the same data but plotted on a lin-log scale

However, a power-law decay of $P(t)$ is seen only for sufficiently large z. When z is too small, we observed that the ecosystem enters the steady state where all ρ_i are positive and there are no extinctions. This is probably due to the fact that the competition among predators is too weak (or rather too sparse). To examine the transition between this two regimes in more detail we measured the averaged time between extinctions τ and the results are seen in Fig. 2. One can see that τ diverges around $z \sim 1.8$ [15]. Such a value of the threshold parameter suggests that this transition might be related with the percolation transition in our network of interspecies interactions. To examine such a possibility we measured the average size of the largest cluster of connected links in the network R (normalized by the number of species N) and the results are shown in Fig. 2. Vanishing of this quantity locates the percolation transition [16]. One can see that the percolation transition takes place at a larger value namely around $z \sim 2.0$. Our results suggest that these two transitions take place at different values of z. However the analysis of finite size effects especially in the estimation of τ is rather difficult and we cannot exclude that these two transitions actually overlap, as might be suggested by their proximity. Such a result would show that a dynamical regime of an ecosystem is determined by the geometrical structure of its interactions.

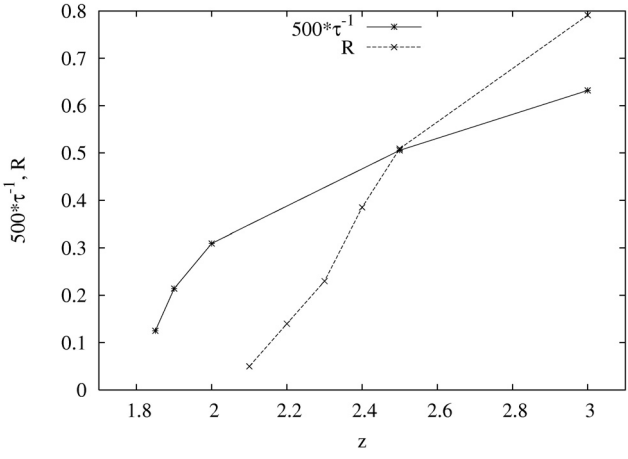


Fig. 2. The inverse average time between extinctions τ^{-1} and the percolation probability R as a function of z. Plotted results are based on calculations for $N = 100$, 200, 300 and 400 and extrapolation $N \to \infty$

3.2 Effect of a Modulated Growth Rate

Now we examine the role of a modulated in time growth rate of preys. Such a modulation is supposed to mimic the influence of an external factor like a

change of a climate. One of the questions that one can ask in this context is how such a change affects the extinction dynamics. We studied model I with $p(k) = 1$. The growth rate of preys we chose as $g(t) = 1 + A\sin(\frac{2\pi t}{T})$, where A and T are parameters. A typical behavior in case of model I with such a growth rate is shown in Fig. 3. One can see that increased growth rate increases the

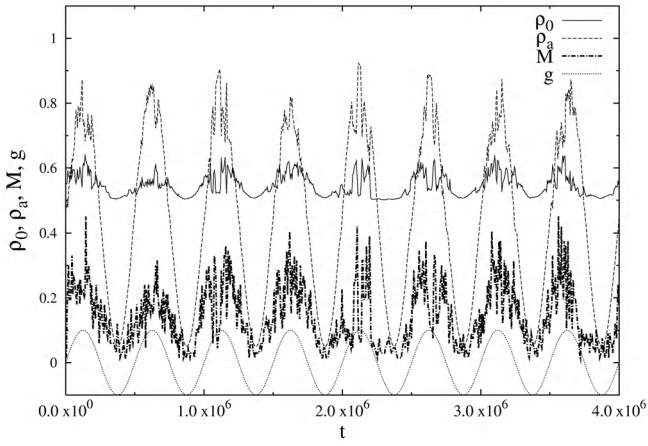


Fig. 3. A time evolution of the density of preys ρ_0, average density of predators $\rho_a = \frac{1}{N}\sum_{i=1}^{N}$, and the number of extinctions M (divided by 20) in the time interval $\Delta t = 10^3$ for the model I with $N = 100$ and $z = 4$. A rescaled modulated growth rate $(g(t) - 1)/10 = 0.09\sin(\frac{2\pi t}{T})$ $(T = 10^5)$ is also shown

density of preys ρ_0 that increases the density of predators. However, it increases also the frequency of extinctions. Such a behavior, namely increased extinction rate during abundance of food, might at first sight look as counterintuitive. This effect is related with the form of environmental capacity terms in in the growth rate in Eq. 2, namely $1-(k_i\rho_i+\sum_j' k_j\rho_j)/(k_i+\sum_j' k_j)$. Such term certainly has a larger variability for increased density of predators ρ_i, and for some species (depending on the distribution of links, coefficients k_i and densities) it causes faster extinction. Let us also notice that since period of modulation T is quite large, there is no retardation effect between density of preys and predators. We observed such retardation for smaller values of $T(\sim 1000)$.

Modulated growth rate of prays affects also the probability distribution of intervals between extinctions $P(t)$ as shown in Fig. 4. One can see that period of modulation T is imprinted in $P(t)$. Let us notice that certain paleontological data do show some signs of periodicity but its origin still remains unclear [12, 11].

It is known that slowly changing ecosystems sometimes undergo catastrophic shifts [17]. As a result, the ecosystem switches to a contrasting alternative stable state. It would be interesting to examine whether multi-species ecosystems, as

described by our model (1)-(2), might also exist in such alternative states. If so, one can ask whether for example structure of the network of interspecies interactions or extinction dynamics are the same in such states.

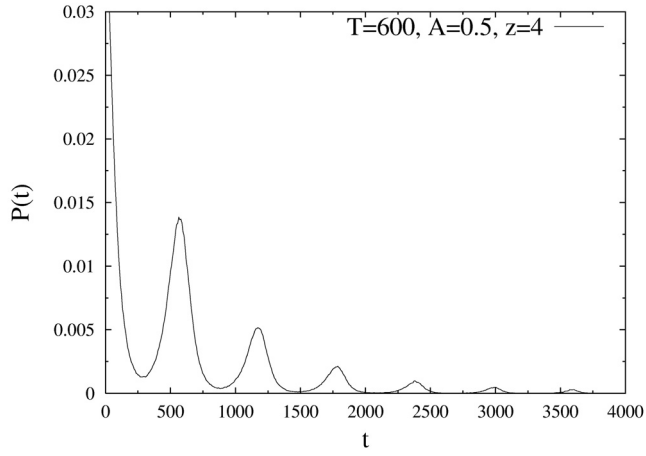


Fig. 4. Probability distribution of intervals between successive extinctions $P(t)$ calculated for model I with modulated growth rate ($N = 100$)

4 Conclusions

In the present paper we studied extinction dynamics of a Lotka-Volterra model of a two-level food web. In our model N species of predators feed on a single species of preys. Competition between predators, that is specified by a certain network of interactions, leads to their extinction and replacement by new species. Distribution of intervals between successive extinctions in some cases has power-law tails and thus resembles extinction pattern of the real ecosystem. However, when the network of interactions between predators is too sparse the ecosystem enters the steady state. We have shown that such a change of behavior might be related with a percolation transition of the network. We also examined an influence of external factors on the evolution of the ecosystem. More specifically, we studied the evolution of our model in case when the growth rate of preys is changing periodically in time. It turns out that such a modulation substantially changes the frequency of extinctions. Counterintuitively, periods with abundance of preys have higher frequency of extinctions than periods with lesser amount of preys.

It would be desirable to examine some extensions of our model. For example one can introduce additional trophic levels or other forms of interspecies interactions. One can also examine a variable number of species that would allow

to create new species using certain mutation mechanism rather than assuming that they appear as soon as a niche becomes empty. Another possibility that is outside the scope of majority of macro-evolutionary models, and that will be discussed in the forthcoming paper [18], is to examine emergent properties of species.

Acknowledgement. This work was partially supported by the Swiss National Science Foundation.

References

1. J. D. Murray, *Mathematical Biology*, (Springer, 1989). J. Hofbauer and K. Sigmund *The Theory of Evolution and Dynamical Systems*, (Cambridge University Press, 1988).
2. P. Bak and K. Sneppen, Phys. Rev. Lett. **71**, 4083 (1993). B. Drossel, Adv. Phys. **50**, 209 (2001).
3. D. Chowdhury, D. Stauffer, and A. Kunvar, Phys. Rev. Lett. **90**, 068101 (2003).
4. P. A. Rikvold and R. K. P. Zia, Phys. Rev. E **68**, 031913 (2003).
5. M. Hall, K. Christensen, S. A. di Collobiano, and H. J. Jensen, Phys. Rev. E **66**, 011904 (2002).
6. G. Abramson, Phys. Rev. E **55**, 785 (1997).
7. B. Drossel and A. J. McKane, e-print: nlin.AO/0202034. D. Chowdhury and D. Stauffer, e-print: q-bio.PE/0311002.
8. C. Quince, P. G. Higgs, and A. J. McKane, in *Biological Evolution and Statistical Physics*, eds. M. Lässig and A. Vallerian (Springer Verlag, Berlin-Heidelberg 2002).
9. R. Albert and A. -L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
10. One of the few macroevolutionary models that takes into account external factors was studied by Roberts and Newman (B. W. Roberts and M. E. J. Newman, J. Theor. Biol. **180**, 39 (1996)). This model, however, is a variant of Bak-Sneppen model with dynamics operating at the level of species.
11. M. E. J. Newman and R. G. O. Palmer, e-print: adap-org/9908002.
12. D. M. Raup and J. J. Sepkoski, Proc. Natl. Acad. Sci. **81**, 801 (1984).
13. Statistical properties of extinctions that we study in this paper are not affected by the precise value of the ε as soon as it has a small value ($\varepsilon \ll 1$).
14. The fact that we remove a species as soon as its density drops below ε might suggest some similarity with Bak-Sneppen (BS) model. Let us notice, however, that in BS model at each time step a species is removed with the lowest fitness. In our model it is the dynamics that determines intervals between extinctions. Morevoer, in our model dynamical variables are densities of species, that in principle are measurable quantities, and not fitness.
15. For noninteger z (i.e., the number of links of newly created species) we used the following procedure: with probability $z - [z]$ we created $[z] + 1$ links and with probability $1 + [z] - z$ we created $[z]$ links. On average such a recipe produces z links.
16. D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London 1982).
17. M. Scheffer, S. Carpenter, J. A. Foley, C. Folks, and B. Walker, Nature **413**, 591 (2001).
18. F. Coppex, M. Droz, and A. Lipowski, in preparation.

Simulation of a Horizontal and Vertical Disease Spread in Population

Maria Magdoń-Maksymowicz

Department of Mathematical Statistics
AR-Agricultural University
Al. Mickiewicza 21
31-120 Krakow, Poland
rrmagdon@cyf-kr.edu.pl

Abstract. The vertical disease spreading from parent to offspring and/or horizontal transmission through infection is discussed, using cellular automata approach implemented on a $N \times N$ lattice. We concentrate on age distribution of the population, resulting from different scenario, such as whether newborns are placed in close vicinity of parents or separated from them. We also include migration aspect in context of disease spreading. Main conclusions drawn are that the vertical version is resistant to manipulations of parameters which control migration. Horizontal version represents self-recovering population unless migration of grown-ups is introduced for the case of offsprings located in vicinity of parents. Then the migration seems to be beneficial for highly infectious and lethally diseases, while it brings more deaths for milder infections.

1 Introduction

Most population evolution models are based on differential equations which describes a statistically significant and representative member of the population and applies deterministic rules to its time evolution. Time is then a continuous variable. However it is not so often that then we may solve the set of differential equations and then we apply discrete time and set rules to predict the $t \to t+1$ transition of the system. Obviously, it may work for the time step sufficiently small, when changes in parameters characterizing the system are small, too. In simulation iterations, time is discrete. The the system often shows elementary interactions between its components on microscopic scale, which is not well described in terms of spatially continuous distribution. In fact, we often get quite different result [1]. Cellular automata [2] is a proper tool for that case, either in the standard deterministic version or for the probabilistic rules. If the system is also vulnerable to some non-deterministic component, it is easy to implement erratic behaviour as a noise or more correlated deviations from the deterministic picture.

The cellular automata technique is often used to describe dynamics of the infection by some viruses [3]. In basic epidemiological models, a disease may be transmitted horizontally through infection (say, due to a direct contact) and/or

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 750–757, 2004.
© Springer-Verlag Berlin Heidelberg 2004

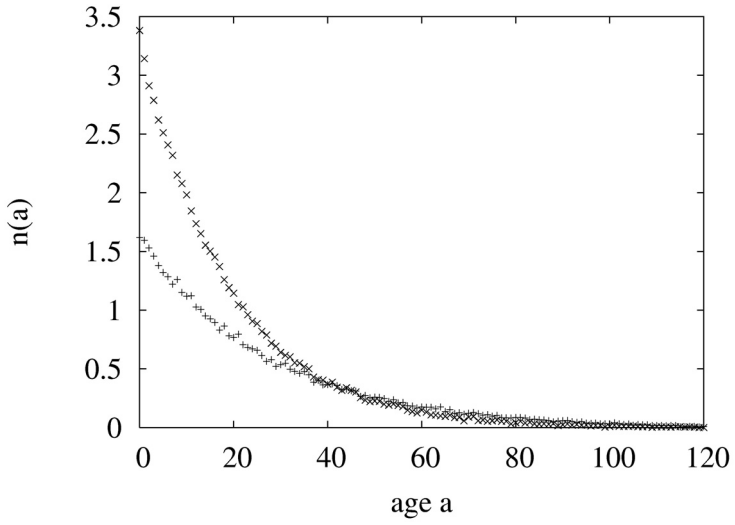


Fig. 1. Simulated age a distribution of population $n(a)$, in arbitrary units, of infection-free population. Upper case corresponds to newborns randomly placed at any lattice cell. Lower case is obtained if babies may be placed only in the nearest neighbours cells of parent's site

vertically, i.e. from parent to offspring. More mathematical approach of parasites as carriers in can be found in [4,5]. However, we do not intend to interpret the typical task of immune system simulation which involves the many different specialized biological cells (B-cells, macrophages, helpers and others), apart from the virus itself, and all related interactions and relations between them. Here we confine ourselves to very simple description of a two dimensional $N \times N$ lattice with cells free or occupied by one item, either infected or free of the virus. The dynamics is controlled by a proposed set of parameters in each evolution step $t \to t+1$ evolution. After the many iterations we mostly concentrate of space or/and age distribution of items, infected or not. Typical age distribution of the non-infected population is shown in figure 1.

We intend to simulate both vertical and horizontal version of the infection pass, and see how it may influence the $n(a)$ distribution.

2 Model

The basic algorithm assumes given number of iteration cycles, for every cycle we scan the $N \times N$ lattice and apply evolution rules to non-empty i-cells, $i = 1..N^2$. Each item is characterized with its age $a(i)$, parameters $c(i)$ responsible for an overall health condition and $v(i)$ indicating the virus infection. At each time step, the i-th individual is verified:

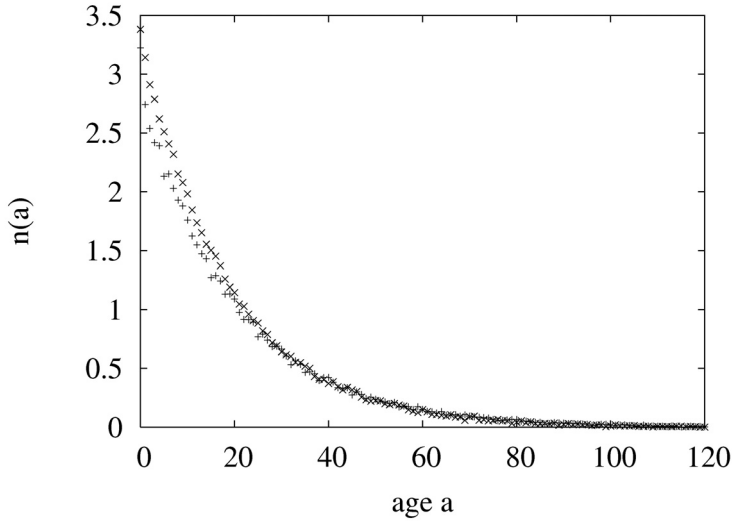


Fig. 2. Age distribution $n(a)$ for vertical transmission of infection. Babies are freely scattered over the lattice. Upper case is for no infection passed, lower case is for 80% offsprings affected and 80% of them later killed by the developed disease

- if its age a is above a biological maximum age $maxA = 120$, the item dies,
- then it also may die with probability $p \propto n/N^2$, the Verhulst factor [6], where n is the current population size,
- if the individual survived, and its age is above minimum reproduction age (here 16) yet still below maximum reproduction age (here 48), it gives birth to $B = 0.2$ babies (It means a baby with a copy of parent's c, is born with conditional probability 0.2, if the proposed destination is an empty cell. It can be chosen either on the whole lattice or limited to the nearest neighbourhood of parents),
- infection may take place according to specification below,
- item's c is corrected according to the nearest neighbours c-values, so that better neighbours pushes its c up, (and *vice versa*)
- also the intensity v of the already infected is up as the disease develops,
- further elimination process continuous - the item is out if its c is below a threshold value $minC$, here $minC = 0$,
- if v is more that a maximum value $maxV$, the item' future is decided: it dies with probability pV, else it is cured and $v = 0$,
- at this stage the individual has survived and enters the next time $t + 1$, perhaps after some movements due to migration process which takes place with probability $pMov$, then getting one year older, $a \to a + 1$ and a little less fit, $c \to c - \Delta c$.

In the vertical transmission version, the baby catches disease $v = 0 \to v = 1$ with probability pI if parent's v is from $v1$ to $v2$. Horizontal transmission is

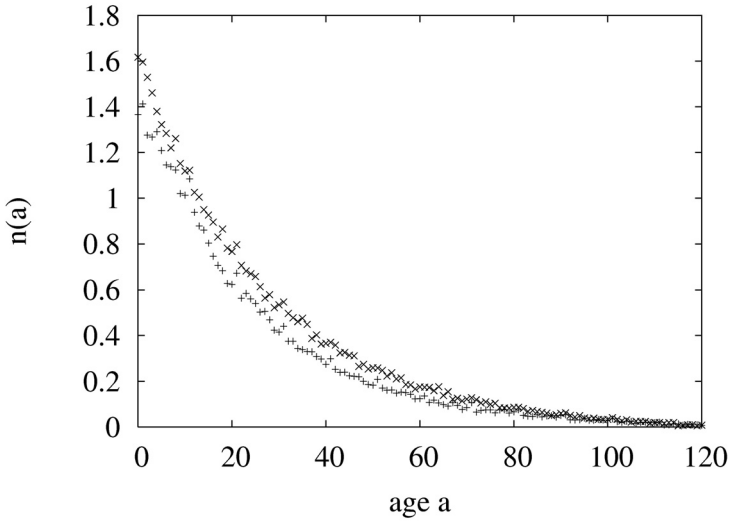


Fig. 3. Age distribution $n(a)$ for horizontal transmission of infection through neighbours. Babies are fixed in vicinity of parents. Upper case is the infection-free population, lower case is for 20% offsprings affected, and 80% of them are later killed by the disease

similar, yet the virus is passed from the nearest neighbour. When the scan all over the lattice is finished, the iteration cycle is completed with standard cyclic boundary conditions.

3 Results and Discussion

The age distribution of the population for vertical transmission of infection to babies is shown on Fig. 2. The two branches of points correspond to (a) no disease passed (upper set), and (b) $pI = 0.8$ fraction of offsprings of ill parents picking up the disease (lower set). When the disease develops and the critical v is reached, the model assumes only 20% of the individuals do recover, $pV = 0.8$. It is seen that a reduction of 10-20% of population is the net result of the vertical disease transmissions, yet the overall characteristics of age distribution is nearly the same. It reflects the long time scale of this type of infection, since the harmless period of the illness development must last long enough for the parent to reach the minimum reproduction age, and more to give chance to produce new items. The discussed case of babies free to sit at any empty lattice site may be verified against the version when babies are kept near the parents. As expected, apart from the general trend of smaller population as result of less room for new members, recall Fig. 1, we observe similar effect of 10-20% further reduction of population and no change in age distribution after it is normalized to cancel out the population size effect.

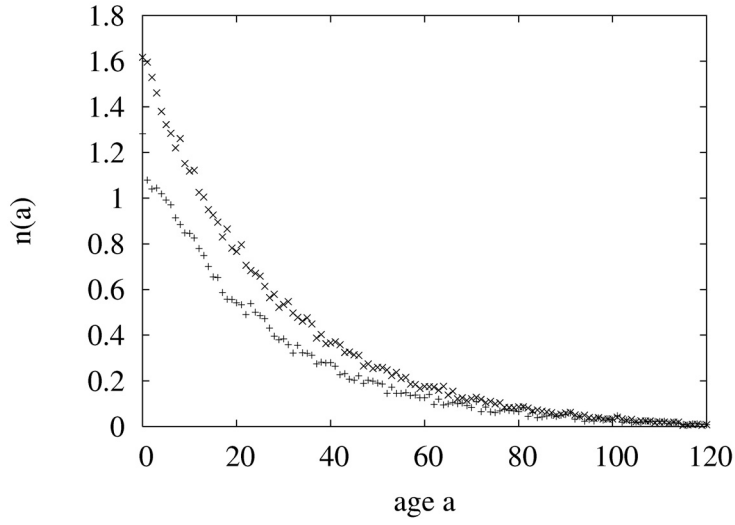


Fig. 4. Age distribution $n(a)$ for simultaneous vertical and horizontal transmission of infection, with babies bound to stay with parents. Upper case is the reference of the infection-free population

Contrary to the vertical case, the horizontal transmission may yield different effects for bounded and unbounded location for children. One can anticipate more significant differences for the bounded case since the nearest neighgours are responsible for the disease spread. Indeed, Fig. 3 for the bounded case shows a decrease in $n(a)$ distribution, in comparison with infection-free population, especially in the middle age fraction of population. The free choice of location for children gives only a tiny smaller $n(a)$ with respects to the disease free reference case. For calculations we used $pI = 0.2$ and $pV = 0.8$. The number 0.2, which replaces 0.8 for vertical case, was chosen as there are 4 nearest neighbours and each of them may infect. Also the time scale is generally much shorter - this time it is not essential that the disease carrier must live long enough to pass the virus. In fact, we applied half of the whole life span as the time for which the vertical disease develops, while only 3% of maximum age limit for the horizontal diseases. If both horizontal and vertical (h&v) mechanisms are present, see Fig. 4, the deviation from the disease-free population is bigger than if only one of the two named mechanisms are active. This is obvious, yet significant difference is the interference which make the net result is not the simple sum of contributions coming from the two contributions. (This may be seen for more detailed analysis when we compare (h&v) data against the independent contributions (h)+(v).) Such interplay is the consequence of the elimination mechanism implemented in the model. We treat the two cases as corresponding to different units, yet weakening by one of the disease makes the item is less resistant to the other sickness and so the death is then more likely. The $n(a)$ distribution may also

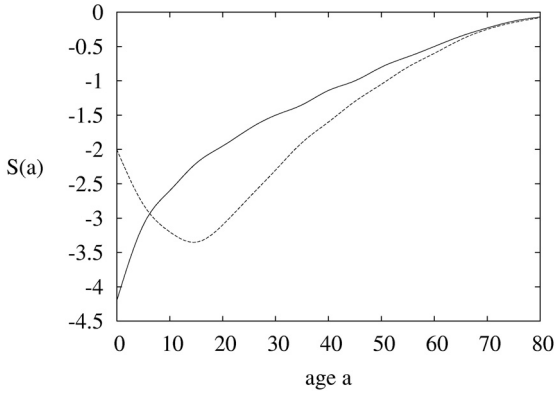


Fig. 5. Results of simulation of age dependence of $S(a) = \log q$, the exponential Gompertz law of mortality $q(a)$ predicts a straight line. The line showing minimum is for the case of simultaneous vertical and horizontal transmission of infection, the other line corresponds to the infection-free population

be seen in terms of the usually discussed mortality $q(a) = 1 - n(a+1)/n(a)$, which is a fraction of the population eliminated at age a. The Gompertz law predicts exponential dependence of $q(a)$, $q(a) \sim e^{b \cdot a}$, which yields a straight line on logarithmic plot of $S(a) = \log q$ against age a. The results of simulations presented on Fig. 5 is the translation of Fig. 4. It shows the general tendency that the disease presence in population changes the mortality distribution. This change leads to a minimum in the $S(a)$ dependence, a feature observed in human population. Deviations from the Gompertz law is of a main interest in many publications, see for example [8] for review.

It is interesting to include migration processes for the horizontal version of disease spread, in the case of babies staying with the parents. The alternative case brings in no effect of migration which is already present in the form of offfsprings dispersed all over. For the new members kept close by, the migration may play dual role. Firstly, for highly infected clusters of local communities, it is a chance to escape from the doom of unavoidably getting infected. Then migration lessen the infection pressure and so the whole picture is shifted towards less infectious environment. This effect is illustrated in Fig. 6. The reference case (upper points) with 30% risk of the disease leading to the death, $pV = 0.3$, is recalculated with higher risk $pV = 0.72$. Not only the population drastically drops (bottom points), but also the distorted age distribution $n(a)$ due to a jump at age $a = 12 \to 13$, is well pronounced. This is so as we applied a sharp disease development limit, $maxV = 12$. If intense migration is allowed, the middle points, some recovery is then observed and the population size is higher.

However, if only a tiny fraction of population is infected, migration helps the infection to spread all over, especially if the pI parameter, indicating of how likely is the virus transmittable, is high.

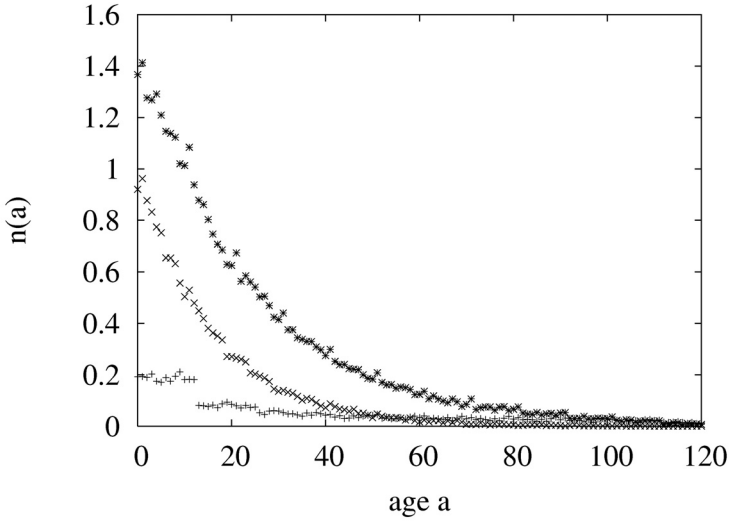


Fig. 6. Babies by parents case. The reference (upper points) set of death risk $pV = 0.3$, is confronted against much higher risk $pV = 0.72$, the bottom points. When high rate of migration $pMov = 0.9$ is switched on, the situation is improved, the set of middle points

4 Conclusions

One can notice some resemblance of the cellular automata approach and the Penna model [7,8,9]. In a way, the parameter c coding the item's health condition, plays similar role as the genome in the Penna model. Activated mutations in Penna model and the threshold maximum value of bad mutations correspond roughly to the decrease in c as time flows. Limited environmental capacity is equivalent to the lattice size $N \times N$, or more precisely $(N \times N - n)$ sites left free as n sites are already occupied. Therefore to some extend the proposed approach may be considered as an alternative.

Main conclusions of the proposed approach were already discussed in the main text. In short, the vertical transmission of infection makes population a little smaller by 10-20%, yet with nearly no effect on age distribution, apart from arbitrary normalization factor. In this case neither migration nor possible fixing of the offsprings make difference in output.

The horizontally passed infections are not important if babies are isolated from parents. It is only during the initial transient period that the population in small isolated clusters suffer and die, or get cured, while babies being free from infection do not carry the disease to their far destination. If babies are in vicinity of parents, they are also vulnerable to the disease. Then the age distribution $n(a)$ differs from the reference sickness-free case, and also possible migration influences the results. The migration itself is helpful for rapidly spreading and

deadly infections. For milder infections, migration is responsible for spreading the disease and also increases the death toll.

Acknowledgements. The work was partly supported by a grant of the Agricultural University, Department of Mathematical Statistics. Main simulations were run on computing power at the Academic Computer Centre CYFRONET-KRAKÓW.

References

1. N. H. Shnerb, Y. Louzoun' E. Bettelheim and S. Solomon, The importance of being discrete: Life always wins on the surface. *Proc. Natl. Acad. Sci. USA* **97** (2000) 10322.
2. T. Toffoli, M. Margolus *Cellular Automata Machines*, MIT Press, Cambridge MA, 1999.
3. H. Atlan, *Bull. Math. Biol.*, **51** (1989) 247.
4. M. S. Lipsitch, S. Silver and M. A. Nowak, The evolution of virulence in pathogens with vertical and horizontal transmission, *Evolution*, **50** (1996) 17729-1741.
5. P. Stadler and R. Happel, The probability of permanence *Math. Biosci.*, **113** (1993) 25-60.
6. D. Brown and P. Rolhery, *Models in Biology: Mathematics, Statistics and Computing*, Wiley, New York, 1993.
7. T. J. P. Penna, A Bit-String Model for Biological Ageing, *J. Stat. Phys.* **78** (1995) 1629.
8. S. Moss de Oliveira, P. M. C. de Oliveira and D. Stauffer, *Evolution, money, war and computers*, Teubner, Stuttgart-Leipzig, 1999.
9. A. T. Bernardes, Monte Carlo Simulations of Biological Ageing, *Ann. Rev. of Computational Physics* **4** (1996) 359.

Evolution of Population with Interaction between Neighbours

Maria Magdoń-Maksymowicz[1] and Andrzej Z. Maksymowicz[2]

[1] Department of Mathematical Statistics
AR-Agricultural University
Al. Mickiewicza 21
31-120 Krakow, Poland
rrmagdon@cyf-kr.edu.pl
[2] Faculty of Physics and Nuclear Techniques
AGH-University of Science and Technology
Al. Mickiewicza 30
30-059 Krakow, Poland

Abstract. Dynamic of population growth on two-dimensional lattice $N_x \times N_y = 1000 \times 1000$ is simulated with rules that involve interaction between neighbours. Biological aging is described by a systematic decrease with each time step of parameter $g(i)$, a measure of the health condition for each individual i. When g drops down below a minimum value, the death of the item occurs. The death toll is also caused by the Verhulst factor which accounts for the limited living space. Newly borns inherit g from parent's, plus some extra value to help them to start the life game. This basic rule is enriched by (a) some corrections of the parameter $g(i)$ according to its current values among the nearest and next nearest neighbours, and/or (b) the possible limitation of newborns to be placed in the close neighbourhood of parents. We discuss influence of the birth rate on the size of the final population and its age distributions, then g-g correlations c of the health condition g on the nearest neighbours. Main conclusion is that interaction between neighbours may play important role modifying the population characteristics, also the restrictions on whether the babies are bound to stay with parents or not yield different results.

1 Introduction

Population evolution is governed by a set of rules which describe how the state of population $n(t)$ at time t yields the population $n(t+1)$ at the next time-step $t+1$. The rules must account for the elimination mechanism of items and also for a new items entering the society. In equilibrium, the number of deaths Δn must be balanced by same number of birth, on the average. Probabilistic approach to the rules may lead to either an analytical description of continuous time sequence or to iterative procedures suitable for computer simulations. Among them, cellular automata may be considered as the tool which we apply in this work. In the simplest model we assume death caused by

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 758–764, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the limited environmental capacity N, so that in each time step we eliminate a fraction $n(t)/N$ of population, known as the Verhulst factor [1]. The deaths are compensated by introducing a fraction B of the remaining individuals, perhaps with the limitation on the minimum reproduction age R. The net balance predicts the equilibrium n and also age distribution $n(a)$. In a more realistic approach we need to account for different deaths mechanisms. Among them, the most often named are bad mutations, oxygen radicals and longevity genes. In the Penna model [2], the population evolution is seen as controled by the accumulation of bad mutations, passed over to offsprings which also may catch additional mutations. The mutations for each individual (i) are represented by bit '1' in the genome - a computer word. The bad mutations are not harmful until they are *activated*. This takes place when we disclose the next bit position at each time step, and the sum of already active mutations is then compared against a threshold value (T). The individual dies on reaching this value. The child genome is inherited from its parent (cloning in the asexual version of the Penna model) with eventually additional mutations spelled at randomly chosen bits of the genome. This Penna model is perhaps the most popular evolution model which simulates biological aging and genetic death (see [3] for review).

Both the rate of born babies, as well as possible deaths scenarios are far more complex then the very simple model presented in this paper. The aim of this work is to test how the very local interactions between neighbouring items may be important. We intend to account for a socializing factor in local communities such as a family or close neighbours. Very often members of such small communities create their own rules and influence each other. This influence may be incorporated into the more general rules that control all population. The first simple idea is to try to adjust parameter $g(i)$, a sort of a fitness parameter of each member i, so that better neighbour improves our condition, and poorer neighbour makes us less fit. In the 2D lattice model, each item occupies a cell and state of all items in the neighbouring cells has some effect on the cell in consideration.

2 Model

The population n is displaced on a two-dimensional $N = N_x \times N_y = 1000 \times 1000$ lattice, each i-th item occupying one cell. Individuals are characterized by a goodness parameter $g(i)$ ranging from a minimum value g_{min}, below which the item is considered dead, up to a certain g_{max} value. At each time step of simulation the age $a(i)$ is icreased by one and the goodness g is diminished by Δg. The Initial population $n(0) < N$ at time $t = 0$ is randomly spread all over the lattice. At each simulation step $t \to t + 1$, all items are scanned. For each item we apply the following rules.

- The item is eliminated with probability $n(t)/N$, the Verhulst factor.
- If it survives and its age a is at least the minimum reproduction age R, B babies may be born. That is, only when randomly chosen cell where we intend to place the newborn is empty, the actual number of offsprings is B;

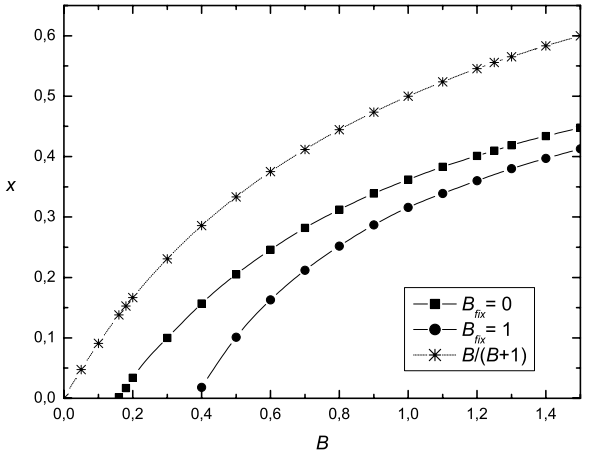


Fig. 1. Normalized population x as a function of the birth rate B. The top curve is the logistic case. Results marked by squares are obtained for the newborns randomly located on the lattice, ($B_{fix} = 0$). Circles correspond to the case when babies are located in the neighbourhood of parents,($B_{fix} = 1$)

otherwise the number is smaller or even zero if all proposed cells happen to be full. In this asexual version, the baby at age $a = 0$ gets initial g of the parent, plus some extra δ value as a beneficial amount.

- Looking around, the i-th item's $g(i)$ is increased (by $g_{nn} = 2$ in case of any of the four nearest neighbours and by $g_{nnn} = 1$ for the 4 next nearest neighbours) if the neighbour has larger g. If the neighbour has smaller g, $g(i)$ is decreased by g_{nn} and/or g_{nnn}, respectively. In other words, the social interaction makes the local society tends to become more uniform.
- Elimination caused by poor health condition takes now place for the weakest members with $g < g_{min}$.
- The last step which completes the transition from t to $t+1$ era is the update, the age a increases by one and goodness g drops by Δg.

It may be noticed some similarities with the Penna model. The systematic decrease in g has similar effect as the activation of bad mutations in the Penna model. The extra amount δ of g value inherited from parent is equivalent to resetting the baby's counter of the active bad mutations to zero.

3 Results and Discussion

We used for the minimum goodness factor $g_{min} = 90$, below which items are eliminated. Biological aging makes $g(i)$ drop down at rate of $\Delta g = 1$ per time step. The follow-the-neighbour shifts of current $g(i)$'s are $g_{nn} = 2$ and $g_{nnn} = 1$,

or none if we switched off this interaction. The minimum reproduction age was assumed $R = 0$. The babies gain credit $\delta = 2$, so that $g(\text{baby})=g(\text{parent})+\delta$ for newborns.

Figure 1 shows population size *vs* birth rate. The logistic game, with the Verhulst factor as the only possible elimination mechanism of an item, offers the most densely populated case. The equilibrium population shows no critical birth rate B_c below which the population is extinct. In the described cellular automata model, the death also occurs for other reasons - below the minimum health condition - and so population x is then smaller.

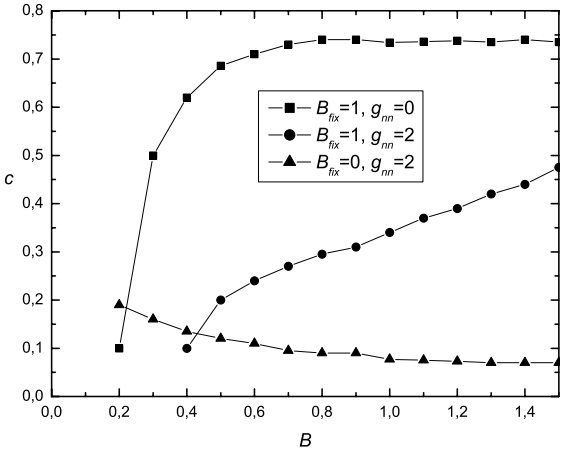


Fig. 2. Correlation c of individuals' health condition against the birth rate B. Results marked by squares or circles are obtained for the case of babies kept close to parents; circles correspond to the follow-the-neighbour-condition rule, while results marked by squares ignore the neighbours. The triangles are for no babies spread over the whole lattice, yet we still mind the neighbours

In particular, we have nonzero B_c. The two lower curves are obtained for different bounds imposed on the available living space by the two facts. Firstly, within the lattice cell approach, each item is subject to the (Pauli) exclusion principle that any cell cannot be occupied by more then one individual. So the effective space is not the lattice $N_x \times N_y = 1000 \times 1000$ size yet it is smaller by factor (1-x). Secondly, the rule of an offspring to be kept in close vicinity of the parent diminishes further the effective space, the lowest curve in Fig. 1.

Results shown on Fig. 2 is the summary of how far the health conditions of the nearest neighbours are correlated. The invisible horizontal $c = 0$ line of uncorrelated values is due to the obvious case of babies placed on the whole

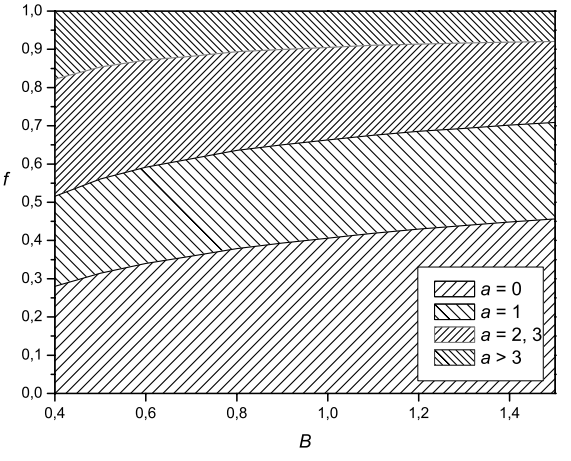


Fig. 3. Age distribution f for babies ($a = 0$), young ($a = 1$), mature ($a = 2, 3$) and old ($a > 3$) members of society, with the follow-the-neighbour-condition rule, *against* the birth rate. Babies are free to choose location

lattice and ignoring our local neighbours. The most striking effect is for babies kept close to parents, the squares in the figure, as the basic mechanism of cloning ensures offsprings become similar to parents. (This also creates a local environment which may differ from the overall environmental characteristics.) However, if we switch on interaction between neighbours, the effect is less correlated population. We may see neighbours and influence neighbours who do not belong to the family. And *vice versa*, the neighbours also affect us. As we said, if babies are placed close by, it plays dominant role for the correlation. The correlations are still present if we accept the influence of neighbours only, and let children to wander all around. This time correlations are relatively smaller and they drop down with the birth rate. This is expected since for larger B, the more randomly placed new items brings in more uncorrelated strangers to the local community, before they become assimilated after the many iteration steps.

The division line between different age groups increases monotically with birth rate B. This is so since for higher reproduction rate there is larger inflow of younger generation members. Also there is less room left for older generation from the point of view of limited environmental space. Figures 3 and 4 are obtained for the newly borned placed randomly or in vicinity of parents. It is seen some, yet not very drastic, influence of this on the age distribution of the population. For small birth rates, with no bonds on where the children go, about 17% of populations are seniors. This fraction diminishes to 10% for large B, according to the mentioned mechanism of repelling older members if we need to accomodate more youngsters. However, if children are forced do stay close by, the corresponding numbers are smaller: 11% and 6%. And, simultaneolusly, the

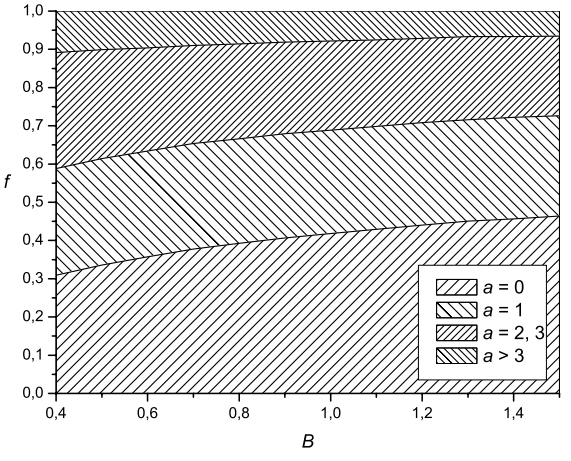


Fig. 4. Age distribution as in 3, yet this time babies are bound to stay near the parents

percentage of youngsters go up. In other words, the age distribution indicates that the whole population becomes younger. It may be interpreted as result of the benificial influence of babies, which gain the extra vital power on birth, and so improve the condition of parents if they are in the neighborhood. At the same time, the babies condition lessens. When growing up, on reaching the old age, they enter it with poorer health condition and so the death toll is higher. This is why percentage of the very old is smaller. However, we must be careful in interpretation of results for given birth rate B. Apart from general tendency, no direct quantitative comparison is possible as the efective birth rate B_{eff} is always smaller then B since no doubly occupied cells are allowed. The point is, this limitation is much more severe for the case when babies stay by parents, $B_{fix} = 1$.

For example, the age distribution $n(a)$ obtained from computer simulation strongly depends on the assumed restriction. The usually discussed characteristics is the mortality $q(a) = 1 - n(a+1)/n(a)$, which is a fraction of the population at age a that will be eliminated before reaching the next $(a+1)$ age. The well known Gompertz law of exponential $q(a)$ dependence, $q(a) \sim e^{b \cdot a}$, predicts a straight line on a plot of $S(a) = \log q$ against age a. The results of simulations are presented on Fig. 5. It is seen that the linear dependence of $S(a)$ is not held. In human populations we observe a minimum and it is only for humans of age above 20 or 30 that the dependence is more or less linear; perhaps with exception of very old individuals when again the deviation from the Gompertz law may be observed. We do not intend to analyse the clear deviations from the Gompertz law, which may stem out from many reasons, and which is currently widely discussed in many publications, also in the already mentioned review [3].

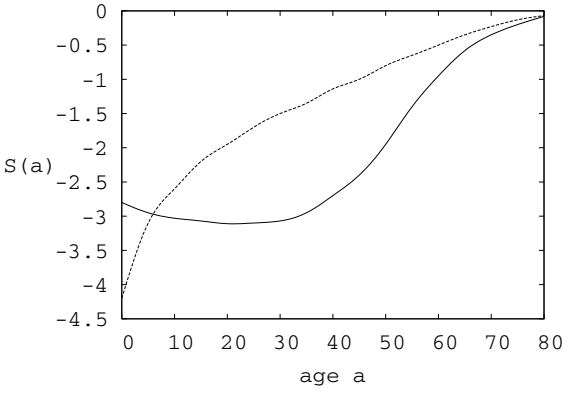


Fig. 5. Simulation results of age dependence of $S(a) = \log q$, the exponential Gompertz law of mortality $q(a)$ predicts a straight line. The line with a minimum is for whole lattice as possible location for newborns, the other line corresponds to offsprings located in close neighborhood to parent

In summary, we only conclude that by suitable choice of the model parameters, corresponding to different proposed mechanisms of evolution, we may try to reproduce population characteristics and try to interpret results in terms of appropriate rules controling the population dynamics.

Acknowledgements. The work was supported by a grant of the Agricultural University, Department of Mathematical Statistics, and by the University of Science and Technology, Faculty of Physics and Nuclear Techniques. Computer calculations were carried out at the Academic Computer Centre CYFRONET-KRAKÓW on HP Exemplar S2000 machine.

References

1. D. Brown and P. Rolhery, *Models in Biology: Mathematics, Statistics and Computing*, Wiley, New York, 1993.
2. T. J. P. Penna, *J. Stat. Phys.* **78** (1995) 1629.
3. S. Moss de Oliveira, P. M. C. de Oliveira and D. Stauffer, *Evolution, Money, War and Computers*, Teubner, Stuttgart-Leipzig, 1999.

The Role of Dominant Mutations in the Population Expansion

Stanisław Cebrat[1] and Andrzej Pękalski[2]

[1] Department of Genomics, Institute of Genetics and Microbiology, University of Wrocław,
ul. Przybyszewskiego 63/77, PL-54148 Wrocław, Poland
cebrat@microb.uni.wroc.pl
[2] Institute of Theoretical Physics, University of Wrocław,
pl Maxa Borna 9, 50-204 Wrocław, Poland
apekal@ift.uni.wroc.pl

Abstract. We have studied the dynamics of expansion of diploid, sexually reproducing populations onto new territories demanding different genetic adaptations. Populations succeeded in the new habitats if the mutations of alleles responsible for adaptation to the new conditions were recessive. If behind the hostile territory another friendly habitat exists, several different scenarios are possible, including one in which the middle, most hostile, habitat remains empty, separating populated two, more friendly, ones.

1 Introduction

Sexual reproduction is one of the most interesting phenomena in Nature. Thus, it is obvious that even physicists are interested in this strategy, though without spectacular successes (negative birth rate is observed in the physics departments all over the world). Some of them even succeeded in proving that the strategy of sexual reproduction is a big misunderstanding of Nature and asexual populations would have much higher reproduction potential than sexual ones [1]. Such conclusions can be reached as the result of an evolution simulation if dominance of defective alleles is assumed. In fact, in Nature genomes posses a lot of redundant information and one of the simplest ways to get redundancy at all genetic loci is to be diploid. Organisms can profit from redundancy if the mutated information (defective gene) is recessive which means that its defect can be complemented by the proper gene - the allele of the defective one in the diploid genome (wild copy of a gene). If the mutated gene dominates and expresses defective phenotype, then increasing the number of copies of such genes is rather a loosing strategy. That is why the number of loci in diploid genomes where mutations lead to dominant deleterious functions is low. In many instances, like mutations in oncogenes, it may be even connected with a special Nature's policy - elimination of old dispensable organisms (professors). Sexual strategy of reproduction usually needs two well balanced sets of information because these sets should be separated into two equivalent pools of genes in gametes which,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 765–770, 2004.
© Springer-Verlag Berlin Heidelberg 2004

joining in pairs, produce new individuals. It seems reasonable that it would still be better to increase the redundancy by forming triploids or tetraploids. But even neglecting the costs of replication of the redundant information it could be a loosing strategy if there were too many loci with potentially dominant mutations in the genomes [2]. Nevertheless, dominance is a rather complicated problem in biology. It is possible that a mutation in the same pair of alleles can be recessive in one condition or for one phenotype while dominant in another one. Let's imagine a pair of alleles producing an enzyme which splits a toxic compound. If one allele of the pair is mutated the toxic compound can still be eliminated from the organism - the mutation is recessive. What would happen if the mutation in the gene produced an enzyme which would transform the other compound into a cancerogen - a product promoting cancerogenesis ? If the organism is exposed to this pre-cancerogenic compound such a mutation behaves as dominant. Thus, the status of mutations could depend on some environmental conditions. Immunology supports some other examples of complicated relations between different alleles of the same locus and environment [3]. The distribution of the main blood group antigens A,B,O in the human population is very uneven on Earth. If an individual has antigen A, it cannot produce antibodies anti-A because it would develop an auto-immune disease. If in the environment there is a pathogen with A-like antigens, organisms cannot develop the immune response against this antigen. Some pathogens are known which can use such a mimicry to find a defenseless host. Some pathogens can be so promiscuous that they use mimicry to invade the host and finally they may induce cancerogenesis [4]. In our simulations we tried to show that redefinition of loci from recessive to dominant is much more dangerous for the survival of the whole population than redefinition of the value of alleles in the locus from positive to negative, keeping the dominance unchanged.

2 Model

We consider a population composed initially of $N(0)$ individuals located at the leftmost part of the square lattice, of dimensions 150×50 lattice sites, forming our system. Each individual, apart from its location on the lattice is characterized by its genotype g and age age. The genotype is a double string of length $L = 32$ of zeros and ones. From the genotype the individual's phenotype,f, also of length L, is constructed as a product of the values at corresponding loci. The rules of taking the product depend whether a zero is declared at the dominant allele, or not. The population is under the influence of the external "field", which could be the climate, represented as the optimal phenotype,F, of length L with components equal either zero or one. The optimal phenotype changes along the X-axis. The system is divided into three regions, labelled later I, II, III, with the corresponding optima F_I, F_{II}, F_{III}. At the leftmost region (I) the optimum is just a string of zeros. In the second region a certain number, Lch, of 0's is switched to 1's. In the third region either there are twice as many changes as in the region II, or the "climate" is the same is in the region I.

At the beginning the population has random values of their genotypes and locations on the lattice. The algorithm goes as follows

1. an individual j is picked at random,
2. its adaptation to the external conditions, i.e. the agreement between its phenotype and the optimal one is calculated from the formula

$$ad_j = \frac{1}{32}\left(1 - \sum_{\alpha=1}^{32} \mathrm{XOR}(f_j^\alpha, F^\alpha)\right), \tag{1}$$

   where the F is taken according to the position of the individual in the system (in the I, II or III region),
3. its survival probability is determined

$$p_j = \exp\left(-s\, age_j/ad_j\right), \tag{2}$$

   where s is the selection pressure. A random number $r \in (0,1)$ is drawn. If $r > p_j$ the individual is removed from the system. Otherwise
4. a search is made in the nearest neighborhood for an empty place to move the individual. If successful and the age of the individuals is greater then one (to eliminate coupling of offspring right after birth)
5. another search is made in the nearest neighborhood of the new position for a partner. Once found
6. the pair produce at most 4 offspring, each of which receives its genotype independently via recombination and crossing over of the parents' genotypes. Each loci of the offspring could be mutated with a probability pm, which is a parameter of the model. Independent search is made in the Moore neighborhood (8 sites nearest to the first parent) to put an offspring. If a search was not successful, the progeny is not born. This is the only place where a kind of Verhulst factor regulates the density of the population.

When, at a time t, as many individuals were picked as first partners, as was the total number of individuals at that time, i.e. $N(t)$, one Monte Carlo step has been made and the age of all individuals is incremented by one. In the following, time is measured always in Monte Carlo steps (MCS). This kind of models has been used before by us in different contexts [5].

3 Results and Discussion

Populations simulated under the chosen selection pressure and mutational pressure parameters are stable and survive the prolonged simulations. In these populations, mutations in all alleles of all loci were recessive. If we re-declare the values of some loci which means that the new environment requires different alleles in these loci, the populations eventually adapt to the new conditions and approximate to the concentration usual for such conditions (Fig. 1).

The situation is much more dangerous for the simulated populations if the redefinition of the loci values is connected with the declared dominance of the

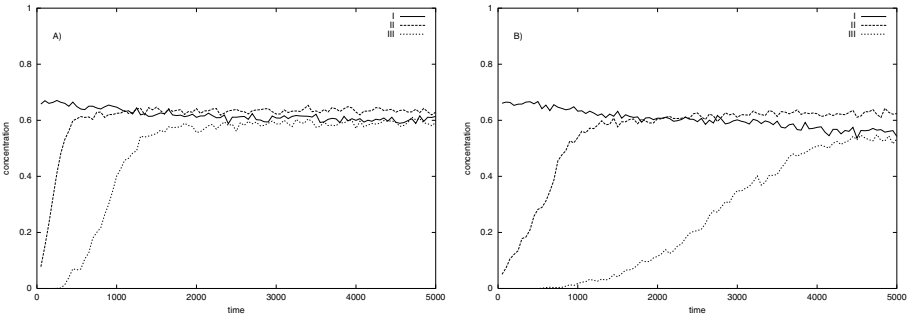


Fig. 1. Expansion of populations on the new territories without changes in dominance conditions. A) II region demands different values in 4 loci, III region demands different values in 8 loci. B) II region demands different values in 7 loci, III region demands different values in 14 loci. Mutations at all loci stay recessive. Average over 10 runs. Three lines correspond to the three regions

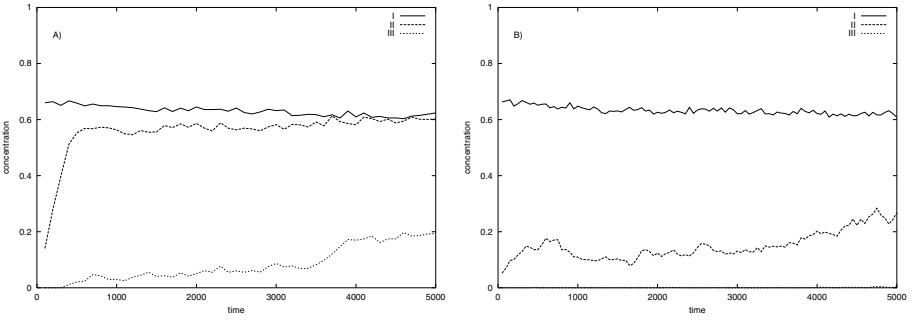


Fig. 2. Expansion of populations on the new territories with changes in dominance conditions. A) II region demands different values in 4 loci, III region demands different values in 8 loci. B) II region demands different values in 7 loci, III region demands different values in 14 loci. Mutations at all loci are dominant. Average over 10 runs. Three lines correspond to the three regions

mutated genes. The chance for survival of the population depends strongly on the number of declared dominant loci (Fig. 2).

In our simulations we have assumed that the territory where the population evolves is divided into three sectors. Initially the population occupies one sector (left) and it is adapted to the environmental conditions in it, with all mutations recessive.

Conditions in the middle sector require different values of some genes and in one case mutations in these loci are recessive, while in the second case these mutations are dominant. In the first case the population invades the second sector, adapts to the new conditions and eventually reaches the third sector which still demands the new adaptations. In the second case the population cannot permanently occupy the middle sector. Some trials are observed of settling the

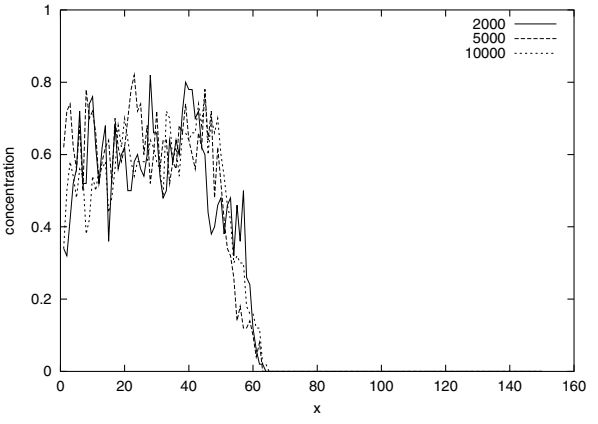


Fig. 3. Spatial distribution of a population at three time steps - 2000, 5000 and 10000 MCS. Zero is the dominant allele and the number of changes in the optimum is 7 in the II and 14 in the III region, respectively

territory close to the border, but concentration of the population drops very quickly with the distance to the friendly maternal territory (Fig. 3).

To test the possibility of crossing the hostile territory we have declared that the third sector (the right one) is as friendly as the first one and it requires the same phenotypes as the first one, but the initial population was settled on the first sector only (Fig. 4). It is obvious that like in the above simulations, if there were no dominant deleterious mutations, populations expanded onto all three sectors with approximately the same concentrations in all of them. But if the mutations in the re-declared loci were dominant, the populations were not able to cross the hostile territory. Only in some rare instances they succeeded in penetrating this sector. Having reached the new territory with the same environmental conditions as the initial one, the population occupied it very fast. The interesting observation is that the middle hostile territory is not desolate any more, though the population is less concentrated in it.

It should be noted that the problem of dominance has been recently raised by the modern eugenics techniques. In the so called "risk cases" the *fertilization in vitro and embryo transfer* followed by preimplantation genetic diagnosis is recommended. It is obvious that heterozygotic embryos are eliminated even in case of recessive mutations. Such a procedure mimics the dominance of mutations. For more discussion of eugenics problem see also [6]. Further studies should answer the question whether the hostile territory in the reach of populations ameliorates the genetic status of the whole population.

4 Conclusion

It has been shown that it is important to consider the dominance of alleles when studying the sexual strategy of reproduction. Populations can expand to new

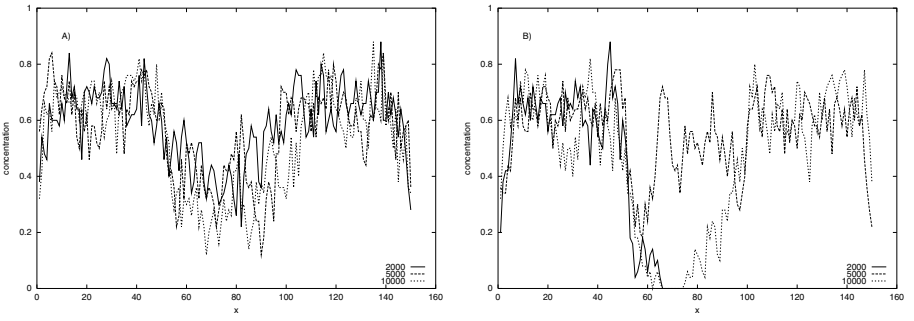


Fig. 4. As in Figure 3, except that now the optimum is the same (more friendly) in the regions I and III. Presented are two populations, starting with similar initial conditions

territories and adapt to new conditions relatively fast if it is not connected with the appearing of new dominant loci. The process of penetration of the hostile territory probably selects very well adapted individuals.

Acknowledgements. Authors thank J. Mleczko and J. Kołodyński for discussions. The work is a part of the program of the COST Action P10.

References

1. Redfield RJ, Nature, 369 (1994) 145.
2. Sousa AO, Moss de Oliveira S, Sa Martins, Evolutionary advantage of diploidal over polyploidal sexual reproduction. Phys.Rev. E 67 (2003) 032903
3. Oldstone MB, Molecular mimmicry and immune mediated diseases. FASEB J. 12 (1998) 1255.
4. Logan RP, *Helicobacter pylori* and gastric cancer. Lancet 344 (1994) 1078.
5. A. Pękalski and K. Sznajd-Weron, Population dynamics with and without selection, Phys.Rev. **E 63**, 031903, 2001
6. S. Cebrat and A. Pękalski, Model of population evolution with and without eugenics, Eur.Phys.J. **B 11**, 687, 1999

On the Efficiency of Simplified Weak Taylor Schemes for Monte Carlo Simulation in Finance

Nicola Bruti Liberati and Eckhard Platen

University of Technology Sydney, School of Finance & Economics and Department of Mathematical Sciences, PO Box 123, Broadway, NSW, 2007, Australia

Abstract. The purpose of this paper is to study the efficiency of simplified weak schemes for stochastic differential equations. We present a numerical comparison between weak Taylor schemes and their simplified versions. In the simplified schemes discrete random variables, instead of Gaussian ones, are generated to approximate multiple stochastic integrals. We show that an implementation of simplified schemes based on random bits generators significantly increases the computational speed. The efficiency of the proposed schemes is demonstrated.

1 Introduction

As described, for instance, in [7] to price an option via simulation, one does not require a pathwise approximation of the solution of the underlying stochastic differential equation (SDE). Only an approximation of its probability distribution has to be considered. Thus the appropriate notion of convergence for such a Monte Carlo simulation should be the weak one, instead of the strong convergence, as described in [6]. It is well known that in order to achieve a certain order of weak convergence one can approximate the random variables in a weak Taylor scheme by appropriate discrete random variables. For instance, instead of a Gaussian increment we can employ in an Euler scheme a much simpler two point distributed random variable. In general, the simplified random variables have to coincide only for certain lower order moments with those of the random variables appearing in the Taylor schemes. In the case of a weak Taylor scheme of second order, to construct a second order simplified method we can use a three point distributed random variable. The aim of this paper is to show that an implementation of such simplified schemes based on random bits generators significantly increases the computational efficiency.
It should be noticed that the simplified Euler method is equivalent to some random walk, which again is approximately equivalent to a binomial tree. The possible states of the tree and of the simplified Euler scheme are approximately the same. Small differences arise only for the level of these states depending on the chosen parametrization of the binomial tree. However, while the tree is a deterministic backward algorithm, the simplified method is a forward method which generates paths. As we will report in Section 4, the numerical behaviour of simplified methods is similar to that of trees. For instance, we will obtain

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 771–778, 2004.
© Springer-Verlag Berlin Heidelberg 2004

an oscillatory convergence in the case of a European call payoff. This is a well-known effect of tree methods, but as will be shown, not limited to this class of methods, see, for instance, [2].
The widespread application of the tree methodology in finance motivates the application of the simplified schemes that will be presented in this note. The similarity between simplified schemes and tree methods is important for the understanding of the numerical properties for both types of methods.
Simplified schemes, being forward algorithms, are not easily suitable to price American options, even that corresponding algorithms have been developed, see for instance [8]. Moreover, with the simplified methods we always have to face the typical statistical error from Monte Carlo simulations. Major advantages of simplified schemes over tree methods are that of flexibility and general applicability in high dimensions.
The implementation of random bits generators will be proposed in this note. It makes simplified methods highly efficient. As shown in [5], implicit simplified methods can overcome certain numerical instability. Most importantly, random bits generators can be efficiently applied to implicit schemes, while tree methods cannot be made implicit. Note that simplified implicit schemes can be understood as being equivalent to implicit finite difference partial differential equation(PDE) methods. However, PDE methods cannot be easily implemented for higher dimensions.
The order of convergence of simplified schemes is independent of the dimension of the problem. As shown in [1], around dimension three or four simulation methods typically become more efficient than tree or PDE methods. It will be shown that simplified methods with random bits generators outperform significantly Taylor schemes, which are based on Gaussian and other random variables. This makes simplified methods with random bits generators efficient tools for high dimensional problems.

2 Weak Taylor Schemes and Simplified Methods

For the dynamics of the underlying security let us consider the following SDE:

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t \tag{1}$$

for $t \in [0, T]$, with $X_0 \in R$. A derivative pricing problem consists in computing an expectation of a payoff function $g(X_T)$ of the solution of the SDE (1). For the numerical approximation of such an expectation we require only an approximation of the probability distribution X_T. Therefore, the appropriate notion of convergence is that of weak convergence, see [6].
Let us assume an equidistant time dicretisation with nth discretisation time $t_n = n\Delta$ for $n \in \{0, 1, \ldots, N\}$ where $\Delta = \frac{T}{N}$ and $N \in \{1, 2, \ldots\}$. As a set of test functions we use the space $\mathcal{C}_P^l$ of the l times continuously differentiable functions g which, together with their partial derivatives of orders up to and including l, have polynomial growth.
We say that a time discrete approximation $Y^\Delta = \{Y_t^\Delta,\, t \in [0, T]\}$ converges weakly to $X = \{X_t,\, t \in [0, T]\}$ at time T with order γ if for each $g \in \mathcal{C}_P^{2(\gamma+1)}$

there exists a positive constant K, which does not depend on Δ, and a $\Delta_0 > 0$ such that $\varepsilon(\Delta) = |E(g(X_T) - E(g(Y_N^{\Delta}))| \leq K\Delta^{\gamma}$ for each $\Delta \in (0, \Delta_0)$.
As explained in [6], based on the Wagner-Platen expansion one can construct the, so called, weak Taylor schemes of any given weak order $\gamma \in \{1, 2, \ldots\}$. The simplest weak Taylor scheme is the *Euler method*, which has the weak order of convergence $\gamma = 1.0$. It is given by the scheme

$$Y_{n+1} = Y_n + a(t_n, Y_n)\Delta + b(t_n, Y_n)\Delta W_n, \tag{2}$$

where $\Delta W_n = W_{t_{n+1}} - W_{t_n}$ is the Gaussian increment of the Wiener process W for $n \in \{0, 1, 2 \ldots, N-1\}$ and $Y_0 = X_0$.
If one uses in the above Euler scheme instead of Gaussian random variables simpler multi-point distributed random variables, then one can still obtain the same weak order of convergence $\gamma = 1.0$, see Theorem 14.5.2 p. 474 in [6]. For the Euler method these simpler random variables have to coincide in their first three moments with those of the Gaussian Wiener process increments. This permits to replace the Gaussian increment ΔW_n in (2), by a two point distributed random variable $\Delta\widehat{W}_n$, where $P(\Delta\widehat{W}_n = \pm\sqrt{\Delta}) = \frac{1}{2}$. We then obtain the *simplified Euler scheme*. Here the first three moments of the Wiener process increments ΔW_n match those of $\Delta\widehat{W}_n$.
The same applies to the *order* 2.0 *weak Taylor scheme*

$$\begin{aligned} Y_{n+1} = Y_n + a\Delta + b\Delta W_n + \frac{1}{2}b'b\left\{\left(\Delta W_n^2\right) - \Delta\right\} + \frac{1}{2}\left(aa' + \frac{1}{2}a''b^2\right)\Delta^2 \\ + a'b\,\Delta Z_n + \left(ab' + \frac{1}{2}b''b^2\right)\left\{\Delta W_n\,\Delta - \Delta Z_n\right\}, \end{aligned} \tag{3}$$

where ΔZ_n represents the double Itô integral $\Delta Z_n = \int_{t_n}^{t_{n+1}} \int_{t_n}^{s_2} dW_{s_1} ds_2$. Here we replace the Gaussian random variables ΔW_n and ΔZ_n by expressions that use a three point distributed random variable $\Delta\widehat{W}_n$ with $P(\Delta\widehat{W}_n = \pm\sqrt{3\Delta}) = \frac{1}{6}$ and $P(\Delta\widehat{W}_n = 0) = \frac{2}{3}$.
Then we obtain the *second order simplified method*

$$\begin{aligned} Y_{n+1} = Y_n + a\Delta + b\Delta\widehat{W}_n + \frac{1}{2}bb'\left\{\left(\Delta\widehat{W}_n\right)^2 - \Delta\right\} + \frac{1}{2}\left(aa' + \frac{1}{2}a''b^2\right)\Delta^2 \\ + \frac{1}{2}\left(a'b + ab' + \frac{1}{2}b''b^2\right)\Delta\widehat{W}_n\Delta. \end{aligned} \tag{4}$$

Since the three point distributed random variable $\Delta\widehat{W}_n$ is such that the first five moments of the increments of the schemes (3) and (4) are matched, the second order simplified scheme (4) can be shown to achieve the weak order $\gamma = 2.0$.
By using four or even five point distributed random variables for approximating the random variables needed, we can obtain simplified weak Taylor schemes of weak order $\gamma = 3$ or 4, respectively, as shown in [6] and in [4].
An important issue for simulation methods for SDEs is their numerical stability. As noticed in [5], when considering test equations with multiplicative noise, the weak schemes described above show narrow regions of numerical stability. In

order to improve the numerical stability one needs to introduce implicitness in the diffusion terms. This leads, for instance, to the *fully implicit Euler scheme*

$$Y_{n+1} = Y_n + \left\{ a\left(t_{n+1}, Y_{n+1}\right) - b\left(t_{n+1}, Y_{n+1}\right) \frac{\partial}{\partial y} b\left(t_{n+1}, Y_{n+1}\right) \right\} \Delta + b\left(t_{n+1}, Y_{n+1}\right) \Delta W_n. \tag{5}$$

Also in this case one can employ the two point distributed random variable $\Delta\widehat{W}_n$ instead of ΔW_n in (5) to obtain the *simplified fully implicit Euler scheme* that still achieves an order $\gamma = 1.0$ of weak convergence.

3 Random Bits Generators

We now demonstrate, for simplified schemes, how to implement highly efficient random bits generators, that exploit the architecture of a digital computer. The crucial part of the resulting simplified schemes, are the random bits generators. These substitute the Gaussian random number generators needed for weak Taylor schemes.

A well known and efficient method to generate a pair of independent standard Gaussian random variables is the polar Marsaglia-Bray method coupled with a linear congruential random number generator, as described in [9]. In our comparative study we use, as our Gaussian random number generator, the routine *gasdev*, see p. 293 of [9].

For the simplified Euler scheme (2) and simplified fully implicit Euler scheme (5) we use a two point distributed random variable in each time step, which is obtained from a random bits generator. This generator is an algorithm that generates a single bit 0 or 1 with probability 0.5. The method implemented is based on the theory of primitive polynomials modulo 2. These are polynomials satisfying particular conditions whose coefficients are zero or one. The important property is that every primitive polynomial modulo 2 of order n defines a recurrence relation for obtaining a new bit from the n preceding ones with maximal length. This means that the period lenght of the recurrence relation is equal to $2^n - 1$. For a study on random number generators based on primitive polynomials modulo 2 we refer to [11].

Since the random number generator for the polar Marsaglia-Bray method has a period of 2^{31} we use a random bits generator based on the following primitive polynomial modulo 2 of order 31: $y(x) = x^{31} + x^3 + 1$. The C++ implementation of this generator is reported in Figure 1, see also [9]. This method is extremely fast and suitable for direct hardware implementation.

On the test computer the CPU time needed to generate 10 million random numbers with the polar Marsaglia-Bray method amounts to 4.7 seconds. The two point random bits generator, described above, is almost 30 times faster using only 0.16 seconds.

For simplified methods of higher order similar multi-point random bits generators can be constructed. For the second order simplified method (4) it is sufficient to use a three point random bits generator. A corresponding code is presented in Figure 2. It produces three bits coupled with an acceptance-rejection method.

```
int irbit1per31(unsigned long & iseed)
{    unsigned long newbit;
     newbit = ((iseed >> 31) & 1)
     ^  ((iseed >> 2) & 1);
     iseed = (iseed << 1) | newbit;
     return int(newbit); }
```

Fig. 1. C++ code of the two point random bits generator.

On the test computer the CPU time needed to generate 10 million random numbers with this generator amounts to 0.8 seconds, which is still 5 times less than the polar Marsaglia-Bray method.

```
int ranbit3per31(unsigned long & iseed)
{    int x1 = 1, x2 = 1, x3 = 0;
     while ( (x1 == 1 && x2 == 1 && x3 == 0)
     || (x1 == 0 && x2 == 1 && x3 == 1))
     {  x1 = irbit1per31(iseed);
        x2 = irbit1per31(iseed);
        x3 = irbit1per31(iseed); }
        return x1 - x3; }
```

Fig. 2. C++ code of the three point random bits generator.

4 Numerical Results

Now, we present some numerical results for the Euler, fully implicit Euler and order 2.0 weak Taylor schemes as well as their simplified versions. As test dynamics we choose an SDE with multiplicative noise of the Black-Scholes type, where

$$dX_t = \mu X_t dt + \sigma X_t dW_t \tag{6}$$

for $t \in [0, T]$. The SDE admits the closed form solution $X_T = X_0 \exp\{(\mu - \frac{\sigma^2}{2})T + \sigma W_T\}$. The CPU times needed to compute 4 million approximate paths with 64 time steps with the Euler, fully implicit Euler and order 2.0 weak Taylor scheme amount to 107, 114 and 110 seconds, respectively. The corresponding approximate simplified versions only require 3.8, 6.2 and 25.6 seconds, respectively. Thus, for the Euler method the simplified version is roughly 28 times faster than the Gaussian one. The simplified fully implicit Euler method is about 18 times

faster than its Gaussian counterpart. For the second order simplified method we found that it is roughly four times more efficient than the order 2.0 weak Taylor scheme.
We analyse now the weak convergence of Monte Carlo simulations when using a smooth payoff function, where we choose the first moment for illustration and consider later on also a non smooth payoff which will be that of a European call option.

4.1 A Smooth Payoff Function

At first, we study the weak error for a fixed number of simulations and time steps. We also compare the CPU time needed to reach a given accuracy. In order to analyse the weak error $\varepsilon(\Delta)$, we run sufficiently many simulations such that the statistical error can be neglected. We use the following parameters: $X_0 = 1$, $\mu = 1.5$, $\sigma = 0.01$, $T = 1$.
An important application of Monte Carlo simulation is the calculation of Value at Risk via the simulation of moments, as applied in Edgeworth expansions and saddle point methods, see [10]. Therefore, as test function we use the first moment $E(X_T)$ of X_T at time T. Other moments give similar numerical results due to the lognormal structure of the Black-Scholes dynamics. We then estimate the weak error of the first moment by comparing the simulated Monte Carlo estimate with the exact expectation $E(X_T) = X_0\, e^{\mu T}$.
In the first plot of Figure 3 we show the logarithm $log_2(\varepsilon(\Delta))$ of the weak error for the Euler, fully implicit Euler, and order 2.0 weak Taylor method versus the logarithm $log_2(\Delta)$ of the time step size. The errors for the corresponding simplified versions are almost identical and therefore omitted. The number of simulated paths amounted to 16 million, which resulted in extremely small confidence intervals that practically do not show up in Figure 3.
We emphasize the important observation that the simplified methods achieve

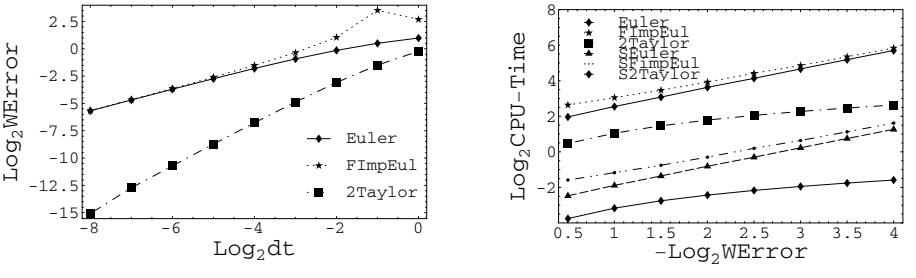


Fig. 3. Log-log plots of weak error versus time step size and CPU time versus the weak error for the Euler, fully implicit Euler and order 2.0 weak Taylor schemes.

almost exactly the same accuracy of their Taylor counterparts. Note in Figure 3 that the Euler and the fully implicit Euler scheme reproduce in the log-log plot the theoretically predicted weak order $\gamma = 1.0$. Furthermore, the order 2.0 weak

Taylor scheme achieves a weak order of about $\gamma = 2.0$, as expected. Moreover, we note in Figure 3 that the fully implicit Euler scheme shows poor results for very large step sizes. However, as shown in [5], the fully implicit method has better stability properties than the explicit schemes once the time step size becomes sufficiently small.
What really matters in practice is the time needed to reach a given level of accuracy. In the second plot of Figure 3 we show the logarithm of the CPU time versus the negative of the logarithm of the weak error observed for the three methods described above and their simplified versions. Since the accuracy for a given time step size is almost identical for the schemes of the same order, the increase in efficiency simply reflects the fact that the simplified schemes are computationally less intensive than their Gaussian counterparts. We recall that, for instance, the simplified Euler scheme is 28 times faster than the Euler scheme.
By comparing all six methods, we conclude that the second order simplified scheme is significantly more efficient for the given example than any other of the considered schemes. This result is rather important in simulations of Black-Scholes dynamics since it points out efficient Monte Carlo simulation algorithms for smooth payoffs.

4.2 An Option Payoff

In option pricing we are confronted with the computation of expectations of non smooth payoffs. To give a simple example, let us compute the price of a European call option. Here we have a continuous but only piecewise differentiable payoff $(X_T - K)^+ = \max(X_T - K, 0)$ with strike price K and the well known Black-Scholes formula as closed form solution for the option price at time $t = 0$. For this non smooth payoff we study the weak error for the Euler and the simplified Euler method, assuming the volatility $\sigma = 0.2$ and the short rate $\mu = 0.1$. We observed no major gain by using higher order methods, which is likely to be due to the non smooth option payoff. Since the second order simplified method (4) is approximately equivalent to a trinomial tree, as discussed in Section 1, this is consistent with an observation in [3]. In [3] it was observed that in option pricing the order of convergence of trinomial trees is not superior to that of binomial trees.
In the first plot of Figure 4 we show the log-log weak error plot for an at the money-forward option, with strike $K = X_0\, e^{\mu T}$. The Euler method generates a weak order $\gamma = 1.0$ with the log error forming a perfect line in dependence on the log time step size. As mentioned earlier, the simplified Euler method is approximately equivalent to a binomial tree. This method still achieves a weak order $\gamma = 1.0$. However, its log-log error plot does not exhibit a perfect line, which is due to the discrete nature of the random variables used. This appears to be the same effect as noticed for tree methods, see [2]. We observed for in the money and out of the money options a similar order of convergence with similar log error patterns.
In the second plot of Figure 4 we show the logarithm of the CPU time versus the negative logarithm of the weak error. For the considered non smooth payoff the increase in computational speed is still about 28 times. The simplified Euler

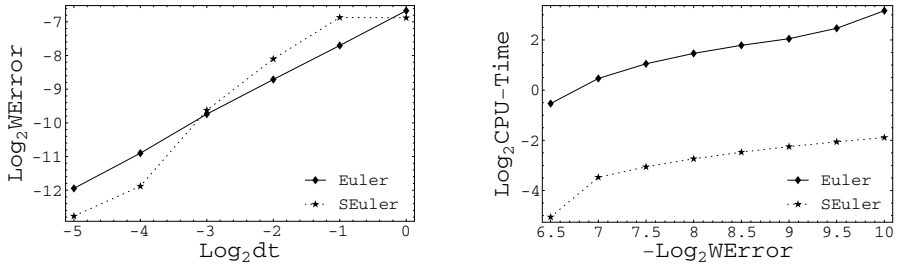


Fig. 4. Log-log plots of weak error versus time step size and CPU time versus the weak error for call option with Euler and simplified Euler scheme.

method is significantly more efficient than the Euler scheme, for every level of accuracy. We observed similar results also for in the money and out of the money options. In summary, one can say that the proposed rather simple random bits generators when combined with simplified schemes can significantly enhance the efficiency of typical Monte Carlo simulations in finance.

References

1. Boyle, P., M. Broadie, & P. Glasserman (1997). Monte Carlo methods for security pricing. *J. Econom. Dynam. Control* **21**(8-9), 1267–1321.
2. Boyle, P. P. & S. H. Lau (1994). Bumping up against the barrier with the binomial method. *J. Derivatives*, 6–14.
3. Heston, S. L. & G. Zhou (2000). On the rate of convergence of discrete-time contingent claims. *Math. Finance* **10**(1), 53–75.
4. Hofmann, N. (1994). *Beiträge zur schwachen Approximation stochastischer Differentialgleichungen*. Ph. D. thesis, Dissertation A, Humboldt Universität Berlin.
5. Hofmann, N. & E. Platen (1996). Stability of superimplicit numerical methods for stochastic differential equations. *Fields Inst. Commun.* **9**, 93–104.
6. Kloeden, P. E. & E. Platen (1999). *Numerical Solution of Stochastic Differential Equations*, Volume 23 of *Appl. Math.* Springer. Third corrected printing.
7. Kloeden, P. E., E. Platen, & H. Schurz (2003). *Numerical Solution of SDE's Through Computer Experiments*. Universitext. Springer. Third corrected printing.
8. Longstaff, F. A. & E. S. Schwartz (2001). Valuing American options by simulations: A simple least-squares approach. *Rev. Financial Studies* **14**(1), 113–147.
9. Press, W. H., S. A. Teukolsky, W. T. Vetterling, & B. P. Flannery (2002). *Numerical Recipes in C++. The art of Scientific Computing* (2nd ed.). Cambridge University Press.
10. Studer, M. (2001). *Stochastic Taylor Expansions and Saddlepoint Approximations for Risk Management*. Ph. D. thesis, Swiss Federal Institute of Technology Zurich.
11. Tausworthe, R. C. (1965). Random numbers generated by linear recurrence modulo two. *Mathematics of Computation* **19**, 201–209.

Time-Scale Transformations: Effects on VaR Models

Fabio Lamantia[1], Sergio Ortobelli[2], and Svetlozar Rachev[3]

[1] University of Calabria,
Ponte Bucci, cubi 3B-3C, 87036 Rende, Italy
lamantia@unical.it
[2] University of Bergamo,
Via dei Caniana 2, 24127 Bergamo, Italy
sol@unibg.it
[3] University of California, Santa Barbara
and University of Karlsruhe,
Postfach 6980, D-76128 Karlsruhe, Germany
rachev@pstat.ucsb.edu

Abstract. This paper investigates the effects of using temporal aggregation rules in the evaluation of the maximum portfolio loss[1]. In particular, we propose and compare different time aggregation rules for VaR models. We implement time-scale transformations for: (i) a EWMA model with Student's t conditional distributions, (ii) a stable sub-Gaussian model, (iii) a stable asymmetric model. All models are subjected to backtest on out-of-sample data in order to assess their forecasting power and to show how these aggregation rules perform in practice.

1 Introduction

Several empirical and theoretical studies on the asymptotic behavior of financial returns (see, among others, [4], [6]) justify the assumption of stable distributed returns. The joint stable sub-Gaussian family is an elliptical family recently used in portfolio theory and risk management (see [8], [10], [11]).

Following these studies, our paper presents and compares some alternative models for the calculation of VaR taking into consideration their time scale transformations. Firstly, we consider EWMA models with conditional elliptical distributed returns and finite variance. Secondly, we describe VaR models in the domain of attraction of stable laws. In particular, we focus our attention on returns either with conditional multivariate Student's t-distributions or with stable Paretian distributions. We describe a time rule for each model and we analyze and compare each performance considering conditional and unconditional coverage tests. We also test a particular temporal rule of VaR for the stable EWMA model in the same way as we did for the elliptical EWMA model with finite variance. In order to consider the asymmetry of financial series, we assume conditional jointly α-stable distributed returns. The asymmetric stable model results from a new conditional version of the stable three fund separation model

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 779–786, 2004.
© Springer-Verlag Berlin Heidelberg 2004

recently proposed in portfolio theory. In this case too, under some regularity conditions, we obtain a time rule of VaR. Finally, we compare the performance of all symmetric and asymmetric VaR time rules proposed. In particular, we evaluate VaR estimates of all models considering different temporal horizons, conditional and unconditional coverage backtesting methods (see, among others, [1]).

The paper is organized as follows: in Section 2 we propose and formalize time rules for elliptical EWMA models with finite variance. Section 3 introduces time rules for returns in the domain of attraction of stable laws. In Section 4 we backtest the proposed VaR models assessing their ability to capture extreme returns. Finally, we briefly summarize the paper.

2 Elliptical EWMA Models with Finite Variance

In some recent approaches (see, among others, [5],[7]) different exponential weighting moving average (EWMA) models were proposed to compute the value at risk of a given portfolio. The EWMA models assume that the conditional distribution of the continuously compounded return is an elliptical law. In particular, the RiskMetrics model is a EWMA model with conditional Gaussian distributed returns. The assumption of conditional elliptical distributed returns simplifies the VaR calculation for those portfolios with many assets. If we denote with $w = [w_1, w_2, \dots, w_n]'$ the vector of the positions taken in n assets forming the portfolio, its return at time t is given by

$$z_{P,t} = \sum_{i=1}^{n} w_i z_{i,t}, \tag{1}$$

where $z_{i,t} = \log\left(\frac{P_{t,i}}{P_{t-1,i}}\right)$ is the (continuously compounded) return of $i-th$ asset during the period $[t-1, t]$, and $P_{t,i}$ is the price of $i-th$ asset at time t. Generally we assume that within a short period of time the expected return is null and that the return vector

$$z_t = [z_{1,t}, z_{2,t}, \dots, z_{n,t}]'$$

follows a conditional joint elliptical distribution. We can distinguish two different types of elliptical EWMA models:

1. models with finite variance,
2. models with infinite variance.

In both cases the conditional characteristic function of the return vector $z_t = [z_{1,t}, ..., z_{n,t}]'$ is given by

$$\Phi_{z_t}(u) = E_t(e^{iu'z_t}) = f\left(u'Q_{t/t-1}u\right),$$

and $Q_{t/t-1} = \left[\sigma^2_{ij,t/t-1}\right]$ is either the variance covariance matrix (if it exists as finite), or another dispersion matrix when the return variance is not finite

(see [9]).That is, every return conditioned by the forecasted volatility level is distributed like a standardized elliptical distribution: $z_{i,t}/\sigma_{ii,t/t-1} \sim E(0,1)$ and any linear combination of the returns is elliptically distributed, $z_{P,t} \stackrel{d}{=} E(0, w'Q_{t/t-1}w)$, where $\sigma^2_{P,t/t-1} = w'Q_{t/t-1}w$ is the dispersion of portfolio $z_{P,t}$ and $Q_{t/t-1} = [\sigma^2_{ij,t/t-1}]$ is the forecasted dispersion matrix. When the elliptical distribution admits a finite variance, then we can estimate the variance and covariance matrix $Q_{t/t-1}$ considering the RiskMetrics' EWMA recursive formulas (see [5]).

The explicit modeling of the volatility series captures the time–varying persistent volatility observed in real financial markets. Under the elliptical assumption for the conditional returns, the Value at Risk of $z_{P,t+1} = w'z_{t+1}$ at $(1-\theta)\%$ (denoted by $VaR_{\theta,t+1}$) is given by simply multiplying the volatility $\sigma_{P,t+1/t}$ forecast in the period $[t, t+1]$, times the tabulated value of the corresponding standard elliptical percentile $k_{1-\theta}$ of $E(0,1)$. Therefore,

$$VaR_{\theta,t+1/t}(z_{P,t+1}) = k_{1-\theta}\sigma_{P,t+1/t}. \tag{2}$$

When both vectors of returns $z_t = [z_{1,t}, \ldots, z_{n,t}]'$ and $Z_{t+T} = [Z_{1,t+T}, ..., Z_{n,t+T}]'$ (where $Z_{i,t+T} = \log\left(\frac{P_{t+T,i}}{P_{t,i}}\right) = \sum_{s=1}^{T} z_{i,t+s}$) follow the Gaussian EWMA model, then, under further regularity assumptions, the $(1-\theta)\%$ VaR in the period $[t, t+T]$ is given by

$$VaR_{\theta,t+T/t} = \sqrt{T}VaR_{\theta,t+1/t}. \tag{3}$$

This time rule simplifies the computation of the maximum loss that could occur for a given level of confidence in a temporal horizon greater than the unity. In addition, among the elliptical EWMA models with finite variance, the RiskMetrics model is the only one for which the temporal rule (3) can be used. As a matter of fact, the Gaussian law is the unique elliptical distribution with finite variance such that the sum of elliptical i.i.d. random variables belongs to the same family of elliptical random variables, that is, vectors $z_t = [z_{1,t}, \ldots, z_{n,t}]'$ and $Z_{t+T} = [Z_{1,t+T}, ..., Z_{n,t+T}]'$ could follow the same elliptical EWMA model only if $Z_{i,t+T} = \sum_{s=1}^{T} z_{i,t+s}$ and $z_{i,m} = \sigma_{ii,m/m-1}\varepsilon_{i,m}$ $(i = 1, ..., n$ and $m = t+1, ..., t+T)$ are conditional Gaussian distributed. Thus, the temporal rule (3) cannot be extended to the EWMA models with conditional elliptical non Gaussian distributed returns and finite variance as well as it cannot be extended to the GARCH-type model (see [2]). However, in [7] it is proved a further time aggregation rule when $z_t = [z_{1,t}, z_{2,t}, \ldots, z_{n,t}]'$ and $Z_{t+T} = [Z_{1,t+T}, ..., Z_{n,t+T}]'$ follow different EWMA models with conditional elliptical returns $z_{i,m} \stackrel{d}{=} E_1(0, \sigma_{ii,m/m-1})$, and $Z_{i,t+T} \stackrel{d}{=} E_2(0, \sigma_{ii,t+T/t})$. Under these assumptions,

$$VaR_{\theta,t+T/t} = \sqrt{T}MVaR_{\theta,t+1/t}, \tag{4}$$

where $M = \frac{\widetilde{k}_{2,1-\theta}}{\widetilde{k}_{1,1-\theta}}$ and $\widetilde{k}_{1,1-\theta}$, $\widetilde{k}_{2,1-\theta}$ are respectively the corresponding $1-\theta$ elliptical $E_1(0,1)$, $E_2(0,1)$ percentiles. Recall that the sum of elliptical i.i.d. random variables is elliptical distributed but it does not necessarily belong to the same elliptical family (see [3]). Then, the sum of q i.i.d. elliptical distributions $E_1(0,1)$ gives another elliptical distribution with variance equal to q, i.e. $\sum_{s=1}^{q} E_1(0,1) \stackrel{d}{=} \sqrt{q} E_2(0,1)$.

A typical multivariate elliptical distribution with null mean and finite variance is the multivariate Student's t-distribution with $v > 2$ degrees of freedom **MV-t**$(\mathbf{0}, \sum_v)$. These distributions were often used in literature in order to justify the leptokurtic behavior of conditional returns (see, among others, [3]). Therefore, we can assume that the return vector $z_s = [z_{1,s}, ..., z_{n,s}]'$ follows a EWMA model with conditional t-distributed returns and $v > 2$ degrees of freedom. Under this assumption every return $z_{i,s}$ admits the following conditional density function

$$t(x/\sigma_{ii,s/s-1}, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma_{ii,s/s-1}\left((v-2)\pi\right)^{1/2}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{\sigma^2_{ii,s/s-1}(v-2)}\right)^{-\frac{v+1}{2}}.$$

We refer to [7] for further details about the properties of the EWMA model with conditional t-distributed returns.

3 Alternative Models with Stable Distributions

In this section we present some alternative models to compute VaR. In particular, we focus our attention on two different stable models for the profit/loss distribution:

1. the stable sub-Gaussian EWMA (SEWMA) model,
2. the stable asymmetric model.

3.1 The SEWMA Model

The SEWMA model assumes that the conditional distribution of the continuously compounded returns vector $z_t = [z_{1,t}, ..., z_{n,t}]'$ is α-stable sub-Gaussian ($\alpha > 1$) with characteristic function

$$\Phi_{z_t}(u) = E_t(e^{iu'z_t}) = \exp\left(-\left(u'Q_{t/t-1}u\right)^{\alpha/2} + iu'\mu_t\right),$$

where $Q_{t/t-1} = \left[\sigma^2_{ij,t/t-1}\right]$ is the conditional dispersion matrix, and $\mu_t = E(z_t)$, even if we assume that within a short period of time the expected return is null. This model is an elliptical EWMA model with infinite variance. In particular,

we observe that for any $i, j = 1, ..., n$ the elements of the dispersion matrix can be defined

$$\sigma^2_{ij,t/t-1} = (A(\alpha, p))^{\frac{2}{p}} f(p, \widetilde{z}_{i,t}, \widetilde{z}_{j,t}) \quad \text{for every } p \in [1, \alpha)$$

where $\widetilde{z}_{i,t} = z_{i,t} - \mu_{i,t}$, $A(\alpha, p) = \frac{\Gamma\left(1-\frac{p}{2}\right)\sqrt{\pi}}{2^p \Gamma\left(1-\frac{p}{\alpha}\right)\Gamma\left(\frac{p+1}{2}\right)}$, $\widetilde{z}_j^{\langle p-1\rangle} = sgn\left(\widetilde{z}_j\right)|\widetilde{z}_j|^{p-1}$ and $f(p, z_{i,t}, z_{j,t}) = E_{t-1}\left(\widetilde{z}_{i,t}\left(\widetilde{z}_{j,t}\right)^{\langle p-1\rangle}\right)\left(E_{t-1}(|\widetilde{z}_{j,t}|^p)\right)^{\frac{2-p}{p}}$. We refer to [7] for further details on the estimation of the elements of the dispersion matrix $Q_{t/t-1}$. Under the assumptions of the SEWMA model, the $(1-\theta)\%$ VaR in the period $[t-1, t]$ is obtained by multiplying the corresponding percentile, $k_{1-\theta,\alpha}$, of the standardized α-stable $S_\alpha(1, 0, 0)$, times the forecast volatility $\sigma_{P,t/t-1} = \sqrt{w' Q_{t/t-1} w}$, that is

$$VaR_{\theta,t/t-1} = k_{1-\theta,\alpha}\sigma_{P,t/t-1}. \tag{5}$$

Moreover, just like in the case of the elliptical EWMA model , we obtain a time rule for the dispersion measure $Q_{t+T/t} = \left[\sigma^2_{ij,t+T/t}\right] = T^{\frac{2}{\alpha}} Q_{t+1/t}$ and under some regularity conditions, it follows the temporal aggregation rule :

$$VaR_{\theta,t+T/t} = T^{\frac{1}{\alpha}} VaR_{\theta,t+1/t}. \tag{6}$$

Observe that among the elliptical distributions, the $\alpha-$stable sub-Gaussian distributions with $\alpha \in (0, 2]$ (where with $\alpha = 2$ we obtain the Gaussian case) are the unique elliptical distributions such that the sum of i.i.d. elliptical random variables belongs to the same family of elliptical random variables. That is, vectors $z_t = [z_{1,t}, ..., z_{n,t}]'$ and $Z_{t+T} = [Z_{1,t+T}, ..., Z_{n,t+T}]'$ could follow the same elliptical EWMA model only if $Z_{i,t+T} = \sum_{s=1}^{T} z_{i,t+s}$ and $z_{i,m} = \sigma_{ii,m/m-1}\varepsilon_{i,m}$ $(i = 1, ..., n$ and $m = t+1, ..., t+T)$ are conditional $\alpha-$stable sub-Gaussian distributed with $\alpha \in (0, 2]$.

3.2 An α-Stable Model with Asymmetric Distributed Returns

As an alternative to the previous model, we can consider the asymmetry of stable distributions generalizing the model proposed in [8]. In particular, we can consider the following three-fund separation model of conditional centered returns:

$$\widetilde{z}_{i,t} = z_{i,t} - \mu_{i,t} = b_{i,t} Y_t + \sigma_{ii,t/t-1}\varepsilon_{i,t}; \;\; i = 1, ..., n, \tag{7}$$

where $\mu_{i,t} = E(z_{i,t})$, the values $b_{i,t}$ will be determined with an OLS method, while the random vector $\varepsilon_t = (\varepsilon_{1,t}, \varepsilon_{2,t}, ..., \varepsilon_{n,t})'$ is α-stable sub-Gaussian distributed with zero mean and it is independent of $Y_t \sim S_\alpha\left(\sigma_{Y_t}, \beta_{Y_t}, 0\right)$. In particular, we assume that the centered return vector $\widetilde{z}_{t+1} = [\widetilde{z}_{1,t+1}, ..., \widetilde{z}_{n,t+1}]'$ is

conditional jointly α-stable distributed with conditional characteristic function

$$\Phi_{\tilde{z}_{t+1}}(u) = E_t(e^{iu'z_{t+1}}) = \exp\left(-\left(\left(u'Q_{t+1/t}u\right)^{\alpha/2} + \left|u'b_{t+1}\sigma_{Y_{t+1}}\right|^{\alpha}\right)\times\right. \tag{8}$$
$$\left.\times\left(1 - i\frac{\left|u'b_{t+1}\sigma_{Y_{t+1}}\right|^{\alpha} sgn(u'b_{t+1})\beta_{Y_{t+1}}}{\left(u'Q_{t+1/t}u\right)^{\alpha/2} + \left|u'b_{t+1}\sigma_{Y_{t+1}}\right|^{\alpha}}\tan\left(\frac{\pi\alpha}{2}\right)\right)\right),$$

where $\sigma_{Y_{t+1}}$ and $\beta_{Y_{t+1}}$ are respectively the dispersion and the skewness of the factor $Y_{t+1} \stackrel{d}{=} S_\alpha\left(\sigma_{Y_{t+1}}, \beta_{Y_{t+1}}, 0\right)$, that is an α-stable asymmetric (i.e. $\beta_{Y_{t+1}} \neq 0$) centered index return. Moreover, just like for the SEWMA model, we obtain the following time rule when the parameters α, β_{Y_t}, σ_{Y_t}, b_t are constant over the time

$$VaR_{\theta,t+T/t} = T^{\frac{1}{\alpha}} VaR_{\theta,t+1/t}. \tag{9}$$

We again refer to [7] for further details on properties of this stable VaR model.

4 Backtest Models

This section presents an analysis through backtest in order to assess the reliability of the models proposed to compute VaR. We propose three different methods for evaluating the Value at Risk estimates of 25 random portfolios and they are: a basic backtest method to verify if the average coverage of the VaR is equal to the nominal coverage; the conditional and the unconditional coverage tests proposed by [1].

During the period 15/11/93–30/01/98 we have examined daily, 10 days, and 60 days returns of Gaussian distribution, Student's t distributions, Stable sub-Gaussian distribution, stable asymmetric distribution and distributions in the domain of attraction of stable laws. We use some of the most representative index returns of the international market (Brent crude, CAC40, Corn n.2, DAX100, Dow Jones Industrial, FTSE all Share, Goldman Sachs, Nikkei 500, S&P500, Reuters) and their relative exchange rates whose values we converted into USD. Over a period of 769 days, we have computed the interval forecasts using the time aggregation rules and considering $\theta = 95\%$ and $\theta = 99\%$.

4.1 The Basic Backtest Method

In the first backtest analysis proposed we determined how many times during the period taken into account the profits/losses fall outside the confidence interval. In particular, for $\theta = 95\%$ and $\theta = 99\%$, the expected number of observations outside the confidence interval must not exceed respectively 5% and 1%.

The first empirical analysis compares the results obtained from the backtest carried out among the elliptical EWMA models and the stable asymmetric model for $\theta = 95\%$ and $\theta = 99\%$. In view of this comparison, we assume the same

parameters of daily models analyzed in [7]. Then, we apply the different time rules (3), (4), (6), and (9) in order to forecast VaR estimates and compare their performance.

Among the alternative models for the VaR calculation, we could observe that the stable and the Student's models and their time rules are more reliable than the RiskMetrics one, in terms of confidence interval $\theta = 99\%$ and 10 days returns particularly. The advantage of using stable models as an alternative to the normal one is reduced when the percentiles are higher than 5% and we consider three months returns.

4.2 Conditional and Unconditional Coverage Tests

Under every distributional hypothesis and for every portfolio $z_{P,t+1}$ we have evaluated daily $VaR_{\theta,t+1/t}(z_{P,t+1})$. Following the interval forecast method proposed in [1], we can propose the following tests:

1. a likelihood ratio test for unconditional coverage LR_{uc} with an asymptotic $\chi^2(1)$ distribution,
2. a likelihood ratio test for conditional coverage LR_{cc} with an asymptotic $\chi^2(2)$ distribution.

These tests partially confirm the previous basic backtest analysis. In particular, we observe that generally the Gaussian time rule does not offer good performance, whilst the time rules (4) and (6) present the best performance. Further tables describing our empirical analysis in details are available by the authors if requested [2].

5 Concluding Remarks

This paper proposes and compares alternative models for the VaR calculation. In the first part we describe several elliptical and stable Paretian exponential weighted moving average models. In the second part, we compare the efficiency of different time aggregation rules to forecast VaR . The empirical comparison confirms that when the percentiles are below 5%, the hypothesis of normality of the conditional return distribution determines intervals of confidence whose forecast ability is low. In particular, the stable Paretian and the Student's t time aggregation rules have shown very good performance to predict future losses when we assume a temporal horizon of 10 days. Whereas, when we consider 60 days returns all the models do not present very good performances.

[2] For tables reporting conditional and unconditional tests on the above time aggregation rules, please refer to the following e-mail address: sol@unibg.it.

References

1. Christoffersen, P.: Evaluating interval forecasts. International Economic Review 39 (1998) 841-862
2. Diebold, F. X., Hickman, A., Inoue, A., Schuermann, T.: Scale models. Risk 11 (1998) 104–107
3. Embrecht, P., Lindskog, F., Mcneil, A.: Modelling dependence with copulas and applications to risk management. In (North Holland eds. S.T. Rachev) Chapter 8 Handbook of Heavy Tailed Distributions in Finance (2003) 329-384
4. Fama, E.: The behavior of stock market prices. Journal of Business 38 (1965) 34-105
5. Longerstaey, J., Zangari, P.: RiskMetrics - Technical Document. J.P. Morgan, Fourth edition, New York (1996)
6. Mandelbrot, B.: New methods in statistical economics. Journal of Political Economy 71 (1963) 421-440
7. Lamantia, F., Ortobelli, S., Rachev, S.: Value at Risk with stable distributed returns. Technical Report, University of Bergamo to appear in Annals of Operation Research (2004)
8. Ortobelli, S., Rachev, S., Schwartz, E.: The problem of asset allocation with stable distributed returns. To appear in Volume (Ed. Marcel Dekker) Stochastic Processes and Functional Analysis (2003)
9. Owen, J., Rabinovitch, R.: On the class of elliptical distributions and their applications to the theory of portfolio choice. Journal of Finance 38 (1983) 745-752
10. Rachev, S., Mittnik, S.: Stable Paretian model in finance. Wiley & Sons, Chichester (2000)
11. Rachev, S., Schwartz, E., Khindanova, I.: Stable modeling of market and credit value at risk. In (North Holland eds. S.T. Rachev) Chapter 8 Handbook of Heavy Tailed Distributions in Finance (2003) 249-328

[1] This work has been partially supported under Murst 40% 2002-3, legge 95/95 CNR 03.00105PF/25.

Environment and Financial Markets

Wojciech Szatzschneider[1], Monique Jeanblanc[2], and Teresa Kwiatkowska[3]

[1] Escuela de Actuaria, Universitat Anahuac,
Mexico City, Huixquilucan, C.P. 52760, Mexico
wojciech@anahuac.mx,
[2] Départment de Mathématiques, Université d'Evry Val d'Essone
rue Jarlan, F-91025 Evry Cedex, France
Monique.Jeanblanc@maths.univ-evry.fr
[3] Departamento de Filosofía, UAM Iztapalapa
Mexico City, San Rafael Atlixco 186 C.P. 09340, Mexico
tkwiatkowska@yahoo.com

Abstract. We propose to put the environment into financial markets. We explain how to do it, and why the financial approach is practically the only one suited for stopping and inverting environmental degradation. We concentrate our attention on deforestation, which is the largest environmental problem in the third world, and explain how to start the project and what kind of optimization problems should be solved to ensure the optimal use of environmental funds. In the final part we analyze the dynamical control for bounded processes and awards partially based on the mean of the underlying value.

1 Introduction

We argue that practical solutions for the environmental degradation are in a short supply. Most of the increasingly complex models set off different opinions about their applicability. Models should be well specified. It means that inputs should be observed or estimated. This requirement is hard to meet in environmental studies. Thus, the efficient global environmental decision–making becomes very difficult. Moreover politicians often tend to justify their decisions by inappropriate theories. This situation leads to proliferation of ineffective studies and waste of resources.

We shall propose to apply the market approach to get solutions of some of the environmental problems. It could result in more transparent transfer of funds and the involvement of everybody concerned. Also we can expect that the transparency could stem in an increment of these funds.

We shall focus on the issue of deforestation due to its importance for the global well–being, and the possibility to assess the number of trees.

This is not a paper about statistics or numbers. We will mention only that the annual deforestation rate is about 60,000 square miles.

Although the choice of any particular model is, at this stage, only of secondary importance, the primary goal being to start the project, in [4] was chosen as a dynamical model for the number of trees in a given region a $\mathrm{BESQ}^{\delta}_{\beta}$ process,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 787–794, 2004.
© Springer-Verlag Berlin Heidelberg 2004

which in financial language is 0th dimensional CIR model for interest rates. The Geometric Brownian Motion can be a candidate for an alternative model, and we use it in this study.

Our approach is based on a positive involvement of holders of "good" options bought or, in the first stage, obtained for free. In the case of the forest "good" means a kind of Asian call option. We hope that it is clearly understood why Asian and why call, "Bad" options are of course put options. We will show that, in a natural way, three kinds of optimization problems crop up, i.e. an individual agent problem, a local optimization problem, and aglobal optimization problem.

The first one concerns the problem how a holder's choice of a good option can possibly contribute to reforestation. The second one is how to choose prices of "good" and eventually "bad" options, to maximize the space mean of the temporal mean of the "asset" in given place. The last one is how to distribute funds into particular projects.

This situation slightly resembles the study of Executive stock options in [1].

However we must work with more difficult to analyze kind of awards based on the temporal mean and also with bounded models that represents more complications as we shall see in Section 3.

To create a market we propose three following stages of actions: 1) – choose a place, and give "good" environmental certificates (which we call options free of charges for the habitants of the community); 2) – sell good options; 3) – create a market with "good" and "bad" options.

The last stage can't be developed without applications of the second one, and could be the object of further challenging studies.

2 Analysis of the Second Stage, Static Approach

2.1 The Model before the Financial Intervention

For modeling a number of trees in a given region a 0th dimensional CIR process was proposed in [4], i.e. a process $\{X(t) : t \geq 0\}$ with negative drift ($\beta < 0$), defined by

$$dX(t) = 2\sqrt{\sigma X(t)}dW(t) + 2\beta X(t)dt, \quad X(0) = x_0 \geq 0.$$

This choice was justified by a heavy traffic approximation of the corresponding Piecewise Deterministic Markov Process as explained in our previous study which contains solutions of relevant mathematical problems within this model. Starting from different assumptions, (and putting for convenience $\sigma = 1$), the heavy traffic approximation leads to the geometric Brownian motion

$$dX(t) = X(t)dW(t) + \beta X(t)dt, \text{where } X(0) = x_0.$$

2.2 Financial Intervention. Local Goal

Suppose that given a fund Σ, the bank sells "good" (a sort of call) options on the number of trees in a given area, which we want to reforest.

Let us assume that one option is sold. We will clarify this assumption soon. While selling good options the bank should choose the optimal (also we will clarify soon what does it mean) price c, and the strike price k_1, in an award (which we call option) $\frac{\Sigma+c}{k-k_1}\int_0^1 (X(s)\wedge k-k_1)_+ds$, where k is the maximal capacity with $k < k_1$, and

$$dX(s) = X(s)dW(s) + BX(s)ds, \quad X(0) = x.$$

With the factor $A = (\Sigma + c)/(k - k_1)$ the bank will never loose money and unused funds will go to other projects.

For example we take one year as a time horizon, though a dynamical approach with several moving horizons could be more appropriate.

Optimal local goal means that the optimal agent's action assures the maximal value of the functional $E\big[\int_0^1 (X(s) \wedge k)ds\big]$.

In agents optimization problem we shall assume their linear utility. In this case the number of options sold is irrelevant in both local and agents optimization goals. Only one comment is needed. This option can be traded and this fact would eventually cause the concentration of capital in the hands of powerful individuals with positive effects. Only powerful individuals are able to face timber barons! In what follows we shall consider exactly one option sold.

2.3 Financial Intervention. Agent's Approach

We assume that an agent acts in the optimal way, maximizing his or her linear utility. It means that an agent can modify (if worthy) the original "asset" into

$$dX(s) = X(s)dW(s) + BX(s)ds, \quad X(0) = x, \quad \text{where} \quad x_0 \le x \le k, \quad B \ge \beta.$$

Assume that the cost of agent's involvement, if worthy, is equal to

$$c + c_1(x - x_0) + c_2(B - \beta)\int_0^1 (X(s) \wedge k)^p ds,$$

where c_1 is clearly the cost of planting (easy to set) and the cost of protection can be written as the product of 2 factors: cost of changes in the tendency and cost of actual state. We set $p = 1$ for further purposes.

Now agent's linear utility can be expressed as

$$A\Big(\int_0^1 E(X_s - k_1)_+ds - \int_0^1 E(X_s - k)_+ds\Big) - c - c_1(x - x_0) - c_2(B - \beta)\Big[k - \int_0^1 E(k - X_s)_+ds\Big].$$

With given optimal B^* and x^* (for any choice of c and k_1), we can choose optimal c and k_1 to maximize $E\int_0^1 (X_s \wedge k)_+ds$. Since $A = (\Sigma + c)/(k - k_1)$, so the parameter c plays a significant role, when solving the problem.

The global goal is to distribute the global environmental fund into particular projects with objective to get the optimal overall reforestation. Some weights can be included to stress importance of predetermined environmental goals.

2.4 An Example

As an example we take the area of 100 km^2. Assume that the agent's gain should be at least $c/3$, being c the price of the option as before. In very first applications

the agent's investment must be very attractive! The agent's return could be lower in the future. We assume $B \leq 0$ and chose the following initial data: $\beta = -1$, $k = 20$, $X(0) = 1$, $\sigma = 30$, $c_1 = 1$, $c_2 = 0.05$, with an additional constraint $c \geq 3$.

The *optimal solution* is: $c = 3$, $k_1 = 2$, $x^* = 14.5$, $B^* = 0$, *local goal* $= 12$, *agent's gain* $= 1.2$. Here k, $X(0)$, σ, c, k_1, x^*, local goal, and agent's gain are expressed in millions of dollars.

Using this example and some crude approximations, we conclude that using our approach the overall deforestation could be stopped with 40 billions of dollars. This amount seems high. However it could easily be bearable if the transfer of money from rich toward poor countries would be not unconditional, often supporting corrupt governments and inefficient local bureaucracy, but instead depend on concrete and predetermined environmental improvements.

Our proposal is that this conditional support should flow through market mechanisms.

In the final stage of good and bad options, these could give forest a kind of market value, which clearly doesn't reflect its real value.

3 Dynamical Approach

Static Agent's optimization problem represents only a crude approach to the reality. Global and local goals are static in their nature (with once fixed horizons) but agents always act dynamically. In this section we assume (for simplicity) that options are given for free, so we are in stage 1.

This reminds the problem of executive options by Cadenillas *et al* in [1] with main two differences: first – at least one part of the award should depend on the temporal mean of the underlying process, and second – options can be traded.

3.1 Elementary Models

Let us formulate first an elementary model. We shall explain later that this cannot be possibly the best choice in our modelling. The number of trees in a given region is represented by

$$dX(t) = \sigma(X(t))dW(t) + (u(t) - \beta)X(t)dt,$$

where $\beta > 0$, and $\sigma = \sigma(x)$ is an arbitrary function satisfying only mild conditions ensuring existence and uniqueness of the solution of this stochastic equation (cf. [7]). Agent's goal is to find

$$\max \mathrm{E}\big(A \int_0^T X(s)ds - c_1 \int_0^T u^2(s)X(s)ds\big).$$

Note that in this formulation there is no local optimization problem.

The Bellman approach (with $c_1 = 1$) leads to the equation

$$V_t(t,x) + x\left(A + \frac{V_x^2(t,x)}{4} - \beta V_x(t,x)\right) + \frac{1}{2}\sigma^2(x)V_{xx}(t,x) = 0,$$

because for optimal u we get $u = V_x(t,x)/2$. We look for the solution of the form $V(t,x) = xf(t)$, $V(T,x) = 0$, and $u = f(t)/2$. Now we have to solve equation

$$f'(t) + \left(\frac{f(t)}{2} - \beta\right)^2 - \beta^2 = -A.$$

Set $f(t) = 4H(t) + 2\beta$. We have: $H'(t) + H^2(t) = (-A + \beta^2)/4$. Let $H(t) = \varphi'(t)/\varphi(t)$, $\varphi(0) = 1$, $\varphi'(T)/\varphi(T) = -\beta/2$. Resulting equation for φ is

$$\frac{\varphi''(t)}{\varphi(t)} = \frac{\beta^2 - A}{4}.$$

This equation has the solution, for any T, only if $\beta^2 \geq A$. If $\beta^2 < A$ the "local" solution is

$$\alpha \sin(t) + (1-\alpha)\cos(t),$$

and $H(t)$ tends to ∞ in a finite interval. This means that agents "push too much" to obtain infinite gain. Therefore we can formulate the problem for short horizons only.

Under the condition $\beta^2 > A$, the solution for φ is $\alpha e^{ct} + (1-\alpha)e^{-ct}$, with $c = \sqrt{\beta^2 - A}/2$, $\alpha = \frac{(c-\beta/2)e^{-cT}}{(c+\beta/2)e^{cT} + (c-\beta/2)e^{-cT}} < 0$.

Finally, $f(0) = 4\left\{[\frac{2(c-\beta/2)e^{-cT}}{(c+\beta/2)e^{cT} + (c-\beta/2)e^{-cT}} - 1]c + 2\beta\right\}$, and $V(0,x) = xf(0)$.

This is a very nice explicite solution. However the model is not adequate because not only we have the natural bound for the number of trees in a given region. Even if not, (considering other environmental problems) the agent's action could cause unpredictable payment and no financial institution would accept the underlying risk. A similar solution (in terms of x instead of t) can be obtained while considering the agent's action until time τ; the first hitting time of k say, in the case of

$$dX(t) = \sqrt{X(t)}dW(t) + (u(t) - \beta)X(t)dt.$$

However what to do after τ (if $\tau < T$) remains unclear, and the hitting time of k in the case of optimal u_t does not have known distribution so the complete analysis becomes extremely difficult. Moreover, cost and drift proportional to $X(s)$ is perhaps not the best choice. The impact of the work implied on the protection and reforestation seems to be more complicated. The impact is smaller for small or large values of x. In the latter case there is not much to do and in the former one much work is to be done to restore the area. On the other hand the rate of deforestation follows a similar pattern. If x is large or small the rate of deforestation is smaller. Large x means no big environmental problems.

3.2 Bounded Models

In this section we give an example how one could deal with restriction imposed in the final part of previous section. We propose as the basic (simplified) model a stochastic equation

$$dY(t) = Y(t)(1-Y(t))dW(t) + \tfrac{1}{2}Y(t)(1-Y(t))(1-2Y(t))dt, \;\; Y(0) = 1/2,$$

with a bounded solution $0 < Y(t) < 1$, given by $Y(t) = 1/(e^{-W(t)} + 1)$, for $t \geq 0$.

As a model for the number of trees under continuous deforestation and the control u we propose the equation

$$dX(t) = X(t)(1-X(t))dW(t)+ \\ \left\{(\delta u(t)-\beta)X(t)(1-X(t)) + \tfrac{1}{2}X(t)(1-X(t))(1-2X(t))\right\}dt,$$

with $X(0) = x$, where $0 < x < 1$. Now defining $Z(t) = e^{\beta W(t)-\beta^2 t/2}$, we get for the agent's cost the expression

$$\int_0^T \tfrac{u^2(s)}{2}X(s)(1-X(s))ds + \int_0^T H(X(s))ds,$$

with some function $H = H(x)$, and for the agent's gain

$$F(X(T))+\tfrac{1}{2}\check{z}\int_0^T Z(s)X(s)(1-X(s))(2X(s)-1)ds+0.005\check{z}T+\int_0^T H(X(s))ds \\ -\left(\int_0^T \tfrac{u^2(s)}{2}X(s)(1-X(s))ds + \int_0^T H(X(s))ds\right),$$

and the goal is to maximize the expectation of this amount with $\check{z}$ playing the role of the Lagrange multipier, to be specified later. The second term is written in a form, which leads to a closed form solution of the problem. We assume that the part of award covers expenses $\int_0^T H(X(s))ds$. An easy application of Girsanov's thorem shows that $Z(s)X(s)(1-X(s))(2X(s)-1) > -0.1$. The term $\check{z}0.005T$ is added to do not have to deal with negative awards. We propose (to make the analysis easier)

$$F(X(t)) = \ln[(X(T-B))_+].$$

Now we are able to impose the financial institution constrain that $Award \leq \sigma$ (initial fund). The function H must be specified. Th function Ln is not the utility of an agent as it was proposed in [1], which remains linear in our approach, but stands here for the *form* of the award. Now

$$M(t) = X(t)Z(t) - \delta\int_0^t u(s)Z(s)X(s)(1-X(s))ds \\ -\tfrac{1}{2}\int_0^t Z(s)X(s)(1-X(s))(1-2X(s))ds$$

is a martingale, with $EM(t) = X(0) = x$.

As in [1], we also define

$$\tilde{F}(z) = \max_{s\geq 0}(F(z)-sz), \quad \tilde{G}(z) = \max_u\left(-\frac{u^2}{2}+\delta uz\right),$$

where the maximum is attained at $\hat{u} = \delta z$, and we will choose $z = \check{z}$ (still to be defined). So,

$$\mathrm{E}\Big[F(X_T)-\int_0^T \tfrac{u^2(s)}{2}X(s)(1-X(s))ds+\tfrac{1}{2}\check{z}\int_0^T Z(s)X(s)(1-X(s))(1-2X(s))ds\Big] \\ \leq \mathrm{E}(\tilde{F}(\check{z}Z_t)) + \int_0^T \tilde{G}(\check{z}Z_s)X(s)(1-X(s))ds + \check{z}x.$$

The optimal $X_T = \frac{1}{\check{z}Z(t)} + B$ on $[0,1]$, and 0 or 1 otherwise.

To avoid negative awards, B must be such that $(X(T)-B)_+ > 1$, so we assume that $B < -1$.

We have $E(M_T) = V_0$, and $u_t = \delta\check{z}Z_t$. From here we can determine $\check{z}$.

The next part becomes more complicated than in [1]. The process
$Z(t)X(t)-\delta\int_0^t\Big(Z(s)\check{z}\delta Z(s)(1-X(s))X(s)-\frac{1}{2}X(s)(1-X(s))(1-2X(s))Z(s)\Big)ds$
is a martingale, and we have

$dX(s)=X(s)(1-X(s))dW(s)+$
$\big\{\delta\check{z}Z(s)X(s)(1-X(s))+\frac{1}{2}X(s)(1-X(s))(1-2X(s))-\beta X(s)(1-X(s))\big\}ds,$

so, we now get

$$x=\mathrm{E}(\tilde{X}_TZ(T))-\delta^2\check{z}\int_0^T Z^2(s)X(s)(1-X(s))ds$$
$$-\tfrac{1}{2}\int_0^T X(s)(1-X(s))(1-2X(s))Z(s)ds.$$

In order to find $\check{z}$ we have to make use of 2 equalities:

$$E(Z^2(t)X(t)(1-X(t)))=E[Z^2(t)u_1(X(t))]$$

and

$$E(Z(t)X(t)(1-X(t))(1-2X(t)))=E[Z(t)u_2(X(t))].$$

To be specific let us fix $\delta=1$, $\beta=1/2$. These expectations can be expressed without Z. Using Girsanov's theorem we have for example

$$E(Z^2u_1(X(t)))=e^{\frac{1}{4}t}E^*(u_1(X(t))),$$

where under P^* law we have

$dX(t)=X(t)(1-X(t))dW^*(t)+\frac{1}{2}X(t)(1-X(t))dt+$
$\check{z}e^{W^*(t)+t/2}X(t)(1-X(t))dt+\frac{1}{2}X(t)(1-X(t))(1-2X(t))dt.$

Using again Girsanov's theorem we get the following, quite complicated formula
$\mathrm{E}(\tilde{X}_TZ(T))-x=$
$\check{z}\int_0^T e^{-s/4}\mathrm{E}\Big\{\Big(\frac{1}{e^{-W(s)}+1}\Big)^2\exp\Big[\int_0^s h(u)dW(u)+\frac{1}{2}\int_0^s h(u)(1-h(u))du\Big]\Big\}ds$
$+\int_0^T E\Big\{\Big(\frac{1}{e^{-W(s)}+1}\Big)^3e^{-W(s)}\left(e^{-W(s)}+1\right)\exp[\int_0^s g(u)dW(u)\quad-$
$\frac{1}{2}\int_0^s g^2(u)du]\Big\}ds,$

$$\text{with}\quad h(u)=\frac{e^{W(u)+u/2}}{\frac{1}{\check{z}}+\int_0^u e^{W(t)+t/2}dt},\quad g(u)=\frac{e^{\frac{1}{2}W(u)+u/8}}{\frac{1}{\check{z}}+\int_0^u e^{\frac{1}{2}W(t)+\frac{1}{8}t}dt}.$$

However we are able to get the solution in the explicit form.

A bounded model that can be controled explicitly is of the following form

$$dX(t)=\sqrt{|X(t)|(1-X(t))}dW(t)+(u(t)-\beta)(1-X(t))dt.$$

This model is bounded analogue of drifted squared Bessel process (CIR) from Section 2.1. As before $\beta>0$, $0\le X(0)<1$. If $u(t)\ge\beta$ for any $t\in T$, then $0\le X(t)\le 1$ and we can skip $|\cdot|$. Our goal is to maximize

$$-A\int_0^T(1-X(s))ds-\int_0^T u^2(s)(1-X(s))ds+\Delta,$$

with $A,\ \Delta>0$. The Bellman approach leads to the equation

$$F_t(x,t)+\left(\frac{F_x^2(x,t)}{4}-\beta F_x(x,t)+A+\frac{1}{2}|x|F_{xx}(x,t)\right)(1-x)=0.$$

Writing $F(x,t) = h(t)(1-x)$, we get

$$h'(t) + \left(\frac{1}{2}h(t) + \beta\right)^2 - \beta^2 + A = 0.$$

Setting $h(t) = 4H(t) - 2\beta$, we obtain

$$H'(t) + H^2(t) = \frac{\beta^2 - A}{4}.$$

As before we ask for $\beta^2 > A$, and $H(t) < 0$ for $0 \leq t \leq T$. This implies that $u(t) = F_x(x,t)/2 = -h(t)/2 > 0$. So, $F(x,t) = (1-x)h(t) < 0$, for $0 \leq t < T$, but for any choice of $u(T) = \gamma \geq 0$ we can make the award positive because of the presence of Δ.

Now the solution for H is as in the previous section: $\frac{\varphi''(t)}{\varphi(t)} = \frac{\beta^2 - A}{4}$, $\varphi(0) = 1$, and $\frac{\varphi'(T)}{\varphi(T)} = \frac{-\gamma + 2\beta}{4} < 0$.

4 Final Comments

So how do we reforest? Particularly what kind of trees are useful for the ecosystem in question? The inclusion of this factor could give us more parameters in our optimization problems.

Many other environmental topics can be treated in a way similar to what we have proposed. We believe that direct involvement in conservation would produce much better results than compensations proposed by James *et al* in [3].

References

1. Cadenillas, A., Civitanic, J. Zapatero, F.: Leverage Decision and Manager Compensation with Choice of Effort and Volatility, working paper, (2003), *http://math.usc.edu/∼cvitanic/papers.html*
2. Fitzimmons, A.: *Defending Illusions*, Rowman & Littlefield, (1999)
3. James A., Gaston K., Balmford A.: Balancing the Earth's Accounts. Nature. Vol 401 (1999) 323–324
4. Jeanblanc, M., Szatzschneider W.: Environment and finance: why we should make the environment a part of the financial markets, *Revista Mexicana de Economía y Finanzas*, Vol. 1, No. 2, (2002) 131–142.
5. Merton, R., Bodie, Z,: A Conceptual Framework for Analyzing the Financial Environment. In the Global Financial System: A Functional Perspective, Harvard Bussiness School Press, (1995)
6. Oates, J.: *Myth and Reality in the Rain Forest*, University of California Press, (1999)
7. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion, A series of comprehensive studies in mathematics. Volume 23. Springer, (1999)

Pricing of Some Exotic Options with NIG-Lévy Input

Sebastian Rasmus[1], Søren Asmussen[2], and Magnus Wiktorsson[1]

[1] Center for Mathematical Sciences, University of Lund,
Box 118, 221 00 Lund, Sweden
{rasmus,magnusw}@maths.lth.se
[2] Theoretical Statistics, University of Aarhus,
Ny Munkegade, 8000 Aarhus C, Denmark
asmus@imf.au.dk

Abstract. We study the problem of pricing barrier options and Russian options driven by exponential NIG Lévy processes by simulation. Simulating at a discrete grid creates a systematic bias because the minimum and maximum in between grid points is neglected. The proposed solution is to simulate the large jumps only and use a Brownian approximation for the rest combined with explicit formulas for Brownian minima and maxima.

1 Introduction

Models for financial data different from the classical Black-Scholes (B&S) geometric Brownian motion (BM) have recently received considerable attention, see for instance Boyarchenko & Levendorskiĭ [8], Chan [9] and Schoutens [16]. The motivation is to a large extent statistical since many log-returns show deviations from normality. The most popular alternative is Lévy process models, and in particular the NIG model (see Barndorff Nielsen [6]) has become popular because of its flexibility.

The consequence for option prices are multiple. One is non-uniqueness of risk-neutral measures, another that the wealth of explicit formulas for BM in the B&S world is no longer available.

For simple options like Europeans call, only marginal distributions are required and pricing may be done by numerical integration. For slightly more complicated examples like Asian options, one may just simulate along a discrete grid. Numerical studies for such problems and comparison with the B&S prices have been carried out for the NIG case in Albrecher & Predota [2]. The present paper continues this work by studying some options (barrier options and Russian options) where the computational aspect is more complicated.

2 Lévy Processes

A Lévy process is a stochastically continuous process with stationary and independent increments. This implies the existence of a *characteristic exponent* ($\kappa(\xi)$) which form is given by the *Lévy triplet* (a, σ, ν) as follows,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 795–802, 2004.
© Springer-Verlag Berlin Heidelberg 2004

$$\mathsf{E}[e^{i\xi X_t}] = e^{t\kappa(i\xi)} = \exp\left\{t\left(ia\xi - \frac{\sigma^2\xi^2}{2} + \int_{\mathbb{R}}(e^{i\xi x} - 1 - i\xi x\ \mathbf{1}_{\{|x|<1\}})\nu(dx)\right)\right\}$$

where $a, \xi \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ and ν a measure on $\mathbb{R}\backslash\{0\}$ which satisfies $\int_{\mathbb{R}}(x^2 \wedge 1)\ \nu(dx) < \infty$. The finite and infinite variation case refer to the integral, $\int_{|x|<1}|x|\ \nu(dx)$ being finite or infinite. See Bertoin [7] and Sato [13] for more details.

2.1 The Esscher Transform

If the Lévy process X_t is a BM or a Poisson process the stock price model based on

$$S_t = S_0\ e^{X_t}$$

is complete. In all other cases the model is incomplete (see Chan [9] and Cherny & Shiryaev [10]) and a there are many martingale measures $\mathbb{Q}$ making X_t a Lévy process and satisfying

$$e^{rt} = \mathsf{E}^{\mathbb{Q}}[e^{X_t}] = \exp\left\{t\kappa^{\mathbb{Q}}(1)\right\} \tag{1}$$

where $\kappa^{\mathbb{Q}}$ is the characteristic exponent of X_t under $\mathbb{Q}$. Within the class of exponentially tilted Lévy processes, i.e. with $\kappa^{\mathbb{Q}}$ of the form

$$\kappa^{\mathbb{Q}}(x) = \kappa(x+\theta) - \kappa(\theta) \tag{2}$$

the equation $\kappa^{\mathbb{Q}}(1) = r$ has either a unique solution θ_0 (this is the typical situation for light tails and the corresponding $\kappa^{\mathbb{Q}}$ is called the *Esscher transform*) or none (typical for heavy tails).

2.2 The *NIG* Lévy Process

The $NIG(\alpha, \beta, \mu, \delta)$ distribution, defined for $\alpha, \delta \geq 0$, $|\beta| \leq \alpha$, $\mu \in \mathbb{R}$ has density

$$f(x) \ = \ \frac{\alpha\delta}{\pi}\ \frac{K_1\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right)}{\sqrt{\delta^2 + (x-\mu)^2}}\ \exp\left\{\delta\gamma + \beta(x-\mu)\right\}$$

where $\gamma^2 = \alpha^2 - \beta^2$ and K_1 is the modified Bessel function of the third kind. The NIG Lévy process is a Lévy process such that X_t has a $NIG(\alpha, \beta, \mu t, \delta t)$ distribution. It is a pure jump process and its Lévy measure has density

$$\nu(x; \alpha, \beta, \delta) = \frac{\delta\alpha}{\pi|x|}\ \exp\{\beta x\}\ K_1(\alpha|x|)$$

With this the characteristic exponent is given by

$$\kappa(\xi) = \mu\xi + \delta\left[(\alpha^2 - \beta^2)^{1/2} - (\alpha^2 - (\beta+\xi)^2)^{1/2}\right]$$

It is easy to see ([2]) that under the Esscher measure X_t is still a NIG Lévy process, with the same α, μ, δ but β replaced by

$$\beta + \theta_0 = -\frac{1}{2} + \sqrt{\frac{(\mu-r)^2}{\delta^2 + (\mu-r)^2}\ \alpha^2 - \frac{(\mu-r)^2}{4\delta^2}}$$

3 Simulating Jumps from the Lévy Measure

The Lévy measure ν contains all information about the jumps. If its total mass λ over $\mathbb{R}$ is finite, the jump times form a Poisson process with intensity λ and the jump size distribution is $g(dx) = \nu(dx)/\lambda$. So, if g can be simulated the path of a Lévy process can be generated as a compound Poisson process: first a random time grid from an $Exp(\lambda)$ distribution is drawn and at each grid point we draw a jump from g.

Since the Lévy measure is allowed to have infinite mass in an open interval containing zero the compound Poisson method may break down. In Asmussen & Rosinski [4] it is shown (under certain conditions valid for the NIG case) that jumps smaller than $\epsilon > 0$ may be approximated by a BM with drift and variance given by,

$$m(\epsilon) = -\int_{\epsilon<|x|<1} x\ \nu(dx), \quad v^2(\epsilon) = \int_{|x|<\epsilon} x^2\ \nu(dx) \tag{3}$$

So, the general idea is to use the compound Poisson process method above for jumps bounded away from zero, say larger than ϵ ($J_{>\epsilon}$). To be explicit we use the notation λ_ϵ and g_ϵ. The rest of the jumps, in $[-\epsilon, \epsilon]$ ($J_{<\epsilon}$), is then approximated by a BM with drift $m(\epsilon)$ and variance $v^2(\epsilon)$.

3.1 Reflecting the Lévy Process

For the Russian option we are interested in the reflected (at the supremum) version R_t of the Lévy process X_t and the hitting time τ_k of the barrier k, i.e.

$$R_t = \sup_{0\le s\le t} X_s - X_t \qquad \tau_k = \inf\{t > 0 : R_t \ge k\} \tag{4}$$

In the infinite variation case the small jump approximation is particularly important. Consider the simulation of the running maximum $M_t^X = \sup_{0\le s\le t} X_s$. On the random grid generated by the compound Poisson process of the large jumps $J_{>\epsilon}$ a maximum between the grid points may be missed. Since there is an explicit formula for the joint distribution of the BM and its maximum (see for instance Asmusen, Glynn & Pitman [3] Algorithm B) the small jump approximation make it possible to compensate for this bias. Note also that since $J_{<\epsilon}$ may be approximated by the well known BM, this compensation is computable.

This aspect becomes even more important simulating τ_k. Here we need the running maximum M_t^R of R_t, which in turn needs M_t^X. To update them both over a grid interval we need the joint distribution the infimum and supremum of a BM over a compact interval. This is unknown to us and we do the following approximation. As time evolve M_t^X and M_t^R are updated independently over each of the exponential intervals between the jumps. This should be a very close approximation since the probability that they both are updated simultaneously over such an interval is negligible: if X_t is close to M_t^X, R_t is close to 0, i.e. far from M_t^R and vice versa.

3.2 The *NIG* Case

In the paper Asmussen & Rosinski [4] it is explicitly shown that the small jump approximation is valid. So, the small jumps are given by the equations (3) and we turn directly to the large jumps.

Using the the asymptote of the Bessel function K_1 in Abramowits & Stegun [1], the corresponding asymptote for the density of the NIG Lévy measure is,

$$\nu(x) = \frac{\delta\alpha}{\pi|x|}\ \exp\{\beta x\}\ K_1(\alpha|x|) \sim \frac{C_2}{x^2} \qquad x \downarrow 0 \quad \text{or } x \uparrow 0$$
$$\nu(x) = \frac{\delta\alpha}{\pi|x|}\ \exp\{\beta x\}\ K_1(\alpha|x|) \sim \frac{C_1}{|x|^{3/2}}\ \exp\{\beta x - \alpha|x|\} \qquad x \to \pm\infty$$

These asymptotes give a hint of how to sample from $g_\epsilon(x)$. Consider first only $x > \epsilon$ (the case $x < -\epsilon$ is analogous)

In this paper we use an Accept and Reject method with a mixed density $f(x)$. For $x \in [\epsilon, \omega]$ we use a quadratic density $f_q(x)$ and in the tail, $x > \omega$, an exponential density $f_e(x)$ is used.

$$f_q(x) = \left(\frac{1}{\epsilon} - \frac{1}{\omega}\right)^{-1} \frac{1}{x^2} \qquad x \in [\epsilon, \omega]$$
$$f_e(x) = (\alpha - \beta)\ \exp\{-(\alpha - \beta)\ (x - \omega)\} \qquad x > \omega$$

This choice majorises the density g_ϵ for the asymptotes $x \downarrow 0$ and $x \to \infty$.

4 Derivative Prices

To evaluate this simulation algorithm we price some derivatives and see how they differ from the B&S prices. The derivatives chosen are listed below and in both cases there exists a closed form pricing formula from the B&S case.

The Barrier Option. We consider the *up-in* Barrier option (with barrier H) on the European call option (with strike K). The payoff function is,

$$\Phi^B(T) = (S_T - K)\ \mathbf{1}\{S_T > K, M_T^S > H\}$$

where M_t^S is the running maximum of S_s up to time t.

The Russian Option. The Russian option is a contract that gives the holder the right to exercise at any ((a.s) finite) stopping time τ yielding the payoff

$$\Phi^R(\tau) = e^{-\eta\tau} \max\left\{M_0, \sup_{0 \le u \le \tau} S_u\right\}, \quad M_0 \ge S_0, \eta > 0$$

where S_t is the underlying asset, M_0 is the *starting maximum* and η is the *punish factor*. In this paper we use $M_0 = S_0$ and vary η.

Some references for the B&S model are Shepp & Shiryaev [14] & [15] and Duffie & Harrison [11]. For the case of a geometric Lévy process see Asmussen, Avram & Pistorius [5]. The computation involve the *first passage function*,

$$v_k(y) = \mathsf{E}_y \left[\exp\{-\eta\tau_k + R_{\tau_k}\}\right]$$

which is maximized over k and R_t and τ_k from equation (4).

5 Results

We present two studies using this algorithm. The first is an evaluation how important the BM compensation is. Secondly, there is a convergence of the NIG towards the B&S prices. We try to evaluate the speed for different derivatives.

5.1 Data

We use the estimations of the Dresdner Bank stock in the paper by Rydberg [12]. The parameters are,

$$(\alpha, \beta, \mu, \delta) = (68.28, 1.81, 0, 0.01)$$

Under the Esscher transform $\beta = 0.8653$ and to compare with the B&S prices we need $Var(X_1) = 1.466\ 10^{-4}$. The following table shows how the jump intensity $\lambda(\epsilon)$ (which may be identified with the computational effort) varies with the truncation at ϵ.

Table 1. Intensities λ_ϵ for different truncations ϵ

ϵ	10^{-2}	$2 \cdot 10^{-3}$	10^{-3}	$2 \cdot 10^{-4}$	10^{-4}	$2 \cdot 10^{-5}$	10^{-5}
$\lambda(\epsilon)$	0.25636	2.6075	5.74728	31.1660	62.989	317.63	635.94

5.2 Grid Discretization

In this section we show some results indicating how well the grid compensation works. The barrier option depends on the hitting of a barrier. In this case it is only the asset that should hit a barrier. In the Russian option case the reflected process should hit the barrier, which actually involve two different running maximums, i.e. updating M^X to get R and M^R to update τ_k.

Barrier option. Here the time to maturity $T = 10$ and the barrier level $H = 1.055$ is fixed and we consider different strike prices K and grid intensities, that is we vary ϵ. We present 0.95% confidence intervals where 0.9 U means upper bound for $K = 0.9$ for instance. The number of simulations are $2 \cdot 10^5$.

The conclusion we make is that the truncation works well for $\epsilon < 10^{-3}$. Without compensation however, it looks like we need $\epsilon < 2 \cdot 10^{-4}$ that is, about 6 times as much computer time, c.f. the Table 1 above. Smaller ϵ give similar confidence bands as $\epsilon = 10^{-4}$ both with or without grid compensation.

Table 2. Barrier option prices for different ϵ with grid compensation

(ϵ, K)	0.9 L	0.9 U	0.95 L	0.95 U	1 L	1 U	1.05 L	1.05 U
$\epsilon = 10^{-2}$	0.0274	0.0280	0.0189	0.0193	0.0104	0.0106	0.00325	0.00335
$\epsilon = 2 \cdot 10^{-3}$	0.0221	0.0226	0.0152	0.0156	0.00840	0.00860	0.00225	0.00233
$\epsilon = 10^{-3}$	0.0218	0.0223	0.0151	0.0154	0.00829	0.00849	0.00217	0.00226
$\epsilon = 2 \cdot 10^{-4}$	0.0216	0.0221	0.0149	0.0152	0.00822	0.00842	0.00214	0.00222
$\epsilon = 10^{-4}$	0.0217	0.0221	0.0149	0.0153	0.00822	0.00842	0.00215	0.00223

Table 3. Barrier option prices for different ϵ without grid compensation

(ϵ, K)	0.9 L	0.9 U	0.95 L	0.95 U	1 L	1 U	1.05 L	1.05 U
$\epsilon = 10^{-2}$	0.0141	0.0145	0.00991	0.0102	0.00579	0.00599	0.00216	0.00225
$\epsilon = 2 \cdot 10^{-3}$	0.0205	0.0210	0.0142	0.0145	0.00792	0.00812	0.00222	0.00230
$\epsilon = 10^{-3}$	0.0211	0.0216	0.0146	0.0149	0.00808	0.00828	0.00216	0.00225
$\epsilon = 2 \cdot 10^{-4}$	0.0215	0.0220	0.0148	0.0152	0.00818	0.00838	0.00214	0.00222
$\epsilon = 10^{-4}$	0.0216	0.0221	0.0149	0.0152	0.00820	0.0084	0.00215	0.00223

Russian option. Here we fix the punish level $\eta = 3 \cdot 10^{-3}$ and vary ϵ, which implies $\tau_k \sim 10$. Again we present 0.95% confidence intervals for the case with compensation and the next two without compensation.

Again we conclude that the truncation works well for $\epsilon < 10^{-3}$ in the compensated case. Without compensation we need $\epsilon < 2 \cdot 10^{-4}$.

Table 4. Russian option prices for different ϵ

ϵ	Comp-Low	Comp-Up	No Comp-Low	No Comp-Up
$\epsilon = 10^{-2}$	1.0229	1.0242	1.0117	1.0170
$\epsilon = 2 \cdot 10^{-3}$	1.0080	1.0084	1.0059	1.0062
$\epsilon = 10^{-3}$	1.0074	1.0077	1.0064	1.0068
$\epsilon = 2 \cdot 10^{-4}$	1.0074	1.0077	1.0072	1.0076
$\epsilon = 10^{-4}$	1.0075	1.0078	1.0074	1.0078

5.3 Convergence to B&S Prices

Since the variance of the NIG variable is finite, Donsker's Theorem

$$\left\{\frac{X_{tT} - atT}{\sqrt{b^2 T}}\right\}_{t\geq 0} \to B_t \text{ as } T \to \infty$$

suggests that NIG and the B&S prices are not too far if the time horizon is large. In this section we give some examples quantifying this.

Barrier option. The closeness measure used in the Barrier option case is

$$Q(K) = \frac{\Pi^{B\&S}(K)}{\Pi^{NIG}(K)}$$

Since we assume that $S_0 = 1$, K represent the moneyness in the option.

Note that for $T = 2$ is $Q < 1$ for all K and close to 1 at $T = 4$ for K small. For larger T will Q grow (see $T = 30$) and then tend to 1 for all $K \in [0.9, 1.05]$ ($T = 80$) as we expected.

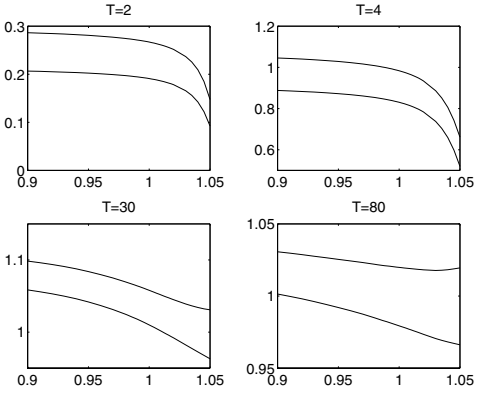


Fig. 1. Confidence intervals for Q against K for different maturities T

Russian option. In the B&S case there is a closed form solution but in the NIG case we present a 95% confidence interval.

The conclusion here is that the relative price quotient above $S_0 = 1$ decreases for smaller punish factor, that is longer time horizon.

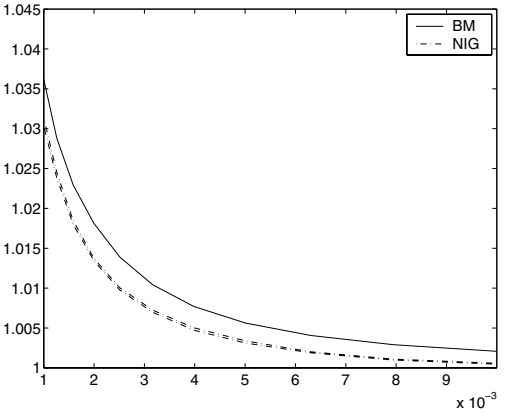


Fig. 2. Price of Russian option against punish factor

References

1. Abramowitz, M. & Stegun, I.A. (editors) (1972). *Handbook of Mathematical Functions.* Dover, New York.
2. Albrecher, H. & Predota, M. (2004). On Asian option pricing for NIG Lévy processes, *Journal of Comp. and Appl. Math.*
3. Asmusen, S., Glynn, P. & Pitman, J. (1995). Discretization error in simulating of one-dimensional reflecting Brownian motion. *Ann. of Appl. Prob.* **5**, 875–896.
4. Asmussen, S. & Rosinski, J. (2001). Approximation of small jumps of Lévy processes with a view towards simulation. *J. Appl. Prob.* **38**, 482–493.
5. Asmussen, S., Avram, F. & Pistorius, M. (2004). Russian and American put options under exponential phase-type Lévy motion, *Stoch. Proc. Appl.* **109**, 79–111.
6. Barndorff-Nielsen, O. (1998). Processes of normal inverse gaussian type. *Finance and Stochastics.* **2**, 41–68.
7. Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press.
8. Boyarchenko, S. & Levendorskiĭ, S. (2002). *Non Gaussian Merton-Black-Scholes Theory*, World Scientific.
9. Chan, T. (1999). Pricing contingent claims with Lévy processes. *Ann. Appl. Prob.* **9**, 502–528.
10. Cherny, A.S. & Shiryaev, A.N. (2001). Change of time and measure for Lévy processes. *MaPhySto Lecture Notes.* **13** (Available on `http://www.maphysto.dk/`).
11. Duffie, D. & Harrison, M. (1993). Arbitrage pricing of Russian options and perpetual Look back options. *Ann. of Appl. Prob.* **3**, 641–651.
12. Rydberg, T.H (1997). The Normal Inverse Gaussian process: simulation and approximation. *Stoch. Models.* **13**, 887–910.
13. Sato, K-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press.
14. Shepp, L. & Shiryaev, A.N. (1993). The Russian option: reduced regret. *Ann. of Appl. Prob.* **3**, 631–640.
15. Shepp, L. & Shiryaev, A.N. (1994). A new look at the "Russian option". *Theory Prob. Appl.* **39**, 103–119.
16. Schoutens, W. (2003). *Lévy Processes in Finance*, Wiley.

Construction of Quasi Optimal Portfolio for Stochastic Models of Financial Market

Aleksander Janicki and Jakub Zwierz

Mathematical Institute, University of Wrocław,
pl. Grunwaldzki 2–4, 50–384 Wrocław, Poland
janicki@math.uni.wroc.pl
http://www.math.uni.wroc.pl/~janicki

Abstract. In the paper we propose a purely computational new method of construction of a quasi–optimal portfolio for stochastic models of a financial market. Here we present the method in the framework of a classical Black–Scholes model of a complete market (see, eg. [4], [6]), considering a well known optimal investment and consumption problem with the HARA type optimization functional. Our method is based on the idea to maximize this functional, taking into account only some subsets of possible portfolio and consumption processes. We show how to reduce the main problem to the construction of a portfolio maximizing a deterministic function of a few real valued parameters but under purely stochastic constraints. It is enough to solve several times an indicated system of stochastic differential equations (SDEs) with properly chosen parametrs.

Results of computer experiments presented here were obtained with the use of the *SDE–Solver* software package. This is our own professional C++ application to Windows system, designed as a scientific computing tool based on Monte Carlo simulations and serving for numerical and statistical construction of solutions to a wide class of systems of SDEs, including a broad class of diffusions with jumps driven by non-Gaussian random measures (consult [1], [2], [5], [8], [9]).

Our method can be easily extended to stochastic models of financial market described by systems of such SDEs.

1 Optimal Investment and Consumption Problem for Black–Scholes Model of a Financial Market

Let us recall that an N dimensional Black–Scholes model of a financial market can be described by the following system of $N+1$ SDEs

$$S_0(t) = S_0(0) + \int_0^t r(s)S_0(s)ds, \tag{1}$$

$$S_n(t) = S_n(0) + \int_0^t \mu_n(s)S_n(s)ds + \sum_{k=1}^{N} \int_0^t \sigma_{n,k}(s)S_n(s)dB_k(s), \tag{2}$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 803–810, 2004.
© Springer-Verlag Berlin Heidelberg 2004

for $n = 1, \ldots, N$ and $t \in (0, T]$, and where we have the money market with a price $S_0(t)$ and N stocks with prices-per-share $S_1(t), \ldots, S_N(t)$, for $t \in [0, T]$.

We assume that processes $r = r(t)$, and $\mu_n = \mu_n(t)$, for $1 \leq n \leq N$, are $\mathbf{L}^1(\Omega \times [0, T])$–integrable, and processes $\sigma_{n,k} = \sigma_{n,k}(t)$ are $\mathbf{L}^2(\Omega \times [0, T])$–integrable on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathrm{P})$ with a filtration $\{\mathcal{F}_t\}$ generated by N–dimensional Brownian motion process $(B_1(t), \ldots, B_N(t))$.

Let the stochastic processes $\eta_0 = \eta_0(t)$ and $\eta_n = \eta_n(t)$, for $n = 1, 2, ..., N$, denote the number of shares of a bond and stocks, respectively. So, the value of the investor's holdings at time t is represented by the *wealth process*, which can be represented by

$$\mathbf{X}(t) \stackrel{\mathrm{df}}{=} \sum_{n=0}^{N} \eta_n(t) S_n(t) = \sum_{n=0}^{N} \pi_n(t), \tag{3}$$

where

$$\pi_n(t) \stackrel{\mathrm{df}}{=} \eta_n(t) S_n(t), \quad n = 0, 1, \ldots, N.$$

Let $\pi(t) = (\pi_1(t), \ldots, \pi_N(t))$. We say that the process

$$(\pi_0, \pi) = \{(\pi_0(t), \pi(t)) : t \in [0, T]\}$$

is the *portfolio process* or simply *portfolio* of an investor.

We define the *gains process* $\{G(t) : t \in [0, T]\}$ as a process which satisfies the following equation,

$$dG(t) = \sum_{n=0}^{N} \eta_n(t) dS_n(t) + \sum_{n=1}^{N} S_n(t) \delta_n(t) dt, \tag{4}$$

where $\delta_n = \delta_n(t)$ is the so called *divident rate payment process* of the nth stock, for all $0 < n \leq N$. By the *consumption process* we understand here any non-negative, regular enough stochastic process $c = \{c(t) : t \in [0, T]\}$. Let $x > 0$ denote the *initial wealth* (or *endowment*, i.e. an amount of money an ivestor has to his disposal at time $t = 0$), what means that we have $X(0) = x$. Let $\Gamma(t) \stackrel{\mathrm{df}}{=} x - \int_0^t c(s) dt$. We say that the portfolio (π_0, π) is *Γ–financed*, when

$$\sum_{n=0}^{N} \pi_n(t) = \Gamma(t) + G(t), \tag{5}$$

with $G(0) = 0, \quad \Gamma(0) = x$.

It is not difficult to check that if conditions (4) and (5) are satisfied, then the wealth process

$$X \equiv X^{x,c,\pi} = \{X^{x,c,\pi}(t) : t \in [0, T]\}, \tag{6}$$

defined by (3), can be obtained as a solution to the following Itô SDE

$$\begin{aligned} dX(t) = (rX(t) - c(t))dt + \sum_{n=1}^{N} (\mu_n(t) + \delta_n(t) - r)\pi_n(t) dt + \\ + \sum_{n=1}^{N} \pi_n(t) \sum_{m=1}^{N} \sigma_{n,m} B_m(t), \end{aligned} \tag{7}$$

with an initial condition of the form $\mathbf{X}(0) = x = \sum_{n=0}^{N} \eta_n(0) S_n(0)$.

From (4) and (5), after application of the classical Itô formula, it follows that the following – very important in our approach – equation must be satisfied

$$\sum_{n=0}^{N} S_n(t)d\eta_n(t) = \sum_{n=1}^{N} S_n(t)\delta_n(t)dt - c(t)dt. \tag{8}$$

In optimization problems utility functions can be chosen in many different ways, however the typical choice for scientific investigations is the HARA model, what means that we chose *utility function* given by

$$U^{(p)}(x) \stackrel{\text{df}}{=} x^p/p, \quad x > 0, \tag{9}$$

for $p \in (-\infty, 0)$ or $p \in (0, 1)$.

The *risk aversion coefficient* is then defined by the formula

$$R \stackrel{\text{df}}{=} -x\frac{d^2}{dx^2}U^{(p)}(x)\Big/\frac{d}{dx}U^{(p)}(x) = 1 - p. \tag{10}$$

Here we are interested in the following optimization problem.

For a given utility function $U^{(p)} = U^{(p)}(c)$ and initial wealth $x > 0$, we look for an *optimal portfolio* $\widehat{\pi}(t) = \{\widehat{\pi}_1(t), \ldots, \widehat{\pi}_N(t)\}$ and an *optimal consumption process* $\widehat{c} = \widehat{c}(t)$, such that for the *value function* of the form

$$V_{c,\pi}(x) \stackrel{\text{df}}{=} \mathrm{E}\left[\int_0^T U^{(p)}(c(t))\mathrm{e}^{-\int_0^t \beta(s)ds}dt\right] \tag{11}$$

the following condition is satisfied

$$V_{\hat{c},\hat{\pi}}(x) = \sup_{(c,\pi)\in\mathcal{A}(x)} V_{c,\pi}(x). \tag{12}$$

Here the condition $(c, \pi) \in \mathcal{A}(x)$ means, that the processes $c = c(t)$ and $\pi = \pi(t)$ are subject to the stochastic constraints, what means that $c = c(t)$ is positive on $[0, T]$ and the corresponding wealth process satisfing SDE (7) is such that

$$X^{x,c,\pi}(t) \geq 0 \ \text{ a.s.} \qquad \text{for } \ t \in [0, T]. \tag{13}$$

2 An Example of a Quasi–optimal Portfolio

An attempt to answer the question how to apply computational methods to solve directly and effectively optimizations problem (12) through analytical methods, e.g. making use of the Malliavin calculus (see [3], [7]) or trying to get a hedging strategy by constructing a relevant replicating portfolio (see eg. [4], [6], etc.) is not an obvious task.

So, our aim is to describe a method of computer construction of a quasi–optimal portfolio solving approximate problem related to (12).

The method is based on the idea to maximize functional (11), taking into account only some subsets of possible portfolio processes derived from equations

(7) and (8), and chosing the class of admissible positive consumption processes arbitrally, in a convenient reasonable way.

We show how to reduce the main problem to the construction of a portfolio maximizing a deterministic function of a few real valued parameters but under purely stochastic constraints.

In order to make the further exposition easier, we restrict ourselves to the classical one dimensional ($N = 1$) Black–Scholes model, which can be described in the following form

$$S_0(t) = S_0(0) + r \int_0^t S_0(s)ds \tag{14}$$

$$S_1(t) = S_1(0) + \mu \int_0^t S_1(s)ds + \sigma \int_0^t S_1(s)dB(s), \tag{15}$$

for $t \in [0, T]$, and initial conditions such that $S_0(0) > 0$, $S_1(0) > 0$.

In the model (14)–(15) all paremeters, i.e. $S_0(0)$, $S_1(0)$, r, μ, σ, are given positive real numbers. So, the processes S_0, S_1 can be described in the explicit closed form

$$S_0(t) = S_0(0)\ \mathrm{e}^{rt}, \tag{16}$$

$$S_1(t) = S_1(0)\ \mathrm{e}^{\{(\mu-\sigma^2/2)t+\sigma B(t)\}}. \tag{17}$$

The portfolio is now given by $\pi = (\pi_0, \pi_1)$, where

$$\pi_0(t) = \eta_0(t)S_0(t), \quad \pi_1(t) = \eta_1(t)S_1(t), \quad t \in [0, T]. \tag{18}$$

In the example we have chosen for computer experiments presented here we reduced the class of admissible portolio process to those which are of the following form

$$\eta_1(t) = p_1 S_1(t), \quad t \in [0, T]. \tag{19}$$

We also restric ourselves to the class of consumption processes defined by

$$c(t) = c_0 S_0(t) + c_1 S_1(t), \quad t \in [0, T], \tag{20}$$

In (19) and (20) parameters p_1, c_0, c_1 are deterministic (real) variables, subject to some stochastic constraints, and which should fixed in an optimal way.

It is not difficult to notice that in such circumstances the wealth process $X(t) = X^{c_0,c_1,p_1}(t)$, defined by (7), solves the following Itô SDE

$$\begin{aligned} dX(t) = \Big(rX(t) + (\mu_1 + \delta_1 - r)\eta_1(t)S_1(t) - \big(c_0 S_0(t) + c_1 S_1(t)\big)\Big)dt \\ +\sigma_1 \eta_1(t)S_1(t)dB(t), \quad t \in (0, T], \quad X(0) = x. \end{aligned} \tag{21}$$

Making use of the equation (8), it is also possible to check that the first component of the portfolio, i.e. the proces $\eta_0 = \eta_0(t)$ solves the following SDE

$$\begin{aligned} d\eta_0(t) = \Big((-\mu_1 + \delta_1)p_1 S_1^2(t) - \big(c_0 S_0(t) + c_1 S_1(t)\big)\Big)/S_0(t)dt - \\ - p_1\sigma_1 S_1^2(t)/S_0(t)dB(t),\ t \in (0, T],\ \eta(0) = (x - p_1 S_1^2(0))/S_0(0). \end{aligned} \tag{22}$$

In this way we arrive at the following problem.

For a given utility function $U^{(p)} = U^{(p)}(c)$ and initial wealth $x > 0$, we look for optimal values of parameters $\widehat{c_0}$, $\widehat{c_1}$, $\widehat{p_1}$, such that for the *value function* of the form

$$V_{c_0,c_1,p_1}(x) \stackrel{\text{df}}{=} \mathrm{E}\left[\int_0^T U^{(p)}(c_0 S_0(t) + c_1 S_1(t))\mathrm{e}^{-\beta t} dt\right] \tag{23}$$

the following condition is satisfied

$$V_{\widehat{c_0},\widehat{c_1}\widehat{p_1}}(x) = \sup_{(c_0,c_1,p_1)\in\mathcal{A}(x)} V_{c_0,c_1,p_1}(x). \tag{24}$$

Now the condition $(c_0, c_1, p_1) \in \mathcal{A}(x)$ means, that the consumption and wealth processes, defined by (20) and (21), are such that we have

$$X^{c_0,c_1,p_1}(t) \geq 0 \text{ a.s.}, \quad c_0 S_0(t) + c_1 S_1(t) \geq 0 \text{ a.s.} \quad \text{for} \quad t \in [0, T]. \tag{25}$$

We see that, having to our disposal stochastic processes solving SDEs (14), (15), and (21), we are able to solve the problem (23)–(25). In order to get values of the value function (23) using the *SDE–Solver* software it is enough to solve the system of two equations

$$dY(t) = \left(c_0 r S_0(t) + c_1 \mu_1 S_1(t)\right)\, dt + c_1 \sigma_1 S_1(t)\, dB(t), \tag{26}$$

$$dZ(t) = U^{(p)}(Y(t))\mathrm{e}^{\beta t} dt, \tag{27}$$

for $t \in (0, T]$, and with initial conditions $Y(0) = c_0 S_0(0) + c_1 S_1(0)$, $Y(0) = 0$, and finally to compute

$$V_{c_0,c_1,p_1}(x) = \mathrm{E}\ Z(T). \tag{28}$$

Then, making use of formulae (22), (19), (18), and (20), one can easily construct quasi optimal portfolio and quasi optimal consumption processes.

3 Results of Computer Experiments

We solved the optimization problem desribed by formulae (14)–(25), with the following fixed values of constant parameters: $T = 1$, $r = 0.2$, $\mu_1 = 0.15$, $\delta_1 = 0.05$, $\sigma_1 = 0.35$, $\beta = 0$ and also $\beta \in \{0, 0.1, 0.2, 0.4, 0.8\}$, $x = 50$, $S_{00} = 50$, $S_{10} = 50$.

The optimal solution for $\beta = 0.0$ is of the following form:

$$\widehat{c_0} = 1.0, \quad \widehat{c_1} = -0.01, \quad \widehat{p_1} = 0.0, \quad V_{\widehat{c_0},\widehat{c_1},\widehat{cp_1}} = 14.8.$$

From a large amount of data obtained in our experiments we present here the optimal solution for $\beta = 0$, including Fig. 1 presenting the convex subset of admissible parameters (c_0, c_1, p_1) in R^3, which can be of independent interest and the Fig. 2 containing the values of the value function V_{c_0,c_1,p_1}.

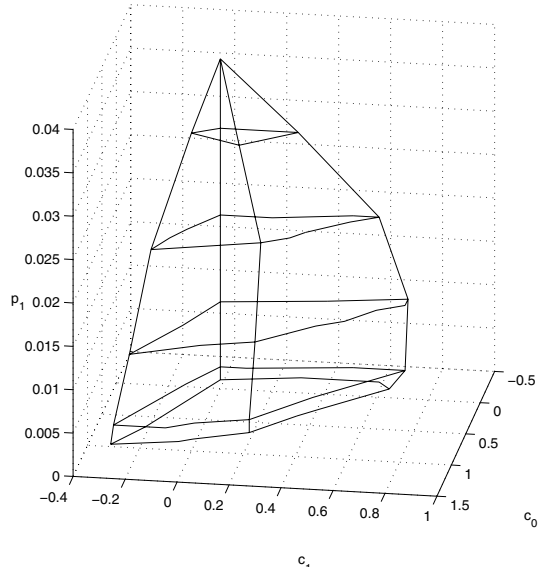


Fig. 1. The convex set in R^3 containing all admissible values of parameters V_{c_0,c_1,p_1}

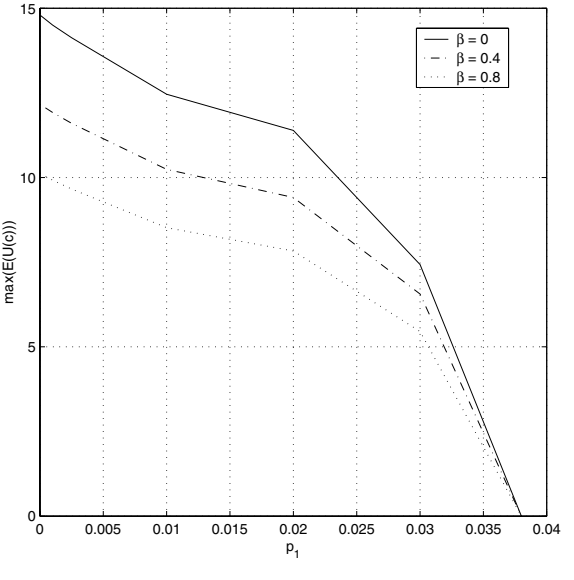


Fig. 2. Dependence of function V_{c_0,c_1,p_1} on parameter p_1 for 3 values of β

In Table 1 below some values of function V_{c_0,c_1,p_1}, indicated in Fig. 2 for 3 different values of β, are presented with corresponding values of parameters c_0, c_1, p_1.

Table 1. Values of V_{c_0,c_1,p_1} from (23)

$\mathbf{V}_{c_0,c_1,p_1}(x)$	$\mathbf{c_0}$	$\mathbf{c_1}$	$\mathbf{p_1}$	β
14.8	**1.00**	**−0.01**	**0.00**	0.00
11.5	0.30	0.30	0.00	0.00
12.5	0.80	−0.10	0.01	0.00
11.5	0.30	0.30	0.01	0.00
11.3	0.30	0.30	0.02	0.00
13.3	1.00	−0.01	0.00	0.20
11.7	0.38	0.38	0.01	0.20
11.3	0.80	−0.10	0.01	0.20
10.3	0.30	0.30	0.02	0.20
12.1	1.00	−0.01	0.00	0.40
10.6	0.80	−0.05	0.01	0.40
9.4	0.30	0.30	0.02	0.40
10.1	1.00	−0.01	0.00	0.80
7.8	0.30	0.30	0.02	0.80

In all runs of the system of equations (14), (15), (21), (26), (27) leading to computation of values of the expression in (28) with the *SDE–Solver*, we had 1000 trajectories of the solution, which were constructed on the grid given by 1000 subintervals of lenght 0.001 of the interval $[0, 1]$. Numerical and statistical approximation methods involved are described in [1], [2], [5], [9].

Another completely solved example (simplier one, with value function depending only on two parameters) of a quasi optimal portfolio and quasi optimal consumption processes can be found in [1]. Instead of (19), (20), and (23), the following conditions describe the optimization problem:

$$\eta_0(t) = p_0 S_0(t), \quad c(t) = c_2 X(t),$$

$$V_{c_2,p_0}(x) \stackrel{\text{df}}{=} \mathrm{E}\left[\int_0^T U^{(p)}(c_2 X(t))\mathrm{e}^{-\beta t} dt\right].$$

Graphical representations, visualizing trajectories and some statistical properties of quasi optimal processes $\eta_0 = \eta_0(t)$, $\eta_1 = \eta_1(t)$, $X = X(t)$, and $c = c(t)$, are included there.

4 Conclusions

We strongly insist that even such rough approximations of the optimal investment and consumption problem as presented here are of important practical interest. One can get quite useful ideas about properties of stochastic processes solving the problem, how they depend on parameters of the stochastic model of financial market, investor preferences, etc. Of course, much more work on improvement of suggested method of construction of quasi optimal portfolio has to be done. It is also quite obvious that the method can be easily extended onto much more sofisticated stochastic models of financial market. There are various questions of mathematical nature, that should be answered in the future, e.g. on the correctness and convergence of the proposed approximate method, when more and more parameters enlargig properly the sets of admissible portfolio and consumption processes, are included.

Our computer experiments indicate to some extent the direction of further development of computational methods and computer software useful in practical solution of such complicated problems as construction of optimal strategies for investors, when stochastic model of the financial market is investigated in the framework of a system of SDEs of Itô or another, more general, type. For example, our approach can be in a very simple and natural way implemented for parallel computing systems.

References

1. Janicki, A., Izydorczyk, A.: Computer Methods in Stochastic Modeling (in Polish). Wydawnictwa Naukowo-Techniczne, Warszawa, (2001)
2. Janicki, A., Izydorczyk, A., Grądalski, P.: Computer Simulation of Stochastic Models with SDE–Solver Software Package. Proceedings of the 3ICCS Sankt Petersburg, Lecture Notes in Computer Science 2657 (2003) 361–370
3. Janicki, A., Krajna, L.: Malliavin Calculus in Construction of Hedging Portfolios for the Heston Model of a Financial Market. Demonstratio Mathematica XXXIV (2001) 483–495
4. Karatzas I., Shreve, S.E.: Methods of Mathematical Finance. Springer-Verlag, Berlin, (1998)
5. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations, 3rd ed. Springer-Verlag, New York, (1998)
6. Musiela, M., Rutkowski, M.: *Martingale Methods in Financial Modelling: Theory and Applications*. Springer-Verlag, New York, (1997)
7. Ocone, D.L., Karatzas, I.: *A generalized Clark representation formula, with application to optimal portfolios*. Stochastics and Stochastics Reports 34 (1991), 187–220
8. Protter, P.: Stochastic Integration and Differential Equations – A New Approach. Springer-Verlag, New York, (1990)
9. Rachev, S. T.: Handbook of Numerical Methods in Finance. Springer-Verlag, Berlin, New York, (2004)

Euler Scheme for One-Dimensional SDEs with Time Dependent Reflecting Barriers

Leszek Słomiński[1]* and Tomasz Wojciechowski[2]

[1] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, ul. Chopina 12/18, 87-100 Toruń, Poland
leszeks@mat.uni.torun.pl

[2] Institute of Mathematics and Physics, University of Technology and Agriculture in Bydgoszcz, ul. Al. Prof. S. Kaliskiego 7, 85-796 Bydgoszcz, Poland

Abstract. We give the rate of mean-square convergence for the Euler scheme for one-dimensional stochastic differential equations with time dependent reflecting barriers. Applications to stock prices models with natural boundaries of Bollinger bands type are considered.

1 Introduction

We consider a market in which fluctuation of stocks prices, and more generally of some economic goods, is given by a stochastic process $S = \{S_t; t \in \mathbb{R}^+\}$ living within the upper- and lower barrier processes $F = \{F_t; t \in \mathbb{R}^+\}$ and $G = \{G_t; t \in \mathbb{R}^+\}$, i.e. $G_t \leq S_t \leq F_t$, $t \in \mathbb{R}^+$. Such models appear for instance if some institutions may want to prevent prices from leaving interval $[G_t, F_t]$ and prices may have some natural boundaries. Recently, in [8] the simplest case of constant boundaries of the form $[l, d]$ was considered. In this case an option pricing formula was obtained under the assumption that S is a solution of an appropriate stochastic differential equation (SDE). Models of prices fluctuation considered in practice by quantitive analysts are much more general: barriers are stochastic processes depending on the process S. Typical examples of such natural boundaries are the so-called Bollinger bands F, G defined by

$$F_t = A_t + \alpha(\frac{1}{M}\sum_{j=1}^{M}(S_{t-\varepsilon j} - A_t)^2)^{1/2}, \quad G_t = A_t - \alpha(\frac{1}{M}\sum_{j=1}^{M}(S_{t-\varepsilon j} - A_t)^2)^{1/2}$$

and trading bands (envelopes) defined by

$$F_t = (1+\alpha)A_t, \qquad G_t = (1-\alpha)A_t,$$

where A is a moving average process $A_t = \frac{1}{M}\sum_{j=1}^{M} S_{t-\varepsilon j}$, $t \in \mathbb{R}^+$, and $\varepsilon, \alpha > 0$, $M \in \mathbb{N}$ are some parameters.

In [11] existence and uniqueness of solutions of SDE with time dependent reflecting barriers driven by a general semimartingale is proved. In the present

* Research supported by Komitet Badań Naukowych under grant 1 P03A 022 26

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 811–818, 2004.
© Springer-Verlag Berlin Heidelberg 2004

paper we restrict ourselves to a one-dimensional SDE with reflecting barriers of the form

$$X_t = X_0 + \int_0^t \sigma(X_s)\,dW_s + \int_0^t b(X_s)\,ds + K_t, \quad t \in \mathbb{R}^+, \tag{1}$$

where $X_0 \in \mathbb{R}$, $W = \{W_t; t \in \mathbb{R}^+\}$ is a standard Wiener process, $\sigma, b : \mathbb{R} \to \mathbb{R}$ are Lipschitz continuous functions and barrier processes F, G are general Lipschitz operators with delayed argument depending possibly on X (for a precise definition see Section 2). Our aim is to define the Euler scheme $\{\bar{X}^n\}$ for the SDE (1) and to give its rate of mean-square convergence.

The main result of the paper says that under mild assumptions on reflecting barrier processes $G(^{\varepsilon}X), F(^{\varepsilon}X)$ with delayed argument for every $q \in \mathbb{R}^+$ there exists $C > 0$ such that

$$E \sup_{t \le q} |\bar{X}_t^n - X_t|^2 \le C\left(\frac{\ln n}{n} + E\omega_{1/n}^2(G(^{\varepsilon}X), q) + E\omega_{1/n}^2(F(^{\varepsilon}X), q)\right), \tag{2}$$

where $\omega_\delta(x, q) = \sup\{|x_t - x_s|,\ 0 \le s < t \le q,\ t - s < \delta\}$, for all $\delta > 0$, $q \in \mathbb{R}^+$ and $x \in \mathbb{D}(\mathbb{R}^+, \mathbb{R})$ ($\mathbb{D}(\mathbb{R}^+, \mathbb{R})$ is the space of all mappings $x : \mathbb{R}^+ \to \mathbb{R}$ which are right continuous and admit left-hand limits).

From (2) we deduce that in both cases of Bollinger and trading bands for every $q \in \mathbb{R}^+$, $\delta > 0$ there exists $C > 0$ such that

$$E \sup_{t \le q} |\bar{X}_t^n - X_t|^2 \le C\left(\frac{1}{n^{1-\delta}}\right).$$

Moreover, in both cases,

$$E \sup_{t \le q} |\bar{X}_t^n - X_t|^2 \le C\left(\frac{\ln n}{n}\right)$$

if σ, b are bounded.

Note that if $G = -\infty$, $F = +\infty$ then $\{\bar{X}^n\}$ is the classical Euler scheme introduced in [6]. In the case $G = 0$, $F = +\infty$ and $G = l$, $F = d$ the rate of mean-square convergence was examined earlier by many authors (see, e.g., [2,4, 5,7,9,10]).

In the paper no attemps has been made to obtain option pricing formulas for markets with dynamics of prices given by (1). This question deserves an independent study.

2 SDEs with Time Dependent Reflecting Barriers

We begin with a definition of the Skorokhod problem with time dependent reflecting barriers.

Definition 1. Let $y, f, g \in \mathbb{D}(\mathbb{R}^+, \mathbb{R})$ with $g \le f$ and $g_0 \le y_0 \le f_0$. We say that a pair $(x, k) \in \mathbb{D}(\mathbb{R}^+, \mathbb{R}^2)$ is a solution of the Skorokhod problem associated with y and barriers f, g (and write $(x, k) = SP(y, f, g)$) if

(i) $x_t = y_t + k_t$, $t \in \mathbb{R}^+$,
(ii) $g_t \le x_t \le f_t$, $t \in \mathbb{R}^+$.
(iii) $k_t = k_t^{(-)} - k_t^{(+)}$, $t \in \mathbb{R}^+$, where $k^{(-)}$, $k^{(+)}$ are nondecreasing, right continuous functions with $k_0 = k_0^{(-)} = k_0^{(+)} = 0$ such that $k^{(-)}$ increases only on $\{t; x_t = g_t\}$ and $k^{(+)}$ increases only on $\{t; x_t = f_t\}$.

Theorem 1. *([11]) Assume that $f, g \in \mathbb{D}(\mathbb{R}^+, \mathbb{R})$ satisfy the condition*

$$\inf_{t \le q}(f_t - g_t) > 0, \quad q \in \mathbb{R}^+.$$

Then for every $y \in \mathbb{D}(\mathbb{R}^+, \mathbb{R})$ with $g_0 \le y_0 \le f_0$, there exists a unique solution (x, k) of the Skorokhod problem associated with y and barriers f, g.

□

Note that in the case of continuous function y similar definitions of the Skorokhod problem were earlier given in [3] and [1]. These papers contain also results on existence and uniqueness of solutions of the Skorokhod problem in the case of continuous y and continuous barriers f, g (see e.g. [3, Lemma 4.1]).

The following theorem, where Lipschitz continuity of solutions of the Skorokhod problem is stated will prove to be very useful in Section 3.

Theorem 2. *([11]) Assume that $y^i, f^i, g^i \in \mathbb{D}(\mathbb{R}^+, \mathbb{R})$, $g_0^i \le y_0^i \le f_0^i$ and*

$$\inf_{t \le q}(f_t^i - g_t^i) > 0, \quad q \in \mathbb{R}^+$$

for $i = 1, 2$. Let $(x^i, k^i) = SP(y^i, f^i, g^i)$, $i = 1, 2$. Then for every $q \in \mathbb{R}^+$

$$\sup_{t \le q} |x_t^1 - x_t^2| \le 3 \sup_{t \le q} |y_t^1 - y_t^2| + \sup_{t \le q} |f_t^1 - f_t^2| + \sup_{t \le q} |g_t^1 - g_t^2|$$

and

$$\sup_{t \le q} |k_t^1 - k_t^2| \le 2 \sup_{t \le q} |y_t^1 - y_t^2| + \sup_{t \le q} |f_t^1 - f_t^2| + \sup_{t \le q} |g_t^1 - g_t^2|.$$

□

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ be a filtered probability space.

Definition 2. Let $\mathcal{D}$ denote the space of all $(\mathcal{F}_t)$ adapted processes with trajectories in $\mathbb{D}(\mathbb{R}^+, \mathbb{R})$. We say that an operator $H : \mathcal{D} \to \mathcal{D}$ is Lipschitz if

(i) $H(X) \in \mathcal{D}$ for any $X \in \mathcal{D}$,
(ii) for any $X, Y \in \mathcal{D}$ and any stopping time τ,
$X^{\tau-} = Y^{\tau-}$ implies $H(X)^{\tau-} = H(Y)^{\tau-}$,
(iii) there exists $L > 0$ such that for any $X, Y \in \mathcal{D}$

$$|H(X)_t - H(Y)_t| \le L \sup_{s \le t} |X_t - Y_t|, \quad t \in \mathbb{R}^+.$$

Given $X \in \mathcal{D}$ and $\varepsilon > 0$ set ${}^{\varepsilon}X = \{X_{t-\varepsilon}; t \in \mathbb{R}^+\}$ (with the convention that $X_t = X_0$ for $t \in [-\varepsilon, 0)$). In what follows barriers of the form $F({}^{\varepsilon}X), G({}^{\varepsilon}X)$, where F, G are Lipschitz operators we will call barrier operators with delayed arguments.

Fix $\varepsilon > 0$. Let W be an $(\mathcal{F}_t)$ adapted Wiener process and let F, G be two Lipschitz operators such that for any $X \in \mathcal{D}$

$$\inf_{t \le q}(F({}^{\varepsilon}X)_t - G({}^{\varepsilon}X)_t) > 0, \quad q \in \mathbb{R}^+. \tag{3}$$

We will say that a pair (X, K) of $(\mathcal{F}_t)$ adapted processes is a strong solution of the SDE (1) with barrier operators $F({}^{\varepsilon}X), G({}^{\varepsilon}X)$ with delayed argument if $(X, K) = SP(Y, F({}^{\varepsilon}X), G({}^{\varepsilon}X))$, where

$$Y_t = X_0 + \int_0^t \sigma(X_s)\, dW_s + \int_0^t b(X_s)\, ds, \quad t \in \mathbb{R}^+. \tag{4}$$

Theorem 3. *([11]) Let $\varepsilon > 0$. Assume that σ, b are Lipschitz continuous functions and F, G are Lipschitz operators satisfying (3) with $G(X_0)_0 \le X_0 \le F(X_0)_0$. Then there exists a unique strong solution (X, K) of the SDE (1).*

□

Theorem 4. *Let $\varepsilon > 0$ and let σ, b, F, G, X_0 satisfy the assumptions of Theorem 3. If*

$$E \sup_{t \le q}(|F(X_0)_t| + |G(X_0)_t|)^{2p} < +\infty, \quad q \in \mathbb{R}^+,\ p \in \mathbb{N}$$

then

(i) $E \sup_{t \le q} |X_t - X_0|^{2p} < +\infty$, $p \in \mathbb{N}$, $q \in \mathbb{R}^+$,

(ii) *if moreover, $F({}^{\varepsilon}X)_s = F({}^{\varepsilon}X^s)_t$ and $G({}^{\varepsilon}X)_s = G({}^{\varepsilon}X^s)_t$ for any $0 \le s \le t$, then for every $p \in \mathbb{N}$, $q \in \mathbb{R}^+$ there exists $C > 0$ such that*

$$E|X_t - X_s|^{2p} \le C(t - s)^p$$

for $s \le t \le q$.

□

Corollary 1. *Under assumptions of Theorems 3 and 4 for any $q \in \mathbb{R}^+$, $\delta > 0$ there exists $C > 0$ such that*

$$E(\omega_{1/n}(F({}^{\varepsilon}X), q))^2 + E(\omega_{1/n}(G({}^{\varepsilon}X), q))^2 \le C(\frac{1}{n^{1-\delta}}). \tag{5}$$

If moreover, σ, b are bounded then

$$E(\omega_{1/n}(F({}^{\varepsilon}X), q))^2 + E(\omega_{1/n}(G({}^{\varepsilon}X), q))^2 \le C(\frac{\ln n}{n}). \tag{6}$$

□

3 Euler Scheme for SDEs with Time Dependent Reflecting Barrier Operators

Let $(\mathcal{F}_t^n)$ denote the discretization of $(\mathcal{F}_t)$, i.e. $\mathcal{F}_t^n = \mathcal{F}_{k/n}$ for $t \in [k/n, (k+1)/n)$ and let F^n, G^n denote the discretizations of operators F, G, i.e. any process X, $F^n(X)_t = F(X)_{k/n}$, $G^n(X)_t = G(X)_{k/n}$ for any $t \in [k/n, (k+1)/n)$.

The Euler scheme for the SDE (1) is given by the following recurrent formula

$$\begin{cases} \bar{X}_0^n & = X_0, \\ \bar{X}_{(k+1)/n}^n & = \max\big[\min\big(\bar{X}_{k/n}^n + \sigma(\bar{X}_{k/n}^n)(W_{(k+1)/n} - W_{k/n}) \\ & \quad + b(\bar{X}_{k/n}^n)\tfrac{1}{n}, F(^{\varepsilon}\bar{X}^{n,k/n})_{(k+1)/n}\big), G(^{\varepsilon}\bar{X}^{n,k/n})_{(k+1)/n}\big], \\ \bar{X}_t^n & = \bar{X}_{k/n}^n, \quad t \in [k/n, (k+1)/n), \end{cases}$$

where ${}^{\varepsilon}\bar{X}_t^{n,k/n} = {}^{\varepsilon}\bar{X}_t^n$ if $t \le k/n$ and ${}^{\varepsilon}\bar{X}_t^{n,k/n} = {}^{\varepsilon}\bar{X}_{k/n}^n$ if $t > k/n$. Set

$$\bar{Y}_t^n = X_0 + \int_0^t \sigma(\bar{X}_{s-}^n)\, dW_s^n + \int_0^t b(\bar{X}_{s-}^n)\, d\rho_s^n, \quad t \in \mathbb{R}^+,$$

where $\rho_t^n = k/n$ for $t \in [k/n, (k+1)/n)$ and W^n is a disretization of Wiener process W, that is $W_t^n = W_{\rho_t^n}$, $t \in \mathbb{R}^+$. Note that $(\bar{X}^n, \bar{K}^n = \bar{X}^n - \bar{Y}^n)$ is a pair of $(\mathcal{F}_t^n)$ adapted processes such that $(\bar{X}^n, \bar{K}^n) = SP(\bar{Y}^n, F^n(^{\varepsilon}\bar{X}^n), G^n(^{\varepsilon}\bar{X}^n))$.

Theorem 5. *Assume that σ, b are Lipschitz continuous functions and F, G are Lipschitz operators satysfying (3) such that $G(X_0)_0 \le X_0 \le F(X_0)_0$*

$$E \sup_{t \le q} (|F(X_0)_t| + |G(X_0)_t|)^{2+\delta} < +\infty, \quad q \in \mathbb{R}^+,$$

for some $\delta > 0$. If (X, K) is a solution of the SDE (1) then for every $q \in \mathbb{R}^+$ there exists $C > 0$ such that (2) holds true.

Proof. We begin by proving that

$$\sup_n E \sup_{t \le q} |\bar{X}_t^n - X_0|^{2+\delta} < \infty. \tag{7}$$

Without loss of generality we may and will assume that

$$\sup_n E \sup_{t \le q-\varepsilon} |\bar{X}_t^n - X_0|^{2+\delta} < \infty.$$

Since $(X_0, 0) = SP(X_0, F^n(X_0)_0, G^n(X_0)_0)$, it follows from Theorem 2 that

$$\begin{aligned} \sup_{t \le q} |\bar{X}_t^n - X_0| \le 3 \sup_{t \le q} \left| \int_0^t \sigma(\bar{X}_{s-}^n)\, dW_s^n + \int_0^t b(\bar{X}_{s-}^n) d\rho_s^n \right| \\ + \sup_{t \le q} (|F^n(^{\varepsilon}\bar{X}^n)_t - F^n(X_0)_t| + |G^n(^{\varepsilon}\bar{X}^n)_t - G^n(X_0)_t|) + H_q, \end{aligned}$$

where $H_q = \sup_{t \le q}(|G^n(X_0)_t - G^n(X_0)_0| + |F^n(X_0)_t - F^n(X_0)_0|)$. Clearly

$$E\sup_{t\le q}(|F^n({}^\varepsilon\bar X^n)_t - F^n(X_0)_t| + |G^n({}^\varepsilon\bar X^n)_t - G^n(X_0)_t|)^{2+\delta} \le CE\sup_{t\le q-\varepsilon}|\bar X^n_t - X_t|^{2+\delta} < +\infty,$$

and $EH_q^{2+\delta} < \infty$. Therefore, by the Burkholder–Davis–Gundy and Hölder inequalities we have

$$E\sup_{t\le q}|\bar X^n_t - X_0|^{2+\delta} \le C\{\int_0^q E\sup_{u\le s}|\bar X^n_{u-} - X_0|^{2+\delta}\,ds + 1\},$$

and hence (7). Since σ, b are Lipschitz continuous (7) yields

$$\sup_n E\sup_{t\le q}|\sigma(\bar X^n)_t|^{2+\delta} < \infty, \quad \sup_n E\sup_{t\le T}|b(\bar X^n)_t|^{2+\delta} < \infty. \tag{8}$$

Set

$$\widehat Y^n_t = X_0 + \int_0^t \sigma(\bar X^n_{s-})\,dW_s + \int_0^t b(\bar X^n_{s-})\,ds, \quad t\in\mathbb{R}^+,$$

and $(\hat X^n, \hat K^n) = SP(\hat Y^n, F^n({}^\varepsilon\bar X^n), G^n({}^\varepsilon\bar X^n))$. Then

$$\hat Y^n_t - \bar Y^n_t = \sigma(\bar X^n_{k/n})(W_t - W_{k/n}) + b(\bar X^n_{k/n})(t - k/n) \tag{9}$$

for $t\in[k/n,(k+1)/n)$. Therefore, by Theorem 2,(8),(9) and [9, Lemma A4]

$$\begin{aligned} E\sup_{t\le q}|\bar X^n_t - \hat X^n_t|^2 &\le 9E\sup_{t\le q}|\bar Y^n_t - \hat Y^n_t|^2 \\ &\le 9\big\{(E\sup_{t\le q}|\sigma(\bar X^n_{t-})|^{2+\delta})^{2/(2+\delta)}(E(\omega_{1/n}(W,q))^{(2\delta+4)/\delta})^{\delta/(\delta+2)} \\ &\quad +(\frac{1}{n})^2 E\sup_{t\le q}|b(\bar X^n_{t-})|^2\big\} \\ &\le C\big(\frac{\ln n}{n} + (\frac{1}{n})^2\big) \le C\big(\frac{\ln n}{n}\big). \end{aligned} \tag{10}$$

Clearly

$$\begin{aligned} E\sup_{t\le q}|\hat X^n_t - X_t|^2 &\le C\big\{E\sup_{t\le q}|\hat Y^n_t - Y_t|^2 \\ &\quad +E(\sup_{t\le q}|F^n({}^\varepsilon\bar X^n)_t - F^n({}^\varepsilon X)_t| + |G^n({}^\varepsilon\bar X^n)_t - G^n({}^\varepsilon X)_t|)^2 \\ &\quad +E(\sup_{t\le q}|F^n({}^\varepsilon X)_t - F({}^\varepsilon X)_t| + |G^n({}^\varepsilon X)_t - G({}^\varepsilon X)_t|)^2\big\} \\ &\le C\big\{\int_0^q E\sup_{u\le s}|\bar X^n_u - X_u|^2\,ds + E\sup_{t\le q-\varepsilon}|\bar X^n_t - X_t|^2 + \delta_n\big\}, \end{aligned} \tag{11}$$

where $\delta_n = E\omega^2_{1/n}(F({}^\varepsilon X), q) + E\omega^2_{1/n}(G({}^\varepsilon X), q)$. Since without loss of generality we may and will assume

$$E\sup_{t\le q-\varepsilon}|\bar X^n_t - X_t|^2 \le C\bigg(\frac{\ln n}{n} + E\omega^2_{1/n}(G({}^\varepsilon X), q-\varepsilon) + E\omega^2_{1/n}(F({}^\varepsilon X), q-\varepsilon)\bigg),$$

from (10) and (11) we obtain

$$E\sup_{t\le q}|\bar{X}_t^n - X_t|^2 \le 2E\sup_{t\le q}|\bar{X}_t^n - \hat{X}_t^n|^2 + 2E\sup_{t\le q}|\hat{X}_t^n - X_t|^2$$
$$\le C\left\{\frac{\ln n}{n} + \int_0^t E\sup_{u\le s}|\bar{X}_u^n - X_u|^2 du + \delta_n\right\}.$$

To complete the proof it suffices now to use Gronwall's lemma. □

Corollary 2. *Assume that σ, b are Lipschitz continuous functions and F, G are $(\mathcal{F}_t)$ adapted processes such that $G_0 \le X_0 \le F_0$, $\inf_{t\le q}(F_t - G_t) > 0$, $q \in \mathbb{R}^+$ and*

$$E\sup_{t\le q}(|F_t| + |G_t|)^{2+\delta} < +\infty, \quad q \in \mathbb{R}^+,$$

for some $\delta > 0$. If (X, K) is a solution of the SDE (1) then for every $q \in \mathbb{R}^+$ there exists $C > 0$ such that

$$E\sup_{t\le q}|\bar{X}_t^n - X_t|^2 \le C\left(\frac{\ln n}{n} + E\omega_{1/n}^2(G, q) + E\omega_{1/n}^2(F, q)\right),$$

□

Corollary 3. *If σ, b are Lipschitz continuous and F, G are Bollinger or trading bands then for every $q \in \mathbb{R}^+$, $\delta > 0$ there exists $C > 0$ such that*

$$E\sup_{t\le q}|\bar{X}_t^n - X_t|^2 \le C\left(\frac{1}{n^{1-\delta}}\right).$$

Proof. Due to Theorem 5 and (5) it is sufficient to prove that the respective barrier operators F, G are Lipschitz.

First we will consider the case of Bolinger bands. We restrict our attention to the operator F. Observe that it has the following form:

$$F(X)_t = B(X)_t + \alpha\Big(\frac{1}{M}\sum_{j=0}^{M-1}(X_{t-\varepsilon j} - B(X)_t)^2\Big)^{1/2},$$

where $B(X)_t = \frac{1}{M}\sum_{j=0}^{M-1} X_{t-\varepsilon j}$, $t \in \mathbb{R}^+$, $X \in \mathcal{D}$. From the above formula it follows immediately that F posseses the properties (i) and (ii) of Definition 2. Moreover, for any $X, Y \in \mathcal{D}$ and $t \le q \in \mathbb{R}^+$,

$$|F(X)_t - F(Y)_t| \le |B(X)_t - B(Y)_t|$$
$$+\alpha\Big|\Big(\frac{1}{M}\sum_{j=0}^{M-1}(X_{t-\varepsilon j} - B(X)_t)^2\Big)^{1/2} - \Big(\frac{1}{M}\sum_{j=0}^{M-1}(Y_{t-\varepsilon j} - B(Y)_t)^2\Big)^{1/2}\Big|$$
$$\le \sup_{t\le q}|X_t - Y_t| + \alpha\Big(\frac{1}{M}\sum_{j=0}^{M-1}(X_{t-\varepsilon j} - Y_{t-\varepsilon j} + B(X)_t - B(Y)_t)^2\Big)^{1/2}$$
$$\le \sup_{t\le q}|X_t - Y_t| + \alpha\max_{0\le j\le M-1}|X_{t-\varepsilon j} - Y_{t-\varepsilon j} + B(X)_t - B(Y)_t|$$
$$\le (1 + 2\alpha)\sup_{t\le q}|X_t - Y_t|,$$

which shows that F is Lipschitz.

In the case of trading bands F has the form:

$$F(X)_t = (1+\alpha)B(X)_t, \quad t \in \mathbb{R}^+.$$

Hence F posseses properties (i) and (ii) of Definition 2 and for $X, Y \in \mathcal{D}$, $q \in \mathbb{R}^+$ we have

$$\sup_{t\le q} |F(X)_t - F(Y)_t| \le (1+\alpha) \sup_{t\le q} |X_t - Y_t|,$$

so F is Lipschitz. □

Corollary 4. *If σ, b are Lipschitz continuous and bounded functions and F, G are Bollinger or trading bands then for every $q \in \mathbb{R}^+$ there exists $C > 0$ such that*

$$E \sup_{t\le q} |\bar{X}^n_t - X_t|^2 \le C\left(\frac{\ln n}{n}\right).$$

Proof. It follows from (6), Theorem 5 and the fact that in both cases F, G are Lipschitz. □

References

1. K. Burdzy, E. Toby, *A Skorokhod-type lemma and a decomposition of reflected Brownian motion*, Ann. Probab., **23** (1995), 584–604.
2. R.J. Chitashvili, N.L. Lazrieva, *Strong solutions of stochastic differential equations with boundary conditions*, Stochastics,**5**, (1981), 225–309.
3. M. Nagasawa, T. Domenig, *Diffusion processes on an open time interval and their time reversal*, Itô's stochastic calculus and probability theory, 261-280, Springer, Tokio 1996.
4. G.N. Kinkladze, Thesis,*Tbilissi*, (1983).
5. D. Lépingle, *Euler scheme for reflected stochastic differential equations*, Mathematics and Computers in Simulations, **38** (1995), 119–126.
6. G. Maruyama, *Continuous Markov processes and stochastic equations*, Rend. Circ. Mat. Palermo, **4**, (1955), 48–90.
7. R. Pettersson, *Approximations for stochastic differential equations with reflecting convex boundaries*, Stochastic Process. Appl. **59**, (1995), 295–308.
8. S. Rady, *Option pricing in the presence of natural boundaries and quadratic diffusion term*, Finance and Stochastics, **1** (1997), 331-344.
9. L. Słomiński, *Euler's approximations of solutions of SDEs with reflecting boundary*, Stochastic Process. Appl., **94**, (2001), 317-337.
10. L. Słomiński, *On approximation of solutions of multidimensional SDEs with reflecting boundary conditions*, Stochastic Process. Appl., **50**, (1994), 197-219.
11. L. Słomiński, T. Wojciechowski, *One-dimensional stochastic differential equations with time dependent reflecting barriers*, submitted (2004).

On Approximation of Average Expectation Prices for Path Dependent Options in Fractional Models*

Bartosz Ziemkiewicz

Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University,
ul. Chopina 12/18, 87-100 Toruń, Poland
bartek@mat.uni.torun.pl

Abstract. Using integral representation of a fractional Brownian motion (fBm) we propose a new method of approximation of its trajectories as well as trajectories of a geometric fBm. We give the rate of convergence in $\mathbb{L}^p$-norm of the approximation. Applications to pricing of path dependent options in a fractional model are considered. Lookback and Asian options are examined in detail.

1 Introduction

Let $(B_t^{\mathbb{H}})_{t\in[0,T]}$ be a fractional Brownian motion (fBm) with Hurst index $\mathbb{H} \in (1/2, 1)$, i.e. $B^{\mathbb{H}}$ is a centered Gaussian process with covariance $EB_t^{\mathbb{H}}B_s^{\mathbb{H}} = \frac{1}{2}\left(t^{2\mathbb{H}} + s^{2\mathbb{H}} - |t-s|^{2\mathbb{H}}\right)$ defined on some probability space $(\Omega, \mathcal{F}, P)$. We denote by $(\mathcal{F}_t)_{t\in[0,T]}$ the filtration generated by $(B_t^{\mathbb{H}})$. We consider a continous-time market model with two assets, the riskless bond $(B_t)_{t\in[0,T]}$, and the risky stock $(S_t)_{t\in[0,T]}$. The dynamics for the prices are

$$B_t = e^{rt} \quad , \quad S_t = S_0 e^{\mu t + \sigma B_t^{\mathbb{H}}} \quad ,$$

where $S_0 > 0$ and $\mu \in \mathbb{R}$, $r, \sigma > 0$. It is convenient to make the assumption that there are no dividends, and no transaction costs. An *option* will be any non-negative $\mathcal{F}_T$-measurable random variable f_T.

Note that if $\mathbb{H} = \frac{1}{2}$, the above model coincides with the well known Black-Scholes model. In this case a *fair price* of the option is the discounted expectation of f_T with respect to the martingale measure of this model.

If $\mathbb{H} \not= \frac{1}{2}$, then $B^{\mathbb{H}}$ is not a semimartingale, and martingale measures do not exist (see e.g. [4]), so we need a different method for option pricing. Instead of martingale measures we will use the so-called *average risk neutral measures*.

Definition 1 ([6]). *If the measure Q is equivalent to the measure P, and*

$$E^Q(S_t) = e^{rt}$$

for all $t \in [0, T]$ then Q is called the average risk neutral measure.

Existence and uniqueness of Q was proved by Valkeila [6].

* Research supported by Komitet Badań Naukowych under grant 1 P03A 022 26

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 819–826, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Definition 2. *The average expectation price $C(f_T)$ of option f_T is defined by*

$$C(f_T) = e^{-rT} E^Q f_T \ . \tag{1}$$

If $\mathbb{H} = \frac{1}{2}$, the price $C(f_T)$ coincides with the classical one for the Black-Scholes model. Definition 2 extends the one given by Valkeila [6], where $C(f_T)$ is defined for options depending only on the price of S at the expiration time T, i.e. $f_T = h(S_T)$ for some $h : \mathbb{R}^+ \to \mathbb{R}^+$. In this case we have the following explicit formula

$$C(f_T) = e^{-rT} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} h\left(S_0 \exp\{\sigma y T^{\mathbb{H}} + rT - \frac{\sigma^2}{2} T^{2\mathbb{H}}\} \right) e^{-\frac{y^2}{2}} dy \ . \tag{2}$$

In the present paper we focus on path dependent options, i.e. options whose payoffs depend on the whole path of S. For example, the payoff of a *lookback option* depends on the minimum or maximum price of S, while the payoff of an *Asian option* depends on the average price of S during the life of option. Evaluation of path dependent options in fractional models is difficult and analytic formulas like (2) are not known. In the paper we will estimate $C(f_T)$ using the Monte Carlo method.

2 Approximation of fBm

In the paper we introduce a new method of approximating of $B^{\mathbb{H}}$ based on the integral representation of fBm as provided in [2]. Define the so-called fundamental martingale (M_t) by

$$M_t = \frac{c_{\mathbb{H}}}{2\mathbb{H}} \int_0^t s^{1/2-\mathbb{H}} dW_s \ , \tag{3}$$

where $c_{\mathbb{H}}$ is a constant depending on $\mathbb{H}$ and W is a standard Brownian motion. M is a Gaussian martingale with independent increments. Now, if we put

$$Y_t = 2\mathbb{H} \int_0^t (t-s)^{\mathbb{H}-1/2} dM_s$$

then the process

$$B_t^{\mathbb{H}} = \int_0^t s^{\mathbb{H}-1/2} dY_s$$

is a fractional Brownian motion (for the proof we refer to [2]).

To implement this method we proceed as follows.

1. Let M^n be a discretization of M, i.e. $M_t^n = M_{\frac{k}{n}}$ for $t \in [t_k, t_{k+1})$, where $t_k = \frac{k}{n}$, $k = 0, 1, \ldots, [nT]$. We can simulate the increments of M_t^n using the fact that $M_{t_k}^n - M_{t_{k-1}}^n \sim \mathcal{N}(0, C(t_k^{2-2\mathbb{H}} - t_{k-1}^{2-2\mathbb{H}}))$.

2. Next, we define the approximation Y^n of Y by

$$Y_t^n = 2\mathbb{H} \sum_{t_k \le t^*} (t^* - t_{k-1})^{\mathbb{H}-1/2} (M_{t_k}^n - M_{t_{k-1}}^n) \ ,$$

where $t^* = \max\{t_k; t_k \le t\}$. It is worth noting that $(t^* - t_k)^{\mathbb{H}-1/2}$ can be written in the form $(\frac{m}{n})^{\mathbb{H}-1/2}$, where $m, n \in \mathbb{N}$, $m \le [nT]$, so to compute Y^n it is sufficient to compute $m^{\mathbb{H}-1/2}$ for $m = 1, \dots, [nT]$.

3. Finally, we approximate $B^{\mathbb{H}}$ by sums of the form

$$B_t^n = \sum_{t_k \le t^*} t_{k-1}^{\mathbb{H}-\frac{1}{2}} (Y_{t_k}^n - Y_{t_{k-1}}^n) \ . \tag{4}$$

Let $\| \cdot \|_p$ denotes the usual $\mathbb{L}^p$-norm.

Theorem 1. *Let (B_t^n) be a process defined by (4). Then for any $p \ge 1$, $\epsilon > 0$ and $T \in \mathbb{R}^+$*

$$\| \sup_{t \le T} |B_t^{\mathbb{H}} - B_t^n| \|_p = \mathcal{O}\left(\frac{1}{n^{\mathbb{H}^* - \epsilon}} \right) \ ,$$

where $\mathbb{H}^* = \min(1 - \mathbb{H}, \mathbb{H} - 1/2)$.

Proof. From (3) it follows that $a_t = [M]_t = Ct^{2-2\mathbb{H}}$, where C is a constant depending on $\mathbb{H}$. Hence $M_t = \overline{W}_{a_t}$ where $\overline{W}$ is a standard Brownian motion. Using the known estimates on the modulus of continuity of $\overline{W}$ (see e.g. [5]) we get

$$\begin{aligned} \| \sup_{t \le T} |M_t - M_t^n| \|_p &\le \| \omega_M(1/n, T) \|_p = \| \omega_{\overline{W}_a}(1/n, T) \|_p \\ &\le \| \omega_{\overline{W}}(a(1/n), T) \|_p \le \mathrm{Const} \left(a(1/n) \ln(a(1/n)^{-1}) \right)^{1/2} \\ &\le \mathrm{Const} \frac{\sqrt{\ln n}}{n^{1-\mathbb{H}}} \ , \end{aligned} \tag{5}$$

where $\omega_x(\delta, T) = \sup\{|x_t - x_s|; s, t \in [0, T], \ |s - t| \le \delta\}$, $\delta > 0$, $T \in \mathbb{R}^+$, $x \in C([0, T], \mathbb{R}^+)$.

We next turn to estimating $\| \sup_{t \le T} |Y_t - Y_t^n| \|_p$. Using the integration by parts formula, we obtain

$$Y_t = 2\mathbb{H} \int_0^t (t - s)^{\mathbb{H}-1/2} dM_s = -2\mathbb{H} \int_0^t M_s d(t - s)^{\mathbb{H}-1/2} \ . \tag{6}$$

The process Y^n can be written as

$$\begin{aligned} Y_t^n &= 2\mathbb{H} \sum_{t_k \le t^*} (t^* - t_{k-1})^{\mathbb{H}-1/2} (M_{t_k}^n - M_{t_{k-1}}^n) \\ &= -2\mathbb{H} \sum_{t_k \le t^*} M_{t_k}^n \left[(t^* - t_k)^{\mathbb{H}-1/2} - (t^* - t_{k-1})^{\mathbb{H}-1/2} \right] \\ &= -2\mathbb{H} \int_0^{t^*} \hat{M}_s^n d(t^* - s)^{\mathbb{H}-1/2} \ , \end{aligned} \tag{7}$$

where $\hat{M}^n_t = M_{\frac{k+1}{n}}$ for $t \in [\frac{k}{n}, \frac{k+1}{n})$. Note that (5) still holds if we replace M^n by $\hat{M}^n$. Now,

$$\begin{aligned}\| \sup_{t\le T} |Y_t - Y^n_t| \|_p &\le \text{Const}(\| \sup_{t\le T} |Y_t - Y_{t^*}| \|_p + \| \sup_{t\le T} |Y_{t^*} - Y^n_t| \|_p) \\ &= I_1 + I_2 \ .\end{aligned}$$

Combining (5) with (6) and (7) gives

$$\begin{aligned}I_2 &\le \text{Const} \left\| \left| \sup_{t\le T} \left| \int_0^{t^*} (M_s - \hat{M}^n_s) d(t^* - s)^{\mathbb{H}-1/2} \right| \right| \right\|_p \\ &\le \text{Const} \| \sup_{t\le T} E|M_t - \hat{M}^n_t| \|_p \\ &\le \text{Const} \frac{\sqrt{\ln n}}{n^{1-\mathbb{H}}} \ . \end{aligned} \tag{8}$$

To estimate I_1, we first prove that for any $t_1, t_2 \in [0, T]$ and $q \ge 1$

$$E|Y_{t_2} - Y_{t_1}|^q \le \text{Const}|t_2 - t_1|^{q(\mathbb{H}-1/2)} \ .$$

Indeed, by the Burkholder-Davis-Gundy inequality we have

$$\begin{aligned}E|Y_{t_2} - Y_{t_1}|^q &= E \left| \int_0^{t_1} \left[(t_2 - s)^{\mathbb{H}-1/2} - (t_1 - s)^{\mathbb{H}-1/2} \right] dM_s \right. \\ &\quad \left. + \int_{t_1}^{t_2} (t_2 - s)^{\mathbb{H}-1/2} dM_s \right|^q \\ &\le \text{Const} \left(E \left| \int_0^{t_1} \left[(t_2 - s)^{\mathbb{H}-1/2} - (t_1 - s)^{\mathbb{H}-1/2} \right]^2 d[M]_s \right|^{q/2} \right. \\ &\quad \left. + E \left| \int_{t_1}^{t_2} (t_2 - s)^{2\mathbb{H}-1} d[M]_s \right|^{q/2} \right) \\ &\le \text{Const} \left((t_2 - t_1)^{q(\mathbb{H}-1/2)} \left| \int_0^{t_1} C s^{1-2\mathbb{H}} ds \right|^{q/2} \right. \\ &\quad \left. + \left| \int_{t_1}^{t_2} C s^{1-2\mathbb{H}} ds \right|^{q/2} \right) \\ &\le \text{Const}(t_2 - t_1)^{q(\mathbb{H}-1/2)} \left| \int_0^{T} C s^{1-2\mathbb{H}} ds \right|^{q/2} \\ &\le \text{Const}(t_2 - t_1)^{q(\mathbb{H}-1/2)} \ . \end{aligned}$$

Let $q > \max(p, (\mathbb{H} - 1/2)^{-1})$. By the Kolmogorov continuity theorem,

$$E \left(\sup_{\substack{t_1, t_2 \le T \\ t_1 \ne t_2}} \frac{|Y_{t_2} - Y_{t_1}|}{|t_2 - t_1|^\alpha} \right)^q < +\infty$$

for $0 < \alpha < (\mathbb{H} - 1/2) - 1/q$. Since $p < q$, from the above it follows that

$$I_1 \le (E \sup_{t\le T} |Y_t - Y_{t^*}|^q)^{1/q} \le \text{Const} \frac{1}{n^{\mathbb{H}-1/2-\epsilon}} \tag{9}$$

for every $\epsilon > 0$. By (8) and (9),

$$\| \sup_{t\le T} |Y_t - Y_t^n| \|_p \le \text{Const} \frac{1}{n^{\mathbb{H}^*-\epsilon}} \ . \tag{10}$$

It remains to show that $\| \sup_{t\le T} |B_t^{\mathbb{H}} - B_t^n| \|_p \le \text{Const} \| \sup_{t\le T} |Y_t - Y_t^n| \|_p$. Using once again the integration by parts formula we obtain

$$B_t^{\mathbb{H}} = \int_0^t s^{\mathbb{H}-1/2} dY_s = t^{\mathbb{H}-1/2} Y_t - \int_0^t Y_s ds^{\mathbb{H}-1/2} \ ,$$

and

$$\begin{aligned} B_t^n = \int_0^t s^{\mathbb{H}-1/2} dY_s^n &= t^{\mathbb{H}-1/2} Y_t^n - \int_0^t Y_s^n ds^{\mathbb{H}-1/2} \\ &- \sum_{t_k \le t} (Y_{t_k}^n - Y_{t_{k-1}}^n)(t_k^{\mathbb{H}-1/2} - t_{k-1}^{\mathbb{H}-1/2}) \ . \end{aligned} \tag{11}$$

Hence

$$\begin{aligned} \| \sup_{t\le T} |B_t^{\mathbb{H}} - B_t^n| \|_p \le \text{Const} \Bigg(& | \sup_{t\le T} |t^{\mathbb{H}-1/2} Y_t - t^{\mathbb{H}-1/2} Y_t^n| \|_p \\ & + \left\| \sup_{t\le T} \left| \int_0^t (Y_s - Y_s^n) ds^{\mathbb{H}-1/2} \right| \right\|_p \\ & + \left\| \sup_{t\le T} \left| \sum_{t_k \le t} (Y_{t_k}^n - Y_{t_{k-1}}^n)(t_k^{\mathbb{H}-1/2} - t_{k-1}^{\mathbb{H}-1/2}) \right| \right\|_p \Bigg) \\ = I_4 + I_5 + I_6 \ . \end{aligned}$$

It is easily seen that I_4 and I_5 are bounded by $\text{Const} \| \sup_{t\le T} |Y_t - Y_t^n| \|_p$. To estimate I_6 we observe first that

$$\begin{aligned} \sup_{t_k \le T} |Y_{t_k}^n - Y_{t_{k-1}}^n| = \sup_{t\le T} |Y_t^n - Y_{t-}^n| &= \sup_{t\le T} \left| (Y_t^n - Y_{t-}^n) - (Y_t - Y_{t-}) \right| \\ &\le 2 \sup_{t\le T} |Y_t - Y_t^n| \ . \end{aligned}$$

Hence

$$I_6 \le T^{\mathbb{H}-1/2} \| \sup_{t_k \le T} |Y_{t_k}^n - Y_{t_{k-1}}^n| \|_p \le \text{Const} \| \sup_{t\le T} |Y_t - Y_t^n| \|_p \ ,$$

which completes the proof. □

3 Approximation of Prices for Path Dependent Options

The method described in the previous section can be used to approximate the price evolution process S. Let S^n be the approximation of S defined by

$$S_t^n = S_0 e^{\mu t + \sigma B_t^n}$$

for all $t \in [0,T]$. The following result is a consequence of Theorem 1.

Theorem 2. *For any $p \geq 1$, $\epsilon > 0$ and $T \in \mathbb{R}^+$*

$$\| \sup_{t \leq T} |S_t - S_t^n| \|_p = \mathcal{O}\left(\frac{1}{n^{\mathbb{H}^* - \epsilon}}\right) \ .$$

Proof. Without loss of generality we may and will assume that $S_0 = 1, \mu = 0, \sigma = 1$. Observe first that by the Schwarz inequality,

$$\begin{aligned} \| \sup_{t \leq T} |S_t - S_t^n| \|_p &= \| \sup_{t \leq T} |e^{B_t^{\mathbb{H}}} [e^{(B_t^{\mathbb{H}} - B_t^n)} - 1]| \|_p \\ &\leq \| \sup_{t \leq T} |e^{B_t^{\mathbb{H}}}| \|_{2p} \ \| \sup_{t \leq T} |e^{(B_t^{\mathbb{H}} - B_t^n)} - 1| \|_{2p} \ . \end{aligned}$$

Since $|e^x - 1| \leq |x| e^{|x|}$, we have

$$\begin{aligned} \| \sup_{t \leq T} |e^{(B_t^{\mathbb{H}} - B_t^n)} - 1| \|_{2p} &\leq \| \sup_{t \leq T} |B_t^{\mathbb{H}} - B_t^n| \cdot e^{|B_t^{\mathbb{H}} - B_t^n|} \|_{2p} \\ &\leq \| \sup_{t \leq T} |B_t^{\mathbb{H}} - B_t^n| \|_{4p} \ \| \sup_{t \leq T} e^{|B_t^{\mathbb{H}} - B_t^n|} \|_{4p} \ . \end{aligned}$$

By Theorem 1,

$$|| \sup_{t \leq T} |B_t^{\mathbb{H}} - B_t^n| \ ||_{4p} = \mathcal{O}\left(\frac{1}{n^{\mathbb{H}^* - \epsilon}}\right) \ ,$$

so it remains to show that $\| \sup_{t \leq T} e^{B_t^{\mathbb{H}}} \|_{8p} < +\infty$ and $\sup_n \| \sup_{t \leq T} e^{B_t^n} \|_{8p} < +\infty$. Novikov and Valkeila [3] proved that $\sup_{t \leq T} |B_t^{\mathbb{H}}| \leq C \sup_{t \leq T} |M_t|$ for some constant $C > 0$. Therefore, by the Doob submartingale inequality,

$$\begin{aligned} \| \sup_{t \leq T} e^{B_t^{\mathbb{H}}} \|_{8p} &= (E e^{\sup_{t \leq T} (8p |B_t^{\mathbb{H}}|)})^{1/(8p)} \\ &\leq (E \sup_{t \leq T} e^{8p\, C |M_t|})^{1/(8p)} \\ &\leq (4\, E e^{8p\, C |M_T|})^{1/(8p)} < +\infty \ , \end{aligned} \tag{12}$$

because M is a Gaussian martingale. Finally, by (11) and (7),

$$\begin{aligned} \sup_{t \leq T} |B_t^n| &\leq T^{\mathbb{H} - 1/2} \sup_{t \leq T} |Y_t^n| + T^{\mathbb{H} - 1/2} \sup_{t \leq T} |Y_t^n| \\ &\quad + T^{\mathbb{H} - 1/2} 2 \sup_{t \leq T} |Y_t^n| \leq 4 T^{\mathbb{H} - 1/2} \sup_{t \leq T} |Y_t^n| \\ &\leq 8 \mathbb{H} T^{2\mathbb{H} - 1} \sup_{t \leq T} |M_t^n| \ . \end{aligned}$$

Therefore, similarly to (12) we get

$$\sup_n \| \sup_{t\le T} e^{B_t^n} \|_{8p} \le \sup_n (4\, E e^{64p\, \mathbb{H} T^{2\mathbb{H}-1} |M_T^n|})^{1/(8p)}$$
$$\le (4\, E e^{64p\, \mathbb{H} T^{2\mathbb{H}-1} |M_T|})^{1/(8p)} < +\infty$$

and the proof is complete. □

Definition 3. *We say that an operator $H : C([0,T], \mathbb{R}^+) \to \mathbb{R}$ is Lipschitz if there exists $L > 0$ such that for any $x, y \in C([0,T], \mathbb{R}^+)$*

$$|H(x) - H(y)| \le L \cdot \sup_{t\le T} |x_t - y_t| \ .$$

From Theorem 1 and Definition 3 we obtain immediately the following corollary.

Corollary 1. *If $H : C([0,T], \mathbb{R}^+) \to \mathbb{R}$ is Lipschitz then for any $p \ge 1$, $\epsilon > 0$*

$$\|H(S) - H(S^n)\|_p = \mathcal{O}\left(\frac{1}{n^{\mathbb{H}^* - \epsilon}}\right) \ .$$

By choosing appropriately a Lipschitz operator H we obtain interesting examples of path dependent options of the form $f_T = H(S)$.

Example 1 (Floating strike lookback call option).
In this case $H(S) = S_T - \inf_{t\le T} S_t$. Note that for any $x, y \in C([0,T], \mathbb{R}^+)$

$$|H(x) - H(y)| = |(x_T - y_T) + (\inf_{t\le T} y_t - \inf_{t\le T} x_t)|$$
$$\le 2 \sup_{t\le T} |x_t - y_t| \ .$$

Example 2 (Fixed strike lookback call option with the strike price K).
Here $H(S) = (\sup_{t\le T} S_t - K)^+$ and for any $x, y \in C([0,T], \mathbb{R}^+)$

$$|H(x) - H(y)| \le |\sup_{t\le T} x_t - \sup_{t\le T} y_t| \le \sup_{t\le T} |x_t - y_t| \ .$$

Example 3 (Asian call option with the strike price K and continuous arithmetic averaging).
For such options

$$H(S) = \left(\frac{1}{T}\int_0^T S_t dt - K\right)^+ \ ,$$

and it is easily seen that

$$|H(x) - H(y)| = \left| \left(\frac{1}{T} \int_0^T x_t dt - K \right)^+ - \left(\frac{1}{T} \int_0^T y_t dt - K \right)^+ \right|$$
$$\leq \left| \frac{1}{T} \int_0^T (x_t - y_t) dt \right| \leq \sup_{t \leq T} |x_t - y_t| \ ,$$

for any $x, y \in C([0,T], \mathbb{R}^+)$.

4 Numerical Results

The method described in Sections 2 and 3 we have applied to evaluate prices of the lookback and Asian options. We have simulated 100000 trajectories of S with step-size 1/1000.

Table 1 gives prices of the floating strike lookback options (LB1), fixed strike lookback options with the strike price $K_1 = 5$ (LB2) and Asian options with the strike price $K2 = 5$ (AS) for different choises of the Hurst index $\mathbb{H}$. The initial asset price S_0 is set equal to 5, parameters r and σ are equal 0.1 and 0.3 respectively.

Table 1. The prices of LB1, LB2 and AS ($S_0 = 5$, $K_1 = 5$, $K_2 = 5$, $\sigma = 0.3$, $r = 0.1$, $T = 1$).

$\mathbb{H}$	0.6	0.7	0.8	0.9
LB1	1.1381	1.0288	0.9399	0.8861
LB2	1.2727	1.1177	1.0013	0.9138
AS	0.4464	0.4358	0.4234	0.4207

References

1. Kwok Y.K.: Mathematical Models of Financial Derivatives. Springer, Singapore (1998)
2. Norros I., Valkeila E., Virtamo J.: An elementary approach to a Grisanov formula and other analytical results on fractional Brownian motions. Bernoulli **5(4)** (1999) 571–587
3. Novikov A., Valkeila E.: On some maximal inequalities for fractional Brownian motions. Statistics & Probability Letters **44** (1999) 47–54
4. Shiryaev A.N.: On Arbitrage and Replication for Fractal Models. Research Report 20, MaPhySto, Centre for Mathematical Physics and Stochastics, September (1998)
5. Słomiński L.: Euler's approximations of solutions of SDEs with reflecting boundary. Stochastics Processes and their Applications **94** (2001) 317–337
6. Valkeila E.: On some properties of geometric fractional Brownian motions. Preprint (1998)

Confidence Intervals for the Autocorrelations of the Squares of GARCH Sequences

Piotr Kokoszka[1], Gilles Teyssière[2], and Aonan Zhang[3]

[1] Mathematics and Statistics, Utah State University,
3900 Old Main Hill, Logan UT 84322-3900, USA,
piotr@stat.usu.edu
http://math.usu.edu/~piotr
[2] NBG Bank (Paris) & ASEF,
gilles@ehess.cnrs-mrs.fr
http://www.gillesteyssiere.net
[3] Mathematics and Statistics, Utah State University,
aonanzhang@cc.usu.edu

Abstract. We compare three methods of constructing confidence intervals for sample autocorrelations of squared returns modeled by models from the GARCH family. We compare the residual bootstrap, block bootstrap and subsampling methods. The residual bootstrap based on the standard GARCH(1,1) model is seen to perform best.

1 Introduction

The paper is concerned with assessing finite sample performance of several methods of finding confidence intervals for autocorrelations of squared returns on speculative assets. While the returns themselves are essentially uncorrelated and most econometric and financial models explicitely imply that they are so, their squares exhibit a rich dependence structure. The sample autocorrelation of squared returns $\hat{\rho}_{n,X^2}(\cdot)$ is a measure of volatility clustering. A large value for $\hat{\rho}_{n,X^2}(\cdot)$ is the evidence of either the presence of long memory in the volatility process, or the inadequacy of a GARCH(1,1) process to fit the data under investigation as the true process might be a non–homogeneous GARCH(1,1) process; see [6].

We compare the performance of the various methods by means of their empirical coverage probability (ECP). Suppose we have a method of constructing, say, a 95% confidence interval $(\hat{l}_n, \hat{u}_n)$ from an observed realization $X_1, X_2, \ldots, X_n$. We simulate a large number R of realizations from a specific GARCH type model from which we construct R confidence intervals $(\hat{l}_n^{(r)}, \hat{u}_n^{(r)})$, $r = 1, 2, \ldots, R$. The percentage of these confidence intervals that contain the population autocorrelation is the ECP, which we want to be as close as possible to the nominal coverage probability of 95%. Our objective is to provide answers to the following questions: Does any method have better ECP than the others? If not, what is the range of optimal applicability of each method? Is it better to use equal-tailed or symmetric confidence intervals (see Section 2.1)? How does the coverage depend

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 827–834, 2004.
© Springer-Verlag Berlin Heidelberg 2004

on the value of γ_{c2} ? For a given series length n, how should one choose the block length b for the block bootstrap and subsampling? For what lengths n do these methods yield useful confidence intervals?

The ultimate goal is to recommend a practical procedure for finding confidence intervals for squared autocorrelations which assumes minimal prior knowledge of the stochastic mechanism generating the returns.

For ease of reference, recall that the sample autocovariances of the squared returns are

$$\hat{\gamma}_{n,X^2}(h) = \frac{1}{n}\sum_{t=1}^{n-h}\left(X_t^2 - \frac{1}{n-h}\sum_{t=1}^{n-h}X_t^2\right)\left(X_{t+h}^2 - \frac{1}{n-h}\sum_{t=h+1}^{n}X_t^2\right), \quad (1)$$

the population autocovariances are $\gamma_{X^2}(h) = E\left[(X_0^2 - EX_0^2)(X_h^2 - EX_0^2)\right]$, while the corresponding autocorrelations (ACF) are

$$\hat{\rho}_{n,X^2}(h) = \frac{\hat{\gamma}_{n,X^2}(h)}{\hat{\gamma}_{n,X^2}(0)}, \quad \rho_{X^2}(h) = \frac{\gamma_{X^2}(h)}{\gamma_{X^2}(0)}. \quad (2)$$

In Section 2, we describe the three methods. Section 3 introduces the various GARCH models we use for the comparison. The results of our simulations are presented in Section 4 with broad conclusions summarized in Section 4.3.

2 Confidence Intervals for ACF of Squared Returns

2.1 Residual Bootstrap

To illustrate the idea, we consider the ARCH(1) model given by

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \omega + \alpha X_{t-1}^2. \quad (3)$$

As we will see in Section 3, the method can be readily extended to any parametric model defined by GARCH type equations by computing the residuals $\hat{Z}_t = X_t/\hat{\sigma}_t$. Since the conditional volatility σ_t^2 is a function of the model parameters, past observations and past innovations, $\hat{\sigma}_t^2$ can be computed recursively once parameter estimates are available. We then proceed as follows:

1. Estimate $\hat{\omega}$ and $\hat{\alpha}$ and compute $\hat{Z}_t = [\hat{\omega} + \hat{\alpha}X_{t-1}^2]^{-1/2}X_t$, with $X_0 = \bar{X}_t$. We use the quasi maximum likelihood estimators (QMLE's) of model parameters with the assumption that the innovations $Z_t \sim N(0,1)$.
2. Form B bootstrap realizations $X_t^2(b) = [\hat{\omega} + \hat{\alpha}X_{t-1}^2(b)]\hat{Z}_t^2(b)$, $t = 1, 2, \ldots, n$, where $\hat{Z}_1^2(b), \ldots \hat{Z}_n^2(b)$, $b = 1, 2, \ldots, B$, are the B bootstrap samples selected with replacement from the squared residuals $\hat{Z}_1^2, \ldots \hat{Z}_n^2$.
3. Calculate the bootstrap autocorrelations $\rho_{n,X^2}^{(b)}(1)$, $b = 1, 2, \ldots, B$ and use their empirical quantiles to find a confidence interval for $\rho_{n,X^2}(1)$.

We now enlarge on step 3). Denote by $F^*_{\rho(1)}$ the EDF (empirical distribution function) of the $\rho^{(b)}_{n,X^2}(1)$, $b = 1, 2, \ldots, B$. The $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of $F^*_{\rho(1)}$ will yield an *equal-tailed* $(1-\alpha)$ level confidence interval. To construct a *symmetric* confidence interval centered at $\hat{\rho}_{n,X^2}(1)$, we need the empirical distribution $F^*_{\rho(1),|\cdot|}$ of the B values $|\rho^{(b)}_{n,X^2}(1) - \hat{\rho}_{n,X^2}(1)|$. Denote by $q_{|\cdot|}(1-\alpha)$ the $(1-\alpha)$ quantile of $F^*_{\rho(1),|\cdot|}$. Then the symmetric confidence interval is

$$\left(\hat{\rho}_{n,X^2}(1) - q_{|\cdot|}(1-\alpha), \quad \hat{\rho}_{n,X^2}(1) + q_{|\cdot|}(1-\alpha)\right).$$

A usual criticism of methods based on a parametric model is that misspecification can lead to large biases. In many applications however these biases have only negligible impact on a statistical procedure of interest. In our setting, we will see that the residual bootstrap confidence intervals based on a misspecified model can produce good coverage probabilities.

2.2 Block Bootstrap

In this section we describe how the popular block-bootstrap of [5] can be used to construct confidence intervals for autocorrelations. This method does not require a model specification, but it relies on a choice of the block size b which is often a difficult task. A good account of block bootstrap is given in [1].

Focusing again on lag one sample autocorrelation of the squared observations, we proceed as follows: having observed the sample $X_1^2, \ldots, X_n^2$, form the vectors $\mathbf{Y}_2 = [X_1^2, X_2^2]', \mathbf{Y}_3 = [X_2^2, X_3^2]', \ldots, \mathbf{Y}_n = [X_{n-1}^2, X_n^2]'$. There are $n-1$ such vectors. Now choose a block length b and compute the number of blocks $k = [(n-1)/b] + 1$ (if $(n-1)/b$ is an integer we take $k = (n-1)/b$). Choose k blocks with replacement to obtain kb vectors. Choosing the k blocks corresponds to generating k observations from the uniform distribution on $\{2, 3, \ldots, n-b+1\}$. Denote these observations $j_1, j_2, \ldots, j_k$. We thus obtained the kb vectors $\mathbf{Y}_{j_1}, \mathbf{Y}_{j_1+1}, \ldots, \mathbf{Y}_{j_1+b-1}, \ldots, \mathbf{Y}_{j_k}, \mathbf{Y}_{j_k+1}, \ldots, \mathbf{Y}_{k_1+b-1}$. If $(n-1)/b$ is not an integer, remove the last few vectors to have exactly $n-1$ vectors. This gives us the bootstrap vector process

$$\mathbf{Y}_2^* = [X_1^{*2}, X_2^{*2}]', \mathbf{Y}_3^* = [X_2^{*2}, X_3^{*2}]', \ldots, \mathbf{Y}_n^* = [X_{n-1}^{*2}, X_n^{*2}]'.$$

The bootstrap sample autocovariances are computed according to (1) with the X_t replaced by the X_t^* defined above. The empirical distribution of $\hat{\rho}^*_{n,X^2}(1)$ is then an approximation to the distribution of $\hat{\rho}_{n,X^2}(1)$. As described in Section 2.1, the quantiles of the empirical distribution of $|\hat{\rho}^*_{n,X^2}(1) - \hat{\rho}_{n,X^2}(1)|$ can be used to construct symmetric confidence intervals.

2.3 Subsampling

The subsampling methodology is described in detail in [7]. [8] investigated subsampling confidence intervals for autocorrelations of linear time series models like ARMA. We adapt their methodology to the squares of GARCH processes.

To lighten the notations, denote $U_t = X_t^2 - \frac{1}{n}\sum_{j=1}^n X_j^2$ and suppress the subscript X^2 in the following formulas in which use definitions (1) and (2). Set

$$s_n^2(h) = \frac{1}{n}\sum_{j=1}^{n-h}\left(U_{j+h} - \hat{\rho}_n(h)U_j\right)^2, \quad \hat{\sigma}_n^2(h) = \frac{s_n^2(h)}{\sum_{j=h}^n U_j^2} \tag{4}$$

and consider the studentized statistic $\hat{\xi}_n = \frac{\hat{\rho}_n(h)-\rho_n(h)}{\hat{\sigma}_n(h)}$. To construct equal-tailed and symmetric confidence intervals, we would need to know the sampling distribution of $\hat{\xi}_n$ and $|\hat{\xi}_n|$, respectively. We use subsampling to approximate these distributions: Consider an integer $b < n$ and the $n-b+1$ blocks of data $X_t, \ldots, X_{t+b-1},\ t = 1, \ldots, n-b+1$. From each of these blocks compute $\hat{\rho}_{b,t}(h)$ and $\hat{\sigma}_{b,t}(h)$ according to respectively (1), (2) and (4), but replacing the original data $X_1, \ldots, X_n$ by $X_t, \ldots, X_{t+b-1}$. Next, compute the subsampling counterpart of the studentized statistic $\hat{\xi}_{b,t}(h) = \dfrac{\hat{\rho}_{b,t}(h) - \hat{\rho}_n(h)}{\hat{\sigma}_{b,t}(h)}$ and construct the EDF

$$L_b(x) = \mathcal{N}_b^{-1}\sum_{t=1}^{n-b+1}\mathbf{1}\left\{\hat{\xi}_{b,t}(h) \le x\right\}, \quad L_{b,|\cdot|}(x) = \mathcal{N}_b^{-1}\sum_{t=1}^{n-b+1}\mathbf{1}\left\{|\hat{\xi}_{b,t}(h)| \le x\right\},$$

with $\mathcal{N}_b = n-b+1$. The empirical quantiles of L_b and $L_{b,|\cdot|}$ allow us to construct, respectively, equal-tailed and symmetric confidence intervals. For example, denoting by $q_{b,|\cdot|}(1-\alpha)$ the $(1-\alpha)$th quantile of $L_{b,|\cdot|}$, a subsampling symmetric $1-\alpha$ level confidence interval for $\rho_n(h)$ is

$$\left(\hat{\rho}_n(h) - \hat{\sigma}_n(h)q_{b,|\cdot|}(1-\alpha), \quad \hat{\rho}_n(h) + \hat{\sigma}_n(h)q_{b,|\cdot|}(1-\alpha)\right).$$

3 GARCH Models

We consider the general framework for GARCH models proposed and studied by [4]; see also references quoted therein. The observations X_t are thus assumed to satisfy $X_t = Z_t\sigma_t$, where Z_t is a sequence of independent identically distributed random variables with zero mean and $\sigma_t^2 = g(Z_{t-1}) + c(Z_{t-1})\sigma_{t-1}^2$. We considered only specifications in which the function $g(\cdot)$ is a constant and the Z_t are standard normal. Denoting $\gamma_{ci} = Ec^i(Z_t)$, [4] proved that under the above assumptions a sufficient and necessary condition for the existence of the $2m$th unconditional moment of X_t is $\gamma_{cm} = Ec_t^m < 1$. Thus, the fourth unconditional moment of X_t exits if and only if $\gamma_{c2} = Ec_t^2 \in [0, 1)$. We considered the following three specific models:

1. The standard GARCH$(1,1)$ model, for which
$$c_{t-1} = \beta + \alpha Z_{t-1}^2, \quad \sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{5}$$
2. The GJR-GARCH$(1,1)$ model, see [3], with
$$c_{t-1} = \beta+(\alpha+\phi I(Z_{t-1}))Z_{t-1}^2, \quad \sigma_t^2 = \omega+(\alpha+\phi I(Z_{t-1}))X_{t-1}^2+\beta\sigma_{t-1}^2, \tag{6}$$
where $I(Z_{t-1}) = 1$ if $Z_{t-1} < 0$, and $I(Z_{t-1}) = 0$ otherwise.

3. The nonlinear GARCH(1,1) model (NL GARCH(1,1,2), see [2], with

$$c_{t-1} = \beta + \alpha(1 - 2\eta \text{ sign}(Z_{t-1}) + \eta^2)Z_{t-1}^2;$$

$$\sigma_t^2 = \omega + \alpha(1 - 2\eta \text{ sign}(Z_{t-1}) + \eta^2)X_{t-1}^2 + \beta\sigma_{t-1}^2. \quad (7)$$

Assuming the errors $Z_t \sim N(0,1)$, the values of γ_{c2} and $\rho_{X^2}(1)$ can be computed in a closed form. If we know the model parameters, we can calculate precisely the population autocorrelation $\rho_{X^2}(1)$ and the value of γ_{c2}.

For each of the three models, we considered five parameter choices, which we labeled as models 1 through 5. The lag one autocorrelations for these choices are, respectively, approximately .15, .22, .31, .4, .5. The corresponding values of γ_{c2} are respectively, approximately .1, .3, .5, .7, .9. To facilitate comparison, models with the same index have similar values of γ_{c2} and $\rho_{X^2}(1)$, e.g. standard GARCH and GJR-GARCH with index 3 both have $\gamma_{c2} \approx .5$ and $\rho_{X^2}(1) \approx .31$.

4 Simulation Results

We investigate the performance of the three methods described in Section 2 by comparing the empirical coverage probabilities (ECP's) for the fifteen data generating processes (DGP's) introduced in Section 3. We generated one thousand replications of each DGP and considered realizations of length $n = 100, 250, 500,$ 1000. We focused on the most commonly used confidence level of 95%. The standard errors in all tables are about 0.5% and are always smaller than 1%.

4.1 Residual Bootstrap

Table 4.1 presents the ECP of the symmetric confidence interval for the three GARCH models. To save space the results for the equal-tailed confidence interval are not presented, but are discussed in the following conclusions. Equal-tailed and symmetric confidence intervals perform equally well for the standard GARCH and GJR-GARCH. However, for the NL GARCH, the symmetric interval is better than the equal-tailed. It is thus seen that the symmetric confidence interval is preferred over the equal-tailed. The ECP decreases as the value of γ_{c2} approaches 1. Recall that $\gamma_{c2} < 1$ is required for the population autocovariances to exist. When $\gamma_{c2} \approx 0.9$, at least 250 observations are needed to ensure reasonable ECP for the standard GARCH and the GJR-GARCH. For the NL GARCH, even series length of 1000, does not produce satisfactory results. For the standard GARCH and the GJR-GARCH increasing the sample size from 500 to 1000 does not improve the ECP. For the NL GARCH a sample size of 1000 observations is needed, except when $\gamma_{c2} \leq 0.3$.

The somewhat worse performance of the residual bootstrap method for the GJR-GARCH which becomes markedly worse for the NL GARCH can be attributed to identification problems, which are particularly acute for the NL GARCH: for the latter model biases of parameter estimates are very large when η in equation (7) is large. Large η corresponds to large γ_{c2}, we omit the details

Table 1. ECP of *symmetric* confidence intervals constructed using *residual bootstrap*.

n	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)
STD GARCH	1	2	3	4	5
100	99.6	85.3	86.0	80.4	77.4
250	92.9	91.3	92.1	89.4	84.4
500	93.4	93.4	94.1	93.7	92.7
1000	95.1	96.8	97.6	97.6	94.4
GJR GARCH	1	2	3	4	5
100	97.7	94.8	92.0	89.5	81.5
250	96.2	96.6	97.0	96.4	92.3
500	98.3	99.2	98.9	99.1	96.5
1000	99.0	99.4	99.6	99.8	98.8
NL GARCH	1	2	3	4	5
100	95.5	83.8	79.8	74.7	66.0
250	91.7	87.3	84.3	81.0	73.6
500	91.7	93.1	88.5	82.1	77.3
1000	96.4	93.3	92.9	87.0	81.0

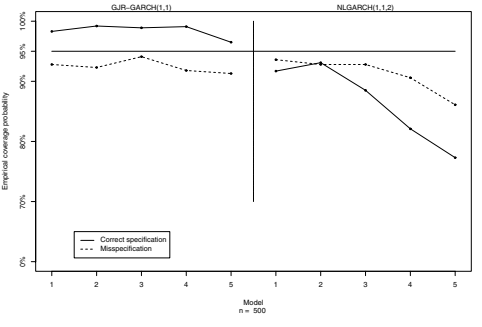


Fig. 1. Comparison of ECP's for symmetric residual bootstrap confidence intervals based on standard GARCH and a correct specification. The nominal coverage of 95% is marked by the solid horizontal line. The series length is $n = 500$.

of the calculation. On the other hand, for the standard GARCH, while they still do exist, the identification problems are much less severe.

Figure 1 shows that estimating the standard GARCH model on all three DGP's might lead to improvements in ECP's, for symmetric confidence intervals and series of length 500. The results for other series lengths look very much the same and are therefore not presented. The residual bootstrap method works best if symmetric confidence intervals are used and the standard GARCH model is estimated. Thus, in our context, misspecifying a model improves the performance of the procedure.

Table 2. ECP of *symmetric* confidence intervals based on the *block bootstrap* method for the five parameter choices in the *GJR-GARCH* model.

Model		1	2	3	4	5
n	b	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)	*e.c.p.* (%)
500	3	87.0	82.0	78.4	65.5	61.4
	5	89.1	83.8	73.4	63.0	58.5
	10	87.9	81.8	71.4	60.6	51.9
	15	84.5	78.7	71.8	63.8	52.7
	30	85.6	79.0	69.6	61.3	50.0
1000	5	87.7	84.4	75.2	67.9	59.6
	10	88.6	85.1	70.8	61.0	52.6
	15	89.7	83.0	72.7	63.6	53.3
	30	87.8	80.9	72.7	59.7	51.2

4.2 Block Bootstrap and Subsampling

The implementation of both methods requires a choice of the block length b. We then have a multitude of cases to explore: 15 models, 2 types of confidence intervals (equal-tailed and symmetric), 4 sample sizes and several choices of b. Since we used 10 values of b in our experiments, we obtained 1,200 ECP's. For space constraints, we describe them and present some typical values in Table 2.

The empirical coverage probabilities are generally too low for all choices of n and b and are in the range of 80% to 90% for $\gamma_{c2} \leq 0.3$ and go down to slightly above 50% for $\gamma_{c2} \approx 0.9$. Irrespective of the value of γ_{c2}, choosing smaller b gives higher coverage. However, extremely small b, like 1 or 2, do not work well. We recommend to use $b = 3$ or $b = 5$. The dependence on b is however not substantial, which is very desirable, as in many other applications choosing optimal b is very difficult. There is not much difference of ECPs between equal-tailed and symmetric confidence intervals. The block bootstrap confidence intervals are generally too short and given that the QML estimates underestimate the true value of the autocorrelation, they are shifted too much to the left what causes the under-coverage.

We observed that the subsampling method is very sensitive to the choice of b. Symmetric confidence intervals have a much better ECP than the equal-tailed. By choosing very short b's, such as 3 or 6, we can obtain ECP's that are quite close to 95% for models with $\gamma_{c2} < 0.6$ and fair coverage for models with greater values of γ_{c2}. Such choice of b is somewhat surprising, as autocovariances are then computed from very short sub-series. The ECP's are generally too low for equal-tailed confidence intervals and are typically in the range of 50-70%. As γ_{c2} approaches 1, the empirical coverage decrease and in some cases may be as low as 10%. Complete tables for ECP's are available at the following site: www.gillesteyssiere.net/ktz_iccs2004.

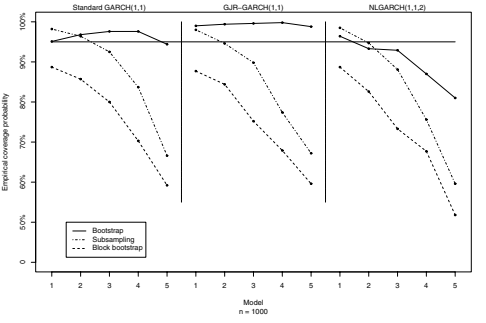


Fig. 2. Comparison of ECP's for symmetric confidence intervals. The nominal coverage 95% is marked by solid horizontal line. The series length is $n = 1000$. For block bootstrap, $b = 5$, for subsampling $b = 3$.

4.3 Conclusions and Practical Recommendations

The best method is residual bootstrap which assumes a standard GARCH(1,1) model. The block bootstrap and subsampling methods do not perform well when γ_{c2} approaches 1. Moreover, these methods require a choice of the block size b. The latter problem is particularly acute for the subsampling method. Except for the NL GARCH, the residual bootstrap method with correct model specification performs reasonably well even for γ_{c2} close to 1. This is probably due to the fact that large values for γ_{c2} correspond to large values of model parameters which are easier to estimate than small values yielding residuals which are close to the unobservable errors. A graphical comparison of symmetric confidence intervals based on the three methods for $n = 1000$ is given below:

References

1. Bühlmann, P.: Bootstrap for time series. Statistical Science **17** (2002) 52–72.
2. Engle, R.F.: Discussion: stock market volatility and the crash of 87. Review of Financial Studies **3** (1990) 103–106.
3. Glosten, L.R., Jagannathan, R. and Runkle, D.: On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance **48** (1993) 1779–1801.
4. He, C. and Teräsvirta, T.: Properties of moments of a family of GARCH processes. Journal of Econometrics **92** (1999) 173–192.
5. Künsch, H.: The jackknife and the bootstrap for general stationary observations. The Annals of Statistics **17** (1989) 1217–1241.
6. Mikosch, T. and Stărică, C.: Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. The Annals of Statistics **28** (2000) 1427–1451.
7. Politis, D.N., Romano, J.P. and Wolf, M.: Subsampling. (1999) Springer Verlag.
8. Politis, D.N., Romano, J.P. and Wolf, M.: Inference for autocorrelations in the possible presence of a unit root. Journal of Time series Analysis (to appear).

Performance Measures in an Evolutionary Stock Trading Expert System

Piotr Lipinski[1, 2] and Jerzy J. Korczak[1]

[1] Université Louis Pasteur, LSIIT, CNRS, Strasbourg, France
[2] University of Wroclaw, Institute of Computer Science, Wroclaw, Poland
{lipinski, jjk}@dpt-info.u-strasbg.fr

Abstract. This paper addresses the problem of investment assessment and selection. A number of various performance measures are evaluated and studied. The goal of these investigations is to compare these performance measures on real-life data and to discover an optimal performance measure for selecting investment strategies in an evolutionary stock trading decision support system. Evaluations have been performed on financial time series from the Paris Stock Exchange.

1 Introduction

A trader on the stock market deals with the problem of selecting one of a number of possible investments. The trader evaluates the performance of investments, and decides to invest in the one whose performance is the highest. The performance measure depends on the trader's profit preferences, risk aversion and utility function.

Traders are interested in the profitability of their investments. Since the future return rates are unknown, traders must estimate them on the basis of a number of past observations. Certainly, traders also estimate the risk related to achieving the desired returns [4]. The expected return rate and the risk constitute the main factors of performance measures.

There are a large number of performance measures [1, 2, 12, 13] coming from several well-known stock market models. Each model makes some assumptions on, among other things, investment return distributions. In practice, the stock market does not perfectly suit any of these models, so the prediction may fail.

Traders use models and their theoretical background as the means to achieve profits. They are particularly interested in obtained return rates. This does not mean that traders are interested in high **expected** return rates – it means that traders are interested in high **achieved** return rates. The **expected** return rate is the rate estimated *a priori*, before a trader begins playing on the stock market over a specific time period. Certainly, it depends on the stock market model assumed. The **achieved** return rate is the rate evaluated *a posteriori*, after a trader completes playing on the stock market over a specific time period. It does not depend on the stock market model assumed because it is evaluated *a posteriori* on the basis of known return rates.

A trader selects an investment according to a chosen performance measure. The pragmatic question is how the performance relates to return rates achieved [8, 9]. Is the performance measure appropriate?

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 835–842, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper, a number of various performance measures are investigated and the relation between the evaluated performance and return rates achieved is studied. The goal of these investigations is to compare these performance measures on real-life data and to discover the optimal performance measure for selecting investment strategies in an evolutionary stock trading decision support system [6, 7]. This is crucial to the efficiency and the quality of elaborated trading expertise. Experiments were performed on financial time series from the Paris Stock Exchange.

This paper is structured in the following manner: Section 2 defines the problem. Section 3 presents first investigations on selecting one of two possible investments. Section 4 discusses applying various performance measure to four financial time series, which include price quotations of four stocks from the Paris Stock Exchange. Section 5 presents an evolutionary stock trading expert system. Section 6 discusses applying various performance measures in the system. Section 7 concludes the paper.

2 Problem Definition

At time t, a trader evaluates performance of all possible financial assets, and decides to invest in an asset A whose performance $P_t(A)$ is the highest. He may also estimate *a priori* a return rate $r^{(e)}_{t,t'}(A)$ of the asset A over the future time period $(t, t']$. Later, at time t', the trader evaluates *a posteriori* a return rate $r_{t,t'}(A)$ of the asset A achieved over the recently-passed time period $(t, t']$ and compares it to the expected return rate estimated *a priori*. Let $\delta_{t,t'}(A) = r_{t,t'}(A) - r^{(e)}_{t,t'}(A)$.

Even though the asset A had been selected because its performance $P_t(A)$ was the highest, sometimes either $r_{t,t'}(A)$ was average in comparison with return rates achieved by other investments, or $\delta_{t,t'}(A)$ was too large with respect to the risk estimated. If this occurs too often, the chosen performance measure may be inappropriate.

In this paper, classic performance measures are investigated, such as the Sharpe ratio (the ratio of the return rate to the standard deviation of the return rate), the Treynor ratio (the ratio of the return rate to the beta coefficient), the Jensen alpha, performance measures with asymmetric preferences, such as the Sortino ratio (the ratio of the return rate to the *SSD*), the ROAS (the ratio of the return rate to the *AS*), the ROPS (the ratio of the return rate to the *PS*), as well as practitioner performance measures, such as the Calmar ratio (the ratio of the return rate to the *MMD*) and the Sterling ratio (the ratio of the return rate to the average *MMD*), where

$$SSD = \sqrt{\text{avg}[(t-r)^2 : t > r)]}\,,\ AS = \text{avg}[(t-r) : t > r)]\,,\ PS = 13/20\,,$$

MMD is the maximum drawdown (i.e. maximum sustained drop over relevant period) and alpha and beta come from the *CAPM* model.

The first goal is to compare these performance measures using the financial time series from the Paris Stock Exchange. The second goal is to discover the optimal performance measure for selecting investment strategies in an evolutionary stock trading decision support system.

3 Selecting One of Two Investments

Let A and B be two available investments. At time t, a trader selects one of them according to a performance measure P. Later, at time t', the trader evaluates *a posteriori* the return rate $r_{t,t'}(A)$ and $r_{t,t'}(B)$ of both investments achieved over the recently-passed time period $(t, t']$ and checks whether the investment with the higher performance has also the higher return rate.

Such experiments were performed using two stocks, AXA and Peugeot, for various t and t' in a period from January 4, 1999 to November 7, 2003, using the Sharpe ratio as the performance measure P.

When AXA had a higher P_t (546 observations), the return rate of AXA was higher than the return rate of Peugeot in 299 observations. When Peugeot had a higher P_t (653 observations), the return rate of Peugeot was higher than the return rate of AXA in 369 observations. That gives the accuracy of the Sharpe ratio $\eta = (299 + 369) / (546 + 653) = 0.5571$. That means that the Sharpe ratio led to the correct investment selection in 55.71% of observations.

In further experiments, the trader selects an investment only if the difference between its performance and the performance of the other investment is significant, i.e. it is greater than a given threshold θ. For $\theta = 0.1$, the accuracy of the Sharpe ratio is $\eta = (137 + 233) / (243 + 384) = 0.5901$. For $\theta = 0.2$, the accuracy of the Sharpe ratio is $\eta = (53 + 146) / (97 + 232) = 0.6049$. Results of the other experiments are presented in Table 1.

Table 1. Accuracy of performance measures η for selecting one of two investments

P	θ	η
Sharpe ratio	0.00	0.5571
Sharpe ratio	0.10	0.5901
Sharpe ratio	0.20	0.6049
Sharpe ratio	0.30	0.6929
Sharpe ratio	0.40	**0.7813**
Treynor ratio	0.00	0.5388
Treynor ratio	0.25	0.5518
Treynor ratio	0.50	0.5780

P	θ	η
Sortino ratio	0.00	0.5588
Sortino ratio	0.10	0.5938
Sortino ratio	0.20	0.6197
Sortino ratio	0.30	**0.8571**
ROAS	0.00	0.4395
ROAS	0.20	0.3786
Sterling ratio	0.00	0.5613
Sterling ratio	0.15	0.6505
Sterling ratio	0.30	**0.7273**

Table 1 shows that relation between certain performance measures and return rates is weak (e.g. the ROAS ratio). However, some performance measures, such as the Sharpe ratio, the Sortino ratio and the Sterling ratio, may be used to select one of two investments if a proper threshold θ is defined.

4 Performance Measures on Financial Time Series

Let P be a specified performance measure. Let A be a financial asset. At time t, a trader evaluates a performance $P_t(A)$ and decides to invest in A. He also estimates *a priori* a return rate $r^{(e)}_{t,t'}(A)$ over the future time period $(t, t']$. Later, at time t', the

trader evaluates *a posteriori* the return rate $r_{t,t'}(A)$ achieved over the recently-passed time period $(t, t']$ and compares it to the expected return rate $r^{(e)}_{t,t'}(A)$ estimated *a priori*. Let $\delta_{t,t'}(A) = r_{t,t'}(A) - r^{(e)}_{t,t'}(A)$.

Such experiments were performed using four stocks from the Paris Stock Exchange, AXA, Credit Lyonnais, Peugeot and STMicroelectronics, for various t and t' in a period from January 4, 1999 to November 7, 2003 using various performance measures. A set of observations, consisting of $P_t(A)$, $r^{(e)}_{t,t'}(A)$, $r_{t,t'}(A)$ and $\delta_{t,t'}(A)$, was obtained for each stock and each performance measure. Since the results were quite similar, only experiments concerning AXA are presented in detail.

Figure 1 shows how $\delta_{t,t'}$ is related to the performance measure values P_t. Although no direct relation is evident, some dependencies may be observed. For instance (Figure 1a), one can notice that investments with a Sharpe ratio of $P_t < 0.02$ generally have a larger $\delta_{t,t'}$ than investments with a Sharpe ratio of $P_t > 0.02$. Thus, traders should rather focus on investments with a Sharpe ratio of $P_t > 0.02$. The remaining question is whether greater values of the Sharpe ratio P_t entail smaller values of $\delta_{t,t'}$. In general, the answer is no because, for instance, the spread of $\delta_{t,t'}$ seems to be quite similar for $P_t = 0.02$ and $P_t = 0.03$.

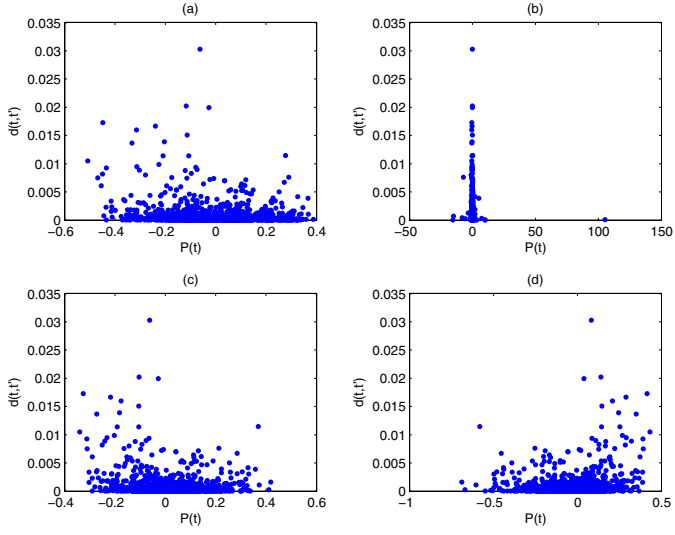


Fig. 1. Dependency of the difference d(t, t') between the achieved return rate and the expected return rate on the performance P(t) (a – Sharpe ratio, b – Treynor ratio, c – Sortino ratio, d - ROAS)

Moreover, experiments similar to those from the previous section were carried out. For each two observations, their performances and their achieved return rates were compared in order to verify whether the investment with the higher performance has also the higher return rate. Results of these experiments are presented in Table 2.

Table 2. Accuracy of performance measures η for selecting one of two investments

P	θ	η
Sharpe ratio	0.00	0.5517
Sharpe ratio	0.20	0.5974
Sharpe ratio	0.40	0.6761
Sharpe ratio	0.60	**0.7676**
Treynor ratio	0.00	0.5447
Treynor ratio	0.25	0.5713
Treynor ratio	0.50	0.5872
Sortino ratio	0.00	0.5509

P	θ	η
Sortino ratio	0.15	0.5930
Sortino ratio	0.30	0.6567
Sortino ratio	0.45	**0.7080**
ROAS	0.00	0.4496
ROAS	0.10	0.4305
Sterling ratio	0.00	0.5512
Sterling ratio	0.15	0.6215
Sterling ratio	0.30	**0.7454**

Table 2 conforms to results presented in the previous section. The Sharpe ratio, the Sortino ratio and the Sterling ratio are more accurate than the other measures.

Although the direct dependence $r_{t'}$ on P_t were not evident, it is interesting to estimate and compare the conditional probabilities $\mathbf{P}[r_{t'} > \theta_r \mid P_t \leq \theta_P]$ and $\mathbf{P}[r_{t'} > \theta_r \mid P_t > \theta_P]$ for certain thresholds θ_P and θ_r. Unfortunately, it requires some assumptions on return rate distributions, which might be hard to verify. However, frequencies of occurrences of observations with high $r_{t'}$ in a set of observations with high P_t were investigated. Let **U** denote the set of all observations. Let **U'** denote a set of observations with high P_t (i.e. $P_t > \theta_P$) and **V** denote a set of observations with high $r_{t'}$ (i.e. $r_{t'} > \theta_r$). Let $\alpha = |\mathbf{V} \cap \mathbf{U'}| / |\mathbf{U'}|$ and $\alpha' = |\mathbf{V} \cap \mathbf{(U\text{-}U')}| / |\mathbf{U\text{-}U'}|$. Table 3 presents results for various performance measures.

Table 3. Frequency of occurrence of observations with high $r_{t'}$

P	θ_P	θ_r	α	α'
Sharpe ratio	0.30	0.01	0.4063	0.3213
Sharpe ratio	0.30	0.02	0.3750	0.1894
Sortino ratio	0.25	0.01	0.5405	0.3167
Sortino ratio	0.25	0.02	0.4595	0.1859
Sterling ratio	0.20	0.01	0.4483	0.3205
Sterling ratio	0.20	0.02	0.3793	0.1897

Experiments prove that observations with high $r_{t'}$ occur much more frequently in the set of observations with high P_t than in the set of observations with average P_t. Certainly, the frequency is different for different performance measure.

To conclude, none of the performance measures considered in this paper constitute perfect criteria for investment selection. However, they significantly do increase the chance of selecting profitable investments. Applications of these performance measures in an evolutionary stock trading decision support system is presented in next sections.

5 An Evolutionary Stock Trading Expert System

Traders on the stock market observe quotations of stock prices with the aim to sell an item if it tends to lose value, to buy an item if it tends to gain value, and to take no

action in the remaining cases. Traders often assume that future values can be, more or less accurately, predicted on the basis of past observations. Many methods that analyze past data behavior were introduced [3, 11, 14]. These methods, herein referred to as trading rules, attempt to detect trends or discover contexts leading to occurrences of particular events, which for instance might cause a rise or fall in stock prices.

Let K_t denote the knowledge available at time t. This knowledge may represent historical data, previous stock quotations, or other information on the considered stock. The concept of a stock market trading rule may be formalized by a function f, which computes a decision $f(K_t) \in \{-1.0 \equiv sell, 0.0 \equiv do\ nothing, 1.0 \equiv buy\}$ on the basis of the knowledge K_t available at time t. Naturally, the function f may be defined in a variety of ways.

Although the number of trading rules commonly used by financial experts and market traders is large, some rules may be effective during one time period and less effective during another time period, some rules may work well only in combination with others. At time t, each trader bases his decision on a certain set of trading rules. The trader chooses the set of rules according to criteria defined by his preferences, concerning, for instance, expected profit rate and risk aversion. Depending on the choice of rules, the trader may receive different advice. The question is which rules the trader should choose.

Let e be a subset of the entire set of trading rules $\{ f_1, f_2, \ldots, f_N \}$. Such a subset will be referred to as a stock market trading expert. In a natural way, the expert e can be presented as a binary vector of length N. The i-th coordinate of the vector corresponds to the i-th rule, where 0 stands for absence of the rule and 1 stands for presence of the rule.

A result $r_e(t)$ of the expert e at time t is defined as the average of results of trading rules included in the expert. The result $r_e(t)$ close to –1 corresponds to advice to sell, close to 1 to advice to buy, and otherwise indicates no recommendation, i.e. is interpreted as advice to do nothing. Let $d_e(t)$ denote the decision of the expert e at time t.

Since the number of trading rules is large ($N = 350$ in these investigations), the number of possible experts is enormous (equal to $2^N = 2^{350}$), making the process of manual selection impractical, so a computer decision support system is proposed to make this selection.

The optimal expert is chosen by an evolutionary algorithm [5, 10] according to a given performance measure. Certainly, different performance measures lead to different optimal experts. The question is which performance measure should be applied in order to obtain the most profitable experts.

6 Performance Measures in the Evolutionary Expert System

This section discusses applying performance measures presented in Section 2 to assessment and selection of experts in the evolutionary expert system.

Experiments were performed on four stocks from the Paris Stock Exchange, AXA, Credit Lyonnais, Peugeot and STMicroelectronics, for specified t and t' in a period

from January 4, 1999 to November 7, 2003 using various performance measures. For each stock and each performance measure, a set of observations, which consist of P_t and $r_{t'}$ evaluated on 8000 trading experts, was obtained. 4000 of experts were randomly generated, so their values P_t were average, and 4000 were optimized by the evolutionary algorithm with a relevant objective function, so their values P_t were high.

Although the direct dependence $r_{t'}$ on P_t were not evident, frequencies of highly profitable expert occurrences in a set of experts with high P_t were investigated. Let **U** denote the entire set of all 8000 experts. Let **U'** denote a set of experts with high P_t (i.e. $P_t > \theta_P$, for a certain threshold θ_P) and **V** denote a set of experts with high $r_{t'}$ (i.e. $r_{t'} > \theta_r$, for a certain threshold θ_r). Let $\alpha = |\mathbf{V} \cap \mathbf{U'}| / |\mathbf{U'}|$ and $\alpha' = |\mathbf{V} \cap \mathbf{(U-U')}| / |\mathbf{U-U'}|$. Table 4 presents results for various performance measures.

Table 4. Frequency of highly profitable expert occurrences

P	θ_P	θ_r	α	α'
Sharpe ratio	0.10	0.045000	0.9514	0.7481
Sharpe ratio	0.10	0.045250	0.9205	0.5912
Sortino ratio	0.06	0.045000	0.9442	0.7280
Sortino ratio	0.06	0.045250	0.9050	0.6131
Sterling ratio	0.06	0.045000	0.9769	0.7479
Sterling ratio	0.06	0.045250	0.9551	0.6006

Experiments prove that highly profitable experts occur much more frequently in the set of experts with high P_t than in the set of experts with average P_t. Although the frequency is different for different performance measures, none of the performance measures may be used itself as perfectly efficient selection criteria. It may be interesting to build a multi-objective performance measure based on performance measures investigated in this paper.

7 Conclusions

In this paper, a problem of investment assessment and selection was considered. A number of various performance measures were evaluated. Relation between the performance evaluated and return rates achieved was investigated. A few relatively efficient performance measures were applied to select investment strategies in an evolutionary stock trading decision support system.

A large number of experiments were carried out in order to compare various performance measures. All experiments were performed on real-life data from the Paris Stock Exchange. Experiments showed that the most efficient performance measures are the Sharpe ratio, the Sortino ratio and the Sterling ratio. However, the efficiency may be slightly different over different time periods. Therefore, the presented hypotheses have to be also tested on other markets with variable time horizon.

Although the direct relation between evaluated performance and future return rates is weak and none of the considered measures may be used itself as a perfectly

efficient selection criterion, several of them significantly increase the chance of selecting a highly profitable investment. Thus, they may be used to build a common multi-objective performance measure. Further research on these issues may lead to valuable results.

Acknowledgements. The authors would like to thank Prof. Patrick Roger from the Laboratoire de Recherche en Gestion et Economie, Université Louis Pasteur, Strasbourg, France for his helpful comments on the preliminary version of this paper.

References

1. Aftalion, F., Poncet, P., "Les Techniques de Mesure de Performance", Economica, 2003.
2. Cantaluppi, L., Hug, R., "Efficiency Ratio: A New Methodology for Performance Measurement", Journal of Investing, 2000, pp.19-25.
3. Colby, W., Meyers, T., "The Encyclopedia of Technical Market Indicators", Down Jones-Irwin, 1990.
4. Jensen, M., "Risk, the pricing of capital assets, and the evaluation of investment performances", Journal of Business, 1968, pp.167-247.
5. Goldberg, D. E., "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley, 1989.
6. Korczak, J., Lipinski, P., Roger, P., "Evolution Strategy in Portfolio Optimization", Artificial Evolution, ed. P. Collet, Lecture Notes in Computer Science, vol. 2310, Springer, 2002, pp.156-167.
7. Korczak, J., Roger, P., "Stock timing using genetic algorithms", Applied Stochastic Models in Business and Industry, 2002, pp.121-134.
8. Lehman, B., Modest, D., "Mutual Fund Performance Evaluation: a Comparison of Benchmarks and Benchmarks Comparison", Journal of Finance, 1987, pp.233-265.
9. Lo, A., W., "The Statistics of Sharpe Ratios", Financial Analysts Journal, 2002, pp.36-52.
10. Michalewicz, Z., "Genetic Algorithms + Data Structures = Evolution Programs", Springer Verlag, New York, 1994.
11. Murphy, J., "Technical Analysis of the Financial Markets", NUIF, 1998.
12. Sharpe, W., F., "Mutual Fund Performance", Journal of Businness, 1966, pp.119-138.
13. Sharpe, W., F., "The Sharpe Ratio", The Journal of Portfolio Management, 1994.
14. Weigend, A., S., Gershenfeld, N., A., "Time Series Prediction: Forecasting the Future and Understanding the Past", Addison-Wesley, 1993.

Stocks' Trading System Based on the Particle Swarm Optimization Algorithm

Jovita Nenortaite and Rimvydas Simutis

Vilnius University Kaunas Faculty of Humanities,
Muitines st. 8,3000 Kaunas, Lithuania
{jovita.nenortaite, rimvydas.simutis}@vukhf.lt

Abstract. One of the central problems in financial markets is to make the profitable stocks trading decisions using historical stocks' market data. This paper presents the decision-making method which is based on the application of neural networks (NN) and swarm intelligence technologies and is used to generate one-step ahead investment decisions. In brief, the analysis of historical stocks prices variations is made using "single layer" NN, and subsequently the Particle Swarm Optimization (PSO) algorithm is applied in order to select "global best" NN for the future investment decisions and to adapt the weights of other networks towards the weights of the best network. The experimental investigations were made considering different number of NN, moving time intervals and commission fees. The experimental results presented in the paper show that the application of our proposed method lets to achieve better results than the average of the market.

1 Introduction

The continuing improvements of computer technologies, telecommunication services' grow make a big influence on globalization of stock markets and more efficient its information processing tools are required. The complexity and "noisiness" of stock markets cause difficulties in making real time analysis of it and forecasting its changes in the future. It was proved that having complex systems a collection of individuals often solves a problem better than an individual - even an expert [1]. Individuals acting within "a swarm" interact with each other in order to solve a global objective in a more efficient manner than a single individual could [8].

The main objective of this paper is to develop the method based on artificial intelligence tools, which includes artificial NN, swarm intelligence technologies (particle swarm optimization), and apply it for the decisions-making in stocks' trading markets.

PSO algorithm is one of the swarm intelligence algorithms. It is a population-based search algorithm based on the simulation of the social behavior among individuals (particles) "moving" through a multidimensional search space. Each particle of the swarm represents a candidate solution to the optimization of the problem. While each particle is "flown" through the multidimensional search

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 843–850, 2004.
© Springer-Verlag Berlin Heidelberg 2004

space it is adjusting its position in search space according to its own experience and that of neighboring particles [3]. The PSO is closely related to evolutionary computation and artificial life (A-life) in general. The same as evolutionary programming it is highly dependent on stochastic processes. The optimizer which is used in the PSO algorithm, while making adjustment towards "local" and "global" best particles, is conceptually similar to the crossover operation used by genetic algorithms [8]. As well PSO includes fitness function, which measures the closeness of the corresponding solution to the optimum. The same function is included in the paradigms of evolutionary computation. The main difference of PSO concept from the evolutionary computing is that flying potential solutions through hyperspace are accelerating toward "better" solutions, while in evolutionary computation schemes operate directly on potential solutions which are represented as locations in hyperspace [7].

The problem of stock markets forecasting was analyzed by many researchers in the past. Considerable efforts have been put into investigation of stock markets changes and creating its forecasting systems. There are not so many examples of Swarm Intelligence applications for the solving of this problem. However, the published examples of swarm intelligence applications seem to be promising and give good results. In paper [10] there is proposed the forecasting methodology for the daily exchange rates of Japanese Yen against the US Dollar and of the US Dollar to the British Pound. The proposed forecasting methodology includes clustering technology, artificial NN and evolutionary computing. In contrast to this paper we focus on the formation of recommendations while making investment decision in stocks' markets. As well we are working with a large data set that lets us to propose more stable investment decision system. In paper [2] the authors are focusing on adapting PSO to dynamic environments. This paper is more focused on the modifications of PSO algorithm, while our goal is to introduce the investment decision-making method where PSO will be only one constituent of it.

The paper is organized as follows: Section 2 presents the problem and describes the proposed method; Section 3 is devoted for the implementation of PSO algorithms; In Section 3 the experimental results are discussed. The main conclusions of the work are presented in Section 4.

2 Description of Proposed Method

The analysis and forecast of stock market variations is stickler, because of its complexity and noisiness. There is not enough to use conventional techniques to conduct the stock markets predictions, as its changes are influenced by stochastic human factors, nonlinear, multivariable and temporal nature of stock price transitions. The use of artificial intelligence had made a big influence on the forecasting and investment decision-making technologies and it was proved that the efficient results can be obtained [4], [6]. In this paper we are proposing a decision making method, which is based on the study of historical data, the use of NN and PSO algorithm.

In experimental investigations we are using 350 stocks, which are taken from the list of SP500 index. The data set represents stocks' price changes for 12 years (01/Oct/91-01/Oct/03). In our method we are using "single layer" NN, with the designation "single layer" referring to the output layer of computation nodes (neurons). The signals (data set) are passed to the cell body. Once they reach it, they are combined additively. The net result is a linear combination of each of the weighted input vectors. The NN' weights are initialized randomly at the beginning of the procedure. The random numbers are kept small, and symmetrical about zero. In the investigations there are considered stocks' prices changes for 5 days. The changes of 5 days were taken making an assumption that such stocks' prices changes have the highest correlation with the next day stocks price change [11]. For example, if stocks' price changes of five days are negative, there is a high possibility that the price of these stocks will increase on the next day and vise versa. Further, for each day, and each stock the recommendations are calculated using different number of "single layer" NN and taking into account prices changes of different number of days. The net result is passed to the hyperbolic tangent function and the recommendations for the stocks' trading are calculated. The recommendations represent the relative rank of investment attraction to each stock in the interval $[-1, 1]$. The values $-1, 0$ and $+1$ represent recommendations: sell, hold and buy respectively. After the recommendations are calculated, all the stocks are sorted in the descending order. Having sorted stocks we are taking into account 3 stocks (for each day) that have the highest recommendations. Further, the observation of these stocks price changes behavior on the next day is made considering different length moving time intervals. This observation lets us to see how good results could we get making the decisions according to the price changes of the stocks with the highest recommendations. The next step in our method is the use of PSO algorithm and the selection of "global best" particles, which will be used for the calculations of expected profit. The more detailed explanation of PSO algorithm is presented in the next section.

3 Application of Particle Swarm Optimization Algorithm

In this paper our objective is to apply the PSO algorithm, which is based on the search of "global best" particle. Here particles are represented as "single layer" NN. So called "global best" particle is chosen for every day taking into account the chosen moving time interval, it means that every day we are comparing the performance (fitness function) of NN and the network with the highest performance is chosen for further experimental investigations. The adaptation of the other particles weights is made towards the weights of "global best" particle. Such adaptation of weights and training of NN let us to move towards the best solution as all the time the trading decision are made using NN that have shown the best performance. Knowing the day and NN, which performance was the best on that day, we are selecting 3 stocks with the highest recommendations. Further the calculation of the prices' change mean of 3 stocks with the highest

recommendations is made. We believe that the investigation of stocks with the highest recommendations will let to achieve the best gain on the next trading day. More detailed explanations of experimental investigations are presented in the next Section.

4 Experimental Investigations

The realization of our proposed method was made using MATLAB software package. All our made experimental investigations could be described in several functional steps:

1. Stocks' prices relative changes (%) for the time period 01/Oct/91–01/Oct/03 were calculated.
2. The recommendations for the purchase of the stock were formed using "single layer" NN.
3. All the recommendations were sorted and for every day three stocks with the highest recommendations were chosen.
4. The price changes of chosen stocks on the next day were explored.
5. There were calculated expected day profits, taking into account the price changes of chosen stocks and the profit estimations was made selecting different moving time intervals. This profit was calculated as a sum of stock prices changes (%).
6. The PSO algorithm was applied and the profit estimation based on the search of "global best" particles was made.
7. In parallel there was estimated the profit, making an assumption that the investments were made into SP500 index.
8. The experiments were run taking into account different number of NN and different size moving time intervals in order to find the situations when the best and the most stable results could be achieved.

4.1 Selection of Moving Time Interval and Number of Neural Networks

The first part of experimental investigations was focused on the choice of optimal number of NN and on the choice of the size of moving time interval. The experiments were run taking into account different size moving time intervals and different number of NN. The number of explored stocks and days was the same for all the cases. In these experimental investigations the commission fee for selling and buying stocks was not considered. The obtained results let us to make several conclusions:

- The bigger number of NN, with different initial weights, let us to achieve more stable results (see Fig. 1 and 2).
- The bigger moving time intervals let to avoid unnecessary variations and to achieve better results (see Fig. 3).

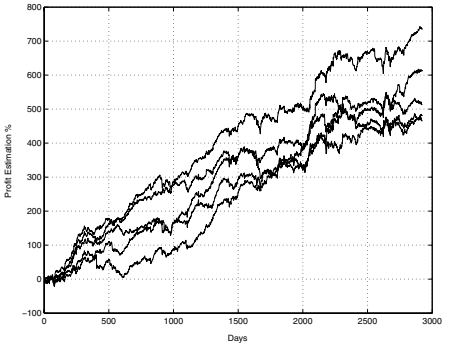


Fig. 1. Profit estimation (5 NN, moving time interval 100 days)

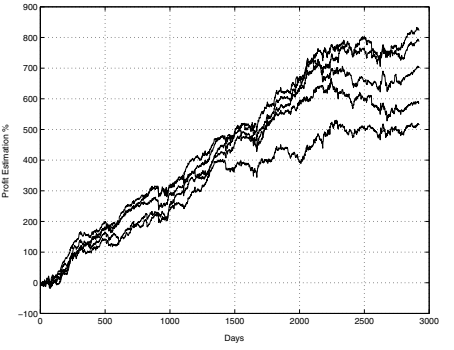


Fig. 2. Profit estimation (30 NN, moving time interval 100 days)

– The experimental investigations showed that the best results are achieved while taking 30 NN and the moving time interval of 100 days.

From the Fig. 1 and 2 it can be seen that the swing of the results is much smaller when there were taken 30 NN. As well the experimental investigations showed that the results are influenced by the choice of moving time interval size. Fig. 3 shows how the profit (% per day) is correlated with the different size moving time intervals while exploring 30 NN. Here numbers 1, 2, ...,10 represent time intervals 10, 20, ..., 100 respectively.

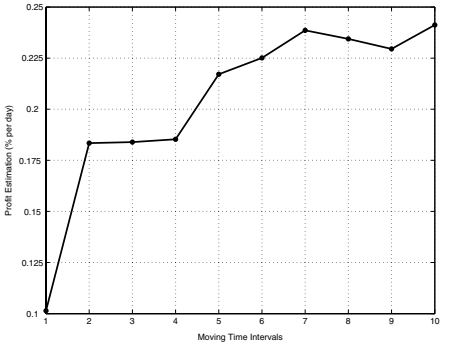


Fig. 3. Correlation between profit and moving time intervals

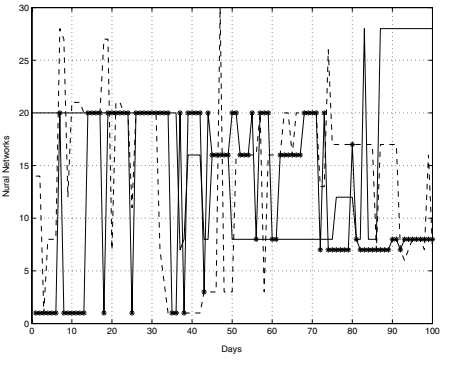


Fig. 4. The swing of NN depending from the moving time interval

As it can be seen from the Fig. 3, the profit is growing while increasing moving time intervals. The variations of the profit become more stable while having moving time intervals from 70 to 100 days (see Fig. 4). The Fig. 4 shows that having moving time interval of 100 days (solid line) the swing between NN (having 30 NN) is much smaller than having the moving time interval of 10

days (dashed line) or 50 days (bold line). Here the swing of NN is meant to be the choice of different "global best" NN depending from the time period. For example: having moving time interval of 100 days (solid line) we can see that starting from the 1st day and until the 38th day the "global best" NN is 20th, while during the same time period, having moving time intervals of 10 and 50 days the "global best" NN are changing almost every 3rd or 5th days. The results could be even more increased taking into account more NN and bigger moving time interval. In this paper there were not made such experimental investigations because of the limited computer capacity.

The comparison of profit estimation results using our proposed method (having 30 NN and moving time interval of 100 days) and results achieved while making investments into SP500 index are presented in Fig. 5.

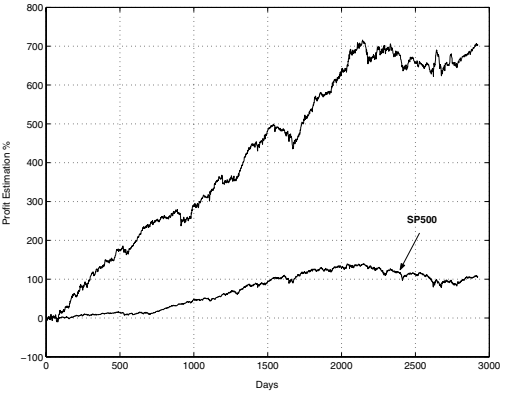


Fig. 5. Comparison of the results

As it can be seen from the Fig. 5 the application of PSO algorithm gives very good results. The profit accumulation results are 5–7 times better than the results achieved making investments into SP500 index. The obtained results confirm that using our method and not paying any commission fee we are able to beat the market. The good results are achieved as every day, while making the desicion, there is made an adaptation of particles' weights to the weights of "global best" and this lets to position all the particles towards the fitness of best particle.

4.2 The Estimation of Profit Considering Account Commission Fee

In the second part of experimental investigations we were focusing on the estimation of the profit considering commission fee. In the first part of experimental investigations we have got results showing, that having 30 NN, moving time interval of 100 days and not considering commission fee we are able to get 0.2344

% average incomes per day. The analysis of the commission fees of different e-brokers showed that commission fee in real trading process is usually between 0.15 % – 0.3 %. For example, such transition fees are provided by the company of Interactive Brokers [5]. Having bigger selling and buying volumes this fee could be even smaller - 0.1 %. Based on that for further investigations we are considering the commission fee which is equal 0.15 %. We are making an assumption that on the first trading day we are investing 1000 $ into the market. Each day we are paying 0.15 % of commission fee for buying new stocks. The value got on the last investigated trading day is considered as the profit. For the trading of SP500 index the commission fee is not considered. The experimental investigations showed that having the commission fee of 0.15 % we are able to earn more than the average of the market (see Fig. 6).

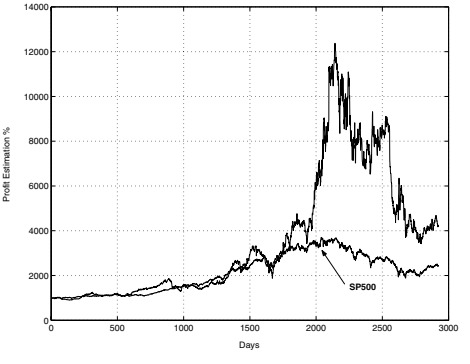


Fig. 6. Profit estimation considering commission fee of 0.15 %

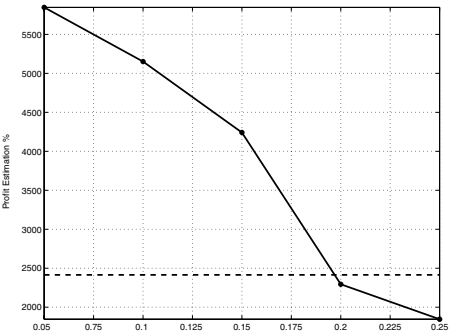


Fig. 7. Profit dependence from the commission fee

The situation is changing while the commission fee is increasing (see Fig. 7). While having the commission fee of 0.2 % we are already slightly loosing compare to the profit that has been got from the investment into SP500 index. The experimental investigations show, that using our proposed method are able to win only big funds that have a possibility to pay small (0.15 %) commission fees for buying and selling stocks. The experimental investigations have showed that the results, achieved while applying our proposed method, are much better as it was got using NN for the investigation of similar time series and its future changes forecasting [9],[12]. However, the proposed method still has to be improved in order to achieve better results while considering bigger commission fees.

5 Conclusions and Future Work

In this paper we proposed the investment method based on PSO algorithm. The method was applied in order to make one-step ahead profit estimation conside-

ring historical data of stocks' price changes. The experimental investigations let as to make the following conclusions:

1. The bigger number of NN and longer moving time interval let us to achieve better and more stable results.
2. The profit accumulation results, while not considering any commission fees, were 5–7 times better than the results achieved making investments into SP500 index.
3. Taking into account the commission fee of 0.15 % the proposed method gives better results than the average of the market.
4. The bigger commission fees (around 0.2 %) reduce the results and the investment into SP500 index becomes more reasonable.
5. The propose method could be useful for big financial funds that are able to pay small commissions fees for buying and selling stocks.

In the future we intend to make more detailed analysis of our proposed method.

References

1. Bartholdson K., Mauboussin J.M.: Thoughts on Organizing for Investing Success. Credit Suisse First Boston Equity Research (2002)
2. Carlisle A., Dozier G.: Adapting Particle Swarm Optimization to Dynamic Environments. 2000 ICAI Proceedings, Las Vegas (2000) 429–434
3. Engelbrecht A.D.: Computational Intelligence (An Introduction). John Wiley & Sons, London (2002)
4. Hellstrom T.: Optimizing the Sharpe Ration for a Rank Based Trading System. Lecture Notes in Artificial Intelligence, LNA 2258. Springer-Verlag, New York (2001)
5. Interactive Brokers: *http://www.interactivebrokers.com*. Current as of February 9th, 2004.
6. Kaastra I., Milton B.: Designing a Neural Network for Forecasting Financial and Economic Time Series. Neurocomputing (1996)
7. Kennedy J., Spears W.M.: Matching Algorithms to Problems: An Experimental Test of the Particle Swarm and Some Genetic Algorithms on the Multimodal Problem Generator. *http://www.aic.nrl.navy.mil/%7Espears/papers/wcci98.pdf*. Current as of December 15th, 2003.
8. Khalil A.S.: An Investigation into Optimization Strategies of Genetic Algorithms and Swarm Intelligence. Artificial Life (2001)
9. Lowe D., Webb A.R.: Time Series Prediction by Adaptive Networks: A Dynamical Systems Perspective. IEEE Computer Society Press (1991)
10. Pavlidis N.G., Tasoulis D.,Vrahatis M.N.: Financial Forecasting Through Unsupervised Clustering and Evolutionary Trained Neural Networks. 2003 Congress on Evolutionary Computation, Canberra Australia (2003)
11. Simutis R.: Stock Trading Systems Based on Stock's Price Ranks (in Lithuanian). Ekonomika (2003)
12. White H.: Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. IEEE International Conference on Neural Networks, San Diego (1988)

Parisian Options – The Implied Barrier Concept

Jasper Anderluh and Hans van der Weide

Delft Institute of Applied Mathematics, Delft University of Technology
Mekelweg 4, 2628 CD, Delft, The Netherlands
j.h.m.anderluh@ewi.tudelft.nl

Abstract. Research into the direction of specific exotic options - like the Parisians - is often driven by the analysis of structured products. These products contain features that are similar to exotic options. Exchange-trading of the pure exotics is very rare. In the period of rising stock markets, investors were less interested in buying bonds. In order to regain their interest, firms added extra features to the bonds they wanted to issue. One of these features is the right of the bond holder to convert the bond into a given number of stocks under certain conditions. Bonds with this feature are called convertible bonds and are nowadays very common. Most convertible bonds can be re-called by the issuer when the convertible trades above some level for some period. Modelling this feature corresponds to valuation of a Parisian option. In this paper we will point out how we quickly can approximate the Parisian option price by using a standard barrier option with a modified barrier. This is common practice for traders; they increase or decrease the barrier a bit. Here we want to argue what that bit should be. First we will introduce the Parisian contract. Thereafter we list the methods of valuing the Parisian, followed by a section about the implied barrier method. Here we will use concepts from the theory on Brownian excursions and exploit them to derive prices for Parisians that are already in the excursion. We will conclude with a numerical example.

1 The Parisian Contract

Let $\{S_t, \mathcal{F}_t; t \geq 0\}$ be a process defined on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. According to the Black-Scholes model we have for the risk neutral price process

$$S_t = s_0 \exp\left((r - 0.5\sigma^2)t + \sigma B_t\right),$$

where $\{B_t, \mathcal{F}_t; t \geq 0\}$ denotes a standard Brownian Motion, s_0 the initial value of the stock, r the interest rate and σ the volatility. We can use this risk-neutral stock price process to calculate the price of a derivative V_Φ with some (path dependent) pay-off $\Phi\left((S_t)_{0\leq t\leq T}\right)$ at time T by,

$$V_\Phi = e^{-rT}\mathbb{E}\left[\Phi\left((S_t)_{0\leq t\leq T}\right)\right].$$

Here Φ is the contract function. A standard barrier option is a derivative that pays off like a put or a call that knocks in or out as soon as the stock price hits

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 851–858, 2004.
© Springer-Verlag Berlin Heidelberg 2004

some level. The Parisian option is like a barrier, but we do not only assume the stock price to hit some level, but also to stay above or below that level for a given time. For a given t let (γ_t, β_t) be the excursion from level L straddling t, i.e.

$$\gamma_t := \sup_{s \leq t} \{S_s = L\} \quad \text{and} \quad \beta_t := \inf_{s \geq t} \{S_s = L\}.$$

Now we can use these random variables to define T_D^-, the first time $(S_t)_{t \geq 0}$ is below L longer than D and T_D^+ the equivalent random time for staying above L by

$$T_D^- := \inf_{t>0} \{t - \gamma_t > D; S_t < L\} \quad \text{and} \quad T_D^+ := \inf_{t>0} \{t - \gamma_t > D; S_t > L\}.$$

For an up and down Parisian we use T_D^+ respectively T_D^-. If it is a knock in we consider the set $\{T_D^{\pm} \leq T\}$ otherwise we consider its complement. With this notation we can write pricing formulae for all kinds of Parisian options, e.g. the Parisian down and in call is given by

$$V_{PDIC} = e^{-rT} \mathbb{E}\left[(S_T - X)^+ 1_{\{T_D^- \leq T\}} \right],$$

where X denotes the strike of the call. By changing the set of the indicator, we can calculate all variations of down/out and up/in. By changing the left part of the pay-off, we can compute prices for Parisian puts instead of calls.

2 Valuation Methods for Parisians

There are two well-known methods for valuing Parisians. One method [6] uses the Black-Scholes PDE with boundary conditions adapted to the Parisian contract. This modified PDE is then solved by numerical methods. The advantage of this method is, that the pricing framework is flexible, i.e. it is possible to deal with discrete dividends and early exercise. The disadvantage is that the convergence is slow. There are methods using some pre-calculated values [5] to increase this convergence. This method is relatively time-consuming in the sense that it is inappropriate for real-time option pricing

Another method calculates the Laplace Transforms of these option prices [2]. There exist quick inversion algorithms [1] to obtain the prices within a given error bound.

Here we propose another method, the so-called implied barrier method. This method consists of coming up with an implied barrier L^*. Then the standard barrier option with barrier L^* and the remaining parameters equal to the Parisian has the same price as the Parisian. This method only uses standard barrier formulae for the European case and obtaining prices is neither difficult nor time-consuming.

3 The Implied Barrier

Explaining the implied barrier method is threefold. At first we define the implied barrier and give a numerical example. Then we come up with heuristics about what the implied barrier should be and use excursion theory to obtain a formula for it. Finally, we will show for a practical example that prices are quite close. Denoting the first time (S_t) hits L by T_L, the price of a call with barrier L^* is given by

$$V_{DIC} := e^{-rT}\mathbb{E}\left[(S_T - X)^+ 1_{\{T_{L^*} \leq T\}}\right].$$

We know that for $L = L^*$ the Parisian option should have a value smaller than this standard barrier, but positive. By decreasing L^* we can have the value of this barrier every value between the $(L^* = L)$-case and 0, so for some particular L^* the standard barrier has the same price as the Parisian. This particular value of L^* is defined as the implied barrier and can formally be stated as:

Definition
Let $\Phi(S_T)$ the vanilla part of the payoff of a Parisian Up contract. Then for $L > S_0$ the implied barrier L^* is defined by

$$\mathbb{E}[\Phi(S_T)1_{\{T^+_{D,L} \leq T\}}] = \mathbb{E}[\Phi(S_T)1_{\{T_{L^*} \leq T\}}].$$

For $L < S_0$ we define the implied barrier L^* by

$$\mathbb{E}[\Phi(ST)1_{\{T^-_{D,L} \leq T\}}] = \mathbb{E}[\Phi(S_T)1_{\{T_{L^*} \leq T\}}].$$

By continuity it is clear that L^* always exists. In the graph in figure 1 we show the implied barrier for a down-and-in Parisian call for different times to maturity and different values of s_0. The Parisian prices we needed for finding this barrier are computed by a numerical inversion of the Laplace Transforms. For this particular option the graph shows that the implied barrier is in the same range for different T and s_0. However, by basic monotonicity arguments we have the following proposition.

Proposition
Let $\Phi(S_T)$ be the vanilla part of a Parisian contract. Suppose for a down contract with $L < S_0$ that for some finite $T > D > 0$ we have

$$\mathbb{E}[\Phi(S_T)1_{\{T^-_{D,L} \leq T\}}] > 0.$$

Then $L^*(t)$, the implied barrier for this contract with maturity t cannot be the same for every $t \in [D, T]$. Moreover $L^*(s_0)$, the implied barrier for this contract with initial stock price s_0 cannot be the same for every $s_0 > L$.

Considering the scale of the vertical axis in the figure on the next page, we will try to approximate this implied barrier by a constant. In the next section we arrive intuitively at the approximation. Using this approximation we come up with a numerical example.

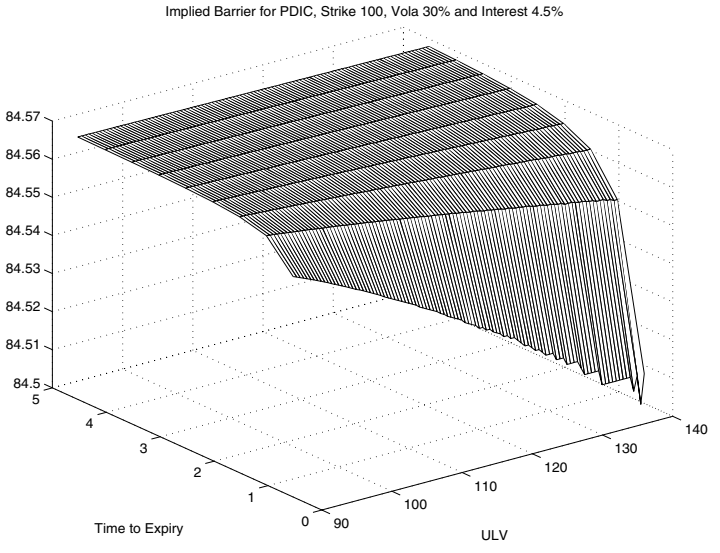


Fig. 1. Implied Barrier vs s_0 and T

4 Approximating the Implied Barrier

In order to arrive at the approximation, we argue that the implied barrier should be - whether the contract is of the up or down type - the minimum or maximum we expect (S_t) to attain, given that the stock follows a path in the Parisian set. For the Parisian contracts we are interested in the set $\{T_D^{\pm} \leq T\}$ and its complement. So, either the realizations of (S_t) do contain an excursion in the right direction w.r.t. to L longer than D or not. If we consider for example the down and in version and divide the stock path into its excursions from L, we can compute the expected minimum that is attained in each separate excursion below L. The minimum the stock path attains up to time T is the minimum of all these expected minima per excursion. We know from excursion theory that the expected minimum of an excursion decreases when the length of the excursion increases. In this down and in case we know that we need to have at least one excursion longer than D, so the expected minimum of the knocking-in stock paths should be below the minimum that we expect to attain during an excursion below L of length D. Using Brownian excursion theory [3,4] we have for the expected minimum m_D of an Brownian excursion with length D

$$\mathbb{E}[m_D] = \sqrt{\frac{D\pi}{2}}.$$

As we are dealing with geometric Brownian motion, this term will appear within an exponential. Furthermore we are considering an excursion around the level L, so we find the following approximation $\tilde{L}^*$ for the implied barrier L^*

$$\tilde{L}^* := L \exp\left(-\sigma\sqrt{D}e^{-\frac{m^2}{2}D}\sqrt{\frac{\pi}{2}}\right) \quad \text{where} \quad m := \frac{r - 0.5\sigma^2}{\sigma}.$$

Here the m is introduced as a consequence of the Girsanov transformation, that is needed in cases where $r \neq 0.5\sigma^2$. The same argumentation holds for the other types of Parisians (up/down, out/in), where the minus sign in the exponent disappears for the up-variant of the Parisian contract. Now we use the $\tilde{L}^*$ to approximate the value $V_{PDIC}(L)$ by $V_{DIC}(\tilde{L}^*)$. In the next section we show how we deal with Parisians that are already in the excursion.

5 Valuing Parisians That Are Already in the Excursion

Here we valuate a Parisian option at time $t > 0$. Suppose we are interested in the $PDIC$ with some barrier L. In the case $S_t < L$, the excursion of the stock path below L has already started and this effects its value. Define d as the time the stock already spent below L by

$$d := t - \gamma_t.$$

Here we are interested in the case where $d < D$ and another excursion longer than D did not happen yet. It is clear that we have two possibilities now. Either we return to the level L too early to knock in, i.e. we are in the set $\{T_L \leq D - d\}$, or, we are in the complement of this set, and so stay long enough to knock in. Here we abuse the notation T_L for the amount of time we have to wait after time t to hit the level L, that we should write as T_L^t given by

$$T_L^t = \inf_{s>0}\{S_{s+t} = L\}.$$

In the same way we will abuse the notation of $T_{D,L}^-$. Considering these two cases we could be in, the value of a PDIC is given by

$$\begin{aligned} V_{PDIC}(d) =& e^{-r(T-t)}\mathbb{E}\left[(S_T - X)^+ 1_{\{T_L > D-d\}}|\mathcal{F}_t\right] + \\ & e^{-r(T-t)}\mathbb{E}\left[(S_T - X)^+ 1_{\{T_{D,L}^- < T-t\}} 1_{\{T_L \leq D-d\}}|\mathcal{F}_t\right]. \end{aligned}$$

So, in the case that we stay long enough below L we get a standard call, otherwise we get a fresh $PDIC$ for the remaining time to expiry. If we write $1_{\{T_L > D-d\}} = 1 - 1_{\{T_L \leq D-d\}}$, we find for $V_{PDIC}(d)$

$$V_{PDIC}(d) = V_{BSC} - e^{-rT'}\mathbb{E}\left[(S_t e^{(r-\sigma^2/2)T' + \sigma B_{T'}} - X)^+ 1_{\{T_{D,L}^- > T'\}} 1_{\{T_L \leq D-d\}}\right].$$

Here V_{BSC} is the value of a standard call with time to expiration $T' := T - t$ and we used the Markov property to remove the conditioning. Now we change our measure to $\mathbb{Q}$ in the same way as above. Conditioning on $\mathcal{F}_{T_L}$ gives

$$\begin{aligned} V_{PDIC}(d) =& V_{BSC} - \\ & e^{-\tilde{r}T'}\mathbb{E}_{\mathbb{Q}}\left[1_{\{T_L \leq D-d\}}\mathbb{E}_{\mathbb{Q}}\left[e^{mZ_{T'}}(S_t e^{\sigma Z_{T'}} - X)^+ 1_{\{T_{D,L}^- > T'\}}|\mathcal{F}_{T_L}\right]\right], \end{aligned}$$

where $\tilde{r}$ is defined by $\tilde{r} := r - 0.5m^2$. Using the Markov property again, we recognize $V_{PDOC}(T' - T_L)$ apart from the proper discounting, so we get

$$V_{PDIC}(d) := V_{BSC} - e^{ml}\mathbb{E}_{\mathbb{Q}}\left[1_{\{T_L \leq D-d\}} e^{-\tilde{r}T_L} V_{PDOC}(T' - T_L)\right],$$

where l is defined by $l = \sigma^{-1} \ln \frac{L}{S_t}$. Using the distribution of the hitting time [3] we can calculate the expectation. If we furthermore approximate $V_{PDOC}(T'-T_L)$ by $V_{DOC}(\tilde{L}^*, T' - T_L)$ we obtain

$$V_{PDIC}(d) := V_{BSC} - \int_0^{D-d} V_{DOC}(L^*, T' - u) \frac{|l| e^{ml - \tilde{r}u - \frac{l^2}{2u}}}{\sqrt{2\pi u^3}} du.$$

This integral should be computed numerically. In the example later on we will show that a relatively small number of steps is sufficient to obtain a value close to the true Parisian price. The same way of reasoning could be used to derive prices for the other types of Parisian contracts.

6 Numerical Examples

As we presented our method as one that can be very useful in practice, we will give some numerical results. Table 1 contains a comparison between implied barrier prices and Laplace prices.

Table 1. Prices for PDIC at $t = 0$

	D = 10/365				D = 20/365				D=200/365			
S_0=100, L=90	Laplace		Imp Bar		Laplace		Imp Bar		Laplace		Imp Bar	
X	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2
80	6.54	11.36	6.55	11.36	4.97	9.42	4.99	9.43	0.26	1.68	0.35	1.77
r=4.5% 90	3.84	8.23	3.84	8.23	2.79	6.69	2.80	6.69	0.08	0.99	0.14	1.06
σ=30% 100	2.18	5.92	2.18	5.92	1.52	4.72	1.52	4.72	0.02	0.58	0.05	0.64
110	1.21	4.25	1.21	4.25	0.81	3.32	0.81	3.32	0.01	0.34	0.02	0.39
ImpBar			84.57	84.57			82.42	82.42			68.13	68.13
80	8.53	14.38	8.54	14.38	6.51	11.98	6.53	11.99	0.35	2.21	0.48	2.35
r=2.5% 90	5.74	11.30	5.75	11.31	4.24	9.27	4.25	9.28	0.15	1.48	0.24	1.60
σ=40% 100	3.84	8.91	3.84	8.91	2.74	7.20	2.75	7.20	0.06	1.00	0.12	1.10
110	2.55	7.04	2.56	7.04	1.77	5.61	1.78	5.62	0.03	0.68	0.06	0.77
ImpBar			85.45	85.45			80.04	80.04			62.22	62.22

Here we see that prices are close to the inverse Laplace prices for D relatively small compared to T. Furthermore we see that the implied barrier that is found

in the upper left corner (i.e. 84.57), is recognized in the implied barrier graph in figure 1 for the same case as shown above. As we would like to use the prices for relatively small D, we also need deltas for D relatively large compared to T. Table 2 shows that the error in the deltas is much less than that in the prices.

Table 2. Deltas for PDIC at $t = 0$

	D = 10/365				D = 20/365				D=200/365			
S_0=100, L=90	Laplace		Imp Bar		Laplace		Imp Bar		Laplace		Imp Bar	
X	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2
80	-0.34	-0.40	-0.34	-0.40	-0.27	-0.35	-0.27	-0.35	-0.03	-0.09	-0.03	-0.09
r=4.5% 90	-0.23	-0.32	-0.23	-0.32	-0.18	-0.28	-0.18	-0.27	-0.01	-0.06	-0.01	-0.06
σ=30% 100	-0.15	-0.26	-0.15	-0.25	-0.11	-0.21	-0.11	-0.21	0.00	-0.04	0.00	-0.04
110	-0.09	-0.20	-0.09	-0.20	-0.06	-0.16	-0.06	-0.16	0.00	-0.02	0.00	-0.02
ImpBar			84.57	84.57			82.42	82.42			68.13	68.13
80	-0.29	-0.32	-0.29	-0.32	-0.24	-0.28	-0.24	-0.28	-0.03	-0.08	-0.03	-0.08
r=4.5% 90	-0.23	-0.28	-0.22	-0.28	-0.18	-0.24	-0.18	-0.24	-0.01	-0.06	-0.02	-0.06
σ=30% 100	-0.17	-0.24	-0.17	-0.24	-0.13	-0.20	-0.13	-0.20	-0.01	-0.04	-0.01	-0.04
110	-0.12	-0.20	-0.12	-0.20	-0.09	-0.17	-0.09	-0.17	0.00	-0.03	0.00	-0.03
ImpBar			85.45	85.45			80.04	80.04			62.22	62.22

We also computed prices for Parisians based on a stock path that is already in the excursion. Table 3 lists these prices.

Table 3. Price for PDIC at $t > 0$ and $d > 0$

r=4.5%, σ=30%		D = 10/365				D = 20/365				D=200/365			
S_0=100, L=90		Laplace		Imp Bar		Laplace		Imp Bar		Laplace		Imp Bar	
X	d	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2
80	0.2D	13.04	18.56	13.04	18.56	11.02	16.40	11.02	16.41	1.40	4.16	1.51	4.24
100	0.2D	5.44	10.77	5.45	10.77	4.34	9.26	4.35	9.26	0.27	1.75	0.31	1.81
80	0.6D	13.98	19.52	13.97	19.52	12.59	18.05	12.59	18.05	2.79	5.92	2.95	5.99
100	0.6D	5.99	11.46	5.99	11.46	5.22	10.43	5.22	10.43	0.82	2.82	0.86	2.88
ImpBar				84.57	84.57			82.42	82.42			68.13	68.13

As these computations involved numerical integration we have to chose a number of integration steps. In this numerical approximation the interval is divided into 25 steps for $D = 10/365$ and $D = 20/365$. We divided it into 250

steps for $D = 200/365$. The results are analogous to the former price table. The prices are quite good for D relative small compared to T. Again we will compute for this particular case the deltas to show that they do not differ that much, as is shown in table 4.

Table 4. Deltas for PDIC at $t > 0$ and $d > 0$

r=4.5%, σ=30%		D = 10/365				D = 20/365				D=200/365			
S_0=100, L=90		Laplace		Imp Bar		Laplace		Imp Bar		Laplace		Imp Bar	
X	d	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2	T=1	T=2
80	0.2D	0.03	0.02	0.02	0.02	-0.25	-0.28	-0.25	-0.28	-0.15	-0.26	-0.14	-0.26
100	0.2D	0.02	0.02	0.02	0.02	-0.12	-0.19	-0.12	-0.19	-0.04	-0.13	-0.03	-0.13
80	0.6D	0.28	0.29	0.28	0.29	-0.19	-0.21	-0.19	-0.21	-0.40	-0.58	-0.39	-0.57
100	0.6D	0.16	0.21	0.16	0.21	-0.09	-0.14	-0.09	-0.14	-0.14	-0.33	-0.13	-0.32
ImpBar				84.57	84.57			82.42	82.42			68.13	68.13

7 Conclusion

Valuation of Parisian options is important as popular investment tools like convertible bonds exhibit the Parisian option feature. Well known methods for calculating its price are either computationally slow (PDE method) or inflexible (Laplace method). Here we provide a way to use standard barrier models with an implied barrier for calculating prices. The advantages of these method for trading firms are threefold. First no new model has to be implemented. Moreover it is possible to value the American options with a Parisian constraint, by using the implied barrier in the standard American barrier option. Discrete dividends can be taken into account. Finally a lot has been done pricing barriers in a non-constant volatility environment. Using the implied barrier again yields an approximation of the Parisian price in such an environment.

References

1. J. Abate, W. Whitt *Numerical Inversion of Laplace Transforms of Probability Distributions*, ORSA Journal on computing **7**(1), 36-43, (1995)
2. M. Chesney, M. Jeanblanc-Pique, M. Yor *Brownian Excursions and Parisian Barrier Options*, Adv. Appl. Prob. **29**, 165-184, (1997)
3. K. L. Chung *Excursions in Brownian Motion*, Ark. Math. **14**, 155-177, (1997)
4. R. T. Durrett, D. L. Iglehart *Functionals of Brownian Meander and Brownian Excursions*, Ann. Probability **5**, 130-135, (1977)
5. A. Grau *Moving Windows*, master thesis at School of Computer Science, University of Waterloo, Canada (2003)
6. R. J. Haber, P.J. Schönbucher, P. Wilmott *An American in Paris*, OFRC Working Papers Series **1999-MF-14**, 24-29, (1999)

Modeling Electricity Prices with Regime Switching Models

Michael Bierbrauer[1], Stefan Trück[1], and Rafał Weron[2]

[1] Chair of Statistics, Econometrics and Mathematical Finance, University of Karlsruhe, Karlsruhe, 76128, Germany
stefan@statistik.uni-karlsruhe.de
http://www.statistik.uni-karlsruhe.de/htm/mitarbeiter/trueck.php
[2] Hugo Steinhaus Center for Stochastic Methods, Wrocław University of Technology, 50-370 Wrocław, Poland
rweron@im.pwr.wroc.pl

Abstract. We address the issue of modeling spot electricity prices with regime switching models. After reviewing the stylized facts about power markets we propose and fit various models to spot prices from the Nordic power exchange. Afterwards we assess their performance by comparing simulated and market prices.

1 Electricity Spot Prices: Markets and Models

The deregulation of the power industry has given way to a global trend toward the commoditization of electric energy. Electricity has transformed from a primarily technical business, to one in which the product is treated in much the same way as any other commodity, with trading and risk management as key tools to run a successful business [2,12,15]. However, we have to bear in mind that electricity is a very unique commodity. It cannot be economically stored, demand of end users is largely weather dependent, and the reliability of the transmission grid is far from being perfect. This calls for adequate models of price dynamics capturing the main characteristics of spot electricity prices.

The spot electricity market is actually a day-ahead market. A classical spot market would not be possible, since the system operator needs advanced notice to verify that the schedule is feasible and lies within transmission constraints. The spot is an hourly (in some markets – a daily) contract with physical delivery. In our analysis we use spot prices from the Nordic power exchange (Nord Pool) covering the period January 1, 1997 – April 25, 2000. The system price is calculated as the equilibrium point for the aggregated supply and demand curves and for each of the 24 hours [14]. Due to limited space, in this paper we restrict the analysis to average daily prices. The averaged time series, however, retains the typical characteristics of electricity prices, including seasonality (on the annual and weekly level), mean reversion and jumps [20,21].

The seasonal character of electricity spot prices is a direct consequence of the fluctuations in demand. These mostly arise due to changing climate conditions,

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 859–867, 2004.
© Springer-Verlag Berlin Heidelberg 2004

like temperature and the number of daylight hours. In the analyzed period the annual cycle can be quite well approximated by a sinusoid with a linear trend [20,21]. The weekly periodicity is not sinusoidal, though, with peaks during the weekdays and troughs over the weekends. Spot electricity prices are also regarded as mean reverting – for time intervals ranging from a day to almost four years the Hurst exponent is significantly lower than 0.5 [18,19]. In addition to seasonality and mean reversion, spot electricity prices exhibit infrequent, but large jumps caused by extreme load fluctuations (due to severe weather conditions, generation outages, transmission failures, etc.). The spot price can increase tenfold during a single hour but the spikes are normally quite short-lived [2,12, 15,21].

Now, that we have discussed the properties of spot electricity prices we can turn to modeling issues. The starting point is the analysis of seasonal components. On the annual level this can be done through approximation by sinusoidal functions [15,21], fitting a piecewise constant function of a one year period [1, 13] or wavelet decomposition [18]. On the weekly (or daily) time scale, the seasonality is usually removed by subtracting an average week (or day) from the data. Once the seasonal components are removed we are left with the stochastic part of the process. In what follows we will analyze the logarithm d_t of the deseasonalized average daily spot prices D_t, see the bottom panel in Figure 1. For details on obtaining d_t from raw data see [20,21].

The stochastic part d_t can be modeled by a diffusion-type stochastic differential equation (SDE) of the form: $dX_t = \mu(X,t)dt + \sigma(X,t)dB_t$, which is the standard model for price processes of stochastic nature. Mean reversion is typically induced into the model by having a drift term $\mu(X,t)$ that is negative if the spot price is higher than the mean reversion level and positive if it is lower, like in the arithmetic Ornstein-Uhlenbeck process:

$$dX_t = (\alpha - \beta X_t)dt + \sigma dB_t = \beta(L - X_t)dt + \sigma dB_t, \tag{1}$$

where $\mu(X,t) = (\alpha - \beta X_t)$ is the drift, $\sigma(X,t) = \sigma$ is the volatility and dB_t are the increments of a standard Brownian motion. This is a one-factor model that reverts to the mean $L = \frac{\alpha}{\beta}$ with β being the magnitude of the speed of adjustment. The equilibrium level L can be also made time dependent to reflect the fact that electricity prices tend to revert to different levels over the year.

The second main feature of electricity spot prices, the "jumpy" character, calls for spot price modeling which is not continuous. One approach is to introduce to eqn. (1) a jump component $J_t dq_t$, where J_t is a random jump size and q_t is a Poisson variate [2,10]. After a spike the price is forced back to its normal level by the mean reversion mechanism or mean reversion coupled with downward jumps. Alternatively, a positive jump may be always followed by a negative jump of the same size to capture the rapid decline – especially on the daily level – of electricity prices after a spike [20,21].

Since spot prices after a jump tend to remain high for several time periods (hours, sometimes even days) there is also need for models that are able to capture this behavior. The so-called regime switching models offer such a possibility and be discussed in the next section.

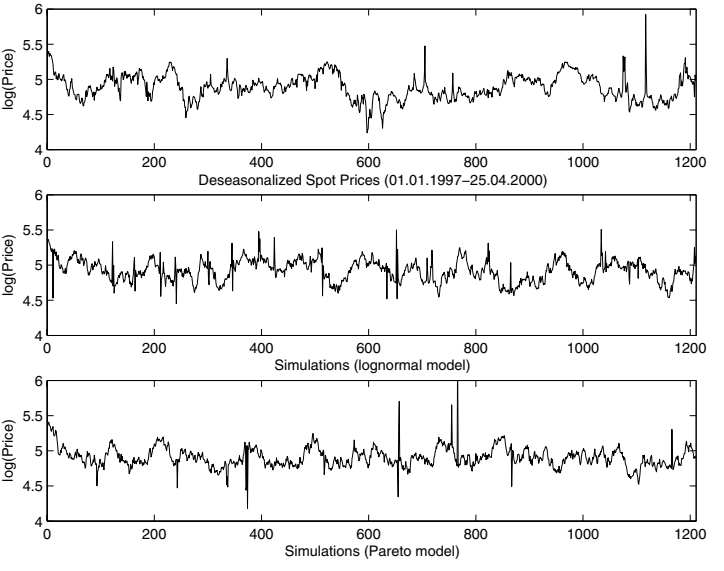


Fig. 1. The deseasonalized log-price process d_t for the time period 01.01.1997-25.04.2000 (*top panel*) and sample simulated price trajectories obtained from the two-regime model with normal (*middle panel*) and Pareto (*bottom panel*) distributions for the spike regime

2 Regime Switching Models

The price behavior of spot electricity prices can be modeled by dividing the time series into separate phases or regimes with different underlying processes. A jump in electricity prices can then be considered as a change to another regime [4,8,9]. The switching mechanism is typically assumed to be governed by a random variable that follows a Markov chain with different possible states. Thus, we have an unobservable variable in the time series that switches between a certain number of states which themselves are driven by independent stochastic processes [5,6,7,16]. Additionally we have a probability law that governs the transition from one state to another.

2.1 Two-Regime Models

To introduce the idea of regime switching models we start with the simplest model with two possible states. The two-regime model distinguishes between a base regime ($R_t = 1$) and a spike regime ($R_t = 2$), i.e. the spot price is supposed to display either mean reverting or jump behavior at each point of time. The price processes $Y_{t,1}$ and $Y_{t,2}$ that are linked to each of the two regimes are assumed to be independent of each other. The variable R_t that determines the current state is a random variable that follows a Markov chain with two possible states,

$R_t = \{1, 2\}$. The transition matrix $\mathbf{P}$ contains the probabilities p_{ij} of switching from regime i at time t to regime j at time $t + 1$:

$$\mathbf{P} = (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}. \tag{2}$$

The current state R_t of a Markov chain depends on the past only through the most recent value R_{t-1}. Thus $P\{R_t = j | R_{t-1} = i\} = p_{ij}$. The probability of being in state j at time $t + m$ starting from state i at time t is given by:

$$\begin{pmatrix} P(R_{t+m} = 1 | R_t = i) \\ P(R_{t+m} = 2 | R_t = i) \end{pmatrix} = (\mathbf{P}')^m \cdot e_i, \tag{3}$$

where $\mathbf{P}'$ denotes the transpose of $\mathbf{P}$ and e_i denotes the ith column of the 2×2 identity matrix.

There are various possibilities for choosing the stochastic processes for the base and the peak regime. However, considering the typical behavior of electricity spot prices described in the previous section, we let the base regime ($R_t = 1$) be governed by a mean-reverting process, eg. given by eqn. (1). In the spike regime ($R_t = 2$) it may be interesting to try different types of distributions for the process $Y_{t,2}$. The Gaussian [9] and lognormal [20] laws were suggested in the literature so far. In the latter case the deseasonalized log-price process d_t is defined by $dY_{t,1} = (c_1 - \beta_1 Y_{t,1})dt + \sigma_1 dB_t$ in the base regime and $\log(Y_{t,2}) \sim \mathrm{N}(c_2, \sigma_2^2)$ in the spike regime. The parameter set $\theta = \{c_1, \beta_1, \sigma_1, c_2, \sigma_2, p_{11}, p_{22}\}$ can be estimated using the so-called EM algorithm [3].

2.2 Alternative Regime Switching Models

Clearly the variety of regime switching models is due to both the possibility of choosing the number of regimes (2, 3, etc.) and different stochastic process for the price in each regime. Especially for the spike regime it may be interesting to choose alternative distributions. Since spikes happen very rarely but usually are of great magnitude the use of heavy-tailed distributions should be considered. We therefore suggest the use of the Pareto distribution (see e.g. [11]) for the spike regime. Also the process that switches between a certain number of states should be chosen in accordance with the typical behavior of spot electricity prices.

Huisman and Mahieu [8] propose a regime switching model with three possible regimes. The idea behind their specification differs significantly from the previous two-state models. They identify three possible regimes: (i) the regime $R_t = 1$ modeling the "normal" electricity price dynamics, (ii) an initial jump regime $R_t = 2$ for a sudden increase or decrease in price, and (iii) a regime $R_t = 3$ that describes how prices move back to the normal regime after the initial jump has occurred. This definition implies that the initial jump regime is immediately followed by the reversing regime and then moves back to the base regime. Thus we get a 3×3 transition matrix with only four non-zero values: p_{11}, $p_{12} = 1 - p_{11}$, $p_{23} = 1$, and $p_{31} = 1$.

Furthermore, Huisman and Mahieu [8] suggest to model the base and reversing jump regimes by a mean reverting process and the initial jump regime by Brownian motion (i.e. a process with increments given by a Gaussian variate). However, we do not see the need for modeling the reversing jump regime with a mean reverting process. The process automatically leaves this regime after one time period and it seems that a Gaussian or a lognormal random variable will do the job as well. The direction of the initial jump is not specified; it can be either an upward or a downward jump. However, we restrict the model so that the reversal jump, on average, is opposite to the initial jump. Hence, our three-regime model is defined by $dY_{t,1} = (c_1 - \beta_1 Y_{t,1})dt + \sigma_1 dB_t$ in the base regime, $\log(Y_{t,2}) \sim \mathrm{N}(c_2, \sigma_2^2)$ in the initial jump regime and $\log(Y_{t,3}) \sim \mathrm{N}(-c_2, \sigma_2^2)$ in the reversing jump regime. In contrast to the two-regime models, the three-regime model does not allow for consecutive spikes (or remaining at a different price level for two or more periods after a jump). In the next section we will compare estimation and simulation results of different regime switching models.

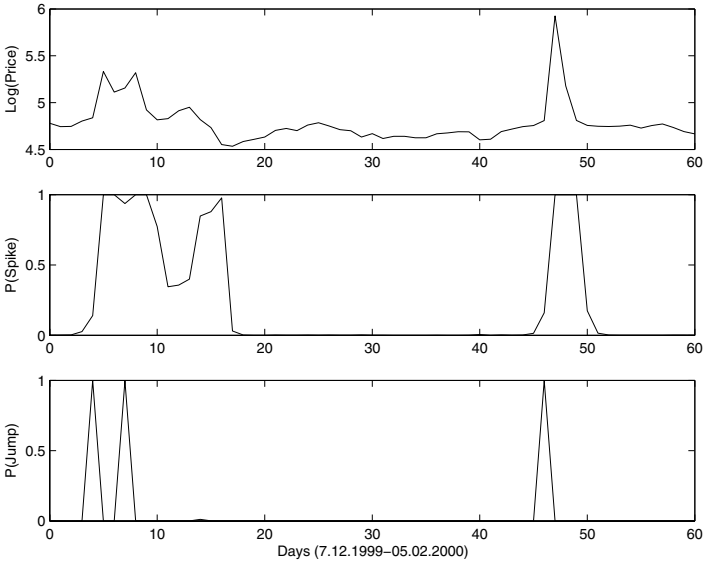


Fig. 2. The deseasonalized log-spot price d_t since December 7, 1999 until February 5, 2000 (*top panel*) together with the probability of being in the spike regime for the estimated two-regime model with lognormal spikes (*middle panel*) and of being in the jump regime for the estimated three-regime model (*bottom panel*)

Table 1. Estimation results for the two-regime model and the deseasonalized log-price d_t for the period January 1, 1997 – April 25, 2000. $E(Y_{t,i})$ is the level of mean reversion for the base regime ($i = 1$) and the expected value of the spike regime ($i = 2$), p_{ii} is the probability of remaining in the same regime in the next time step, and $P(R = i)$ is the unconditional probability of being in regime i

	β_i	c_i	σ_i^2	$E(Y_{t,i})$	p_{ii}	$P(R=i)$
Two-regime model with Gaussian spikes						
Base regime ($i = 1$)	0.0426	0.2078	0.0018	4.8807	0.9800	0.9484
Spike regime ($i = 2$)	—	—	0.0610	4.9704	0.6337	0.0512
Two-regime model with lognormal spikes						
Base regime ($i = 1$)	0.0426	0.2078	0.0018	4.8807	0.9800	0.9484
Spike regime ($i = 2$)	—	1.6018	0.0600	4.9678	0.6325	0.0516
Two-regime model with Pareto spikes						
Base regime ($i = 1$)	0.0459	0.2241	0.0020	4.8822	0.9860	0.9699
Spike regime ($i = 2$)	—	—	0.8294	4.9980	0.5497	0.0301
Three-regime model						
	c_i	mu_i	σ_i^2	$E(Y_{t,i})$	p_{ii}	$P(R=i)$
Base regime ($i = 1$)	0.2328	—	0.0024	4.8731	0.9924	0.9851
Init. Jump regime ($i = 2$)	—	0.0839	0.0697	-	0	0.0075
Rev. Jump regime ($i = 3$)	—	-0.0839	0.0697	-	0	0.0075

3 Empirical Analysis

In this section we analyze and model the logarithm d_t of the deseasonalized average daily spot prices from the Nord Pool power exchange since January 1, 1997 until April 25, 2000. For details on obtaining d_t from raw data see [20,21]. As we can see in Figure 1, the data exhibits several extreme events that can be considered as spikes. While most spikes only last for one day there are periods where the prices exhibit three or more extreme events in a row, a behavior that could be considered as consecutive spikes, see the top panel in Figure 2. This is the motivation for fitting the two-regime models with the base regime dynamics given by $dY_{t,1} = (c_1 - \beta_1 Y_{t,1})dt + \sigma_1 dB_t$ and the dynamics in the spike regime following a normal, a lognormal or a Pareto distribution, see Section 2.1. For comparison, we also fit the three-regime model described in the previous section.

The estimation results are summarized in Table 1. In all models, the probability of remaining in the base regime is quite high: in the two-regime model we get $p_{11} = 0.9800$ for the normal and lognormal model specifications and $p_{11} = 0.9860$ for the Pareto specification. For the three-regime model we get an extremely low probability of leaving the base regime $p_{11} = 0.9924$. However, while in the three-regime model the price level immediately returns to the mean-reversion process after a jump, estimating the two-regime model we find $p_{22} = 0.6325$ for the normal, $p_{22} = 0.6337$ for the lognormal, and $p_{22} = 0.5497$ for the Pareto model. Thus, in all three models the probability of staying in the spike regime is quite high, see also Figure 2. The data points with a high probability of being in the

Table 2. Measures of the goodness-of-fit (Mean Squared Error – MSE, Mean Absolute Error – MAE, Loglikelihood – LogL) for the estimated regime switching models. Performance of the models is also assessed by comparing the number of spikes, the return distributions quantiles ($q_{0.99}$ and $q_{0.995}$), and the extreme events

	MSE	MAE	LogL	spikes	$q_{0.99}$	$q_{0.995}$	max	min
Real Data	–	–	–	9.00	0.1628	0.2235	1.1167	-0.7469
2-regime (normal)	0.0047	0.0403	1890.28	17.26	0.3310	0.4523	0.7580	-0.8038
2-regime (lognormal)	0.0047	0.0403	1890.67	18.05	0.3353	0.4648	0.7937	-0.7875
2-regime (Pareto)	0.0047	0.0402	1866.11	33.32	0.5410	0.7851	2.1688	-2.2602
3-regime model	0.0048	0.0398	1854.56	13.72	0.3087	0.4144	0.7347	-0.6883

jump regime ($P\{R_t = 2\} > 0.5$) tend to be grouped in blocks in the two-regime models. Due to model specifications, in the three-regime model the probability of remaining in the second regime is zero.

Considering the unconditional probabilities we find that there is a 5.16%, 5.12% and 3.01% probability of being in the spike regime for the Gaussian, lognormal and Pareto two-regime models, respectively. This value is substantially larger than the probability of a jump in the three-regime model which is approximately equal to $P(R = 2) = P(R = 3) = 0.75\%$. Surprisingly, the normal and lognormal distributions produce almost identical results. A closer inspection of the parameter estimates uncovers the mystery – with such a choice of parameter values the lognormal distribution very much resembles the Gaussian law. However, using a heavy-tailed distribution, like the Pareto law, gives lower probabilities for being and remaining in the spike regime and a clearly higher variance.

Simulated price trajectories were used to check for similarity with real prices and stability of results. Reestimating the models with simulated data led to only slightly biased estimates for the parameters. Sample trajectories for the two-regime model with lognormal and Pareto spikes can be found in Figure 1. The trajectories of both models show strong similarity to real price data. We also checked the simulation results considering spikes as the most particular feature of electricity spot prices, see Table 2. Defining a spike as a change in the log-prices that is greater than 0.3 – either in positive or negative direction – we find that the regime switching models produce more spikes than there could be observed in real data. Especially in the two-regime model the number of spikes in simulations is about twice the number that was observed in the considered period. While the number of extreme events are overestimated in all models (see the values of $q_{0.99}$ and $q_{0.995}$ in Table 2), the magnitude of the largest spike in either direction is underestimated in the normal and lognormal models and overestimated by the Pareto distribution. This may suggest the use of alternative heavy-tailed distributions, e.g. a truncated Pareto or a stable distribution with parameter $\alpha > 1$ [17] for the spike regime.

4 Conclusions

This paper addresses the issue of modeling spot electricity prices. For the deseasonalized log-prices d_t we propose different regime switching models, which exhibit mean reversion and jump behavior.

We find that the models produce estimates for transition probabilities that can be interpreted according to market behavior. Simulated trajectories show high similarity with real price data. However, we find that the number of price spikes or extreme events produced by simulations of the estimated models is higher than what could be observed in real price data. This is especially true for the two-regime models where consecutive spikes have a higher probability than in the three-regime model.

References

1. K. Bhanot, Behavior of power prices: Implications for the valuation and hedging of financial contracts. The Journal of Risk 2 (2000) 43-62
2. L. Clewlow, C. Strickland, Energy Derivatives – Pricing and Risk Management, Lacima Publications, London, 2000
3. A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc. 39 (1977) 1-38
4. R. Ethier, T. Mount, Estimating the volatility of spot prices in restructured electricity markets and the implications for option values, Cornell University Working Paper 12/1998
5. S. Goldfeld, R. Quandt, A Markov Model for Switching Regressions, J. Econometrics 1 (1973) 3-16
6. J.D. Hamilton, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica 57 (1989) 357-384
7. J.D. Hamilton, Time Series Analysis, Princeton University Press, 1994.
8. R. Huisman, R. Mahieu, Regime jumps in electricity prices, Energy Economics 25 (2003) 425-434
9. R. Huisman, C. de Jong, Option pricing for power prices with spikes, Energy Power Risk Management 7.11 (2003) 12-16
10. B. Johnson, G. Barz, Selecting stochastic processes for modelling electricity prices, in: Risk Publications, Energy Modelling and the Management of Uncertainty, Risk Books (1999) 3-21
11. N. Johnson, S. Kotz and Narayanaswamy Balakrishnan, Continuous Univariate Distributions, Wiley, New York, 1995
12. V. Kaminski (ed.) Managing Energy Price Risk, Risk Books, London, 1999
13. J.J. Lucia, E.S. Schwartz, Electricity prices and power derivatives: Evidence from the Nordic Power Exchange, Rev. Derivatives Research 5 (2002) 5-50
14. Nord Pool, Nord Pool Annual Report, Nord Pool ASA, 2002
15. D. Pilipovic, Energy Risk: Valuing and Managing Energy Derivatives, McGraw-Hill, New York, 1998
16. R. Quandt, The Estimation of the Parameters of a linear Regression System Obeying two Separate Regimes, J. Amer. Statist. Assoc. 55 (1958) 873-880
17. C. Mugele, S.T. Rachev and S. Trück, Analysis of Different Energy Markets with the alpha-stable distribution, University of Karlsruhe Working Paper 12/2003

18. I. Simonsen, Measuring anti-correlations in the Nordic electricity spot market by wavelets, Physica A 322 (2003) 597-606
19. R. Weron, B. Przybyłowicz, Hurst analysis of electricity price dynamics, Physica A 283 (2000) 462-468
20. R. Weron, M. Bierbrauer, S. Trück, Modeling electricity prices: jump diffusion and regime switching, Physica A (2004) to appear
21. R. Weron, I. Simonsen, P. Wilman, Modeling highly volatile and seasonal markets: evidence from the Nord Pool electricity market, in H. Takayasu (ed.), The Application of Econophysics, Springer, Tokyo, 2004, 182-191

Modeling the Risk Process in the XploRe Computing Environment

Krzysztof Burnecki and Rafał Weron

Hugo Steinhaus Center for Stochastic Methods, Wrocław University of Technology,
Wyspiańskiego 27, 50-370 Wrocław, Poland
{burnecki,rweron}@im.pwr.wroc.pl
http://www.im.pwr.wroc.pl/~hugo

Abstract. A user friendly approach to modeling the risk process is presented. It utilizes the insurance library of the XploRe computing environment which is accompanied by on-line, hyperlinked and freely downloadable from the web manuals and e-books. The empirical analysis for Danish fire losses for the years 1980-90 is conducted and the best fitting of the risk process to the data is illustrated.

1 Introduction

The simulation of risk processes is vital for insurance companies. It boils down to generating aggregated claims for the calculation of losses that may occur. Since claims arrive at random times, the number of claims up to a given time is typically assumed to be driven by a stochastic process. In its simplest form the risk process utilizes the homogeneous Poisson process as the claim arrival process, however, more general processes like the non-homogeneous Poisson, the mixed Poisson, the Cox (or doubly stochastic Poisson) and the renewal process are also considered.

The risk process $\{R_t\}$ of an insurance company can be approximated by the sum of the initial capital of the company and the so-called premium function (representing income from sold insurance policies) minus the aggregated claim process (expressing liabilities resulting from claims covered by the previously sold insurance policies) [3,8,9,14]. The latter is typically modeled by a sum of random length, governed by the claim arrival point process, where the summands – representing the claim severities – form an independent sequence of positive i.i.d. random variables.

Since parameter estimation and simulation schemes can be tedious a number of computer packages have been written to automate the process. In this paper we want to present a novel solution which takes the form of a library of procedures of the XploRe system combined with on-line, hyperlinked manuals and e-books.

XploRe is a computing environment which offers a combination of classical and modern statistical procedures, in conjunction with sophisticated, interactive graphics. XploRe is also a high level object-oriented programming language. With all the standard features like recursion, local variables, dynamic data

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 868–875, 2004.
© Springer-Verlag Berlin Heidelberg 2004

structures, loops, and conditional execution it provides a platform for advanced statistical and econometric analysis, research, as well as teaching [10,11].

The statistical methods of XploRe are provided by various procedures and scripts called quantlets. Quantlets are combined into libraries called quantlibs. Among other these include: finance, econometrics, wavelets, generalized (partial) linear models, time series analysis and filtering, neural networks, non- and semiparametric methods, and teachware. Recent additions to this family comprise the stable distributions and insurance libraries [6]. Proprietary methods can be incorporated into XploRe, enabling the user to easily extend the environment.

One of the most outstanding features of the XploRe environment is the unique combination of computing capabilities and on-line, hyperlinked manuals and books. A variety of electronic statistical and econometric volumes are available from the XploRe web page (www.xplore-stat.de) in the html and pdf file formats. All books contain a large number of quantlets which illustrate the theoretical content. One of the newest additions is the "Statistical Tools for Finance and Insurance" e-book [6], which offers the reader over a hundred methods and procedures related to insurance. Some of these quantlets will be utilized in this paper.

The paper is organized as follows. In section 2 we briefly recall the methods of simulating two possible choices for the claim arrival process, namely the homogeneous Poisson process (HPP) and the non-homogeneous Poisson process (NHPP). In section 3 we discuss a number of claim severities distributions and present the methods for judging the goodness-of-fit. Finally, in section 4 we conduct the empirical analysis for Danish fire losses for the years 1980-90 and illustrate the best fitting risk process for two choices of the claim severities distribution.

2 Claim Arrival Process

In this section we focus on the efficient simulation of the claim arrival point process. Typically this process is simulated via the arrival times $\{T_i\}$, i.e. moments when the ith claim occurs, or the inter-arrival times (or waiting times) $W_i = T_i - T_{i-1}$, i.e. the time periods between successive claims.

2.1 Homogeneous Poisson Process

A continuous-time stochastic process $\{N_t : t \geq 0\}$ is a (homogeneous) Poisson process with intensity (or rate) $\lambda > 0$ if (i) $\{N_t\}$ is a point process, and (ii) the times between events are i.i.d. exponential random variables with intensity λ, i.e. with mean $1/\lambda$. Therefore, successive arrival times $T_1, T_2, \ldots, T_n$ of the Poisson process can be generated by a simple algorithm consisting of generating independent exponential random variables with intensity λ and taking a cumulative sum of them [3,15]. In the insurance library of XploRe this procedure is implemented in the `simHPP.xpl` quantlet.

Since the expected value of the homogeneous Poisson process $\mathrm{E}N_t = \lambda t$, it is natural to define the premium function as a linear function of the form $c(t) = (1+\theta)\mu\lambda t$. Here $\mu = \mathrm{E}X_k$ is the expected value of the claim size and $\theta > 0$ is the relative safety loading which "guarantees" survival of the insurance company. With such a choice of the risk function we obtain the classical form of the risk process [3,8,9]:

$$R_t = u + (1+\theta)\mu\lambda t - \sum_{i=1}^{N_t} X_i \; . \tag{1}$$

The nonnegative constant u represents the initial capital of the company and $\{X_i\}$ is the i.i.d. claim severities sequence.

2.2 Non-homogeneous Poisson Process

In real life situations the homogeneous Poisson process may be too simplistic. In order to obtain a more reasonable description of reality we might want to include the possibility of a variable (eg. increasing) size of the portfolio of insurance contracts or seasonality in the number of claims. For modeling such phenomena the non-homogeneous Poisson process (NHPP) is much better. To distinguish it from the HPP we denote it by M_t. The NHPP can be thought of as a Poisson process with a variable intensity defined by the deterministic intensity (rate) function $\lambda(t)$. Note that the increments of a NHPP do not have to be stationary. In the special case when $\lambda(t)$ takes the constant value λ, the NHPP reduces to the homogeneous Poisson process with intensity λ.

The simulation of a NHPP is slightly more complicated than in the homogeneous case. The first approach, the so-called "integration method", is based on the observation that for a NHPP with rate function $\lambda(t)$ the increment $M_t - M_s$, $0 < s < t$, is distributed as a Poisson random variable with intensity $\widetilde{\lambda} = \int_s^t \lambda(u)du$ [9]. Hence, the distribution function F_s of the waiting time W_s is given by:

$$\begin{aligned} F_s(t) &= P(W_s \le t) = 1 - P(W_s > t) = 1 - P(M_{s+t} - M_s = 0) = \\ &= 1 - \exp\left(-\int_s^{s+t} \lambda(u)du\right) = 1 - \exp\left(-\int_0^t \lambda(s+v)dv\right) \; . \end{aligned} \tag{2}$$

If the function $\lambda(t)$ is such that we can find an explicit formula for the inverse F_s^{-1} then for each s we can generate a random quantity X with the distribution F_s by using the inverse transform method. Otherwise, we have to apply numerical schemes and the algorithm becomes relatively slow.

The second approach, known as the "thinning" or "rejection method", is based on the following observation [2,15]. Suppose that there exists a constant $\overline{\lambda}$ such that $\lambda(t) \le \overline{\lambda}$ for all t. Let $T_1^*, T_2^*, T_3^*, \ldots$ be the successive arrival times of a HPP with intensity $\overline{\lambda}$. If we accept the ith arrival time with probability $\lambda(T_i^*)/\overline{\lambda}$, independently of all other arrivals, then the sequence $T_1, T_2, \ldots$ of the accepted arrival times (in ascending order) forms a sequence of the arrival times

of a NHPP with rate function $\lambda(t)$. The resulting algorithm [3] is implemented in the insurance library of XploRe in the `simNHPP.xpl` quantlet.

Since the mean-value function, i.e. the expected value of the process M_t, is given by:

$$\mathrm{E}M_t = \int_0^t \lambda(s)ds \ , \tag{3}$$

in the non-homogeneous case the premium function is typically defined as $c(t) = (1+\theta)\mu \int_0^t \lambda(s)ds$ and the risk process takes the form:

$$R_t = u + (1+\theta)\mu \int_0^t \lambda(s)ds - \sum_{i=1}^{M_t} X_i \ . \tag{4}$$

3 Claim Severities Distribution

The derivation of claim size distributions from the loss data could be considered to be a separate discipline in its own [7]. The objective is to find a distribution function F which fits the observed data in a satisfactory manner. The approach most frequently adopted in insurance is to find a suitable analytic expression which fits the observed data well and which is easy to handle, see e.g. [5].

The claim distributions, especially describing property losses, are usually heavy-tailed. Note, that in the actuarial literature such distributions are defined as having tails heavier than exponential. The lognormal, Pareto, Burr and Weibull distributions are typical candidates considered in applications [13]. Let us briefly recall them.

If the random variable X is normally distributed with mean μ and variance σ^2, then the distribution of $Y = e^X$ is lognormal. Its distribution function (d.f.) is given by $F_{LN}(x) = \Phi([\ln x - \mu]/\sigma)$, where $\Phi(x)$ is the standard normal d.f., and in the insurance library of XploRe is implemented in the `cdfln.xpl` quantlet. The lognormal law is very useful in modeling of claim severities. It has a thick right tail and fits many situations well.

One of the most frequently used analytic claim size distributions is the Pareto law which is defined by $F_P(x) = 1 - (\lambda/[\lambda + x])^\alpha$ with $\alpha, \lambda > 0$, see the `cdfPareto.xpl` quantlet. The first parameter controls the thickness of the tail: the smaller the α, the heavier the tail. Empirical experience has shown that the Pareto formula is often an appropriate model for the claim size distribution, particularly when exceptionally large claims may occur [4,7]. However, there is a need to find heavy tailed distributions which offer yet greater flexibility. Such flexibility is offered by the Burr distribution, which is just a generalization of the Pareto law. Its distribution function is given by $F_B(x) = 1 - (\lambda/[\lambda + x^\tau])^\alpha$ with all three parameters (α, λ, and τ) being positive real constants, see the `cdfBurr.xpl` quantlet.

Another frequently used analytic claim size distribution is the Weibull distribution which is defined by $F_W(t) = 1 - e^{-\lambda t^\alpha}$ with $\alpha, \lambda > 0$. Observe, that the Weibull distribution is a generalization of the exponential law. In XploRe it is implemented in the `cdfWeibull.xpl` quantlet.

Once the distribution is selected, we must obtain parameter estimates. In what follows we use the moment and maximum likelihood estimation approaches. The next step is to test whether the fit is adequate. This is usually done by comparing the fitted and empirical distribution functions. More precisely, by checking whether values of the fitted distribution function at sample points form a uniform distribution. In the next section we apply the χ^2, the Kolmogorov-Smirnov (KS), the Cramer-von Mises (CM), and the Anderson-Darling (AD) test statistics [1,5,16]. Generally, the smaller the value of the statistics, the better the fit. These test statistics are implemented in the quantlets `chi2stat.xpl`, `kstat.xpl`, `cmstat.xpl`, and `adstat.xpl`, respectively [6].

4 Empirical Analysis

We conducted empirical studies for Danish fire losses recorded by Copenhagen Re. The data, see Fig. 1, concerns major Danish fire losses in Danish Krone (DKK), occurred between 1980 and 1990 and adjusted for inflation. Only losses in profits connected with the fires were taken into consideration. In order to calibrate the risk process we had to fit both the distribution function F of the incurred losses $\{X_i\}$ and the claim arrival process. First we studied the

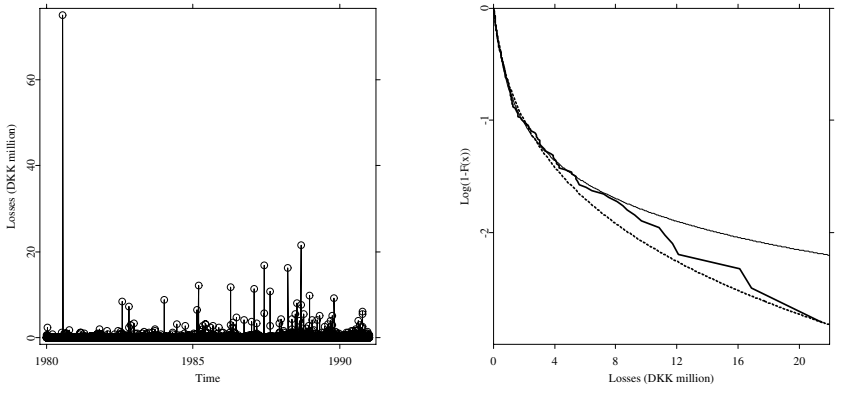


Fig. 1. *Left panel*: Illustration of the major Danish fire losses in millions of Danish Krone (DKK), occurred between January 1, 1980 and December 31, 1990 and adjusted for inflation. *Right panel*: Logarithm of the right tails of the empirical claim sizes distribution function $F(x)$ (thick solid line) together with lognormal (dotted line) and Burr (thin solid line) fits

loss sizes. We fitted lognormal, Pareto, Burr, and Weibull distributions using the `estln.xpl`, `estPareto.xpl`, `estBurr.xpl`, and `estWeibull.xpl` quantlets, respectively. The results of the parameter estimation and test statistics are

Table 1. Parameter estimates and test statistics for the fire loss amounts

d.f.:	Lognormal	Pareto	Burr	Weibull	Gaussian
Parameters:	$\mu = 12.704$ $\sigma = 1.4271$	$\alpha = 2.4189$ $\lambda = 1.0261e6$	$\alpha = 0.8935$ $\lambda = 1.1219e7$ $\tau = 1.2976$	$\alpha = 0.6963$ $\lambda = 8.9740e\text{-}5$	$\mu = 4.7332e5$ $\sigma = 6.7224e5$
χ^2	56.109	73.879	**48.493**	129.24	592.98
KS	**0.0373**	0.0397	0.0413	0.0783	0.2433
CM	0.1687	0.2878	**0.1438**	1.5245	6.3655
AD	1.0533	2.7712	**0.8221**	10.638	78.410

presented in Table 1. For illustration purposes we also added parameter estimates for the Gaussian distribution. The lognormal distribution with parameters $\mu = 12.7036$ and $\sigma = 1.4271$ and the Burr distribution with $\alpha = 0.8935$, $\lambda = 1.1219 \cdot 10^7$ and $\tau = 1.2976$ produced the best results, see Fig. 1. Hence, we chose them for further analysis. Next, we fitted the claim arrival process.

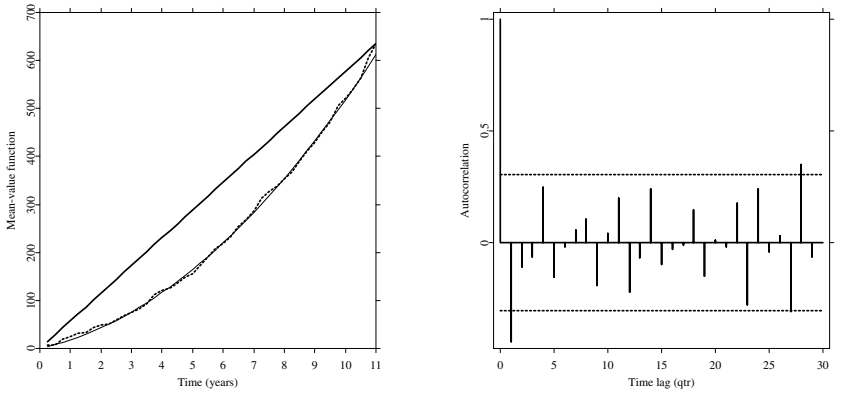


Fig. 2. *Left panel*: The aggregate number of losses of the fire data (dotted line), the HPP (thick solid line), and the NHPP (thin solid line). *Right panel*: Autocorrelation function of the differenced quarterly fire losses data revealing no seasonality or significant dependencies. Horizontal dashed lines represent the 95% confidence intervals for white noise

We started the analysis with the homogeneous Poisson process N_t with intensity λ_1. Studies of the quarterly numbers of losses and the interoccurence times of the catastrophes led us to the conclusion that the homogeneous Poisson process

with the annual intensity $\lambda_1 = 57.72$ gave the best fit. However, as we can see in Fig. 2, the fit is not very good suggesting that the HPP is too simplistic and causing us to consider the NHPP.

The data reveals no seasonality but a clear increasing trend can be observed in the number of quarterly losses, see Fig. 2. We tested different exponential and polynomial functional forms, but a simple linear intensity function $\lambda_2(s) = c + ds$ yielded the best fit. Applying a least squares procedure we arrived at the following values of the parameters: $c = 13.97$ and $d = 7.57$. Both choices of the intensity function, λ_1 and $\lambda_2(s)$, are illustrated in Fig. 2, where the accumulated number of fire losses and mean-value functions for all 11 years of data are depicted. The simulation results are presented in Fig. 3. We consider a

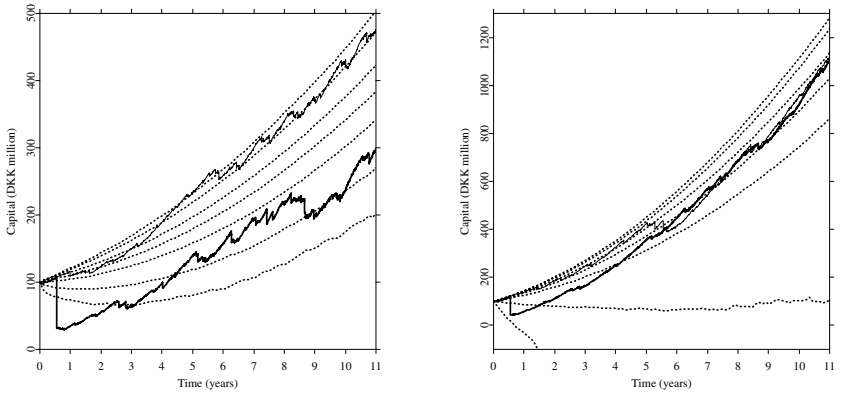


Fig. 3. Simulation results for a non-homogeneous Poisson process with lognormal (*left panel*) and Burr (*right panel*) claim sizes. The thick solid line is the "real" risk process, i.e. a trajectory constructed from the historical arrival times and values of the losses. The thin solid line is a sample trajectory. The dotted lines are the sample $0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99$-quantile lines. Clearly, the Burr distribution describes the extreme losses much better

hypothetical scenario where the insurance company insures losses resulting from fire damage. The company's initial capital is assumed to be $u = 100$ million kr and the relative safety loading used is $\theta = 0.5$. We chose two models of the risk process whose application is most justified by the statistical results described above: a non-homogeneous Poisson process with lognormal claim sizes and a non-homogeneous Poisson process with Burr claim sizes.

In both subplots of Fig. 3 the thick solid line is the "real" risk process, i.e. a trajectory constructed from the historical arrival times and values of the losses. The thin solid line is a sample trajectory. The dotted lines are the sample $0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99$-quantile lines based on 10000 trajectories of

the risk process. Recall that the function $\hat{x}_p(t)$ is called a sample p-quantile line if for each $t \in [t_0, T]$, $\hat{x}_p(t)$ is the sample p-quantile, i.e. if it satisfies $F_n(x_p-) \leq p \leq F_n(x_p)$, where F_n is the sample distribution function. Quantile lines are a very helpful tool in the analysis of stochastic processes. For example, they can provide a simple justification of the stationarity (or the lack of it) of a process, see [12]. In Fig. 3 they visualize the evolution of the density of the risk process. Clearly, if claim severities are Burr distributed then extreme events are more probable to happen than in the lognormal case, for which the historical trajectory falls even outside the 0.01-quantile line. This suggests that Burr distributed claim sizes are more adequate for modeling the "real" risk process.

Acknowledgements. The authors thankfully acknowledge the support of the State Committee for Scientific Research (KBN) Grant No. PBZ-KBN 016/P03/99.

References

1. D'Agostino, R.B., Stephens, M.A.: Goodness-of-Fit Techniques. Marcel Dekker, New York (1986)
2. Bratley, P., Fox, B.L., Schrage, L.E.: A Guide to Simulation. Springer-Verlag, New York (1987)
3. Burnecki, K., Härdle, W., Weron, R.: An Introduction to Simulation of Risk Processes. In: Teugels, J., Sundt, B. (eds.): Encyclopedia of Actuarial Science. Wiley, Chichester (to appear)
4. Burnecki, K., Kukla, G.: Pricing of Zero-Coupon and Coupon CAT Bonds. Applicationes Mathematicae **30(3)** (2003) 315–324
5. Burnecki, K., Kukla, G., Weron, R.: Property Insurance Loss Distributions. Physica A **287** (2000) 269–278
6. Cizek, P., Härdle, W., Weron, R. (ed.): Statistical Tools for Finance and Insurance. Springer, Heidelberg (2004)
7. Daykin, C.D., Pentikainen, T., Pesonen, M.: Practical Risk Theory for Actuaries. Chapman&Hall, London (1994)
8. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events. Springer, Berlin (1997)
9. Grandell, J.: Aspects of Risk Theory. Springer, New York (1991)
10. Härdle, W., Hlavka, Z., Klinke, S. (ed.): XploRe Application Guide. Springer, Heidelberg (2000)
11. Härdle, W., Klinke, S., Müller, M. (ed.): XploRe Learning Guide. Springer, Heidelberg (2000)
12. Janicki, A., Weron, A.: Simulation and Chaotic Behavior of α-Stable Stochastic Processes. Marcel Dekker, New York (1994)
13. Panjer, H.H., Willmot, G.E.: Insurance Risk Models. Society of Actuaries, Schaumburg (1992)
14. Rolski, T., Schmidli, H., Schmidt, V., Teugels, J.L.: Stochastic Processes for Insurance and Finance. Wiley, Chichester (1999)
15. Ross, S.: Simulation. 3rd edn. Academic Press, San Diego (2001)
16. Stephens, M.A.: EDF Statistics for Goodness-of-Fit and Some Comparisons. Journal of the American Statistical Association **69** (1974) 730–737

A Dynamic Stochastic Programming Model for Bond Portfolio Management

Liyong Yu[1], Shouyang Wang[1], Yue Wu[2], and K.K. Lai[3]*

[1] Institute of Systems Science, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, P.R. China
[2] School of Management, University of Southampton, Southampton, UK
[3] Department of Management Sciences, City University of Hong Kong, Hong Kong, P.R. China

Abstract. In this paper we develop a dynamic stochastic programming model for bond portfolio management. A new risk measurement-shortfall cost is put forward. It allows more tangible expression of the risks that the decision makers face than does the traditional risk measure-variance of terminal wealth. We also adopt the interest rate model of Black et al. to generate scenarios of riskless short rates at future periods. An example of bond portfolio management is presented to illustrate that our model dominates the usual fixed-mix model.

Keywords: Bond portfolio management; Stochastic programming; Scenario generation

1 Introduction

The bond portfolio management can be viewed as a multi-period dynamic decision problem. Stochastic programming with recourse is used in the formulation of the problem.

There are numbers of literature significantly contributed in this field. A fundamental contribution was made by Bradley and Crane(1972)[1], who proposed a multi-period model for bond portfolio management. This kind of models involve enormous computation work that exceeds the computation ability of that time. Over these years, progress in computational methods is impressive and large scale problems can be efficiently solved with high reliability. Moreover, powerful computers can also easily be accessed to conduct optimization searches. The obstacles for applying stochastic programming models are quickly receding. A successful application of this kind of models was reported by Carino et al.(1994)[2] for asset/liability management problems in insurance firms. Golub et al.(1995)[3] and Zenios et al.(1998)[4] applied them to the portfolio management of fixed-income securities. Extensive validation experiments are carried out to establish the effectiveness of the model in dealing with uncertainty. Empirical tests shown that this kind of models outperform classical models based

* The corresponding author. Email address: mskklai@cityu.edu.hk

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 876–883, 2004.
© Springer-Verlag Berlin Heidelberg 2004

on portfolio immunization and single period models. The latest progress in applying stochastic programming models was made by Kouwenberg(2001)[5]. The author proposed a multi-period model for a Dutch pension fund. The development of portfolio management models is fully discussed in the book of Ziemba and Mulvey(1998)[6], and that of Wang and Xia(2002)[7].

In the formulation of the model, one of the most important issues is how to character the risk attitude of the investors. One natural alternative is the variance of terminal wealth, as has been adopted by lots of literatures. However, the dynamic investment decisions involve the realization of investment goals at the end of every period. Usually investors attach different importance to the yields of each period, rather than consider the terminal wealth only. In this paper we formulate a multi-period stochastic programming model for bond portfolio management, in which risk attitude is measured by shortfall cost. So that the preferences of investors to the yields of different periods can be incorporated.

The paper is organized as follows: Section 2 presents the model formulation. Section 3 provides the scenario generation process, in which interest rates scenarios are generated on the basis of Black-Derman-Toy model. In Section 4, an application of the model is illustrated and the result is compared with that of fixed-mix model. We conclude this paper with some insights and comments.

2 Model Formulation

2.1 Variables and Parameters

In the model we use variables to represent the buying and selling of bonds, investment in the riskless asset and holdings of bonds in the portfolio. Investment decisions are of face value. The bid and ask prices of a kind of bond are assumed to be the same for convenience. We define the following notations.

Parameters of the Model

S_t: set of scenarios anticipated at time t, for $t = 0, 1, ..., T$. These could be scenarios of interest rates, prices and so on. We use s_t to index scenarios from the set S_t.

l_t: paths of scenarios that are resolved (i.e., all information becomes known) until period t, where $t = 1, 2, .., T$.

J: set of available bonds, with cardinality m.

c_0: riskless asset available at $t = 0$.

$(b_{0j})_{j=1}^{m}$: vectors denoting the initial holdings of bonds.

$(p_{0j})_{j=1}^{m}$: vectors of bond prices at $t = 0$. These prices are known with certainty.

$(p_{tj}(l_t))_{j=1}^{m}$ for all $l_t \in P_t$ and $t = 1, 2, ..., T$: vectors of bond prices realized at t. These prices depend on the path of scenarios followed from 0 to t.

$(k_{tj}(l_t, s_t))_{j=1}^{m}$ for $t = 0, 1, ..., T$: vectors of cash accrual factors during the interval from t to $t + 1$.

$r_t(l_t, s_t)$ for $t = 0, 1, ..., T$: short term riskless reinvestment rates during the interval from t to $t + 1$. These rates depend on the path followed up to t, and are conditioned on the scenario to be realized during the interval from t to $t + 1$.

First-stage Variables at $t = 0$.

$(x_{0j})_{j=1}^m$: vector denoting the face value bought of each bond.
$(y_{0j})_{j=1}^m$: vector denoting the face value sold of each bond.
$(z_{0j})_{j=1}^m$: vector denoting the face value held of each bond.
v_0: amount invested in the riskless asset.

Time-staged Variables.

$(x_{tj}(l_t))_{j=1}^m$: vector denoting the face values bought of each bond.
$(y_{tj}(l_t))_{j=1}^m$: vector denoting the face value sold of each bond.
$(z_{tj}(l_t))_{j=1}^m$: vector denoting the face values held in the portfolio.
$v_t(l_t)$: amount invested in the riskless asset.

2.2 Model Formulation

There are two basic kind of constraints in stochastic programming models for portfolio optimization. One expresses *cashflow accounting* for the riskless asset, and the other is an *inventory balance* equation for each bond at all time periods.

First-stage Constraints. At the first stage(i.e., at time $t = 0$)all prices are known with certainty. The *cashflow accounting* equation deals with conservation of the initial holdings of the riskless asset(i.e., cash):

$$c_0 + \sum_{j=1}^{m} p_{0j} y_{0j} = \sum_{j=1}^{m} p_{0j} x_{0j} + v_0 \ . \tag{1}$$

For each bond $j \in J$ in the portfolio we have an *inventory balance* Constraint:

$$b_{0j} + x_{0j} = y_{0j} + z_{0j} \ . \tag{2}$$

Time-staged Constraints. Decisions made at any time period t, after $t = 0$, depend on the path l_t and are conditioned on the scenarios S_t anticipated at t. Hence at each time instance t we have one constraint for each path in P_t and each scenario in S_t. These decisions also depend on the investment decisions made at previous periods.

Cashflow accounting constraints:

$$r_{t-1}(l_{t-1}, s_{t-1}) v_{t-1}(l_{t-1}) + \sum_{j=1}^{m} k_{(t-1)j}(l_{t-1}, s_{t-1}) z_{(t-1)j} + \sum_{j=1}^{m} p_{tj}(l_t) y_{tj}(l_t) = v_t(l_t) + \sum_{j=1}^{m} p_{tj}(l_t) x_{tj}(l_t) \ . \tag{3}$$

Inventory balance constraints:

$$z_{(t-1)j}(l_{t-1}) + x_{tj}(l_t) = y_{tj}(l_t) + z_{tj} \ . \tag{4}$$

Objective Function. In some papers about dynamic portfolio management, the objective functions adopt the form of utility optimization (Hakansson and Ziemba (1995)[9], Zenios et al.(1998) [4]). The nonlinear objective functions result in too complex optimization problems which are difficult to resolve, especially when the number of the variables and constraints is large. In our model, we use shortfall costs to character the risk attitude of the portfolio manager. For each time period, we set a goal wealth on the account of the manager's judge. The realized wealth may be greater than or less than them. The shortfalls are penalized in the form of costs. To represent the importance that we attach to different time periods, we set different weights to the costs. The objective of the problem is to maximize the expected wealth at the end of the investment horizon, minus the weighted sum of the shortfall costs (All are discounted to the present value). It can be given by

$$\max \ \frac{\sum_{l_T} \pi_{l_T} W(l_T)}{\prod_{i=0}^{T} r_i} - \sum_{t=1}^{T} \lambda_t \frac{\sum_{l_t} \pi_{l_t} c_{t,l_t}}{\prod_{i=0}^{t-1} r_i} \ . \tag{5}$$

π_{l_t} denotes the probability of scenario path l_t. $W(l_T)$ is the terminal wealth at the end of period T if the scenario path l_T is realized. It can be denoted as follow

$$W(l_T) = r_{T-1}(l_{T-1}, s_{T-1}) v_{T-1}(l_{T-1})$$
$$+ \sum_{j=1}^{m} k_{(T-1)j}(l_{T-1}, s_{T-1}) z_{(T-1)j} + \sum_{j=1}^{m} P_{Tj}(l_T) y_{Tj}(l_T) \ .$$

λ_t is the weight we set to the shortfall cost of time period t. It represents the importance of the goal wealth we attach to that period. The nearer the periods to the beginning, the more the cost weights are that correspond to them. c_{t,l_t} denotes shortfall cost of period t when l_t happened.

3 Scenario Generation

Scenarios of uncertain economic factors play fundamental role in stochastic programming models for portfolio optimization, which determine whether the models are valid. Here we generate scenarios of interest rates using the interest model of Black et.al[8], then show how prices can be resolved at different states.

In the one-factor interest rate model of Black et.al, the current structure of long rates and their estimated volatilities are used to construct a tree of possible future short rates. This tree can then be used to value interest-rate-sensitive securities. Now we explain how to price the bond at some future time period τ. Possible state of short rates σ at τ can be obtained using binomial lattice. Let

S_σ denote the set of interest rate scenarios that emanate from state σ of the binomial lattice at some future time period τ. Let also r_t^s be the short rate at time period $t(\tau \leq t \leq T)$ associated with scenario $s \in S_\sigma$, and C_t^s be the cash flow generated by the security at period t under the same scenario s. The price of bond is given by the expected present value of the cash flows C_t generated by the security from $t = \tau$ until maturity T, as

$$P_\tau^\sigma = \frac{1}{|S_\sigma|} \sum_{s \in S_\sigma} \sum_{t=\tau+1}^{T} \frac{C_t^s}{\prod_{i=\tau}^{t-1} (1 + \rho r_i^s)} \quad . \tag{6}$$

In this formulation, ρ is the *option adjusted premium*(OAP), a factor that incorporates the the risk premia due to defaults, illiquidity, prepayments, lapse and other risks that are relevant to the security; see Babbel and Zenios (1992)[10]. The OAP for a given security is the solution of the following nonlinear equation in ρ:

$$P_0 = \frac{1}{|S_0|} \sum_{s \in S_0} \sum_{t=1}^{T} \frac{C_t^s}{\prod_{i=0}^{t-1} (1 + \rho r_i^s)} \quad . \tag{7}$$

Here, S_0 denote the set of scenarios that emanate from the root of the binomial lattice, and P_0 is the present price of the security considered.

Making the assumption that this risk premium remains constant until τ, we compute the *option adjusted price* of the security P_σ^τ as

$$P_\tau^\sigma = \frac{1}{|S_\sigma|} \sum_{s \in S_\sigma} \sum_{t=\tau+1}^{T} \frac{C_t^s}{\prod_{i=\tau}^{t-1} (1 + \rho_0 r_i^s)} \quad . \tag{8}$$

where ρ_0 denotes the option adjusted premium computed by solving (7).

4 An Application

We apply our model to the following bond portfolio management problem in the China market. Three bonds are selected to invest in after analysis of the bond market. They are two treasury bonds, 96(6) and 99(5), one corporate bond, 98 Oil (see Table 1.). The beginning of the investment horizon is July 1st, 2003, end in July 1st, 2006. The compositions of the portfolios are determined at the beginning of each year period based on scenarios of riskless short rates and bond prices.

First we calculate short rates using the structure of long rates and their estimated volatilities. According to the data from China Government Securities Depository Trust and Clearing Corporation(*www.chinabond.com.cn*), we obtain the treasury bond yield (adopting the term structure of July 1st, 2003) and the yield volatilities, as in Table 2.

Using the method presented by Black et al., we can find the short rates on tree nodes step by step. Figure 1. displays the full tree of short rates at one year intervals that matches the term structure of Table 1.

Table 1. Features of the Bonds Selected

Type of Bond	Par Value ¥	Coupon Rate %	Maturity (years)	Day of Maturity (mm/dd/yy)
Treasury Bond 96(6)	100	11.83	10	06/14/2006
Treasury Bond 99(5)	100	3.28	8	08/20/2007
98 Oil	100	4.5	8	09/08/2007

Table 2. Term Structure of Treasury Bond Yield of China (July 1, 2003)

Maturity (years)	Yield %	Yield Volatility %
1	2.08	39.5
2	2.21	32
3	2.33	27
4	2.44	24
5	2.55	23

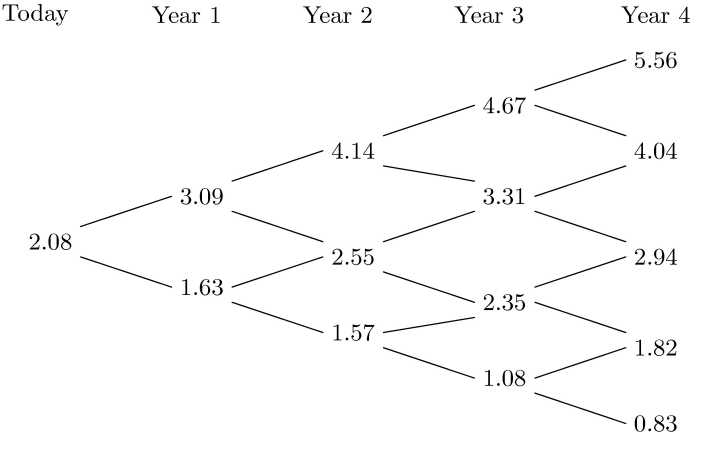


Fig. 1. Short Rates that Match the Term Structure of Table 1

The prices of the bonds at the beginning of the horizon are known(In July 1st, 2003, the prices of 96(6), 99(5) and 98 Oil are 127.37, 102.94 and 109.38 respectively.). They are uncertain in subsequent years which are mainly determined by short interest rates. To get the future prices, the *option adjusted premium*(OAP) of the bonds must be known. Using (7), we know the OAP of 96(6), 99(5) and

98 Oil is 0.9710, 0.9836, and 0.9925, respectively. Then the future prices of the bonds under different scenario paths can be obtained.

Having known the riskless short rates and prices of the bonds at different periods, we can solve the model. Assume the investment at initiate period is ¥10,000, and goal wealth of each period are 10300, 10600, 11000 respectively. Setting the weights of shortfall costs of each period is 50, 30, and 20 in turn. Then the expected wealth that we can get at the end of each period are 10316.23, 10695.20, and 10979.94, respectively.

Comparing the result of the dynamic stochastic programming model with that of a fixed-mix model, which is commonly used in the market, we can identify the effect of our model. There are previous literature reported the results of comparing the performance of stochastic programming models with others. The comparison is a complex task, as it involves solving a large number of stochastic programs.

Now we apply the fixed mix model to the bond portfolio management problem presented above. It has the same objective function with the dynamic stochastic programming model, for the reason to guarantee that decision maker has the same risk attitude.

The performance of fixed mix model is listed in Table 3, which is compared with that of stochastic programming model. Comparing the wealth at the end of every period under different models, we can see the dynamic stochastic programming approach weakly dominates the fixed mix approach. It is also illustrated by other tests over a large number of realistic scenarios created by means of simulation. This indicates the former has better ability to adapt to the information in the scenario tree. We expect that the degree of dominance would increase if the number of periods in the decision model is increased.

Table 3. Performance of two approaches

Type of model	T=1	T=2	T=3
stochastic programming	10316.23	10695.20	10979.94
fixed mix	10309.45	10667.18	10894.79

5 Conclusions

Dynamic stochastic programming models are extremely versatile. They can incorporate transaction costs, cash infusions or withdrawals, risk of default and do not depend on assumptions for temporal independence or normality of returns. They are powerful tools to deal with complex bond portfolio management.

When using stochastic programming models, scenario generations play important roles, which determine the validity of the models. A few methods have

been presented to model economic factors and asset returns, such as statistical modelling with the Value-at-Risk approach(Jamshidian and Zhu (1997)[11], Consiglio and Zenios(2001)[12]), vector autoregressive models (Boender (1997)[13]), etc.. Future research may consider to evaluate different methods in generating scenarios of the uncertainties in the multi-period models.

References

1. Bradley, S.P. and Crane, D.B., "A dynamic model for bond portfolio management". *Management Science*, 19: 139-151, 1972.
2. Carino, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K. and Ziemba, W.T., "The Russell-Yasuda Kasai model: an asset/liability model for a Japanese insurance company using multistage stochastic programming". *Interfaces*, 24(1): 29-49, 1994.
3. Golub, B., Holmer, M., Mc Kendall, R. and Zenios, S.A., "Stochastic programming models for money management". *European Journal of Operational Research*, 85: 282-296, 1995.
4. Zenios, S.A., Holmer, M.R., McKendall, R., and Vassiadou-Zeniou, C., "Dynamic models for fixed-income portfolio management under uncertainty". *Journal of Economic Dynamics and Control*, 22: 1517-1541, 1998.
5. Kouwenberg, R., "Scenario generation and stochastic programming models for asset liability management". *European Journal of Operational Research*, 134: 279-292, 2001.
6. Ziemba, W.T. and Mulvey, J.M., *Worldwide Asset and Liability Modlling*, Cambridge University Press, 1998.
7. Wang, S.Y. and Xia, Y.S., *Portfolio Selection and Asset Pricing*, Berlin: Spring-Verlag, 2002.
8. Black, F., Derman, E. and Toy, W., "A one-factor model of interest rates and its application to treasury bond options". *Financial Analysts Journal*, 33-39, 1990.
9. Hakansson, N.H. and Ziemba, W.T., "Capital growth theory". In: Jarrow, R.A., Maksimovic, V., Ziemba, W.T. (Eds.), *Finance*. North-Holland, Amsterdam, 65-86, 1995.
10. Babbel, D.F. and Zenios, S.A., "Pitfalls in the analysis of option-adjusted spreads". *Financial Analysts Journal*, 65-69, 1992.
11. Jamshidian F. and Zhu Y., "Scenario simulation: theory and methodology". *Finance and Stochastics*, 43-67, 1997.
12. Consiglio A. and Zenios S.A., "Integrated simulation and optimization models for tracking international fixed income indices". *Mathematical Programming*, Series B, 89: 311-339, 2001.
13. Boender G.C.E., "A hybrid simulation/optimisation scenario model for asset/liability management". *European Journal of Operational Research*, 99: 126-135, 1997.

Communication Leading to Nash Equilibrium without Acyclic Condition
– S4-Knowledge Model Case –

Takashi Matsuhisa*

Department of Liberal Arts and Sciences, Ibaraki National College of Technology
Nakane 866, Hitachinaka-shi, Ibaraki 312-8508, Japan.
mathisa@ge.ibaraki-ct.ac.jp

Abstract. A pre-play communication-process is presented which leads to a Nash equilibrium of a strategic form game. In the communication process each player predicts the other players' actions, and he/she communicates privately his/her conjecture through message according to a protocol. All the players receiving the messages learn and revise their conjectures. We show that after a long round of the communication the profile of players' conjectures in the revision process leads a mixed strategy Nash equilibrium of the game.

Keywords: Knowledge, Information, Modal logic **S4**, Nash equilibrium, Communication, Mutual learning, Protocol, Conjecture, Non-corporative game.

1 Introduction

The purpose of this paper is to present the pre-play communication-process leading to a mixed strategy Nash equilibrium of a strategic form game.

The stage sets up as follows: The players start with the same prior distribution on a finite state-space. In addition they have the private information given by the non-partition structure corresponding to the modal logic **S4**. Each player communicates privately his/her conjecture about the other players' actions through messages according to a protocol, and the receiver of the message updates her/his conjecture. When a player communicates with another, the other players are not informed about the contents of the message. Suppose that all players are expected utility maximizers. Then

Main Theorem. *The players' predictions about the other players' actions regarding as the future conjectures converge in the long run, and those convergent conjectures constitute a mixed strategy Nash equilibrium of the game.*

The concept of Nash equilibrium has become central in game theory, economics and its related fields. Yet a little is known about the process by which

* Partially supported by the Grant-in-Aid for Scientific Research(C)(2)(No.14540145) in the Japan Society for the Promotion of Sciences.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 884–891, 2004.
© Springer-Verlag Berlin Heidelberg 2004

players learn if they do. Recent papers by J. S. Jordan [4] and E. Kalai and E. Lehrer [5] indicate increasing interest in the mutual learning processes in Bayesian games which leads to Bayesian equilibrium.

As for as Nash's fundamental notion of strategic equilibrium (J. F. Nash [8]) is concerned, R.J. Aumann and A. Brandenburger [1] gives epistemic conditions for Nash equilibrium. However it is not clear just what learning process lead to Nash equilibrium. The present paper aims to fill this gap. The pre-play communication process according to a protocol is proposed which is a mutual learning process leading to a Nash equilibrium of a strategic form game as a cheap talk. The emphasis is on that any topological assumption on the communication graph is not required. T. Matsuhisa [6] proved the theorem under the assumption that the graph contains no cycle.

This paper organizes as follows. Section 2 presents the communication process for a game according a protocol. In Section 3 we give the explicit statement of the main theorem (Theorem 1) and we shall prove it.

2 The Model

Let Ω be a non-empty set called a *state-space*, N a set of finitely many *players* $1, 2, \ldots n$ $(n \geq 2)$, and let 2^Ω be the family of all subsets of Ω. Each member of 2^Ω is called an *event* and each element of Ω called a *state*. Let μ be a probability measure on Ω which is common for all players.

2.1 Information and Knowledge[1]

An *information structure* $(P_i)_{i \in N}$ is a class of mappings P_i of Ω into 2^Ω. It is called an *RT-information structure* if for every player i the two properties are true: For each ω of 2^Ω,

Ref $\omega \in P_i(\omega)$;

Trn $\xi \in P_i(\omega)$ implies $P_i(\xi) \subseteq P_i(\omega)$.

Given our interpretation, player i for whom $P_i(\omega) \subseteq E$ knows, in the state ω, that some state in the event E has occurred. In this case we say that in the state ω the player i knows E. By i's *knowledge operator* we mean the mapping $K_i : 2^\Omega \to 2^\Omega$ defined by $K_i E = \{\omega \in \Omega | P_i(\omega) \subseteq E\}$. This is the set of states of Ω in which i knows that E has occurred. We note that K_i satisfies the following properties[2]: For every E, F of 2^Ω,

N $K_i \Omega = \Omega$ and $K_i \emptyset = \emptyset$; **K** $K_i(E \cap F) = K_i E \cap K_i F$;

T $K_i F \subseteq F$; **4** $K_i F \subseteq K_i K_i F$.

[1] See Bacharach [2], Samet [10], Binmore [3]

[2] According to these we can say the structure $\langle \Omega, (K_i)_{i \in N} \rangle$ is a model for the multi-modal logic $\mathbf{S4}_n$.

The set $P_i(\omega)$ will be interpreted as the set of all the states of nature that i believes to be possible at ω, and K_iE will be interpreted as the set of states of nature for which i believes E to be possible. We will therefore call P_i i's *possibility operator* on Ω and also will call $P_i(\omega)$ i's *possibility set* at ω. An event E is said to be i's *truism* if $E \subseteq K_iE$ We should note that the RT-information structure P_i is uniquely determined by the knowledge operator K_i such that $P_i(\omega)$ is the minimal truism containing ω; that is, $P_i(\omega) = \bigcap_{\omega \in K_iE} E = \bigcap_{\omega \in T = K_iT} T$.

2.2 Game and Knowledge[3]

By a *game* G we mean a *finite* strategic form game $\langle N, (A_i)_{i \in N}, (g_i)_{i \in N} \rangle$ with the following structure and interpretations: N is a finite set of players $\{1, 2, \ldots, i, \ldots n\}$ with $n \geq 2$, A_i is a finite set of i's *actions* (or i's pure strategies) and g_i is an i's *payoff-function* of A into $\mathbb{R}$, where A denotes the product $A_1 \times A_2 \times \cdots \times A_n$, A_{-i} the product $A_1 \times A_2 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_n$. We denote by g the n-tuple $(g_1, g_2, \ldots g_n)$ and denote by a_{-i} the $(n-1)$-tuple $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ for a of A .

A probability distribution ϕ_i on A_{-i} is said to be i's *overall conjecture* (or simply i's *conjecture*). For each player j other than i, this induces the marginal on j's actions; we call it i's *individual conjecture* about j (or simply i's *conjecture about* j.) Functions on Ω are viewed like random variables in a probability space (Ω, μ). If $\mathbf{x}$ is a such function and x is a value of it, we denote by $[\mathbf{x} = x]$ (or simply by $[x]$) the set $\{\omega \in \Omega | \, \mathbf{x}(\omega) = x\}$.

An RT-information structure $(P_i)_{i \in N}$ with a common-prior μ yields the distribution defined by $\mathbf{q}_i(a, \omega) = \mu([\mathbf{a} = a] | P_i(\omega))$; and i's overall conjecture defined by the marginal $\mathbf{q}_i(a_{-i}, \omega) = \mu([\mathbf{a}_{-i} = a_{-i}] | P_i(\omega))$ which is viewed as a random variable of ϕ_i. Where we assume that $[a_i] := [\mathbf{a}_i = a_i]$ is i's truism for every a_i of A_i. The pay-off functions $g = (g_1, g_2, \ldots, g_n)$ is said to be *actually played* at a state ω if $\omega \in [\mathbf{g} = g] := \bigcap_{i \in N} [\mathbf{g}_i = g_i]$. i's action a_i is said to be *actual* at a state ω if $\omega \in [\mathbf{a}_i = a_i]$.

Player i is said to be *rational* at ω if each i's actual action a_i maximizes the expectation of his actually played pay-off function g_i at ω when the other players actions are distributed according to his conjecture $\mathbf{q}_i(*, \omega)$: Formally, letting $g_i = \mathbf{g}_i(\omega)$ and $a_i = \mathbf{a}_i(\omega)$, $\mathbf{Exp}(g_i(a_i, \mathbf{a}_{-i}); \omega) \geq \mathbf{Exp}(g_i(b_i, \mathbf{a}_{-i}); \omega)$ for every b_i in A_i.[4] Let R_i denote the set of all the states at which player i is rational, and $R = \cap_{j \in N} R_j$.

2.3 Protocol[5]

We assume that the players in N communicate by sending *messages*. Let T be the time horizontal line $\{0, 1, 2, \cdots t, \cdots\}$. A *protocol* among N is a mapping Pr of the set of non-negative integers into the Cartesian product $N \times N$ that assigns

[3] See Aumann and Brandenburger [1]

[4] We denote $\mathbf{Exp}(g_i(b_i, \mathbf{a}_{-i}); \omega) := \sum_{a_{-i} \in A_{-i}} g_i(b_i, a_{-i}) \; \mathbf{q}_i(a_{-i}, \omega)$.

[5] See Parikh and Krasucki [9]

to each t a pair of players $(s(t), r(t))$ such that $s(t) \neq r(t)$. Here t stands for *time* and $s(t)$ and $r(t)$ are, respectively, the *sender* and the *receiver* of the communication which takes place at time t. We consider the protocol as the directed graph whose vertices are the set of all players M and such that there is an edge (or an arc) from i to j if and only if there are infinitely many t such that $s(t) = i$ and $r(t) = j$.

A protocol is said to be *fair* if the graph is strongly-connected; in words, every player in this protocol communicates directly or indirectly with every other player infinitely often. It is said to contain a *cycle* if there are players $i_1, i_2, \ldots, i_k$ with $k \geq 3$ such that for all $m < k$, i_m communicates directly with i_{m+1}, and such that i_k communicates directly with i_1. The *period* of the protocol is the minimal number of all the natural number m such that $\Pr(t + m) = \Pr(t)$ for every t.

2.4 Pre-play Communication

By this we intuitively mean the learning process such that each player communicates privately his/her conjecture about the other players' actions through messages according to a protocol, and she/he updates her/his conjecture according to the message received. In addition, at every stage each player communicates privately not only his/her conjecture about the others' actions but also his/her rationality as messages, the receivers update their private information and revise their conjecture. When a player communicates with another, the other players are not informed about the contents of the message. Formally,

Definition 1. A *pre-play communication-process* according to a protocol among N for a game G with revisions of players' conjectures is a tuple

$$\langle \Pr, (P_i^t)_{i \in N}, (\phi_i^t)_{i \in N} |\ t \in T \rangle$$

with the following structures: the players have a common-prior μ on Ω, the protocol Pr among N, $\Pr(t) = (s(t), r(t))$, is fair and it satisfies the conditions that $r(t) = s(t+1)$ for every t and that the communications proceed in *rounds*.[6] The information structure P_i^t at time t is the mapping of Ω into 2^Ω for player i that is defined inductively as follows: If $i = s(t)$ is a sender at t, he/she sends the message W_i^t defined as below to $j = r(t)$ at t.

- Set $P_i^0(\omega) = P_i(\omega)$.
- Assume that P_i^t is defined. It yields the overall conjecture $\mathbf{q}_i^t(a_{-i}, \omega) = \mu([\mathbf{a}_{-i} = a_{-i}] | P_i^t(\omega))$; whence
  - R_i^t denotes the set of all the state ω at which i is *rational* according to his conjecture $\mathbf{q}_i^t(\omega)$:[7]

[6] That is, there exists a time m such that $\Pr(t) = \Pr(t + m)$ for all t.

[7] That is, each i's actual action a_i maximizes the expectation of his pay-off function g_i being actually played at ω when the other players actions are distributed according to his conjecture $\mathbf{q}_i^t(\omega)$ at time t. Formally, letting $g_i = \mathbf{g}_i(\omega)$, $a_i = \mathbf{a}_i(\omega)$, the

- $\mathcal{Q}_i^t$ denotes the partition induced by $\mathbf{q}_i{}^t$ on Ω, which is decomposed into the components $\mathcal{Q}_i^t(\omega)$ consisting of all the states ξ such that $\mathbf{q}_i{}^t(\xi) = \mathbf{q}_i{}^t(\omega)$.
- $\mathcal{G}_i$ denotes the partition $\{[\mathbf{g}_i = g_i],\ \Omega \setminus [\mathbf{g}_i = g_i]\}$ of Ω, and $\mathcal{R}_i^t$ the partition $\{R_i^t, \Omega \setminus R_i^t\}$.
- W_i^t denotes the join $\mathcal{G}_i \vee \mathcal{Q}_i^t \vee \mathcal{R}_i^t$ the partition of Ω generated by $\mathcal{G}_i$, $\mathcal{Q}_i^t$ and $\mathcal{R}_i^t$.[8]

– Then P_i^{t+1} is defined as follows:
- If i is a receiver of a message at time $t+1$ then $P_i^{t+1}(\omega) = P_i^t(\omega) \cap W_{s(t)}^t(\omega)$.
- If not, $P_i^{t+1}(\omega) = P_i^t(\omega)$. □

It is of worth noting that $(P_i^t)_{i\in N}$ is an RT-information structure for every $t \in T$. We require that the pre-play communication-process satisfies the following two conditions: Let K_i^t be the knowledge operator corresponding to P_i^t;[9]

A-1 For each $i \in N$ and every $t \in T$, $[\phi_i^t] \subseteq K_i^t([\phi_i^t])$ and $R_i^t \subseteq K_i^t(R_i^t)$;

A-2 For every $t \in T$, $\bigcap_{i\in N} K_i^t([g_i] \cap [\phi_i^t] \cap R_i^t) \neq \emptyset$.

The specification of **A-1** is that each player's conjecture and his/her rationality are truism, and the specification of **A-2** is that each player knows his/her pay-off, rationality and conjecture at every time t.

Remark 1. For $i \in N$, the sequence of correspondences $\{P_i^t \mid t = 0, 1, 2, \ldots\}$ is stationary in finitely many rounds. Furthermore so is the sequence of i's conjectures $\{\mathbf{q}_i^t \mid t = 0, 1, 2, \ldots\}$ in finitely many rounds. That is, there is a sufficiently large time $\tau \in T$ such that for every i, for all $\omega \in \Omega$ and for all $t \geq \tau$, $P_i^t(\omega) = P_i^\tau(\omega)$, and therefore $\mathbf{q}_i^t = \mathbf{q}_i^\tau$.

Notation: For sufficient large $t \geq \tau$ we denote τ by ∞. Hence we can write $\mathbf{q}_i^\tau$ by $\mathbf{q}_i^\infty$ and ϕ_i^τ by ϕ_i^∞

3 Proof of Main Theorem

We now state the main theorem in Introduction as below and we shall prove it:

Theorem 1. *Suppose that the players in a strategic form game have a common-prior. In a pre-play communication process according to a protocol among all players in the game with revisions of their conjectures* $\{(\phi_i^t)_{i\in N} \mid t = 0, 1, 2, \ldots\}$, *there exists a time* τ *such that for each* $t \geq \tau$, *the* n*-tuple* $(\phi_i^t)_{i\in N}$ *induces a mixed strategy Nash equilibrium of the game.*

expectation at time t, $\mathbf{Exp}^t$, is defined by

$$\mathbf{Exp}^t(g_i(b_i, \mathbf{a}_{-i}); \omega) := \sum_{a_{-i}\in A_{-i}} g_i(b_i, a_{-i})\ \mathbf{q}_i^t(\omega)(a_{-i}).$$

Player i is said to be rational according to his conjecture $\mathbf{q}_i^t(\omega)$ at ω if for all b_i in A_i, $\mathbf{Exp}^t(g_i(a_i, \mathbf{a}_{-i}); \omega) \geq \mathbf{Exp}^t(g_i(b_i, \mathbf{a}_{-i}); \omega)$.

[8] Therefore the component $W_i^t(\omega) = [g_i] \cap [\phi_i^t] \cap R_i^t$ if $\omega \in [g_i] \cap [\phi_i^t] \cap R_i^t$.

[9] That is, K_i^t is defined by $K_i^t E = \{\omega \in \Omega | P_i^t(\omega) \subseteq E\}$.

A non-empty event H is said to be P_i- *invariant* if for every ξ of H, $P_i(\xi)$ is contained in H.

Fundamental lemma *Let $(P_i)_{i\in N}$ be an RT-information structure with μ a common-prior. Let X be an event and q_i i's posterior of X; that is, $q_i = \mu(X|P_i(\omega))$. If there is an event H such that the following two conditions* (a), (b) *are true:*

(a) *H is non-empty and it is P_i-invariant,*
(b) *H is contained in $[q_i] := \{\omega \in \Omega \,|\, \mu(X|P_i(\omega)) = q_i\}$,*

then we obtain that $\mu(X|H) = q_i$.

Proof. See Matsuhisa and Kamiyama [7].

Let $\omega_\infty \in \bigcap_{i\in N} K_i^\infty([g_i] \cap [\phi_i^\infty] \cap R_i^\infty) \subseteq \bigcap_{i\in N}([g_i] \cap [\phi_i^\infty] \cap R_i^\infty)$. The following result is the another key to proving Theorem 1:

Proposition 1. *In a pre-play communication-process among all the players in a game with revisions of their conjectures $\{(\phi_i^t)_{i\in N} \,|\, t = 0, 1, 2, \ldots\}$, both the marginals of the conjectures ϕ_i^∞ and ϕ_j^∞ on A_{-i-j} must coincide for all $i, j \in N$; that is, $\phi_i^\infty(a_{-i-j}) = \phi_j^\infty(a_{-i-j})$ for all $a \in A$.*

Proof: It suffices to verify that $\mathbf{q}_i^\infty(a;\omega) = \mathbf{q}_j^\infty(a;\omega)$ for all $(a;\omega) \in A\times\Omega$. Let us first consider the case that $(i,j) = (s(\infty), t(\infty))$. We denote by Π the partition on Ω with which each component is defined by $\Pi_i^\infty(\omega) = \{\xi \in \Omega \mid P_i^\infty(\xi) = P_i^\infty(\omega)\}$ In view of the construction of $\{P_i^t\}_{t\in T}$ we can observe that $P_i^\infty(\omega)$ is P_j^∞-invariant; i.e., $P_j^\infty(\xi) \subseteq P_i^\infty(\omega)$ for all $\xi \in P_i^\infty(\omega)$. It immediately follows that $P_i^\infty(\omega)$ is decomposed into a disjoint union of components $\Pi_j^\infty(\xi)$ for $\xi \in P_i^\infty(\omega)$; $P_i^\infty(\omega) = \bigcup_{k=1,2,\ldots,m} \Pi_j^\infty(\xi_k)$ where $\xi_k \in P_i^\infty(\omega)$. It can be observed that $\mu([\mathbf{a} = a]|\, P_i^\infty(\omega)) = \sum_{k=1}^m \lambda_k \mu([\mathbf{a} = a]|\, \Pi_j^\infty(\xi_k))$ for some $\lambda_k > 0$ with $\sum_{k=1}^m \lambda_k = 1$.[10] By Fundamental lemma we note that $\mu([\mathbf{a} = a]|\, \Pi_j^\infty(\xi_k)) = \mathbf{q}_i^\infty(a;\xi_k)$ and thus by the above equation it can be observed that for all $\omega \in \Omega$ there is some $\xi_\omega \in P_i^\infty(\omega)$ such that $\mathbf{q}_i^\infty(a;\omega) \leq \mathbf{q}_j^\infty(a;\xi_\omega)$. Continuing this process according to the *fair* protocol the below facts can be plainly verified: For each $\omega \in \Omega$

1. For every $i \neq j$, $\mathbf{q}_i^\infty(a;\omega) \leq \mathbf{q}_j^\infty(a;\xi)$ for some $\xi \in \Omega$; and
2. $\mathbf{q}_i^\infty(a;\omega) \leq \mathbf{q}_i^\infty(a;\xi) \leq \mathbf{q}_i^\infty(a;\zeta) \leq \cdots$ for some $\xi, \zeta, \cdots \in \Omega$.

Since Ω is finite the equation $\mathbf{q}_i^\infty(a;\omega) = \mathbf{q}_j^\infty(a;\omega)$ can be obtained for every a, ω and for all i, j, in completing the proof.

Proof of Theorem 1. We denote by $\Gamma(i)$ the set of all the players who directly receive the message from i on N; i.e., $\Gamma(i) = \{\, j \in N \mid (i,j) = \Pr(t)$ for some $t \in T\}$. For any subset I of N denote $a_{-I} := (a_i)_{i\in N\setminus I}$.

For each $i \in N$, we denote $[g_i] \cap [\phi^\infty] \cap R^\infty$ by F_i. It is noted that F_i is a non-empty P_i-invariant set because $\emptyset \neq \bigcap_{i\in N}([g_i] \cap [\phi_i^\infty] \cap R_i^\infty) \subseteq F_i$ and

[10] This property is called *convexity* in Parikh and Krasucki [9].

because $P_i^\infty(\omega) \subseteq F_i$ for every $\omega \in F_i$ by the definition. We observe the first point that for each $i \in N$, $j \in \Gamma(i)$ and for every $a \in A$,

$$\mu([\mathbf{a}_{-j} = a_{-j}] \,|\, F_i \cap F_j) = \phi_j^\infty(a_{-j}) : \tag{1}$$

For, we note that $F_i \cap F_j \subseteq [\phi_j^\infty(a_{-j})]$ and $F_i \cap F_j$ is P_j^∞-invariant because $j \in \Gamma(i)$. Hence by Fundamental lemma (3), we plainly obtain (1) as required. Then summing over a_{-i}, we can observe that

$$\mu([\mathbf{a}_i = a_i] \,|\, F_i \cap F_j) = \phi_j^\infty(a_i) \ \text{ for any } a \in A. \tag{2}$$

In view of Proposition 1 it can be observed that $\phi_j^\infty(a_i)$ is independent of the choices of every $j \in \Gamma(i)$.

We set the probability distribution σ_i on A_i by $\sigma_i(a_i) := \phi_j^\infty(a_i)$, and the profile $\sigma = (\sigma_i)$. We observe the second point that for every $a \in \prod_{i \in N} \mathrm{Supp}(\sigma_i)$,

$$\phi_i^\infty(a_{-i}) = \sigma_1(a_1) \cdots \sigma_{i-1}(a_{i-1}) \sigma_{i+1}(a_{i+1}) \cdots \sigma_n(a_n) : \tag{3}$$

In fact, viewing the definition of σ_i we shall show that $\phi_i^\infty(a_{-i}) = \prod_{k \in N \setminus \{i\}} \phi_i^\infty(a_k)$. To verify this it suffices to show that for every $k = 1, 2, \cdots, n$, $\phi_i^\infty(a_{-i}) = \phi_i^\infty(a_{-I_k}) \prod_{k \in I_k \setminus \{i\}} \phi_i^\infty(a_k)$: We prove by induction on k. For $k = 1$ the result is immediate. Suppose it is true for $k \geq 1$. On noting the protocol is fair, we can take the sequence of sets of players $\{I_k\}_{1 \leq k \leq n}$ with the following properties:

(a) $I_1 = \{i\} \subset I_2 \subset \cdots \subset I_k \subset I_{k+1} \subset \cdots \subset I_m = N$:

(b) For every $k \in N$ there is a player $i_{k+1} \in \bigcup_{j \in I_k} \Gamma(j)$ with $I_{k+1} \setminus I_k = \{i_{k+1}\}$.

We let take $j \in I_k$ such that $i_{k+1} \in \Gamma(j)$. Set $H_{i_{k+1}} := [\mathbf{a}_{i_{k+1}} = a_{i_{k+1}}] \cap F_j \cap F_{i_{k+1}}$. We note that $H_{i_{k+1}}$ is not empty because $\sigma_i(a_i) = \phi_j^\infty(a_i) = \mu([\mathbf{a}_i = a_i] \,|\, F_i \cap F_j) > 0$ in viewing of (2), and we note that $H_{i_{k+1}}$ is $P_{i_{k+1}}^t$- invariant which is included in $[\phi_{i_{k+1}}^\infty(a_{-j-i_{k+1}})]$. It immediately follows from Fundamental lemma (3) that $\mu([\mathbf{a}_{-j-i_{k+1}} = a_{-j-i_{k+1}}] \,|\, H_{i_{k+1}}) = \phi_{-j-i_{k+1}}^\infty(a_{-j})$. Dividing $\mu(F_j \cap F_{i_{k+1}})$ yields that $\mu([\mathbf{a}_{-j} = a_{-j}] \,|\, F_j \cap F_{i_{k+1}}) = \phi_{i_{k+1}}^\infty(a_{-j}) \mu([\mathbf{a}_{i_{k+1}} = a_{i_{k+1}}] \,|\, F_j \cap F_{i_{k+1}})$. In viewing of (1) and (2) it follows $\phi_j^\infty(a_{-j}) = \phi_{i_{k+1}}^\infty(a_{-j-i_{k+1}}) \phi_j^t(a_{i_{k+1}})$; then summing over a_{I_k} it immediately follows from Proposition 1 that $\phi_i^\infty(a_{-I_k}) = \phi_i^\infty(a_{-I_k - i_{k+1}}) \phi_i^\infty(a_{i_{k+1}})$. Viewing (4) we have just observed that $\phi_i^\infty(a_{-i}) = \phi_i^\infty(a_{-I_{k+1}}) \prod_{k \in I_{k+1} \setminus \{i\}} \phi_i^\infty(a_k)$, as required. Furthermore we can observe from (4) that all the other players i than j agree on the same conjecture $\sigma_j(a_j) = \phi_i^\infty(a_j)$ about j.

Each action a_i with $\phi_i^\infty(a_i) > 0$ for some $j \in \Gamma(i)$ maximizes g_i against ϕ_i^∞ because $a_i = \mathbf{a}_i(\omega_i)$, $g_i = \mathbf{g}_i(\omega_i)$ and $\phi_i^\infty = \mathbf{q}_i^\infty(\omega_i)$ at some state ω_i of $H_i = [a_i] \cap F_i \cap F_j$. Viewing (3) we conclude that each action a_i appearing with positive probability in σ_i maximizes g_i against the product of the distributions σ_l with $l \neq i$. This implies that the profile $\sigma = (\sigma_i)_{i \in N}$ is a Nash equilibrium of G. This establishes the theorem.

4 Concluding Remarks

Our real concern is with what learning process leads to a mixed strategy Nash equilibrium of a finite strategic form game from the epistemic point view. As we have observed, in the pre-play communication process with revisions of players' conjectures about the other actions, their predictions induces a Nash equilibrium of the game in the long run. Where the players privately communicate each other through message according to any non-acyclic graph, and they are required neither to have the common-knowledge assumption about their conjectures nor to have a partition information structure.

The communication process treated in this article will give a new aspect of the algorithms converging to Nash equilibrium from the epistemic point of view. This issue needs to be sorted out at a more fundamental level, and it has not been discussed at all. There is a research agenda of potential interest which we hope to pursue further.

References

1. Aumann, R. J. and Brandenburger, A.: Epistemic conditions for Nash equilibrium, Econometrica **63** (1995) 1161–1180.
2. Bacharach, M.: Some extensions of a claim of Aumann in an axiomatic model of knowledge, Journal of Economic Theory **37** (1985) 167–190.
3. Binmore, K. *Fun and Games.* D. C. Heath and Company, Lexington, Massachusetts USA (1992) xxx+642pp.
4. Jordan, J. S.,: Bayesian learning in normal form games, Games and Economic Behavior **3** (1991) 60–81.
5. Kalai, E. and Lehrer, E.: Rational learning to Nash equilibrium, Econometrica **61** (1993) 1019–1045.
6. Matsuhisa, T.: Communication leading to Nash equilibrium, in T. Maruyama (Editor): *Mathematical Economics*, Surikaiseki-kenkyusyo Kokyuroku **1165** (2000) 245-256.
7. Matsuhisa, T. and Kamiyama, K.: Lattice structure of knowledge and agreeing to disagree, Journal of Mathematical Economics **27** (1997) 389–410.
8. Nash, J. F. (1950) Equilibrium points in n-person games, Proceedings of the National Academy of Sciences of the United States of America, **36** 48-49.
9. Parikh, R. and Krasucki, P.: Communication, consensus, and knowledge, Journal of Economic Theory **52** (1990) 178–189.
10. Samet, D.: Agreeing to disagree in infinite information structures, International Journal of Game Theory **21** (1992) 213–218.

Support Vector Machines Approach to Credit Assessment*

Jianping Li [1,2], Jingli Liu[1,2], Weixuan Xu[2], and Yong Shi[3]

[1] University of Science & Technology of China, Hefei, 230026, P.R.China
[2]Institute of Policy and Management,
Chinese Academy of Sciences, Beijing 100080, P.R.China
ljp@mail.casipm.ac.cn, jianpingli@yahoo.com

[3] Graduate School of Chinese Academy of Sciences, Beijing 100039, P.R.China
yshi@unomaha.edu

Abstract. Credit assessment has attracted lots of researchers in financial and banking industry. Recent studies have shown that Artificial Intelligence (AI) methods are competitive to statistical methods for credit assessment. This article applies support vector machines (SVM), a relatively new machine learning technique, to the credit assessment problem for better explanatory power. The structure of SVM has many computation advantages, such as special direction at a finite sample and irrelevance between the complexity of algorithm and the sample dimension. A real credit card data experiment shows that SVM method has outstanding assessment ability. Compared with the methods that are currently used by a major Chinese bank, the SVM method has a great potential superiority in predicting accuracy.

Keywords: Credit assessment; Classification; Support vector machines

1 Introduction

Credit risk management has played a key role in financial and banking industry. Inferior credit risk assessment tool is the primary reason of enterprise bankruptcy. Generally speaking, credit risk management includes credit risk analysis, assessment (measurement) of enterprise credit risk and how to manage the risk efficiently, while credit risk assessment is the basic and critical factor in credit risk management. The main purpose of credit risk assessment is to measure the default possibility of borrowers and provide the loaner a decision-aid by conducting qualitative analysis and qualitative computation to the possible factors that will cause credit risk. At present, the classification is the most popular method used in credit risk assessment. That is, according to the financial status of the borrowers, we can use a credit scoring system to estimate the corresponding risk rate so that the status can be classified as

* This research has been partially supported by National Excellent Youth Fund under 70028101, and the President Fund of Chinese Academy of Sciences (CAS) (2003).

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 892–899, 2004.
© Springer-Verlag Berlin Heidelberg 2004

normal or default. Thus the problem can be transformed to some kind of classification.

The Multivariate Statistical Analysis technique especially Multiple Discriminant Analysis (MDA) has been widely used in credit risk classification. This basic idea is to use the historical sample as training set to build a discriminant function and use the function to classify the new sample. The outstanding advantage of Multivariate Statistical Analysis is that it is simple and easy to interpret. But the requiring of multiple normal distributed data and the equality of covariance matrix is conflict with the real data. Thus it brings many questions [1]. In order to solve this problem, researchers modified MDA from different aspects, for example, adopting logarithmic transformation, QDA model, Logit analysis model, neural network model and decision tree [2, 3]. These techniques can partially solve the problem, but they are not perfect. In general, the modification of MDA is not ideal. Because of the complexity of credit risk and various data structure, the theoretical advantage of many methods doesn't work well in practice. In China's situation, lacking of samples and high dimension are the notable characters of Chinese credit data. Because the historical sample data is small and the data character used in credit risk classification is not steady, using MDA in credit risk classification will not produce practical results [4].

This paper applies Support Vector Machine (SVM) to the field of credit assessment. SVM is a novel learning machine based on mathematical programming theory, which has yielded excellent performance on a wide range of problems. The remainder of the paper is structured as follows. The basic principle of SVM follows first the introduction. Then, some ideas how to use SVM in credit risk management is provided. In the experimental study, we provide descriptions of the data sets, the experiment results and analysis. We also conduct a comparison study to the current method used in a major Chinese commercial bank. Finally, we discuss the future research directions.

2 Analytical Methods

2.1 Review of Support Vector Machines

Since the mid of 1990, SVM appeared with the continuous development and maturation of machine learning theory. It has exhibited performance which superior to the other existing methods.

SVM is a novel learning machine introduced first by Vapnik [5]. It is a theory of machine learning focusing on small sample data based on the structural risk minimization principle from computational learning theory. Hearst et al. [6] positioned the SVM algorithm at the intersection of learning theory and practice: ''it contains a large class of neural nets, radial basis function (RBF) nets, and polynomial classifiers as special cases. Yet it is simple enough to be analyzed mathematically, because it can be shown to correspond to a linear method in a high dimensional feature space nonlinearly related to input space.'' In this sense, support vector machines can be a good candidate for combining the strengths of more theory-driven

and easy to be analyzed conventional statistical methods and more data-driven, distribution free and robust machine learning methods.

The main advantages of SVM can be summarized as follows:

1) SVM is used in the situation of finite sample data. It aims to get the optimal solution based on the present information rather than the optimal value when the number of sample tends to be infinite.

2) The algorithm is finally transformed into the optimization of quadratic program. Theoretically, it will get a global optimization value, which solves the unavoidable local optimization problem while using neural network.

3) The algorithm performs a nonlinear mapping from the original data space into some high dimension feature space, in which it constructs a linear discriminant function to replace the nonlinear functions in the original data space. This special character assures that SVM has good generalization ability. At the same time, it solves the problem of dimension disaster because its computation complexity is independent to the sample dimension.

A brief description of the SVM algorithm is provided here, for more details please refer to Refs [7, 8].

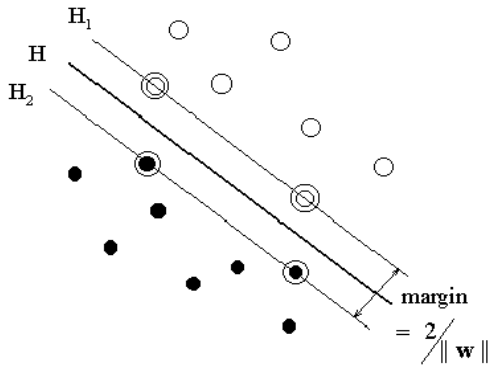


Fig. 1. Optimal Separating Hyperplane

Consider the problem of separating the set of training vectors belonging to two separate classes. D= $\{(x^1, y^1),\ldots, (x^l, y^l)\}$, $x \in R^n$, $y \in \{-1,1\}$, with a hyper plane, $(\omega \cdot x) + b = 0$. Figure 1 is a simple linearly separatable case.

Solid points and circle points represent two kinds of sample separately. H is the separating line. H_1 and H_2 are the closest lines parallel to the separating line of the two-class sample vectors. The distance between H_1 and H_2 is called margin. The separating hyperplane is said to be optimal if it classifies the samples into two classes without error (training error is zero) and the margin is maximal. The sample vectors in H_1, H_2 are called support vectors. The separating hyperplane equation is $(\omega \cdot x) + b = 0$, where the sample vectors (x_i, y_i), i=1,…, n, should satisfy

$$y_i[(\omega \cdot x) + b] \geq 1, \quad i = 1,\ldots, n. \tag{2.1}$$

The distance of point x to the hyperplane (ω, b) is $d(\omega, b; x) = \frac{|(\omega \cdot x^l) + b|}{\|\omega\|}$. The optimal hyperplane is given by maximizing the margin d, subject to equation (2.1).

The margin can be given by $\rho(\omega,b) = \frac{2}{\|\omega\|}$. Hence the hyperplane that optimally separates the data is the one that minimizes

$$\phi(\omega) = \frac{\|\omega\|}{2} \tag{2.2}$$

The Lagrange function of (2.2) under constraints (2.1) is,

$$\phi(\omega,b,a) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l}\alpha_i(y^i[(\omega,x^i)+b]-1), \tag{2.3}$$

The optimal classification function, if solved, is $f(x) = \text{sgn}\ (\omega^* \cdot x) + b^*$).

For nonlinear case, we map the original space into a high dimension space by a nonlinear mapping, in which an optimal hyperplane can be sought. The inner product function enables the classification in the new space, however, the computation complexity will not increase. Thus the corresponding program is,

$$Q(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)^* \tag{2.4}$$

The corresponding separating function is,

$$f(\boldsymbol{x}) = \text{sgn}(\sum_{i=1}^{n}\alpha_i^* y_i K(\boldsymbol{x}_i,\boldsymbol{x}) + b^*) \tag{2.5}$$

This is the so- called SVM.

SVM provides a method to solve the possible dimension disaster in the algorithm: when constructing a discriminant function, SVM does not obtain solution in the feature space after mapping the original sample space into a high dimension space by nonlinear mapping. Instead, it compares the sample vectors (for example, computes sample vectors' inner product or some kinds of distance) in the input space, then it performs nonlinear mapping after the comparison [9]. Function *K* is called the kernel function of dot product. In [10], it is defined as a distance between sample vectors. The method above can assure all training samples are accurately classified. That is, on condition that the empirical risk is zero, SVM can get the best generalization ability by maximizing the margin. However, different kinds of dot product in SVM can accomplish such works as Polynomial approaching, Bayesian classifier, and Radial Basic Function. How to define the dot product is critical to the classification result.

With these advantages of SVM, we attempt to apply SVM to the credit risk management because the feature of credit database can be potentially attacked by SVM.

2.2 Applications in Credit Assessment

In the last few years, there have been substantial developments in different aspects of support vector machine. These aspects include theoretical understanding, algorithmic strategies for implementation and real-life applications. SVM has yielded high performance on a wide range of problems including pattern recognition [10, 11, 12], function approach [13], data mining [14] and nonlinear system control [15] etc. These application domains typically have involved high-dimensional input space, and the

performance is also related to the fact that SVM's learning ability can be independent of the dimensionality of the feature space. The SVM approach has been applied in several financial applications recently, mainly in the area of time series prediction and classification [16, 17]. A recent study closely related to our work investigated the use of the SVM approach to predict credit rating analysis with a market comparative study. They reported that SVM achieved accuracy comparable to that of back-propagation neural networks [18]. In this study, we are interested in evaluating the performance of the SVM approach in credit assessment in comparison with that of the method currently being used by a major Chinese commercial bank.

The standard SVM formulation solves only the binary classification problem. While in credit risk assessment, it is not enough to classify the evaluation object into two classes. Hsu and Lin's recent paper [19] compared several methods for multi-class SVM and concluded that 'one-against-one' and DAG are more suitable for practical uses. This result offers a good solution to method selection in credit risk assessment.

In practical application, we use a sample consists of n variables to assess the personal credit risk, for each variable value of training sample, we want to find an interpretive vector $\mathbf{x}_i \in R^d$ and a symbol y_i that describes which class the indicator value belongs to. Then we predict new sample classification after training.

3 Experiment Results and Analysis

3.1 Data Sets

We have used real life credit card data to conduct the experiment. We selected one thousand sample data from a Chinese commercial bank, of which two classes have been defined: good and bad credit. We separate the credit applicants into two classes: good and bad. There are 245 bad records and 755 good records in the total sample. 14 variables are selected for personal credit assessment. The variables are listed in table 1.

3.2 Experiment Results and Analysis

We use Matlab6.1 and Osusvms3.00 toolbox developed by Junshui Ma, Yi Zhao and Stanley Ahalt [20] to conduct the computation. We divide the total sample into two parts: one is used for training and the other for test. Table 2 lists the predicting accuracy results of different number of training sample.

The result shows that the mean predicting accuracy is all above 70%. If we use less than 1/3 of the total sample as training sample and the rest as test sample, the results will be better, where the predicting accuracy will be above 80%. The preliminary result indicates that the application of SVM can improve the classification accuracy. However, we observe that the result of the predicting accuracy on training sample is better than that of test sample, which shows the predicting ability is not very good and needs further study.

Table 1. The assessment variables

Index	Indicators
1	Year of birth
2	Number of children
3	Number of other dependents
4	Is there a home phone
5	Spouse's income
6	Applicant's employment status
7	Applicant's income
8	Residential status
9	Value of Home
10	Mortgage balance outstanding
11	Outgoings on mortgage or rent
12	Outgoings on Loans
13	Outgoings on Hire Purchase
14	Outgoings on credit cards

Table 2. Predicting accuracy

The ratio of training set to total sample	Predicting accuracy of training sample	Predicting accuracy of test sample	Mean Predicting accuracy
10%	0.9098	0.7153	0.8125
20%	0.9430	0.7102	0.8266
25%	0.9085	0.7095	0.8090
33.33%	0.9093	0.7062	0.8077
50%	0.8873	0.6900	0.7886
66.67%	0.8873	0.6455	0.7664
75%	0.8715	0.6352	0.7533
90%	0.8648	0.6098	0.7373

4 A Comparison Result

In order to further verify the practical effect, we conducted a comparative study. We compare the SVM method with the basic grade criterion (mainly used for the grant of credit card to applicants) which is presently used by the Chinese bank for personal credit bound. The indicators used in that method is almost the same as those in SVM. We use that criterion to get a grade for each sample, by which to classify the sample into two classes: good and bad. An applicant would get credit if his/her grade is no

less than the threshold value determined by the bank (in this criterion the lowest grade is 110). The results are listed in table 3:

Table 3. The predicting accuracy of the present method used by a Chinese bank

Total number	False classification number	Correct classification number	Accuracy
1000	449	551	55.1%

The predicting accuracy of current method is 55.1%. Compared with the result outlined in table 2, this accuracy is much lower than that of SVM. The predicting accuracy of SVM can exceed current method result by 50%. This can illustrate that the assessment classification result got by using SVM has obvious superior to the current method of the bank. At the same time, the accuracy of 55.1% shows that the current method of the bank has big problem in credit card risk management and needs to be improved urgently. Further discussion about this situation can be referred to [21].

5 Conclusion and Future Research

This paper has applied the SVM approach to credit assessment and reported a comparison study with the current credit assessing method of a major Chinese commercial bank. The preliminary experiment results show that the SVM method turns out to be an effective classification tool for credit assessment. The comparison analysis indicates the SVM method is better than the current method of the Chinese commercial bank, which the predicting accuracy can increase 50%.

Our future directions of the research would focus on how to improve the predicting accuracy especially in the testing sample and compare the SVM method with other well-known methods, such as the back-propagation neural networks and decision tree. Inspired by the preliminary results, we believe that deeper data processing and more suitable kernel function selection will contribute to increase the predicting accuracy. Extending the two-class classification to multi-class classification is also our future research work.

References

1. Eisenbeis R A., Pitfalls in the application discriminant analysis in business and economics. Journal of Finance, (1977) 32: 875-900.
2. Tam K Y, Kiang M. Managerial applications of neural networks: the case of bank failure predictions. Management Sciences, (1992) 38(1): 926- 947.
3. Frydm an H, Altman E I, Kao Duen Li. Introducing recursive partitioning for financial classification: the case of financial distress. Journal of Finance, (1985) 40 (1): 269-291.
4. Chunfeng Wang, Research on small sample data credit risk assessment. Journal of Management Science in China, (2001)4(1): 28-32 (in Chinese).

5. V.Vapnik. Nature of Statistical Learning Theory. New York, Springer-Verlag,. (1995)
6. M.A Hearst, S.T. Dumais, E.Osman, J. Platt, B.Scholkopf. Support Vector Machines, IEEE Intelligent System, (1998)13(4):18-28
7. N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge Univ. Press, Cambridge, NewYork (2000).
8. K.-R. Mu¨ller, S. Mika, etc. An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks(2001) 12 (2) , 181– 201.
9. Corinna Cortes, V.Vapnik. Support-Vector Network. Machine Learning, (1995) 20.273-297
10. J.C.Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Bell Laboratories, Lucent Technologies(1997).
11. Roobacert D. Hulle M M Van. View-based 3d Object Recognition with Support Vector Machines: an Application to 3d Object Recognition with Cluttered Background. In Proc. SVM Workshop at IJCAI'99, Stockholm, Sweden(1999)
12. Scholkopf B, et al. Face Pose Discrinination Using Support Vector Machines, in: Proceedings of CVPR 2000, Hilton Head Island, (2000)430-437.
13. Sola A J, Scholkopf B. A Tutorial on Support Vector Regression [J], NeuorCOLT TR NC-TR-98-030, Royal Holloway College, University of London, UK. (1998)
14. Bradley P. Mathematical Programming Approaches to Machine Learning and Data Mining. Ph.D thesis. University of Wisconsin, Computer Science Department, Madison, WI, USA, TR-98-11 (1998).
15. Suykens J A K, et al. Optimal Control by Least Squares Support Vector Machines. Neural Networks, (2001)14(1): 23-25
16. F.E.H. Tay, L.J. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing, (2002)48: 847– 861.
17. T. Van Gestel, J.A.K. Suykens, etc, Financial time series prediction using least squares support vector machines within the evidence framework, IEEE Transactions on Neural Networks, (2001)12 (4):809– 821.
18. Zan Huanga, Hsinchun Chen, etc. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems (In press).
19. C.W.Hsu, C.J.Lin. A Comparison of Methods for Multi-class Support Vector Machines, Technical Report, National Taiwan University, Taiwan (2001).
20. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
21. Jianping Li, Jingli Liu, etc. An improved credit scoring method for Chinese commercial bank. Working paper (2004).

Measuring Scorecard Performance

Zheng Yang, Yue Wang, Yu Bai , and Xin Zhang

Colledge of Economics, Sichuan University
Chengdu, Sichuan, 610064, China
Yangzheng9@126.com

Abstract. In this paper, we look at ways to measure the classification performance of a scoring system and the overall characteristics of a scorecard. We stick to the idea that we will measure the scoring system by how well it classifies, which are still problems in measuring its performance. This is because there are different ways to define the misclassification rate mainly due to the sample that we use to check this rate. If we test how good the system is on the sample of customers we used to build the system, the results will be better than that we did the test on another sample. This idea is illustrated in this paper. Two measures, Mahalanobis distance and KS score, are used in the paper.

Keywords: credit-scoring systems, measuring scorecard, classification, holdout, mahalanobis distance, KS score.

1 Introduction

Having built a credit or behavioral scorecard, the obvious question is, ”How good is it ?” this begs the question of what we mean by good. The obvious answer is in distinguishing the good from the bad because we want to treat these groups in different ways in credit-scoring systems---for example, accepting the former for credit and rejecting the latter. Behavioral scoring systems are used in a more subtle way, but even if we stick to the idea that we will measure the scoring system by how well it classifies, there are still problems in measuring its performance. This is because there are different ways to define the misclassification rate, mainly due to the sample that we use to check this rate. If we test how good the system is on the sample of customers we used to build the system, the results will be must better than if we did the test on another sample. This must follow because built into the classification system are some of the nuances of that data set that do not appear in other data sets. Thus section 2 looks at how to test the classification rate using a sample, called the holdout sample, separate from the one used to build the scoring system. This is a very common thing to do in the credit-scoring industry because of the availability of very large samples of past customers, but it is wasteful of data in that one does not use all the information available to help build the best scoring system. There are times, however, when the amount of data is limited, for example, when one is building a system for a completely new group of customers or products. In the case, one can test

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 900–906, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the validity of the system using methods that build and test on essentially the same data set but without causing the misclassification errors to be optimistically biased. There are a number of standard ways to describe how different two populations are in their characteristics. These can be used in the case of scorecard-based discrimination methods to see how different the scores are for the two groups of good and bad. These measure how well the score separates the two groups, and in section 3, we look at two such measures of separation-the Mahalanobis distance and Kolmogorov-Smirnov statistics. These give a measure of what is the best separation of the groups that the scorecard can make.

2 Error Rates Using Holdout Samples and 2 × 2 Tables

We defined the decision making in a credit-scoring system as follows: X=($X_1,X_2,...,Xp$) are the application variables, and each applicant gives a set of answers **x** to these variables and thereafter is found to be either good(G) or bad(B). Assuming that the application characteristics are continuous (a similar analysis works for the discrete case), let f(x) be the distribution of application characteristics, $f(G/\mathrm{x})$ be the probability of being a good if the application characteristics were **x**, and $f(B/\mathrm{x})=1-f(G/\mathrm{x})$ be the probability of being a bad if those are the characteristics.

The optimal or Bayes error rate is the minimum possible error rate if one knew these distributions completely over the whole population of possible applicants.

$$e(Opt) = \int \min\left\{f(B|x), f(G|x)\right\} f(x)dx \tag{1}$$

Any given credit-scoring system built on a sample S of n consumers estimates the functions $f(G/\mathrm{x})$and $f(B/\mathrm{x})$and in light of this defines two regions of answers A_G and A_B ,in which the applicants are classified as good or bad. The actual or true error for such a system is then defined as

$$e_S(Actual) = \int_{A_G} f(B|x)f(x)dx + \int_{A_B} f(G|x)f(x)dx \tag{2}$$

This is the error that the classifier built on the sample S would incur when applied to an infinite test set. The difference occurs because one has used only a finite sample S to build the estimator. One is usually interesting in estimating e_S (Actual), but if one had to decide what system to use before seeing the data and adapting the system to the data, then one would take the expected error rate e_n(Expected),which is the expectation of e_S(Actual) over all samples of size n.

The difficulty in calculating e_S(Actual)is that one does not have an infinite test set on which to calculate it and so one has to use a sample S^* to estimate it on. Hence one calculates

$$e_S(S^*) = \int_{A_G} f(B|x)fs^*(x)dx + \int_{A_B} f(G|x)fs^*(x)dx \tag{3}$$

where fs^*(x) is the distribution of characteristics **x** in the sample S^*.

The obvious thing is to check the error on the sample on which the classifier was built and so calculate $e_S(S)$. Not surprisingly, this underestimates the error considerably, so $e_S(S) < e_S$(Actual). This is because the classifier has incorporated in it

all the quirks of the sample S even if these are not representative of the rest of the population. Hence it is far better to test the data on a completely independent sample S^*. This is called the holdout sample, and it is case that the expected value of $e_S(S^*)$ over all samples excluding S is e_S(Actual). This means that this procedure gives an unbiased of the actual error.

In credit scoring, the cost of the two errors that make up the error rate are very different. Classifying a good as a bad means a loss of profit L, while classifying a bad as a good means an expected default of D, which is often considerably higher than L. Thus instead of error rates one might look at expected loss rates, where the optimal expected loss l(Opt) satisfies

$$l(Opt) = \int \min \{Df(B|x), Lf(G|x)\} f(x)\,dx \tag{4}$$

In analogy with (2) and (3), the actual or true loss rate for a classifier based on a sample S is

$$l_S(Actual) = \int_{A_G} Df(B|x) f(x)dx \int_{A_B} Lf(G|x) f(x)dx \tag{5}$$

While the estimated rate using a test sample S^* is

$$l_S(S^*) = \int_{A_G} Df(B|x) f_{S^*}(x)dx \int_{A_B} Lf(G|x) f_{S^*}(x)dx \tag{6}$$

Bayes's theorem will confirm that the expression for the actual loss rate in (5).

So how does one calculate $e_S(S^*)$ and $ls(S^*)$, having built a classifier on a sample S and having a completely independent holdout sample S^* available? What we do is to compare the actual class G or B of each customer in the sample S^* with the class that the scorecard predicts. The results are presented in a 2×2 table called the confusion matrix (see Table 1) which gives the numbers in each group.

For example, we might have the confusion matrix given in Table 2. In this sample, n, n_G, and n_B (1000,750,250) are fixed as they describe the sample chosen. Thus really only two of the four entries g_G, b_G, g_B, and b_B are independent. The actual error rate is calculated as $(b_G+g_B)/n$=250/1000=0.25. If the losses are L=100, D=500, then their actual loss per customer is $(Lb_G+Dg_B)/n$=(100*150)+(500*100)/1000=65. One can use confusion matrices to compare systems or even to try to decide on the best cutoff score when a scorecard had been developed. In the latter case, changing the scorecard in the example in Table 2 may lead to the confusion matrix in Table 3.

Table 1. General confusion matrix

		True Class		
		G	B	
Predicted Class	G	g_G	g_B	g
	B	b_G	B_B	b
		n_G	n_B	n

Table 2. Example of a confusion matrix

		True Class		
		G	*B*	
Predicted Class	*G*	600	100	700
	B	150	150	300
		750	250	1000

Table 3. Confusion matrix with a different cutoff

		True Class		
		G	*B*	
Predicted Class	*G*	670	130	800
	B	80	120	200
		750	250	1000

The actual error rate is (130+80)/1000=0.21, which suggests this is a better cutoff. However, the actual expected loss rate is (100*80)+(500*130)/1000=73, which is higher than the expected loss with the other cutoff. So perhaps the other system was superior. This difference between the two ways of trying to decide which system to go for is very common. Sometimes in credit scoring, one will find that the error rate is minimized by classifying everyone as good and so accepting them all. It is a brave---and foolish---credit analyst who suggests this and thus makes himself redundant! One approach that is useful in comparing the difference between the way two credit-scoring systems perform on a holdout sample is to look at the swap sets. This is the group of people in the holdout sample, who are classified differently by the two systems. The swap sets for Table 2 and 3 might look like Table 4.

Table 4. Example of a swap set analysis

	True	
	G	*B*
Predicted *G* by Table 2, *B* by Table 3	50	10
Predicted *B* by Table 2, *G* by Table 3	120	40

This table cannot be calculated from the two confusion matrices, although from these we can see that 100 more in the sample were predicted *B* by Table 2 and *G* by Table 3 than were predicted *G* by Table 2 and *B* by Table 3. There could be in this case a number of people who move a different way to the overall trend. The fact that (50+10+120+40)/1000=0.22 of the population change between the two scorecard suggests that the scorecards are quite different. If one looks only at the swap sets caused by changing the cutoff, then obviously no customers will have their own classifications changed against the movement. Hence one of the rows in the swap sets will be 0.

3 Separation Measures: Mahalanobis Distance and Kolmogorov-Smirnov Statistics

A number of measures used throughout statistics describe how far apart the characteristics of two populations are. If one has a scoring system that gives a score to each member of the population, then one can use these measures to describe how different the scores of the good and the bad are. Thus these approaches can be used only for credit-scoring systems that actually give a score, like the regression approaches or linear programming [7,8,9]. They cannot be used for the credit-scoring systems that group, like classification trees, or where a score is not explicit, like neural networks. Moreover, they describe the general properties of the scorecard and do not depend on which cutoff score is used. This is useful in that these measures give a feel for the robustness of the scorecard if the cutoff score is changed and may be useful in determining what the cutoff score should be. However, when it comes to it, people will want to know how well the scorecard will predict, and to know that one needs to have chosen a specific cutoff score so that one can estimate the error rates and confusion matrices of section 2. For accurate estimates, we should calculate these measures on a holdout sample, which is independent of the development sample on which the scorecard was built. However, often speed and the fact that in many statistical packages it is much easier to calculate the measures on the development sample than on the holdout sample mean that one calculates them first on the development sample. Since they are only indicators of the relative effectiveness of different scorecards, the assumption is made that if one is much better than the other on the development sample, it will remain so on the holdout sample.

The first measure, the Mahalanobis distance, appeared earlier in the discussion on the Fisher approach to using linear regression as a classification method. There we found the linear combination Y of the application variables so that M, the difference between the sample mean of Y for the good and the sample mean Y of the bad, divides by the standard deviation in Y for each group was as large as possible. This M is the Mahalanobis distance and is a measure of by how much the scores of the two groups of the good and the bad differ. Formally, if $n_G(s)$ and $n_B(s)$ are the numbers of good and bad with score s in a sample of n, where there are n_G good and n_B bad, the $p_G(s) = n_G(s)/n_G$ ($p_B(s) = n_B(s)/n_B$) are the probabilities of a good (and bad) having a score s. Then $m_G = \Sigma s p_G(s)$ and $m_B = \Sigma s p_B(s)$ are the mean scores of the good and bad. Let σ_G and σ_B be the standard deviation of the scores of the good and the bad, calculated as

$$\sigma_G^2 = \left(\sum_s s^2 pG(s) - m_G^2 \right)^{\frac{1}{2}} . \sigma_B^2 = \left(\sum_s s^2 pG(s) - m_B^2 \right)^{\frac{1}{2}} \quad (7)$$

Let σ be the pooled standard deviation of the good and the bad from their respective means: it is calculated as follow:

$$\sigma = \left(\frac{n_G \sigma_G^2 + n_B \sigma_G^2}{n} \right)^{\frac{1}{2}} \quad (8)$$

The Mahalanobis distance M is then the difference between the mean score of the two groups, suitably standardized

$$M = (m_G - m)/\sigma \tag{9}$$

This is indifferent to any linear scaling of the score, and as Figure 1 suggest, one would assume that if a scorecard has a large Mahalanobis distance, it will be a better classifier. In Figure 1, the dotted lines represent possible cutoff scores, and the errors in the figure on the left with the smaller M are much greater than those on the right.

The Mahalanobis distance measures how far apart the means of the good score and the bad score are. The Kolmogorov-Smirnov statistics measures how far apart the distribution functions of the scores of the good and the bad are. Formally, if $P_G(s) = \Sigma_{x \le s} p_G(x)$ and $P_B(s) = \Sigma_{x \le s} p_B(x)$ (or the sums replaced by integrals if the scores are continuous), then the Kolmogorov-Smirnov (KS) statistic is

$$KS = \max_s \left| P_G(s) - P_B(s) \right| \tag{10}$$

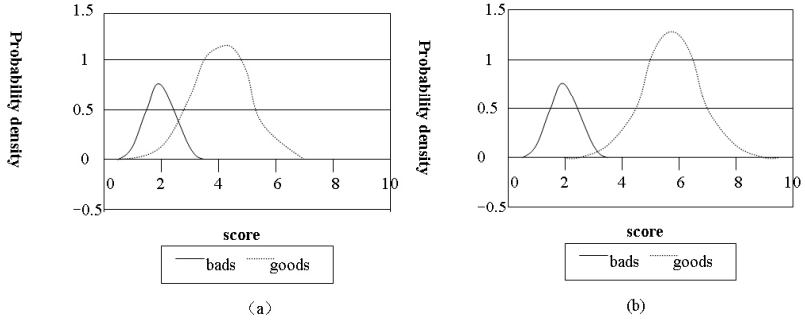


Fig. 1. a) Good and bad similar, b) good and bad different

In Figure 2, the Kolmogorov-Smirnov statistic is the length of the dotted line at the score that maximizes the separation in the distribution function. If the distribution functions are sufficiently regular, then the Kolmogorov-Smirnov distance occurs at the score where the good and bad histograms in Figure 1 cross. The Kolmogorov-Smirnov statistic for an attribute, rather than the score, is used to find the best splits in a classification tree.

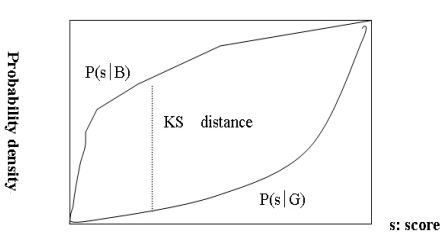


Fig. 2. Kolmogorov–Smirnov distance

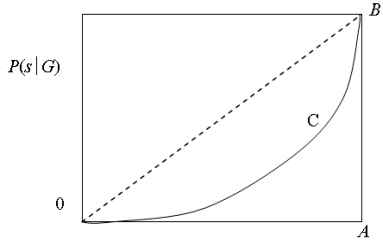


Fig. 3. ROC *curve*

4 Concluding Remarks

Kolmogorov-Smirnov statistics can be displayed, as in Figure 2, by plotting two curves, but it is possible to display the same information on one curve by plotting $P_G(s)$ against $P_B(s)$.The result is a curve as in Figure 3, where each point on the curve represents some score s, and its horizontal distance is $P_B(s)$ and its vertical value is $P_G(s)$. This is the ROC curve, sometimes called the Lorentz diagram. It describes the classification property of the scorecard as the cutoff score varies. The best possible scorecard would have an ROC curve that goes all the way along the horizontal axis before going up the vertical axis. Thus point A would correspond to a score s^*, where $P_B(s^*) =1$ and $P_G(s^*) = 0$; i.e., all of the bad have scores less than s^* and none of the good do. An ROC curve along the diagonal $O\ B$ would correspond to one where at every score $P_G(s) = P_B(s)$, so the ratio of good to bad is the same for all score ranges. This is no better than classifying randomly given that one knows the ratio of good to bad in the whole population. Thus the further from the diagonal the ROC curve is, the better the scorecard. If one scorecard has a ROC curve that is always further from the diagonal than the ROC curve of another scorecard, then the first scorecard dominates the second and is a better classifier at all cutoff score. More common is to find ROC curves that cross so one scorecard is a better classifier in one score region and the other is better in the other region.

References

1. Capon, N.: Credit scoring systems: A critical analysis, J .Marketing, 46 (1982) 82-91.
2. Churchill, G.A., Nevin, J.R., Watson, R.R.: The role of credit scoring in the loan decision, Credit World, March (1997) 6-10.
3. Crook, J.N.: Credit constraints and U.S. households, Appl. Financial Economy, 6 (1996) 477-485.
4. Crook, J.N.: Consumer credit and business cycles. In: D.J. Hand and S.D. Jack (eds.): Statistics in Finance. Arnold, London (1998).
5. Edelman, D.B.: Building a model to forecast arrears and provisions. In Proceedings of Credit Scoring and Credit Control VI. Credit Research Center, University of Edinburgh Edinburgh, Scotland (1999).
6. Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit. J. Roy. Statist. Soc. Ser. A, 160 (1997) 523-541.
7. Haykin, S.: Statistical Aspects of Credit Scoring, PH. D thesis, Open University, Milton Keynes, U. K. (1995).
8. Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M., Xu, W.: Multiple Criteria Linear Programming to Data Mining: Models, Algorithm Designs and Software Developments. Optimization Methods and Software, 18 (2003) 453-473.
9. Shi, Y., Wise, M., Luo, M. and Lin, Y.: Data mining in credit card portfolio management: a multiple criteria decision making approach, in M. Koksalan and S. Zionts, eds., Multiple Criteria Decision Making in the New Millennium, Springer, Berlin (2001) 427-436.
10. Shi, Y., Peng, Y., Xu, W., Tang, X.: Data Mining via Multiple Criteria Linear Programming: Applications in Credit Card Portfolio Management. International Journal of Information Technology and Decision Making. 1 (2002) 131-151.

Parallelism of Association Rules Mining and Its Application in Insurance Operations*

Jinlan Tian, Lin Zhu, Suqin Zhang, and Gang Huang

Department of Computer Science and Technology
Tsinghua University., Beijing, 100084, PR China
tjlan@mail.tsinghua.edu.cn

Abstract. Association rules mining is a basic method in data mining.This paper first introduces the basic concepts of association rules mining and Apriori algorithm. It also provides a parallel association rules model scheme for improving the mining efficiency when treating large numbers of data sets as well as the analyse of the scheme effect. In conclusion we discuss how to apply association rules mining to insurance data sets, find out the knowledge hidden behind the data sets, and provide powerful decision-making support for people.

Keywords: data mining; association rules mining; parallel algorithm; insurance

1 Introduction

Data Mining is an advanced process to find and extract confident, novel, effective and comprehensible patterns which are hidden behind a mass of data sets [1][2]. Along with the prevalence of database, many large-scale companies collect the data as many as billions of or ever more. The huge data sets should be well understood for providing information for company directors. The ability of dealing with large data sets is crucial for data mining tools.

Association rules mining is an important pattern of knowledge patterns of data mining. The concept of association rules mining,which is simple and practical, is provided by Agrawal£Imielinski and Swami[1]. The patterns of association rules mining belong to descriptive patterns, and the algorithms to discover rules belong to unsupervised learning. In this paper, we first introduce the definition and attributes of association rules mining, and Apriori, the algorithm which has close relationship to it. Then we present a design of parallel scheme of association rules mining. At last, we discuss how to utilize MineSet, a data mining tool of SGI, into insurance data sets to dig out association rules.

* This research is supported by a joint research grant from National Science Foundation of China (project No.60131160743) and Hong Kong Research Grant Council.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 907–914, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 The Definition and Attributes of Association Rules Mining

Just pay attention to some transactions that relate to many items: item X appears at transaction 1, item Y appears at transaction 2, and item X and Y appear at transaction 3 simultaneously. Therefore, are there any rules of item X and Y appear at transactions? In the area of data mining, association rules act as such knowledge patterns which describe the rules of items appearing at the same transaction simultaneously. More exactly, association rules describe how the appearance of item X affects the appearance of item Y through quantification. Some basic concepts of association rules are rendered below[4].

Let $R = \{I_1, I_2, \cdots, I_m\}$ be a set of items. Let W, the task-relevant data, be a set of database transaction where each transaction T is a set of items such that $T \subseteq R$. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I, B \subset I$, and $A \cap B = \emptyset$.Here are the four basic attributes of associate rules[3][8].

Confidence. The rule $A \Rightarrow B$ has confidence c in the transaction set W if c is the percentage of transactions in W containing A that also contain B. Confidence is the scale of veracity of association rules.

Support. The rule $A \Rightarrow B$ holds in the transaction set W with support s, where s is the percentage of transactions in W that contain $A \bigcup B$ (i.e., both A and B). Support is the scale of weightiness of association rules. The bigger the support, the more important the rule is.[5]

Expected confidence. The rule $A \Rightarrow B$ has expected confidence e in the transaction set W if e is the percentage of transactions in W that contain B.

Lift. Lift is the ratio of confidence and expected confidence. Lift describes how the appearance of A affects the appearance of B. Generally speaking, available association rules should have the lift that is bigger than 1.

3 Parallelism of Association Rules Mining Algorithm

3.1 Algorithm of Association Rules Mining

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called "strong"[6]. "How are association rules mined from large databases?" Association rule mining is a two-stop process:

- **Find all frequent itemsets:** By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.
- **Generate strong association rules from the frequent itemsets:** By definition, these rules must satisfy minimum support and minimum confidence.

Additional interestingness measures can be applied, if desired. The second step is the easier of the two. The overall performance of mining association rules is determined by the first step.

Aprior [7][8] is a basic algorithm for mining frequent itemsets for Boolean association rules. Aprior employs an iterative approach known as a *level-wise* search, where k-itemsets are used to explore $(k+1)$-itemsets. First, the set of frequent 1-itemsets is found. This set is denoted L_1. L_1 is used to find L_2,the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the database.The pseudocode for the Apriori algorithm is shown below.

```
procedure AprioriAlg()
begin
   L1:= { frequent 1-itemsets};    // find the frequent 1-itemsets
   for ( k:=2;Lk ≠ ∅ ; k++)  do  {
      Ck:=apriori-gen(Lk−1; // generate candidates Ckbased on
                        // Lk−1  in order to find Lk
      For all transactions t in the dataset do  {
                       //scan the dataset for counts
         for all candidates c ∈ Ck contained in t do
            c.count ++;
      }
      Lk:={c ∈ Ck|c.count ≥ minsupport}
   }
   Answer:=⋃k L[K];
end;
```

The database is scanned repetitious times in Apriori algorithm. The first scan of Apriori finds the frequent 1-itemsets. The every following scan, for example, the kth scan, firstly use L_{k-1} to generate candidates C_k,and then scan the whole database to find grequent itemsets in C_k and put them into L_k.

3.2 Parallel Scheme Design

Data mining always treats very large members of data, often millions of records in some data sets. It will cost insufferably long time for data mining tools to deal with these records. So the ability of dealing with huge numbers of data sets is crucial to data mining tools. They need parallel technologies.

Most job of the Apriori algorithm is about the scanning and statistics of data. However, the statistics is independent transversely (among association rules) and longitudinally (among records). Therefore, the data should be distributed to multi-processors. Two schemes here.

1) Longitudinal partition

Every processor runs an integrated algorithm under such partition. But each processor treats with different columns of data records. For more detail, count combinative modes among columns of records, and then every processor treats with one kind of combinative mode to build association rules of the data treated with. For example, there are three columns (F_1, F_2, F_3) in records. There are three kinds of combinative modes according to every two combination. Assign

them into three separate processors, and then we will get the related association rules.

However, the efficiency of this partition method is far from satisfying. Because:

- There are many repetitious statistics because some different records involve the same columns.
- The bottleneck of the algorithm we discussed is the scan of database. But the longitudinal partition does not solve this problem effectively.

2) Transverse partition

Divide the data into n parts in such partition methods, and assign the n parts into n processors. Then collect the statistical results at last[9].

It will spend most of time in Apriori algorithm to scan databases and count the frequency that the items appear. So it is the key issue to make this module into parallel.

There are two characters in this kind of parallel:

- Regard the record as unit, so the data are independent of each other. Different records will not affect each other.
- Different records share the same handle process.

Therefore we could use SIMD model[10] to construct out parallel algorithm. First, divide the data into n parts, assign them into n processor to parallel, and synthesize the results at last. The processors are not required to be synchronous.

Figure 1 shows the flow chart of transverse partition method:

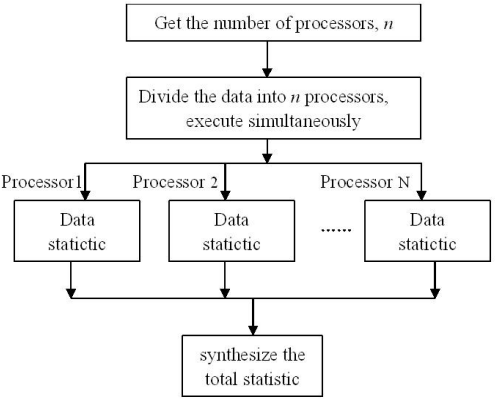


Fig. 1. Flow chart of transverse partition method

In conclusion, we do not need to consider the data-correlation problem and the synchronization of processors. This scheme is so effective that it can reduce the execution time to the ratio of almost $1/n$ towards the primary algorithm.

4 Application in Insurance Operations

Association rules can be used in many fields, such as merchandise retail, finance, insurance and so on. There are many data mining tools implementing the association rules mining method at present. Now we will discuss how to utilize MineSet[11], a data mining tool provided by SGI Company, to operate insurance data with association rules.

1) Data Preparation

Here are the records in the data data resource.

Individual Insurance ID	Age	Total Salary per Year	Company ID	Company Name	Company Type	Company Area Code	Compensation Times
3502043808264031	60	7,051	0000000664	Oil company of city A	3	3	0
3502042104054011	77	7,382	0000000664	Oil company of city A	3	3	1
1324274512206811	53	17,617	0000000663	The third Middle school of city A	2	1	0
3101107203033631	26	15,485	0000000662	Government of city A	2	1	0
.....							

Each record includes the basic information of a policy holder and the times he has claimed for compensation. We want to find out the characters of people who have claimed form compensation and those whose have not. The data should be prepared before mining. As we just concern the *Compensation Times* and its related information, the redundant information in the dataset, such as *individual insurance ID, company ID and company name*, should be removed.

Take notice of age and total salary per year, they are continuous. But when using association rules, the association rules generated algorithms cannot deal with continuous data and should be discretized. Divide the attribute age into five groups (...40], (40...50], (50...60], (60...70] and (70...]; *total salary per year* into three groups £...6,000], (6,000...10,000] and (10,000...].

As for the attribute of compensation times, we only concern whether a policy holder claimed or not, and ignore the concrete number of *compensation times*. Thus attribute compensation times are converted to *If Compensating*, 1 represents compensated and 0 represents not. Here are the data after convertion.

Age	Total Salary per Year	Type of Company	Company Area Code	If Compensating
(50...60)	(6,000...10,000]	3	3	0
(70...)	(6,000...10,000]	3	3	1
(50...60)	(10,000...]	2	1	0
(...40)	(10,000...]	2	1	0
.....				

2) Mine Association Rules

We need to specify the minimum support and minimum confidence. Take the minimum support as 1%. As minimum support rises, the rules we can find reduce, so the time used in mining will reduce. But if minimum support is too high, we may miss some rules which should be found. We take the minimum credibility as 50%, before using MineSet.

3) Visualized Display and Comprehend the Rules

When MineSet find out Association rules, it will use Scatter Visualize [11] to display the mining result. Figure 2 illustrates the result.

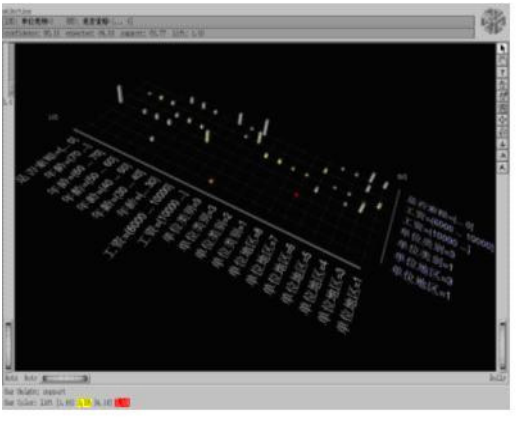


Fig. 2. Association rules mining result

It is a three-dimensional picture. The first is LHS (left-hand side) to represents the items set A. The second is RHS (right-hand side) to represent the items set B. Every oblong in the figure represents an association rule $A \Rightarrow B$,as A is the value projected on LHS, B is the value projected on RHS.

The third dimension is the height of an oblong which represents the confidence of the rule. The color of every oblong shows the lift of the rule. (each color stands for an area of lift, you can find the corresponding relations at the bottom left corner of the figure). When the mouse is on an oblong, the information of corresponding association rule will be displayed at top left of the screen. For example, the information included in the oblong in the figure above is:

LHS Company Type=3 RHS If Compensating=0

confidence: 85.18 expected: 84.00 support: 60.77 lift: 1.03

So the association rule which the oblong in the picture stands for is: Company Type=3 ⇒ If Compensating=0. The value of the four parameters of this rule is: Support= 60.77%, Confidence= 85.18%, Expected Confidence= 84.00%, Lift= 1.03.

We can see, 84.00% of all policy holders have not claimed for compensation, 60.77% have done and their company type is 3. Among the policy holders whose company type is 3, 85.13% have not claimed for compensation. “Lift = 1.03”

tells us that "Company Type= 3" does not have a considerable affection on whether the policy holders has claimed or not. It is because that without this term the compensation rate of policy holders does not has an obvious change.

Some of the association rules gained are not useful to insurance business, such as the association rules between *total salary per year and company type*, and ignored.

In the association rules we have just mentioned, LHS and RHS include only one item. They are one-to-one single rules. If more items for LHS and RHS are permitted, multi-corresponding association rules[12] can be obtained. They are shown with Record Visualize:

Support	Confidence	Expected	Lift	LHS	RHS
1.9997	84.77	84	1.0091	Company Type=2 and Age=(40 ... 50]	If Compensating =0
3.4213	92.8	84	1.1047	Company Area Code =5	If Compensating =0
1.6091	86.55	84	1.0304	Company Area Code=6 and Age =(30 ... 40]	If Compensating =0
15.1695	93.1	84	1.1083	Company Type=3 and Age =(40 ... 50]	If Compensating =0
2.2653	87.88	84	1.0461	Company area Code =4 and Total Salary per Year=(6000 ... 10000]	If Compensating =0
.....					

A row represents an association rule. The columns show LHS, RHS and other four parameters. We can see that among the people whose *company type* is 3 and *age* is between 40 and 50 years old, 93.1% have not claimed for compensation, much higher than expected confidence. However, among the people whose *company type* is 2 and *age* is between 40 and 50 years old, 84.77% have not claimed for compensation. If these results are based on some external reason, (For example, the companies with the *company type* 3 do not bring too much pressure to their employees, so the employees are not so tired to catch ill easily.) insurance company then may pay more attention on those customers who meet such conditions above for reducing the investment venture and increase operation earnings. When those rules are applied in practice, they should be adjusted with the time goes by. Then we can gain the most appropriate and best results that they are the closest to the original requirements.

5 Conclusion

In this paper we present the parallel scheme design of association rules mining and its application on insurance data. Parallel data mining algorithms are effec-

tive methods to solve the performance problem when treating large numbers of data sets. Association rules mining can be utilized into many areas other than insurance, such as credit card company, stock exchange and bank. Markets could use association rules to decide what to buy and how to put goods. In addition, association rules mining also has great affects in the application on communication industry, it can be used to analyze the factors of customer loss, adjust preferential action in time and reduce the loss of important customers. In conclusion, association rules mining can be applied at various areas widely, discover the knowledge hidden behind the data, and provide powerful decision-making support for people.

References

1. Peatetsky-Shapiro, G., Fayyad, U., and Smyth, P., From Data Mining to Knowledge Discovery: An Overview, Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., (eds), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, pp.1-35,1996
2. Introduction to data mining and knowledge discovery, Third Edition, by Two Crows Corporation
3. K.Decker and S.Focardi, Technology Overview: A Report on Data Mining, Technical Report CSCR TR-95-02, Swiss Scientific Computing Center, 1995
4. Heikki Mannnila, Hannu Toivonen and A. Inkeri. Verkamo, Efficient algorithms for discovering association rules, AAAI Workshop on Knowledge Discovery in Databases, pages 181-192, July 1994
5. Tony Xiaohua Hu, Knowledge Discovery in Databases: An Attribute-Oriented Rough Set Approach, Ph.D. thesis, Regina university, 1995
6. Gao Wen, KDD Knowledge Discovery in Databases, Computer World, vol. 37, 1998
7. Rakesh Agrawal , Manish Mehta , John Shafer and Ramakrishnan Srikant, The Quest Data Mining System, AAAI/MIT Press, pp.244-249,1996
8. R.Feldman, A.Amir, Y.Auman, A.Zilberstien and H.Hirsh, Incremental Algorithms for Association Generation, AAAI/MIT Press,pp.227-241,1995
9. Z.Zhang Y.Lu and B.Zhang, An Effective Partitioning-Combining Algorithm for Mining Quantitative Association Rules, AAAI/MIT Press,pp.241-252,1995
10. Chen Guoliang, Design and Analysis of parallel algorithms, Higher Education publishing company, China, 2002
11. SGI Company, MineSet2.0 Tutorial
12. David W.Cheung, Vincent T.Ng and Benjamin W.Tam, Maintenance of Discovered Knowledge: A Case in Multi-level Association Rules, AAAI/MIT Press, pp.307-309,1996

No Speculation under Expectations in Awareness

Ken Horie[1] and Takashi Matsuhisa[2⋆]

[1] Advanced Course of Electronic and Computer Engineering, Ibaraki National College of Technology
[2] Department of Liberal Arts and Sciences, Ibaraki National College of Technology, Nakane 866, Hitachinaka-shi, Ibaraki 312-8508, Japan.
mathisa@ge.ibaraki-ct.ac.jp

Abstract. We present the extended notion of pure exchange economy under uncertainty, called an economy with awareness structure, where each trader makes decision under his/her awareness and belief and has strictly monotone preferences. We show the no speculation theorem: If the initial endowment is ex-ante Pareto optimal then there exists no other expectations equilibrium in awareness for any price.

Keywords: Pure exchange economy, Awareness, Belief, No speculation, Expectations equilibrium, Ex-ante Pareto optimum.

1 Introduction

This article relates economies and distributed knowledge. The purposes are the two points: First we present an extended notion of economy under uncertainty, called an *economy with awareness structure*, where each trader makes decision in his/her awareness and belief under incomplete information. Secondly we show the following 'no trade theorem' under generalized expectations equilibrium in the extended economy:

No speculation theorem. *In a pure exchange economy under uncertainty, the traders are assumed to have an awareness structure and to have strictly monotone preferences. If the initial endowment is ex-ante Pareto optimal then there exists no other expectations equilibrium in awareness.*

Recently researchers in such fields as economics, AI, and computer science have become interested in reasoning of belief and knowledge. There are pragmatic concerns about the relationship between knowledge (belief) and actions. Of most interest to us is the emphasis on situations involving the distributed knowledge (belief) of a group of agents rather than that of a single agent. At the heart of any analysis of such situations as a conversation, a bargaining session or a

⋆ Corresponding author. Partially supported by the Grant-in-Aid for Scientific Research(C)(2)(No.14540145) in the Japan Society for the Promotion of Sciences. *Dedicated to Professor Shoji Koizumi on the occasion of his 80th birthday*

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 915–922, 2004.
© Springer-Verlag Berlin Heidelberg 2004

protocol run by processes is the interaction between agents. A agent in a group must take into account not only events that have occurred in the world but also the knowledge of the other agents in the group.

As for as concerned in economic situations, many authors have shown that there can be no speculation in an economy under uncertainty (e.g Kreps [7]; Milgrom and Stokey [12]; Morris [13] and others[1]). The serious limitations of their analysis are to assume 'partition' structure as information the traders receive, and to assume common-knowledge on traders' willingness to trade. From the epistemic point of view, the partition structure represents the traders' knowledge satisfying the postulates: 'Truth' **T** (what is known is true), the 'positive introspection' **4** (that we know what we do) and the 'negative introspection' **5** (that we know what we do not know). The postulate **5** is indeed so strong that describes the hyper-rationality of traders, and thus it is particularly objectionable. Also is the common knowledge assumption because the notion of common knowledge is defined by an infinite recursion of all of individual knowledge.

This raises the question to what extent results as the no speculation theorem depend on both common knowledge and the information partition structure (or the equivalent postulates of knowledge). The answer is the results which strengthen the Milgrom and Stokey's theorem. In this article we weaken the conditions: symmetry, reflexivity, transitivity and common-knowledge. As has already been pointed out in Geanakoplos (1989), this relaxation can potentially yield important results in a world with imperfectly Bayesian agents.

This article is organized as follows: In Section 2 we present our formal model: First we recall belief and awareness together with the associated information structure. Secondly we give economies with awareness structure, and we extend the notion of rational expectations equilibrium for economies with incomplete information to the economies with awareness structure. In Section 3 we explicitly state the no speculation theorem as above and prove it. Finally we discuss related results and conclude by giving some remarks on the assumptions of our results.

2 The Model

Let Ω be a no-empty *finite* set called a *state space*, T a set of finitely many at least two *traders* $\{1, 2, \ldots, t, \ldots n\}$ $(n \geq 2)$, and let 2^Ω denote the field 2^Ω that is the family of all subsets of Ω. Each member of 2^Ω is called an *event* and each element of Ω called a *state*.

2.1 Awareness and Belief

We present a model of awareness according to Matsuhisa and Usami [11][2]. *A belief structure* is a tuple $\langle \Omega, (B_t)_{t \in N} \rangle$ in which $B_t : 2^\Omega \to 2^\Omega$ is trader t's

[1] The references cited in Fudenberg and Tirole [5], footnote 3, p.543

[2] This model follows from E. Dekel, B. L. Lipman and A. Rustichini [3]. A different approach of awareness models is discussed in R. Fagin, J.Y. Halpern, Y. Moses and M.Y. Vardi [4].

belief operator. The interpretation of the event B_tE is that 't believes E.' An *awareness structure* is a tuple $\langle \Omega, (A_t)_{t\in T}, (B_t)_{t\in T}\rangle$ in which $\langle \Omega, (B_t)_{t\in T}\rangle$ is a belief structure and A_t is t's *awareness* operator on 2^Ω such that Axiom **PL** (axiom of plausibility) is valid:

PL $\quad B_tE \cup B_t(\,\Omega \setminus B_tE\,) \subseteq A_tE \quad$ for every E in 2^Ω.

The interpretation of A_tE is that 't is aware of E.' The axiom **PL** says that t is aware of E if he believes it or if he believes that he does not believe it. An event E is said to be *t's evident belief* if $T \subseteq B_tT$. We can think of it as embodying the essence of what is involved in t making his/her direct observations.

2.2 Associated Information Structure

M. Bacharach [1] introduces the *strong epistemic model* that coincides with the Kripke semantics corresponding to the modal logic **S5**.[3] Further he defines the information partition induced from the knowledge operator of the epistemic model.[4] Following his line we generalize the notion of information partition as follows.

Definition 1. The *associated information structure* $(P_t)_{t\in T}$ with an awareness structure $\langle \Omega, (A_t), (B_t)\rangle$ is the class of t's *associated information function* P_t of Ω into 2^Ω defined by $P_t(\omega) = \bigcap_{E\in 2^\Omega}\{E \,|\, \omega \in E \subseteq B_tE\}$. (If there is no event E for which $\omega \in E \subseteq B_tE$ then we take $P_t(\omega)$ to be no-defined.) The *domain of* P_t denoted by $\mathrm{Dom}(P_t)$ is the set of all the states at which P_t is defined.

We call $P_t(\omega)$ t's *evidence set* at ω, which is interpreted as the basis for all t's evident beliefs. This is because each t's evident belief E is decomposed into a union of all evidence sets contained in E.

We note that each mapping P_t is *reflexive* on its domain in the below sense: The mapping $P_t : \Omega \to 2^\Omega$ is said to be *reflexive* if

Ref $\quad \omega \in P_t(\omega) \quad$ for every $\omega \in \mathrm{Dom}(P_t)$,

and it is said to be *transitive* if:

Trn $\quad \xi \in P_t(\omega)$ implies $P_t(\xi) \subseteq P_t(\omega)$ for any $\xi, \omega \in \mathrm{Dom}(P_t)$.

Furthermore P_t is said to be *symmetric* if:

[3] The strong epistemic model is a tuple $\langle \Omega, (K_t)_{t\in T}\rangle$, in which t's knowledge operator $K_t : 2^\Omega \to 2^\Omega$ satisfies the five postulates: For every E, F of 2^Ω,

N $K_t\Omega = \Omega$; **K** $K_t(E \cap F) = K_tE \cap K_tF$; **T** $K_tF \subseteq F$;
4 $K_tF \subseteq K_tK_tF$; **5** $\Omega \setminus K_tF \subseteq K_t(\Omega \setminus K_tF)$.

[4] t's information partition P_t induced from the knowledge operator K_t is defined by $P_t(\omega) = \bigcap_{T\in 2^\Omega}\{T \in 2^\Omega \,|\, \omega \in K_tT\}$.

Sym $\xi \in P_t(\omega)$ implies $P_t(\xi) \ni \omega$ for any $\xi, \omega \in \mathrm{Dom}(P_t)$.

Remark 1. The strong epistemic model can be interpreted as the awareness structure $\langle \Omega, (A_t), (B_t) \rangle$ such that (Ω, B_t) is the Kripke semantics corresponding to the logic **S5**. In this situation it is easily verified that A_t must be the *trivial* operator; i.e. $A_t(F) = \Omega$ for every $F \in 2^\Omega$. We note that for each t, the associated information function P_t is a partition of the state space Ω with $\mathrm{Dom}(P_t) = \Omega$ in the strong epistemic model.

2.3 Economy with Awareness Structure

A *pure exchange economy under uncertainty* is a tuple $\langle T, \Omega, \mathbf{e}, (U_t)_{t\in T}, (\pi_t)_{t\in T} \rangle$ consisting of the following structure and interpretations: There are l commodities in each state of the state space Ω, and it is assumed that Ω is *finite* and that the consumption set of trader t is $\mathbb{R}^l_+$; $\mathbf{e}(t, \cdot) : T \times \Omega \to \mathbb{R}^l_+$ is t's *initial endowment*; $U_t : \mathbb{R}^l_+ \times \Omega \to \mathbb{R}$ is t's von-Neumann and Morgenstern utility function; π_t is a subjective prior on Ω for a trader $t \in T$. For simplicity it is assumed that (Ω, π_t) is a finite probability space with π_t *full support*[5] for all $t \in T$.

Definition 2. A *pure exchange economy with awareness structure* is a structure $\mathcal{E}^A = \langle \mathcal{E}, (A_t)_{t\in T}, (B_t)_{t\in T}, (P_t)_{t\in T} \rangle$, in which $\mathcal{E}$ is a pure exchange economy such that $\langle \Omega, (A_t)_{t\in T}, (B_t)_{t\in T}, (P_t)_{t\in T} \rangle$ is an awareness structure with $(P_t)_{t\in T}$ the associated information structure. By the *domain* of the economy $\mathcal{E}^A$ we mean $\mathrm{Dom}(\mathcal{E}^A) = \cap_{t\in T} \mathrm{Dom}(P_t)$. We always assume the below condition:

A-0 $\mathrm{Dom}(\mathcal{E}^A) \neq \emptyset$.

Remark 2. An economy under asymmetric information is an economy $\mathcal{E}^A$ in which $\langle \Omega, (A_t)_{t\in T}, (B_t)_{t\in T} \rangle$ is given by the strong epistemic model by Bacharach [1].

We denote by $\mathcal{F}_t$ the field generated by $\{P_t(\omega) |\ \omega \in \Omega\}$ and denote by $\Pi_t(\omega)$ the atom containing $\omega \in \Omega$. We denote by $\mathcal{F}$ the join of all $\mathcal{F}_t (t \in T)$; i.e. $\mathcal{F} = \vee_{t\in T} \mathcal{F}_t$, and denote by $\{\Pi(\omega) |\ \omega \in \Omega\}$ the set of all atoms $\Pi(\omega)$ containing ω of the field $\mathcal{F} = \vee_{t\in T} \mathcal{F}_t$.

We shall often refer to the following conditions: For every $t \in T$,

A-1 t's endowment $\mathbf{e}(t, \cdot)$ is $\mathcal{F}_t$-measurable with $\sum_{t\in T} \mathbf{e}(t, \omega) > 0$ for all $\omega \in \Omega$.
A-2 For each $x \in \mathbb{R}^l_+$, the function $U_t(x, \cdot)$ is $\mathcal{F}_t$-measurable.
A-3 For each $\omega \in \Omega$, the function $U_t(\cdot, \omega)$ is continuous, increasing, strictly quasi-concave and *non-saturated*[6] on $\mathbb{R}^l_+$.
A-4 For all $t \in T$, $\mathrm{Dom}(\mathcal{E}^A) = \mathrm{Dom}(P_t)$

[5] That is, $\pi_t(\omega) \neq 0$ for every $\omega \in \Omega$.
[6] That is, for any $x \in \mathbb{R}^l_+$ there exists an $x' \in \mathbb{R}^l_+$ such that $U_t(x', \omega) > U_t(x, \omega)$.

An *assignment* $\mathbf{x}$ is a mapping from $T \times \Omega$ into $\mathbb{R}^l_+$ such that for every $\omega \in \Omega$ and for each $t \in T$, the function $\mathbf{x}(t, \cdot)$ is at most $\mathcal{F}$-measurable. We denote by $\mathcal{A}ss(\mathcal{E}^A)$ the set of all assignments for the economy $\mathcal{E}^A$. By an *allocation* we mean an assignment $\mathbf{a}$ such that for every $\omega \in \Omega$, $\sum_{t \in T} \mathbf{a}(t, \omega) \leq \sum_{t \in T} \mathbf{e}(t, \omega)$. We denote by $\mathcal{A}lc(\mathcal{E}^A)$ the set of all allocations.

We introduce the revised notion of trader's expectation of utility in $\mathcal{E}^A$. By t's *ex-ante* expectation we mean $\mathbf{E}_t[U_t(\mathbf{x}(t, \cdot)] := \sum_{\omega \in \mathrm{Dom}(P_t)} U_t(\mathbf{x}(t, \omega), \omega)\pi_t(\omega)$ for each $\mathbf{x} \in \mathcal{A}ss(\mathcal{E}^A)$.[7] The *interim* expectation $\mathbf{E}_t[U_t(\mathbf{x}(t, \cdot)|P_t]$ is defined by

$$\mathbf{E}_t[U_t(\mathbf{x}(t, \cdot)|P_t](\omega) := \sum_{\xi \in \mathrm{Dom}(P_t)} U_t(\mathbf{x}(t, \xi), \xi)\pi_t(\{\xi\} \cap A_t(\{\xi\})|P_t(\omega)).^8$$

An allocation $\mathbf{x}$ in an economy $\mathcal{E}^A$ is said to be *ex-ante Pareto-optimal* if there is no allocation $\mathbf{a}$ such that $\mathbf{E}_t[U_t(\mathbf{a}(t, \cdot)] \geq \mathbf{E}_t[U_t(\mathbf{x}(t, \cdot)]$ for every $t \in T$, with at least one inequality strict.

2.4 Expectations Equilibrium in Awareness

A *price system* is a non-zero $\mathcal{F}$-measurable function $p : \Omega \to \mathbb{R}^l_+$. We denote $\Delta(p)$ the partition induced by p, and denote by $\sigma(p)$ the field generated by $\Delta(p)$. The *budget set* of a trader t at a state ω for a price system p is defined by $B_t(\omega, p) := \{\ x \in \mathbb{R}^l_+ \mid p(\omega) \cdot x \leq p(\omega) \cdot \mathbf{e}(t, \omega)\ \}$.

Let $\Delta(p) \cap P_t : \mathrm{Dom}(P_t) \to 2^\Omega$ be defined by $(\Delta(p) \cap P_t)(\omega) := \Delta(p)(\omega) \cap P_t(\omega)$; it is plainly observed that the mapping $\Delta(p) \cap P_t$ satisfies **Ref**. We denote by $\sigma(p) \vee \mathcal{F}_t$ the smallest σ-field containing both the fields $\sigma(p)$ and $\mathcal{F}_t$, and denote by $\Pi_t(p)(\omega)$ the atom containing ω. We note that $\Pi_t(p)(\omega) = (\Delta(p) \cap \Pi_t)(\omega)$. We shall give the extended notion of rational expectations equilibrium.

Definition 3. An *expectations equilibrium* for an economy $\mathcal{E}^A$ under reflexive information structure is a pair $(p, \mathbf{x})$, in which p is a price system and $\mathbf{x}$ is an allocation satisfying the following conditions:

RE 1 For all $t \in T$, $\mathbf{x}(t, \cdot)$ is $\sigma(p) \vee \mathcal{F}_t$-measurable.

RE 2 For all $t \in T$ and for every $\omega \in \Omega$, $\mathbf{x}(t, \omega) \in B_t(\omega, p)$.

RE 3 For all $t \in T$, if $\mathbf{y}(t, \cdot) : \Omega \to \mathbb{R}^l_+$ is $\sigma(p) \vee \mathcal{F}_t$-measurable with $\mathbf{y}(t, \omega) \in B_t(\omega, p)$ for all $\omega \in \Omega$, then $\mathbf{E}_t[U_t(\mathbf{x}(t, \cdot))|\Delta(p) \cap P_t](\omega) \geq \mathbf{E}_t[U_t(\mathbf{y}(t, \cdot))|\Delta(p) \cap P_t](\omega)$ pointwise on $\mathrm{Dom}(P_t)$.

RE 4 For every $\omega \in \Omega$, $\sum_{t \in T} \mathbf{x}(t, \omega) = \sum_{t \in T} \mathbf{e}(t, \omega)$.

The allocation $\mathbf{x}$ in $\mathcal{E}^A$ is called an *expectations equilibrium allocation* for $\mathcal{E}^A$.

[7] Where it is noted that $\xi \in \Omega$ in the summation runs over the domain of P_t.

[8] It should be noted that we use not the usual notion of posterior $\pi_t(\{\xi\}|P_t(\omega))$ but the revised one $\pi_t(\{\xi\} \cap A_t(\{\xi\})|P_t(\omega))$. For the discussion why this improvement of the notion of posterior is needed, see Matsuhisa and Usami [11] (Section 4).

We denote by $RE(\mathcal{E}^A)$ the set of all the rational expectations equilibria of a pure exchange economy $\mathcal{E}^A$ with awareness structure, and denote by $\mathcal{R}(\mathcal{E}^A)$ the set of all the expectations equilibrium allocations for the economy. We can establish the existence theorem of the expectations equilibrium for a pure exchange economy $\mathcal{E}^A$ with awareness structure.

Theorem 1. *Let $\mathcal{E}^A$ be a pure exchange economy with awareness structure satisfying the conditions* **A-1**, **A-2**, **A-3** *and* **A-4**. *Then there exists an expectations equilibrium for the economy; i.e., $RE(\mathcal{E}^A) \neq \emptyset$.*

Proof. Can be given in the same line of Matsuhisa [8]. □

3 No Speculation Theorem

We can now state the main theorem explicitly and prove it.

Theorem 2. *Let $\mathcal{E}^A$ be a pure exchange economy with awareness structure satisfying the conditions* **A-1**, **A-2**, **A-3** *and* **A-4**. *Suppose that the initial endowment* $\mathbf{e}$ *is ex-ante Pareto optimal in $\mathcal{E}^A$. If $(p, \mathbf{x})$ is an expectations equilibrium for $\mathcal{E}^A$ for some price system p then $\mathbf{x} = \mathbf{e}$.*

Before proceeding with the proof we shall show the below propositions: Let $\mathcal{E}^A(\omega)$ denote the economy with complete information $\langle T, (\mathbf{e}(t,\omega))_{t\in T}, (U_t(*,\omega))_{t\in T}\rangle$ for each $\omega \in \Omega$. We set by $W(\mathcal{E}^A(\omega))$ the set of all the competitive equilibria for $\mathcal{E}^A(\omega)$, and we denote by $\mathcal{W}(\mathcal{E}^A(\omega))$ the set of all the competitive equilibrium allocations for $\mathcal{E}^A(\omega)$.

Proposition 1. *Let $\mathcal{E}^A$ be a pure exchange economy with awareness structure satisfying the conditions* **A-1**, **A-2**, **A-3** *and* **A-4**. *The set of all expectations equilibrium allocations for $\mathcal{R}(\mathcal{E}^A)$ coincides with the set of all the assignments* $\mathbf{x}$ *such that $\mathbf{x}(\cdot,\omega)$ is a competitive equilibrium allocation for the economy with complete information $\mathcal{E}^A(\omega)$ for all $\omega \in \mathrm{Dom}(\mathcal{E}^A)$; i.e.,*

$$\mathcal{R}(\mathcal{E}^A) = \{\mathbf{x} \in \mathcal{A}lc(\mathcal{E}^A) | \quad \textit{There is a price system } p \textit{ such that } (p(\omega), \mathbf{x}(\cdot,\omega)) \in W(\mathcal{E}^A(\omega)) \textit{ for all } \omega \in \mathrm{Dom}(\mathcal{E}^A)\}.$$

Proof. [9] In view of the conditions **A-1**, **A-2** and **A-3**, it follows from the existence theorem of a competitive equilibrium that for each $\omega \in \Omega$, there exists a competitive equilibrium $(p^*(\omega), \mathbf{x}^*(\cdot,\omega)) \in W(\mathcal{E}^A(\omega))$ (C.f.; Debreu [2]). We take a set of strictly positive numbers $\{k_\omega\}_{\omega\in\Omega}$ such that $k_\omega p^*(\omega) \neq k_\xi p^*(\xi)$ for any $\omega \neq \xi$. We define the pair $(p, \mathbf{x})$ as follows: For each $\omega \in \Omega$ and for all $\xi \in \Pi(\omega)$, $p(\xi) := k_\omega p^*(\omega)$ and $\mathbf{x}(t,\xi) := \mathbf{x}^*(t,\omega)$. It is noted that $\mathbf{x}(\cdot,\xi) \in W(\mathcal{E}^A(\omega))$ because $\mathcal{E}^A(\xi) = \mathcal{E}^A(\omega)$, and we note that $\Delta(p)(\omega) = \Pi(\omega)$.

We shall verify that $(p, \mathbf{x})$ is an expectations equilibrium for $\mathcal{E}^A$: In fact, it is easily seen that p is $\mathcal{F}$-measurable with $\Delta(p)(\omega) = \Pi(\omega)$ and that $\mathbf{x}(t,\cdot)$ is

[9] The proof is given in the similar line of Matsuhisa, Ishikawa and Hoshino [10] (Proposition 2).

$\sigma(p) \vee \mathcal{F}_t$-measurable, so **RE 1** is valid. Because $(\Delta(p) \cap P_t)(\omega) = \Pi(\omega)$ for every $\omega \in \mathrm{Dom}(\mathcal{E}^A)$, it can be plainly observed that $\mathbf{x}(t,\cdot)$ satisfies **RE 2**, and it follows from **A-2** that for all $t \in T$, $\mathbf{E}_t[U_t(\mathbf{x}(t,\cdot))|\Delta(p) \cap P_t](\omega) = U_t(\mathbf{x}(t,\omega),\omega)$

On noting that $\mathcal{E}^A(\xi) = \mathcal{E}^A(\omega)$ for any $\xi \in \Pi(\omega)$, it is plainly observed that $(p(\omega),\mathbf{x}(t,\omega)) = (k_\omega p^*(\omega),\mathbf{x}^*(t,\omega))$ is also a competitive equilibrium for $\mathcal{E}^A(\omega)$ for every $\omega \in \mathrm{Dom}(\mathcal{E}^A)$, and it can be observed by the above equation that **RE 3** is valid for $(p,\mathbf{x})$, in completing the proof. □

Proposition 2. *Let $\mathcal{E}^A$ be a pure exchange economy with awareness structure satisfying the conditions* **A-1**, **A-2**, **A-3** *and* **A-4**. *Then an allocation* $\mathbf{x}$ *is ex-ante Pareto optimal if it is an expectations equilibrium allocation for $\mathcal{E}^A$ relative to a price system.*

Proof. Let $(p,\mathbf{x})$ be an expectations equilibrium for $\mathcal{E}^A$. It follows from Proposition 1 that $(p(\omega),\mathbf{x}(\cdot,\omega))$ is a competitive equilibrium for the economy $\mathcal{E}^A(\omega)$ at each $\omega \in \mathrm{Dom}(\mathcal{E}^A)$. Therefore in viewing the well known fundamental theorem of welfare in the economy with complete information $\mathcal{E}^A(\omega)$, we can plainly observe that for each t and for all $\omega \in \mathrm{Dom}(\mathcal{E}^A) = \mathrm{Dom}(P_t)$, $\mathbf{x}(\cdot,\omega)$ is Pareto optimal in $\mathcal{E}^A(\omega)$, and we conclude that $\mathbf{x}$ is ex-ante Pareto optimal. □

Proof of Theorem 2. Let $(p,\mathbf{x}) \in RE(\mathcal{E}^A)$. It follows from Proposition 2 that $\mathbf{x}$ is ex-ante Pareto optimal in $\mathcal{E}^A$. Suppose to the contrary that $\mathbf{x} \neq \mathbf{e}$. Since $\mathbf{e}$ is ex-ante Pareto optimal in $\mathcal{E}^A$ it can be observed that there exist an $s \in S$ such that $\mathbf{E}_s[U_s(\mathbf{e}(s,\cdot))] > \mathbf{E}_s[U_s(\mathbf{x}(s,\cdot))]$. Therefore, it can be plainly verified that for some $\omega_0 \in \mathrm{Dom}(P_s)$, $U_s(\mathbf{e}(s,\omega_0),\omega_0) > U_s(\mathbf{x}(s,\omega_0),\omega_0)$. On the other hand, it follows from Proposition 1 that $(p(\omega_0),\mathbf{x}(\cdot,\omega_0)) \in W(\mathcal{E}^A(\omega_0))$, thus $U_s(\mathbf{x}(s,\omega_0),\omega_0) > U_s(\mathbf{e}(s,\omega_0),\omega_0)$ in contradiction. This completes the proof. □

4 Related Results and Remarks

We shall discuss related results: Welfare in economics and Core equivalence. Finally we give some remarks about the conditions **A-1** to **A-4**.

The converse of Propositions 2 can be proved by the similar way in Matsuhisa and Ishikawa [9] (Proposition 4). Therefore we can characterize welfare under the expectations equilibrium for the economy $\mathcal{E}^A$: *Let $\mathcal{E}^A$ be a pure exchange economy with awareness structure satisfying the conditions* **A-1**, **A-2**, **A-3** *and* **A-4**. *An allocation is ex-ante Pareto optimal if and only if it is an expectations equilibrium allocation for $\mathcal{E}^A$ relative to some price system.*

Matsuhisa [8] introduces the notion of ex-post core in the economy for modal logic **KT**. Based on the article of Matsuhisa, Ishikawa and Hoshino [10] he establishes the core equivalence theorem in the economy equipped with non-atomic measure on the traders space: The ex-post core in the economy for **KT** coincides with the set of all its rational expectations equilibria. We can extend

the core equivalence theorem into the economy with awareness structure, and we shall report it in near future.

It well ends some remarks about the auxiliary assumptions **A-1** to **A-4** in results in this article. Could we prove the theorems without four conditions **A-1**, **A-2** and **A-3**. The answer is no vein. The suppression of any of these assumptions renders the existence theorem of expectations equilibrium for $\mathcal{E}^A$ (Theorem 1) vulnerable to the discussion and the example proposed in Remarks 4.6 of Matsuhisa and Ishikawa [9]. Could we prove the theorems removing out **A-4**? The answer is no again. If t's associated information function P_t does not satisfy **Ref** then his/her expectation with respect to a price cannot be defined at a state because it is possible that $\Delta(p)(\omega) \cap P_t(\omega) = \emptyset$ for some $\omega \in \Omega$.

References

1. Bacharach,M.: Some extensions of a claim of Aumann in an axiomatic model of knowledge. Journal of Economic Theory **37** (1985) 167-190.
2. Debreu, G.: Existence of competitive equilibrium. In Arrow, K.J. and Intriligator, M.D., (eds): Handbook of Mathematical Economics, Volume 2. North-Holland Publishing Company, Amsterdam (1982) 697-744.
3. Dekel,E., Lipman,B.L. and Rustichini,A.: Standard state-space models preclude unawareness. Econometrica **66** (1998) 159-173.
4. Fagin,R., Halpern,J.Y., Moses,Y. and Vardi, M.Y.: *Reasoning about Knowledge.* The MIT Press, Cambridge, Massachusetts, London, England, 1995.
5. Fudenberg, D. and Tirole, J.,*Game Theory.* MIT Press, Cambridge USA, xxii+579, 1991
6. Geanakoplos, J.: Game theory without partitions, and applications to speculation and consensus, Cowles Foundation Discussion Paper No.914 (1989) (Available in http://cowles.econ.yale.edu)
7. Kreps, D.: A note on fulfilled expectations equilibrium, Journal of Economic Theory **14** (1977) 32-44.
8. Matsuhisa, T.: Core equivalence in economy for modal logic, in P.M.A. Sloot, D. Abramson et al. (eds): Computational Science-ICCS 2003, Proceedings, Springer Lecture Notes of Computer Science, Vol. 2658 (2003) 74-83.
9. Matsuhisa, T. and Ishikawa, R.: Rational expectations can preclude trades. Preprint, Hitotsubashi Discussion Paper Series 2002-1 (2002) (Available in http://wakame.econ.hit-u.ac.jp/).
10. Matsuhisa, T., Ishikawa, R. and Hoshino, Y., 2002. Core equivalence in economy under generalized information. Working paper. Hitotsubashi Discussion Paper Series No.2002-12 (2002) (Available in http://wakame.econ.hit-u.ac.jp/)
11. Matsuhisa,T. and Usami,S.-S.: Awareness, belief and agreeing to disagree. Far East Journal of Mathematical Sciences **2** (2000) 833-844.
12. Milgrom, P. and Stokey, N.: Information, trade and common knowledge, Journal of Economic Theory 26 (1982) 17-27.
13. Morris, S.: Trade with heterogeneous prior beliefs and asymmetric information, Econometrica **62** (1994) 1327-1347.

A Method on Solving Multiobjective Conditional Value-at-Risk*

Min Jiang[1], Qiying Hu[2], and Zhiqing Meng[3]

[1] School of Economics and Management,
Xidian University, Xi'an, 710071,China
j_yx@263.net
[2] College of International Business & Management,
Shanghai University, Shanghai, 201800,China
huqiying@sina.com
[3] College of Business and Administration,
Zhejiang University of Technology, Hangzhou, 310032,China
mengzhiqing@xtu.edu.cn

Abstract. This paper studies Conditional Value-at-Risk (CVaR) with multiple losses. We introduce the concept of $\boldsymbol{\alpha}$-CVaR for the case of multiple losses under the confidence level vector $\boldsymbol{\alpha}$. The $\boldsymbol{\alpha}$-CVaR indicates the conditional expected losses corresponding to the $\boldsymbol{\alpha}$-VaR. The problem of solving the minimal $\boldsymbol{\alpha}$-CVaR results in a multiobjective problem (MCVaR). In order to get Pareto efficient solutions of the (MCVaR), we introduce a single objective problem (SCVaR) and show that the optimal solutions of the (SCVaR) are the Pareto efficient solutions of (MCVaR).

Keywords: Credit risk, Loss functions, $\boldsymbol{\alpha}$-CVaR, Pareto efficient solutions

1 Introduction

Value-at-Risk (VaR) is a measure for the potential loss in the financial market. With respect to a specified probability level α, the α-VaR of a portfolio is the lowest amount y such that, with probability α, the loss will not exceed y. VaR has achieved its great success in practice. However, research shows that VaR has undesirable properties both in theory and in practice. The main undesirable characteristics are as follows [1]. (a) There are various methodologies for modeling VaR, such as Historical Simulation, Monte Carlo, Extreme Value Theory, but the results using these methodologies are discrepant. (b) It is difficult to solve directly the problem of optimizing VaR. (c) VaR dissatisfies sub-additivity, which leads that financial organizations cannot obtain the whole VaR from those of their branches. (d) It is computed the loss exceeding VaR with likehood $(1-\alpha)$,

* The project was supported by the National Natural Science Foundation of China with grant 72072021.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 923–930, 2004.
© Springer-Verlag Berlin Heidelberg 2004

but no information is provided on the amount of the excess loss,which may be significantly greater.

The concept of Conditional Value at Risk (CVaR) ([2]) was presented mainly to solve the problem of sub-additivity. With a specified probability α, the α-CVaR is the conditional expectation of losses above the VaR. CVaR overcomes several limitations of VaR and has good properties, especially its good computability. Recently, there are several studies on CVaR area. Chernozhukov and Umantsev in [3] stressed the important aspects of measuring the external and intermediate conditional risk and gave an empirical application characterizing the key economic determinants of various levels of conditional risk. Andersson, Mausser, Rosen *et al.* in [4] presented an efficient algorithm to minimize CVaR. Rockafellar and Uryasev [5] derived some fundamental properties of conditional value-at-risk (CVaR) for loss distributions in Finance that can involve discreetness.

However, the losses that should be considered in practical risk management are often multiple, such as those due to interest risk, exchange risk, shares risk, commercial risk. So, the risk management problems are of multiobjective. Krokhmal, Palmquist and Uryasev [6] and Wang and Li [7] solve efficient frontier problems with three losses under the framework of CVaR. However, they did not concern in theoretical studies.

This paper will study the multiple conditional VaR. After introducing the concept of $\boldsymbol{\alpha}$-VaR with confidence level vector $\boldsymbol{\alpha}$ of a portfolio for multi losses based on the paper [2], we define $\boldsymbol{\alpha}$-CVaR as the conditional expected losses above the $\boldsymbol{\alpha}$-VaR. We then present the optimization problem of minimizing the multiple $\boldsymbol{\alpha}$-CVaR as a multiple programming. We prove that under some conditions the multiple $\boldsymbol{\alpha}$-CVaR problem can be transformed into a single objective nonlinear programming problem, which can be solved relative simply.

The remainder of this paper is organized as follows. In Section 2, we introduce the concept of α-CVaR and the main results of paper [2]. Then, in Section 3, we define the multiple $\boldsymbol{\alpha}$-CVaR. In Section 4, we show the main results of solving the multiple $\boldsymbol{\alpha}$-CVaR. Section 5 gives the conclusions.

2 CVaR with Single Loss

Rockfellar and Uryasev [2] discussed the CVaR model with single loss. Their main concepts and results are as follows.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ be the portfolio, which is also called the decision vector, with $\boldsymbol{x} \in X$ and $X \subset R^n$. Let $f(\boldsymbol{x}, \xi) : R^n \times R^m \to R^1$ be the loss associated with the decision vector $\boldsymbol{x}$, where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)^T$ is a random vector representing the uncertainties that affect the loss. For simplicity, it is assumed that ξ is a continuously random variable with p.d.f. $p(z)$. The cumulative distribution function of $f(\boldsymbol{x}, \xi)$ is denoted by

$$\Psi(\boldsymbol{x}, y) = P\{f(\boldsymbol{x}, \xi) \leq y\} = \int_{f(\boldsymbol{x}, z) \leq y} p(z) dz,$$

which is nondecreasing and right continuous in y for each $\boldsymbol{x} \in X$.

Definition 2.1 Given $\alpha \in (0,1)$ and $\boldsymbol{x} \in X$, the α-VaR (of the loss) associated with the decision $\boldsymbol{x}$ under the confidence level α is defined by

$$y_\alpha(\boldsymbol{x}) = \min\{y \in R : \Psi(\boldsymbol{x}, y) \geq \alpha\}, \tag{1}$$

which is the lowest amount y such that, with probability α, the loss will not exceed y. Moreover, α-CVaR (of the loss) associated with $\boldsymbol{x}$ under α is defined by

$$\phi_\alpha(\boldsymbol{x}) = (1-\alpha)^{-1} \int_{f(\boldsymbol{x},z) \geq y_\alpha(\boldsymbol{x})} f(\boldsymbol{x}, z) p(z) dz. \tag{2}$$

$\phi_\alpha(\boldsymbol{x})$ is the conditional expectation of the loss associated with $\boldsymbol{x}$ relatively to that loss being the α-VaR $y_\alpha(\boldsymbol{x})$ or greater. It is difficult to work directly with the α-CVaR because of unknowing $y_\alpha(\boldsymbol{x})$. Rockfellar and Uryasev introduced a simpler function:

$$F_\alpha(\boldsymbol{x}, y) = y + (1-\alpha)^{-1} \int_{z \in R} [f(\boldsymbol{x}, z) - y]^+ p(z) dz, \tag{3}$$

where $t^+ = \max(0, t)$.

Lemma 2.1 As a function of y, $F_\alpha(\boldsymbol{x}, y)$ is convex, continuously differentiable and

$$\frac{\partial}{\partial y} F_\alpha(\boldsymbol{x}, y) = (1-\alpha)^{-1} [\Psi_\alpha(\boldsymbol{x}, y) - \alpha].$$

Under the condition of

$$P\{f(\boldsymbol{x}, \xi) = y\} = \int\limits_{f(\boldsymbol{x},z)=y} p(z) dz = 0, \tag{4}$$

the following result is shown.

Theorem 2.1 Minimizing the α-CVaR $\phi_\alpha(\boldsymbol{x})$ over $\boldsymbol{x} \in X$ is equivalent to minimizing $F_\alpha(\boldsymbol{x}, y)$ over $(\boldsymbol{x}, y) \in X \times R$, in the sense that

$$\min_{\boldsymbol{x} \in X} \phi_\alpha(\boldsymbol{x}) = \min_{(\boldsymbol{x}, y) \in X \times R} F_\alpha(\boldsymbol{x}, y).$$

Furthermore, $F_\alpha(\boldsymbol{x}, y)$ is convex with respect to $(\boldsymbol{x}, y)$ and $\phi_\alpha(\boldsymbol{x})$ is convex with respect to $\boldsymbol{x}$ when $f(\boldsymbol{x}, y)$ is convex with respect to $\boldsymbol{x}$. In this case, if X is a convex set, the above joint minimization is a convex programming.

According to Theorem 2.1, for the purpose of determining an $\boldsymbol{x}$ that yields the minimum α-CVaR, it is not necessary to minimize directly the function $\phi_\alpha(\boldsymbol{x})$ which may be hard to do. Instead, one can operate on the far simpler function $F_\alpha(\boldsymbol{x}, y)$.

Theorem 2.1 opens the door to minimize α-CVaR over $\boldsymbol{x} \in X$ and so we can use it to solve practical risk management problems.

3 CVaR with Multiple Losses

In this section, we generalize the concepts of α-VaR and α-CVaR into the case of multiple losses.

Let $f_i(\boldsymbol{x}, \boldsymbol{\xi}) \in R^n \times R^m \to R^1 (i = 1, 2, \cdots, I)$ be the loss functions which depends upon the decision vector $\boldsymbol{x} \in X \subset R^n$ and the random vector $\boldsymbol{\xi} \in R^m$. X is the set of possible portfolios. For simplicity, we assume that ξ is a random variable with the probability density function $p(z)$. However, this assumption is not critical for our discussions in the following. Denote by $\Psi_i(\boldsymbol{x}, \cdot)$ the distribution function of the loss $f_i(\boldsymbol{x}, \xi)$, i.e.,

$$\Psi_i(\boldsymbol{x}, y) = P\{f_i(\boldsymbol{x}, \xi) \le y\} = \int\limits_{f_i(\boldsymbol{x}, z) \le y} p(z)dz, \qquad i = 1, 2, \cdots, I. \tag{5}$$

We give the following definition.

Definition 3.1 Given $\alpha_i \in (0, 1)$, $i = 1, 2, \cdots, I$, and $\boldsymbol{x} \in X$, the $\boldsymbol{\alpha}$-VaR (of the loss) associated with $\boldsymbol{x}$ under the confidence level vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_I)$ is defined by

$$y^*(\boldsymbol{x}) = \min\{y \mid \Psi_i(\boldsymbol{x}, y) \ge \alpha_i, \quad i = 1, 2, \cdots, I\}. \tag{6}$$

When $I = 1$, the $\boldsymbol{\alpha}$-VaR is exactly the α-VaR for one loss defined by Definition 2.1. For $i = 1, 2, \cdots, I$, $\Psi_i(\boldsymbol{x}, y) \ge \alpha_i$, or $P\{f_i(\boldsymbol{x}, \xi) \le y\} \ge \alpha_i$, means that, with probability α_i, the ith loss will not exceed y for the portfolio $\boldsymbol{x}$. So,

$$y_i^*(\boldsymbol{x}) = \min\{y \mid \Psi_i(\boldsymbol{x}, y) \ge \alpha_i\}$$

is the α_i-VaR of x for the ith loss. Therefore, the $\boldsymbol{\alpha}$-VaR $y^*(x)$ for a portfolio $\boldsymbol{x}$ is the lowest amount y such that ith loss will not exceed y with probability α_i for each i. Clearly, when $\Psi_i(\boldsymbol{x}, y)$ is nondecreasing and continuous in y, $y_i^*(\boldsymbol{x})$ is the smallest root of the equation $\Psi_i(\boldsymbol{x}, y) = \alpha_i$.

Obviously, given the confidence level vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_I)$, $y^*(\boldsymbol{x})$ is the maximum of the α_i-VaR, i.e., $y^*(\boldsymbol{x}) = \max\{y_1^*(\boldsymbol{x}), y_2^*(\boldsymbol{x}), \cdots, y_I^*(\boldsymbol{x})\}$.

When we discuss $\boldsymbol{\alpha}$-VaR of $\boldsymbol{x}$, we want to know how much the expected losses of VaR is. For each $i = 1, 2, \cdots, I$, we define

$$\phi_{\alpha_i}(\boldsymbol{x}, y) = (1 - \alpha_i)^{-1} \int\limits_{f_i(\boldsymbol{x}, z) \ge y} f_i(\boldsymbol{x}, z)p(z)dz. \tag{7}$$

Then, we have

$$\phi_{\alpha_i}(\boldsymbol{x}, y) = (1 - \alpha_i)^{-1} P\{f_i(\boldsymbol{x}, \xi) \ge y\} E\{f_i(\boldsymbol{x}, \xi) \mid f_i(\boldsymbol{x}, \xi) \ge y\}.$$

When $y = y_i^*(\boldsymbol{x})$, $P\{f_i(\boldsymbol{x}, \xi) \ge y_i^*(\boldsymbol{x})\} = \alpha_i$ and so $\phi_{\alpha_i}(\boldsymbol{x}, y_i^*(\boldsymbol{x})) = E\{f_i(\boldsymbol{x}, \xi) \mid f_i(\boldsymbol{x}, \xi) \ge y_i^*(\boldsymbol{x})\}$ is exactly the conditional expected ith loss for the portfolio $\boldsymbol{x}$. Thus we use $\phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x}))$ to approximate the conditional expected ith loss for $\boldsymbol{x}$ and α_i under the loss $y^*(\boldsymbol{x})$. It describes the risk of the portfolio $\boldsymbol{x}$ for the ith loss. If $\phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x}))$ is small, the risk will be also small.

Hence, $(\phi_{\alpha_1}(\boldsymbol{x}, y^*(\boldsymbol{x})), \phi_{\alpha_2}(\boldsymbol{x}, y^*(\boldsymbol{x})), \cdots, \phi_{\alpha_I}(\boldsymbol{x}, y^*(\boldsymbol{x})))$ describes the conditional expected losses under the $\boldsymbol{\alpha}$-VaR $y^*(\boldsymbol{x})$ for the portfolio $\boldsymbol{x}$ and the confidence level vector $\boldsymbol{\alpha}$.

Definition 3.2 For given $\alpha_i \in (0,1)$, $i = 1, 2, \cdots, I$, and $\boldsymbol{x} \in X$, we call the vector $(\phi_{\alpha_1}(\boldsymbol{x}, y^*(\boldsymbol{x})), \phi_{\alpha_2}(\mathbf{x}, y^*(\boldsymbol{x})), \cdots, \phi_{\alpha_I}(\boldsymbol{x}, y^*(\boldsymbol{x})))$ the $\boldsymbol{\alpha}$-CVaR associated with $\boldsymbol{x}$ under the confidence level vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_I)$.

It is not easy to compute $\phi_{\alpha_i}(\boldsymbol{x}, y)$. Thus we introduce another loss function as in [2]. For $i = 1, 2, \cdots, I$, we define

$$F_{\alpha_i}(\boldsymbol{x}, y) = y + (1 - \alpha_i)^{-1} \int\limits_{z \in R} [f_i(\boldsymbol{x}, z) - y]^+ p(z) dz. \tag{8}$$

By Lemma 2.1, we have the following result under the condition that all the loss functions satisfy (4), i.e., $P\{f_i(x, \xi) = y\} = 0$ for $i = 1, 2, \cdots, I$.

Lemma 3.1 For each $i = 1, 2, \cdots, I$, $F_{\alpha_i}(\boldsymbol{x}, y)$ is a continuous differential and convex function, and

$$\min_{y \in R} F_{\alpha_i}(\boldsymbol{x}, y) = \phi_{\alpha_i}(\boldsymbol{x}, y_i^*(\boldsymbol{x})), \tag{9}$$

$$\frac{\partial F_{\alpha_i}(\boldsymbol{x}, y)}{\partial y} = (1 - \alpha_i)^{-1} [\Psi_{\alpha_i}(\mathbf{x}, y) - \alpha_i]. \tag{10}$$

4 Main Results

We need to consider the minimal $\boldsymbol{\alpha}$-CVaR of $\boldsymbol{x} \in X$. That is, we face the following multiobjective problem:

$$\begin{aligned} \text{(MCVaR)} \quad & \min \ (\phi_{\alpha_1}(\boldsymbol{x}, y^*(\boldsymbol{x})), \phi_{\alpha_2}(\boldsymbol{x}, y^*(\boldsymbol{x})), \cdots, \phi_{\alpha_I}(\boldsymbol{x}, y^*(\boldsymbol{x}))) \\ & \text{s. t. } \boldsymbol{x} \in X. \end{aligned}$$

We first introduce the concept of efficient solutions in mutliobjective problems. Consider a multiobjective problem ([8])

$$\begin{aligned} \text{(MP)} \quad & \min \ (h_1(y), h_2(y), \cdots, h_I(y)) \\ & \text{s. t. } y \in Y, \end{aligned}$$

with $h_i : R^m \to R, 1 = 1, 2, \cdots, I$ and $Y \subset R^m$. For $y^* \in Y$, if there is no $y \in Y$ such that

$$h_1(y) \le h_1(y^*), h_2(y) \le h_2(y^*), \cdots, h_I(y) \le h_I(y^*)$$

where at least one inequality is strict, then y^* is called a Pareto efficient solution to (MP).

For the given $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_I)$, if $\boldsymbol{x}^*$ is a Pareto efficient solution to (MCVaR), then $(\phi_{\alpha_1}(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*)), \phi_{\alpha_2}(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*)), \cdots, \phi_{\alpha_I}(\boldsymbol{x}^*, y^*(\boldsymbol{x}^*)))$ is called

a Pareto-$\boldsymbol{\alpha}$-CVaR and $\boldsymbol{x}^*$ is called a Pareto-$\boldsymbol{\alpha}$-CVaR efficient solution. The set of all Pareto-$\boldsymbol{\alpha}$-CVaR efficient solutions is denoted by $E(\boldsymbol{\alpha})$. In the following, we want to find out such an efficient solution.

It is difficult to compute $\phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x}))$ from (7), so do (MCVaR). Therefore, we should consider other problems. First, we have the following lemma.

Lemma 4.1 For each $\boldsymbol{x} \in X$ and $y \in R$, we have that

$$\phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x})) \leq F_{\alpha_i}(\boldsymbol{x}, y), \qquad i = 1, 2, \cdots, I. \tag{11}$$

Proof. Since that $\phi_{\alpha_i}(\boldsymbol{x}, y)$ is nonincreasing in y, we have that

$$\phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x})) \leq \phi_{\alpha_i}(\boldsymbol{x}, y_i^*(\boldsymbol{x})). \tag{12}$$

Then by Lemma 3.1, we obtain (11). □

For the given weight $0 < \lambda_i < 1, i = 1, 2, \cdots, I$, we consider the problem $\min\limits_{y \in R} \sum\limits_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y)$, for which we have the following theorem.

Theorem 4.1 Let $\boldsymbol{x} \in X$. Suppose that $\overline{y}$ is an optimal solution to the problem $\min\limits_{y \in R} \sum\limits_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y)$ and satisfies the condition

$$P\{f_i(\boldsymbol{x}, \xi) = \overline{y}\} = 0, \qquad i = 1, 2, \cdots, I. \tag{13}$$

Then

$$\sum_{i=1}^{I} \lambda_i \phi_{\alpha_i}(\boldsymbol{x}, \overline{y}) = \min_{y \in R} \sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y). \tag{14}$$

Proof. Let $F(\boldsymbol{x}, y) = \sum\limits_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y)$. By Lemma 3.1, we have that

$$\sum_{i=1}^{I} \lambda_i (1 - \alpha_i)^{-1} (\Psi_i(\boldsymbol{x}, \overline{y}) - \alpha_i) = 0. \tag{15}$$

With this and (7), we have

$$\sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, \overline{y}) = \sum_{i=1}^{I} \lambda_i \phi_{\alpha_i}(\boldsymbol{x}, \overline{y}) + \overline{y} \sum_{i=1}^{I} \lambda_i (1 - \alpha_i)^{-1} (\Psi_i(\mathbf{x}, \overline{y}) - \alpha_i),$$

which together with (15) implies (14). □

When $f_i(\boldsymbol{x}, \xi)$ is a continuous random variable, the condition (13) must be true. But, when $f_i(\boldsymbol{x}, \xi)$ is a discrete type random variable, (13) holds if and only if that $\overline{y}$ is not in the range of $f_i(\boldsymbol{x}, \xi)$.

Now, suppose that $\Psi_i(\boldsymbol{x}, y)$ is strictly increasing in y. By Lemma 4.1 and Theorem 4.1, we have the following results.

Corollary 4.1 Under the conditions given in Theorem 4.1, $\overline{y} = y^*(\boldsymbol{x})$ if and only if

$$\Psi_i(\boldsymbol{x}, \overline{y}) - \alpha_i = 0, \qquad i = 1, 2, \cdots, I. \tag{16}$$

Corollary 4.2 For any $\boldsymbol{x} \in X$,

$$\sum_{i=1}^{I} \lambda_i \phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x})) \leq \min_{y \in R} \sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y). \tag{17}$$

Moreover, if the problem $\min\limits_{y \in R} \sum\limits_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y)$ has an optimal solution $\overline{y}$ that satisfies (13) and (16), then the inequality (17) becomes an eqality.

Based on the above discussions, we now consider the following single objective problem.

$$\begin{aligned} &\min \sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y), \\ &\text{s. t. } y \in R, \boldsymbol{x} \in X. \end{aligned} \tag{18}$$

We have the following theorem to relate this single objective problem to the multiobjective problem (MCVaR).

Theorem 4.2 If $(\overline{\boldsymbol{x}}, \overline{y})$ is an optimal solution to (18) and $\overline{y}$ satisfies (13) and (16) for $\boldsymbol{x} = \overline{\boldsymbol{x}}$, then $\overline{\boldsymbol{x}}$ is a Pareto-α-CVaR efficient solution to (MCVaR).

Proof. If $(\overline{\boldsymbol{x}}, \overline{y})$ is an optimal solution to (18), then due to Corollary 4.2, $\overline{\boldsymbol{x}}$ is also an optimal solution to $\min\limits_{\boldsymbol{x} \in X} \sum\limits_{i=1}^{I} \lambda_i \phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x}))$. Therefore, by the theory of mutliobjective programming, $\overline{\boldsymbol{x}}$ is also a Pareto efficient solution to (MCVaR) ([8]). Hence, the theorem is true. □

The remained problem is that the condition (16) is too serious. For this, we consider the following single objective problem:

$$\begin{aligned} (\text{SCVaR}) \min &\sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y), \\ \text{s. t. } &y \in R, \boldsymbol{x} \in X, \\ &0 \leq \lambda_i \leq 1, \sum_{i=1}^{I} \lambda_i = 1. \end{aligned}$$

The difference between this programming with (18) is that $\{\lambda_i, i = 1, 2, \cdots, I\}$ are give in (18) while are decision variables here. The function by letting λ_i be decision variables is to delete condition (16). Let

$$\Lambda = \{\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_I) \mid \lambda_i \in [0, 1], i = 1, 2, \cdots, I, \sum_{i=1}^{I} \lambda_i = 1\}$$

We have the following better result.

Theorem 4.3 Suppose that $\Psi_i(\boldsymbol{x}, y)$ is strictly increasing in y. If $(\overline{\boldsymbol{x}}^*, \overline{y}^*, \overline{\lambda}^*)$ is an optimal solution to (SCVaR) and satisfies (13), then $\overline{\boldsymbol{x}}^*$ is a Pareto-$\boldsymbol{\alpha}$-CVaR efficient solution to (MCVaR).

Proof. For $\lambda \in \Lambda$, let

$$y^*(\boldsymbol{x}, \lambda) = \min\{y \mid \sum_{i=1}^{I} \lambda_i (1-\alpha_i)^{-1} \Psi_i(\boldsymbol{x}, y) \geq \sum_{i=1}^{I} \lambda_i \alpha_i (1-\alpha_i)^{-1}\}.$$

We show that

$$\min_{\boldsymbol{x} \in X} \sum_{i=1}^{I} \lambda_i^* \phi_{\alpha_i}(\boldsymbol{x}, y^*(\boldsymbol{x})) = \min_{\boldsymbol{x} \in X} \min_{\lambda \in \Lambda} \min_{y \in R} \sum_{i=1}^{I} \lambda_i F_{\alpha_i}(\boldsymbol{x}, y). \tag{19}$$

From (19), if $(\overline{\boldsymbol{x}}^*, \overline{y}^*, \overline{\lambda}^*)$ is an optimal solution to (SCVaR), we know from Theorem 4.2 that $\overline{\boldsymbol{x}}^*$ is a Pareto-α-CVaR efficient solution to (MCVaR). □

Therefore, in order to solve the multiobjective problem (MCVAR), we need only to solve the single objective problem (SCVAR), which can be solved clearly not too difficult.

5 Conclusion

This paper discusses the CVaR problem with multiple losses. We introduce the concepts of $\boldsymbol{\alpha}$-VaR and $\boldsymbol{\alpha}$-CVaR with the confidence level vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_I)$. It is shown that obtaining Pareto efficient solutions of the minimal $\boldsymbol{\alpha}$-CVaR can be transformed into solving a single objective problem under some mild conditions. This paper resolves the problem of multiobjective CVaR in theory, and provides the theory base to numerical calculate and help us to study theory and solution of multiobjective CVaR problems under others definitions of $\boldsymbol{\alpha}$-VaR furthermore.

References

1. Chen, J.-L. and Zhang, W.: CVaR and the unifying model of portfolio optimization CVaR (in Chinese), Systems Enginnering-Theory, Methodolgy, and Applications, **1**(2002)68-71.
2. Rockafellar, R. T. and Uryasev, S.: Optimization of Conditional Value-at-Risk, Journal of Risk **2**(2000)21-41.
3. Chernozhukov, V. and Umantsev, L.: Conditional Value-at-Risk: aspects of modeling and estimation, Empirical Economics, **26**(2001) 271-292.
4. Andersson, F., Mausser, H., Rosen, D. and Uryasev, S.: Credit risk optimization with Conditional Value-at-Risk criterion, Math. Program. **89**(2001)273-291.
5. Rockafellar R. T. and Uryasev, S.: Conditional Value-at-Risk for general loss distributions, Journal of Banking & Finance **26**(2002) 1443-1471.
6. Krokhmal, P., Palmquist, J. and Uryasev, S.: Portfolio optimization with Conditional Value-at-Risk objectives and constraints, Journal of Risk, **2**(2002)124-129.
7. Wang J. H. and Li C. L., New method of measurement and control finance risk (in Chinese), Joural of Wuhan University of Techology, 2002, 24:2, pp. 60-63.
8. Sawragi, Y. , Nakayama, H. and Tanino, T.: Theory of multiobjective optimization, Academic Press, New York (1985).

Cross-Validation and Ensemble Analyses on Multiple-Criteria Linear Programming Classification for Credit Cardholder Behavior*

Yi Peng[1], Gang Kou[1], Zhengxin Chen[1], and Yong Shi [1, 2**]

[1]College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68182, USA
{ypeng,gkou,zchen,yshi}@mail.unomaha.edu
[2]Graduate School of Chinese Academy of Sciences, Beijing 100039, China

Abstract. In credit card portfolio management, predicting the cardholders' behavior is a key to reduce the charge off risk of credit card issuers. As a promising data mining approach, multiple criteria linear programming (MCLP) has been successfully applied to classify credit cardholders' behavior into two or multiple-groups for business intelligence. The objective of this paper is to study the stability of MCLP in classifying credit cardholders' behavior by using cross-validation and ensemble techniques. An overview of the two-group MCLP model formulation and a description of the dataset used in this paper are introduced first. Then cross-validation and ensemble methods are tested respectively. As the results demonstrated, the classification rates of cross-validation and ensemble methods are close to the rates of using MCLP alone. In other words, MCLP is a relatively stable method in classifying credit cardholders' behavior.

Keywords: Credit Card Portfolio Management, Data Mining, Classification, Multi-criteria Linear Programming, Cross-Validation, and Ensemble

1 Introduction

Mining useful information or discovering knowledge from large databases becomes a cutting-edge information technology tool in today's competitive business world. Data mining techniques help organizations and companies to discover previously unknown and actionable information from various and large databases for decision-making. Recently, Multiple-Criteria Linear Programming (MCLP) has been applied to credit cardholders' behavior classification for business decision [1, 2, 3, 4]. However, the stability of MCLP classification in credit card portfolio management remains unexplored. In order to respond this challenge, this paper conducts cross-validation and ensemble analysis using real-life credit card data from a large US bank. In this study, we intend to investigate whether the classification results generated by MCLP method can be improved. The k-fold cross-validation is first applied on the credit card data to acquire a number of global optimal solutions of MCLP method.

* This research has been partially supported by a grant of US Air Force Research Laboratory (PR No. E-3-1162) and a grant from the K.C. Wong Education Foundation (2003), Chinese Academy of Sciences.

** The corresponding Author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 931–939, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Then, part of these global optimal solutions is aggregated to form a single ensemble. The output ensemble is used to justify and improve the MCLP classification. Our findings indicate that the classification rates of cross-validation and ensemble analyses are close to that of MCLP alone. In other words, MCLP is a relatively stable method in classifying credit cardholders' behavior.

This paper is organized as follows. Next section is an overview of two-group MCLP model formulation. The third section describes the characteristics of the credit card dataset. The fourth section discusses the process and results of cross validation. The fifth section describes the procedure of ensemble analysis and examination of the resulting classification. The last section concludes the paper with some remarks.

2 Two-Group Multiple-Criteria Linear Programming Model

This section describes the two-group MCLP model briefly. For more details of MCLP and MCLP model formulation, please refer to [1, 2].

Often linear classification models use a linear combination of the minimization of the sum of overlapping (represented by α_i) and maximization of the sum of distance (represented by β_i) to reduce the two criteria problem into a single criterion. A two-criterion linear programming model is stated as:

(Model 1) *Minimize* $\Sigma_i \alpha_i$ and *Maximize* $\Sigma_i \beta_i$

Subject to:

$A_i X = b + \alpha_i - \beta_i, A_i \in G1,$

$A_i X = b - \alpha_i + \beta_i, A_i \in G2,$

Where A_i are given, X and b are unrestricted, and α_i and $\beta_i \geq 0$. The advantage of this conversion is to easily utilize all techniques of LP for separation, while the disadvantage is that it may miss the scenario of trade-offs between these two separation-criteria.

Applying the techniques of MCLP and the compromise solution, we want to minimize the sum of α_i and maximize the sum of β_i simultaneously. We assume the "ideal value" of $-\Sigma_i \alpha_i$ be $\alpha^* > 0$ and the "ideal value" of $\Sigma_i \beta_i$ be $\beta^* > 0$. Then, if $-\Sigma_i \alpha_i > \alpha^*$, we define the regret measure as $-d_\alpha^+ = \Sigma_i \alpha_i + \alpha^*$; otherwise, it is 0. If $-\Sigma_i \alpha_i < \alpha^*$, the regret measure is defined as $d_\alpha^- = \alpha^* + \Sigma_i \alpha_i$; otherwise, it is 0. Thus, we have (i) $\alpha^* + \Sigma_i \alpha_i = d_\alpha^- - d_\alpha^+$, (ii) $|\alpha^* + \Sigma_i \alpha_i| = d_\alpha^- + d_\alpha^+$, and (iii) $d_\alpha^-, d_\alpha^+ \geq 0$. Similarly, we derive $\beta^* - \Sigma_i \beta_i = d_\beta^- - d_\beta^+$, $|\beta^* - \Sigma_i \beta_i| = d_\beta^- + d_\beta^+$, and $d_\beta^-, d_\beta^+ \geq 0$ (see Figure 1). A two-group MCLP model has been gradually evolved as:

(Model 2) *Minimize* $d_\alpha^- + d_\alpha^+ + d_\beta^- + d_\beta^+$

Subject to:

$\alpha^* + \Sigma_i \alpha_i = d_\alpha^- - d_\alpha^+,$

$\beta^* - \Sigma_i \beta_i = d_\beta^- - d_\beta^+,$

$A_i X = b + \alpha_i - \beta_i, A_i \in G1,$

$A_i X = b - \alpha_i + \beta_i, A_i \in G2,$

where A_i, α^*, and β^* are given, X and b are unrestricted, and α_i, β_i, d_α^-, d_α^+, d_β^-, $d_\beta^+ \geq 0$.

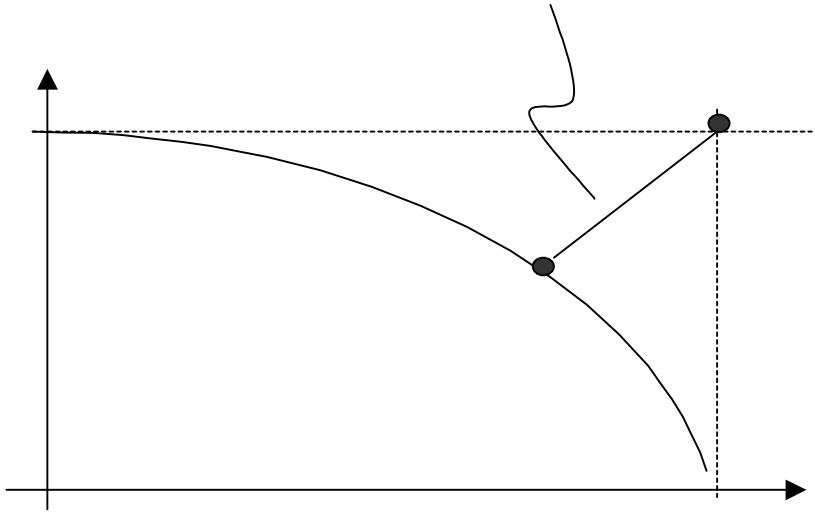


Fig. 1. Compromise Formulation

3 Credit Card Dataset Description

In order to understand the data mining process, it is important to comprehend the dataset. This section presents the nature and structure of the credit card data in details.

One of important data mining applications in banking industry is credit card bankruptcy analysis. Given a set of attributes, such as monthly payment, balance, purchases, and cash advance, the purpose is to find the optimized classifier through a training set and then use the classifier to predict future customers' spending behaviors [1, 3]. The common practice in credit card portfolio management is to separate credit cardholders' behaviors into two classes: bankruptcy or current. This is also known as the method of making "black list." Popular methods include Behavior Score, Credit Bureau Score, FDC Bankruptcy Score, and Set Enumeration Decision Tree Score [3]. These methods were developed by either statistics or decision tree.

The 5000 credit card records used in this paper were randomly selected from 25,000 real-life credit card records of a major US bank. Each record has 113 columns or variables (38 original variables and 65 derived variables) which are used to describe cardholders' behaviors. The 38 original variables can be divided into five categories: balance, purchase, payment, cash advance, and related variables. Balance, purchase,

payment, and cash advance categories each have six variables that represent raw data of six consecutive months. Other related variables include interest charges, date of last payment, times of cash advance, account open date, and so on. The 65 derived variables (CHAR01-CHAR65) are derived from original 38 variables using simple arithmetic methods to reinforce the comprehension of cardholders' behaviors, such as times of over limit in last two years, calculated interest rate, cash as percentage of balance, purchase as percentage to balance, payment as percentage to balance, purchase as percentage to payment, and so forth. These variables are not static; rather, they are evolving. New variables which are considered important can be added and variables which are proved to be trivia or irrelative in separating can be removed. Within the 5000 records, 815 accounts are bankrupted and 4185 are current.

After the discussion of MCLP model formulation and dataset structure, the basis of classification of credit cardholders' behaviors has established. The next section will introduce the concept of cross-validation and report the experimental steps and results.

4 Empirical Studies of Cross-Validation

Cross-validation is frequently used for estimating generalization error, model selection, experimental design evaluation, training exemplars selection, or pruning outliers [5]. By definition, cross-validation is the practice of partitioning a sample of data into sub samples such that analysis is initially performed on a single sub sample, while further sub samples are retained "blind" in order for subsequent use in confirming and validating the initial analysis [6]. The basic idea is to remove some of the data before training. After training is done, the data that was removed is used to test the performance of the model.

Three kinds of cross validation forms: holdout cross validation, k-fold cross validation, and leave-one-out cross validation are widely used. In this paper, k-fold cross validation is used. In the k-fold method, the data set is divided into k subsets and the holdout method is repeated k times. Each time, one of the k subsets is used for testing and the other k-1 subsets are used for training. The advantage is that all the examples in the dataset are eventually used for both training and testing. The error estimate is obtained as the average error rate on test examples. The variance of the resulting estimate is reduced as k is increased [7]. The disadvantage of this method is that it required high computation cost.

One of the important questions of cross validation is to decide the number of folds. With a large number of folds, the bias of the true error rate estimator will be small and the computational time will be larege. On the other hand, when the number of folds is small, the bias of the estimator will be large and the computional time will be reduced. In practice, the choice of the number of folds depends on the size of the dataset. A common choice for k-fold cross-validation is around 10 [7]. Due to the structure of our dataset, k is decided to be 7. As stated in section 3, the total bankruptcy accounts are 815 and current accounts are 4185. Since bankruptcy class has smaller number of records, records of each class in the training dataset should be

calculated using bankruptcy data: 815×6/7=699. For easier computation, 700 is used instead of 699. Thus, the training dataset is formulated with 1400 records (700 bankruptcy and 700 current).

The procedure to select training datasets is described as follows: first, the bankruptcy dataset (815 records) is divided into 100 intervals (each interval has 8 records). Within each interval, 7 records are randomly selected. Thus the total of 700 bankruptcy records are obtained after this selection was repeated 100 times. Second, the current dataset (4185 records) is divided into 100 intervals (each interval has 41 records). Within each interval, 7 records are randomly selected. Thus the total of 700 current records are obtained after this selection was repeated 100 times. Third, the 700 bankruptcy and 700 current records are combined to form a single training dataset. Finally, the remaining 115 bankruptcy and 3485 current accounts become the testing dataset. According to this procedure, the total possible combination of this selection equals to $(C_8^7 \times C_{41}^7)^{100}$, which is infinitely large. That is, we can consider the possibilty to get identical training or testing datasets is approximately zero. Considering the limited data availability in this study, we set the across-the-board threshold of 65% for absolute catch rate of the Bankruptcy class and 70% for the Current class to select the experimental results from training and test processes. This criterion is also applied to the ensemble analysis. Under the conditions stated above, the following steps are designed to carry out cross-validation:

Algorithm 1 (Cross Validation)

Step1 Generate the Training set (700 Bankruptcy data+700 Current Data) and Testing set (115 Bankruptcy data+3485 Current Data) from the credit card data set.

Step2 Apply the two-group MCLP model to compute the compromise solution $X^* = (x_1^*, x_2^*, \ldots, x_{65}^*)$ as the best weights of all 65 variables with given values of control parameters (b, α^*, β^*).

Step3 The classification score $MCLP_i = A_i X^*$ against of each observation has been calculated against the boundary b to check the performance measures of the classification.

Step4 If the classification result of Step 2 is acceptable (i.e., the given performance measure is larger or equal to the given threshold), go to Step 5. Otherwise, choose different values of control parameters (b, α^*, β^*) and go to Step 1.

Step5 Use $X^* = (x_1^*, x_2^*, \ldots, x_{65}^*)$ to calculate the MCLP scores for all A_i in the test set and conduct the performance analysis. If it produces a satisfying classification result, go to the next step. Otherwise, go back to Step 1 to reformulate the Training Set and Testing Set.

Step6 Repeat the whole process until a preset number (e.g. 999) of different X^* are generated.

Some samples of the Cross-Validation tests based on Algorithm 1 are summarized in Tables 1.

The training and testing datasets have been computed using the above procedure. A part (20 out of the total 467 cross-validation results) of the results against the threshold is summarized. The worst and best classification catch rates for training set are 78.00% and 81.29% for Bankruptcy, 77.14% and 81.14% for the Current. The worst and best classification catch rates for testing set are 65.22% and 73.04% for Bankruptcy, 71.68% and 76.13% for the Current. As shown in table 1, the absolute catch rates of the Bankruptcy class are all above 65% and the absolute catch rates of the Current class are all above 70% for the selected experimental results. The result indicates that a good separation of Bankruptcy and Current is observed with this method.

Table 1. A Sample of the Cross-Validation Results

Cross Validation	Training Set (700 Bankruptcy data + 700 Current data)				Testing Set (115 Bankruptcy data + 3485 Current data)			
	Bank-ruptcy	Catch Rate	Cur-rent	Catch Rate	Bank-ruptcy	Catch Rate	Cur-rent	Catch Rate
DataSet 1	563	80.43%	557	79.57%	78	67.83%	2575	73.89%
DataSet 2	546	78.00%	546	78.00%	75	65.22%	2653	76.13%
DataSet 3	564	80.57%	560	80.00%	75	65.22%	2550	73.17%
DataSet 4	553	79.00%	553	79.00%	78	67.83%	2651	76.07%
DataSet 5	548	78.29%	540	77.14%	78	67.83%	2630	75.47%
DataSet 6	567	81.00%	561	80.14%	79	68.70%	2576	73.92%
DataSet 7	556	79.43%	548	78.29%	77	66.96%	2557	73.37%
DataSet 8	562	80.29%	552	78.86%	79	68.70%	2557	73.37%
DataSet 9	566	80.86%	557	79.57%	83	72.17%	2588	74.26%
DataSet 10	560	80.00%	554	79.14%	80	69.57%	2589	74.29%
DataSet 11	548	78.29%	540	77.14%	79	68.70%	2592	74.38%
DataSet 12	554	79.14%	546	78.00%	79	68.70%	2521	72.34%
DataSet 13	571	81.57%	568	81.14%	83	72.17%	2498	71.68%
DataSet 14	560	80.00%	552	78.86%	77	66.96%	2598	74.55%
DataSet 15	549	78.43%	535	76.43%	77	66.96%	2637	75.67%
DataSet 16	569	81.29%	563	80.43%	75	65.22%	2586	74.20%
DataSet 17	560	80.00%	555	79.29%	75	65.22%	2580	74.03%
DataSet 18	562	80.29%	557	79.57%	80	69.57%	2619	75.15%
DataSet 19	564	80.57%	560	80.00%	84	73.04%	2572	73.80%
DataSet 20	550	78.57%	550	78.57%	81	70.43%	2575	73.89%

5 An Ensemble Analysis

An ensemble consists of a set of individually trained classifiers whose predictions are combined when classifying novel instances. There are two fundamental elements of ensembles: a set of properly trained classifiers and an aggregation mechanism that organizes these classifiers into the output ensemble. Normally, the aggregation process will be an average or a simple majority vote over the output of the ensembles [8]. Previous research has shown that an ensemble can help to increase accuracy and stability [9, 10, 11].

There are two basic criterion based on which the ensemble is chosen: first, voters of an ensemble have to satisfy the across-the-board threshold of 65% for absolute catch rate of the Bankruptcy class and 70% for the Current class as mentioned previously. Second, as the majority vote method is used here, the number of voters in any ensemble must be odd.

From previous k-fold cross-validation, we have computed some optimal solutions of MCLP. Part of these optimal solutions is selected to form an ensemble. Each solution will have one vote for each credit card record and final classification result is determined by the majority votes. The reason for choosing the specific numbers of voters to form ensembles is that we have about 700 optimal solutions available from cross-validation studies. In order to utilize these results, number of voters is determined to be 9, 199, 299, and 399. The test did not go further because we observed that the catch rates remain stable when the number of voters equals to 399. Actually, the number of voters in an ensemble should be determined by the particular context.

The following steps describe the whole process:

Algorithm 2 (Ensemble Analysis)

Step1 A committee of certain odd number (e.g. 9, 99, 199) of classifiers X^* is formed.

Step2 The classification score $MCLP_i = A_i X^*$ against of each observation has been calculated against the boundary b by every member of the committee. The performance measures of the classification will be decided by majorities of the committee. If more than half of the committee members succeed in the classification, then the prediction for this observation is successful, otherwise, the prediction is failed.

Step3 The catch rate for each group will be computed by the percentage of successful classification in all observations.

Several results of Algorithm 2 for different ensemble committee size are summarized in Table 2. The results point out three findings: (1) the classification rates are close to cross-validation process; (2) the number of voters does not affect the classification results significantly; (3) although the catch rates of ensembles do not outperform the best results of cross-validation, they are more steady than cross-validation.

Table 2. Ensemble Results

No. of Voters	Bank-ruptcy	Catch Rate	Cur-rent	Catch Rate	Bank-ruptcy	Catch Rate	Cur-rent	Catch Rate
9	553	79.00%	544	77.71%	79	68.70%	2605	74.75%
99	552	78.86%	542	77.43%	78	67.83%	2595	74.46%
199	553	79.00%	544	77.71%	78	67.83%	2596	74.49%
299	552	78.86%	545	77.86%	78	67.83%	2597	74.52%
399	553	79.00%	544	77.71%	78	67.83%	2600	74.61%

6 Conclusions

Classification of credit cardholders' behavior is an important data mining application in banking industry. According to prior research, MCLP method exhibits promising results in credit cardholders' behavior classification. This paper explored the stability of MCLP method on credit card dataset by using cross-validation and ensemble analyses. The experimental results have shown that there is little effect on the MCLP classification with cross-validation and ensemble techniques. In other words, this indicates that MCLP is a reasonably stable classification method in this specific application. However, the comprehensive understanding on the general impact of using cross-validation and ensemble on MCLP performance needs to be further investigated. We shall report the related findings in the near future.

References

1. Shi, Y., Wise, M., Luo, M. and Lin, Y. (2001), Data mining in credit card portfolio management: a multiple criteria decision making approach, in M. Koksalan and S. Zionts, eds., *Multiple Criteria Decision Making in the New Millennium*, Springer, Berlin, 427-436.
2. Shi, Y., Peng, Y., Xu, W. and Tang, X.: Data Mining via Multiple Criteria Linear Programming: Applications in Credit Card Portfolio Management, *International Journal of Information Technology and Decision Making*. 1 (2002) 131-151.
3. Peng, Y. (2002). Data Mining in Credit Card Portfolio Management: Classifications for Card Holder Behavior. Master Thesis, College of Information Science and Technology, University of Nebraska at Omaha.
4. Kou, G., X. Liu, Y. Peng, Y. Shi, M. Wise and W. Xu, "Multiple Criteria Linear Programming to Data Mining: Models, Algorithm Designs and Software Developments" *Optimization Methods and Software*, Vol. 18, 453-473, 2003.
5. Plutowski, M.E. (1996). "Survey: Cross-Validation in Theory and in Practice." Unpublished manuscript. Available online at: `http://www.emotivate.com/CvSurvey.doc`.
6. From Wikipedia, the free encyclopedia, available online: `http://en2.wikipedia.org/wiki/Cross-validation`.

7. Ricardo Gutierrez-Osuna, "Selected topics in computer science", Texas A&M University. Available online:
http://faculty.cs.tamu.edu/rgutier/courses/cs790_wi02/.
8. Gabriele Zenobi and Pádraig Cunningham, An Approach to Aggregating Ensembles of Lazy Learners That Supports Explanation, *Lecture Notes in Computer Science*, Vol. 2416, p. 436-447, 2002.
9. David Opitz and Richard Maclin, "Popular ensemble methods: an empirical study", *Journal of Artificial Intelligence Research 11* (1999) 169-198.
10. Dietterich, T.G. (2000). Ensemble methods in machine learning. First international workshop on multiple classifier systems. pp.1-15. New York, 2000. Springer Verlag.
11. Jinseog Kim, Ensemble methods for data mining, Probability and data mining lab, Feb 4, 2002. Available online:
http://srccs.snu.ac.kr/VerII/Activity/Tutorial/ensemble.pdf

A Cache Mechanism for Component-Based WebGIS

Yingwei Luo, Xiaolin Wang, and Zhuoqun Xu

Dept. of Computer Science and Technology, Peking University, Beijing, P.R.China, 100871
lyw@pku.edu.cn

Abstract. First, a component-based WebGIS system Geo-Union is introduced, then the architecture and the web application mode of Geo-Union are discussed. After that, a spatial cache framework is designed and adopted in Geo-Union to improve its performance in network environment. There are three levels of spatial cache: spatial database cache, network spatial cache and spatial data proxy server, and they play different roles in different situations to improve spatial data access performance of Geo-Union.

1 Introduction

Component affords a new software construction mode, which is more efficient, agile and powerful than existing technique of object-orientation or traditional modularization. For example, developers could combine and reuse binary modules that independently exploited by different individual or groups, which dramatically simplify and expedite the development. Component is independent of language and hardware, and can run on Web. Moreover, this technique provides much more flexibility for application development [1][2].

Component technology provides a practicable way for develop WebGIS. In fact, there are too many WebGIS products based on component technology in market now, such as MapObject from ESRI [3] and MapX from MapInfo [4]. Assigning functions in reason and improving performance are two key issues for making WebGIS more practicable. In this paper, we analyze the modeling technique of component-based WebGIS, construct a multi-level WebGIS system Geo-Union, and explore its architecture, composition and functional partition of components. Also, a spatial cache framework used to improve performance in Geo-Union is given.

2 Component-Based WebGIS: Geo-union

Geo-Union has a multi-level Client/Server architecture, which is implemented by principle of ORDB and component techniques. Geo-Union provides an object-oriented, extensible GIS component library for further GIS application developers. Geo-Union can be used in both stand-alone environment and network environment.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 940–947, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2.1 The Architecture of Geo-union

Component model is a primary approach to deepen the functions of WebGIS. Geo-Union is divided into four layers: application layer, component layer, service layer and storage layer, where service layer has different units to provide both client services and server services. Figure1 shows the architecture [5][6]. Hierarchical spatial component object model can distribute GIS functions in network reasonably and make the system reusable, as well as provide efficiency approach for further development and integration with other systems.

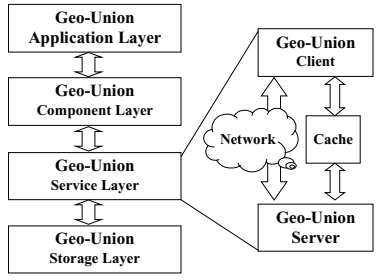


Fig. 1. Geo-Union Architecture

(1) Storage layer is responsible for storage and management of both spatial and non-spatial data based on ORDB. The main problems solved at this layer are how to represent and store spatial data, and how to maintain relationships among spatial data.

(2) Service layer is in charge of spatial data access and process, and includes two parts: Geo-Union client provides data access and process services to component layer, and Geo-Union server provides data access and process services to Geo-Union client through interacting with storage layer. Geo-Union server can manage different spatial data resources, and also can reply to different spatial data requests from different clients. Geo-Union server provides the services of data access, spatial index, basic spatial query, transaction process, data share and so on. Geo-Union client provides different GIS tools and further development functions to component layer based on the services from Geo-Union server. Cache is an important unit of service layer, which is imported to reduce network load and improve response speed of the system. Using Geo-Union client, we can develop a simulation server, which can reduce network load and improve response speed through its cache (see section 3).

(3) Component layer provides a rich set of services (components) to develop domain-oriented GIS applications for further developers. Component layer provides interface of GIS functions to users, but the implementation details are completed in service layer. Component layer exists as a component library, and servers as a bridge between users and service layer. Component layer provides function-explicit and reusable interface components for users according to the functions of service layer.

(4) The work in application layer is to exploit application systems for different special domains by assembling and integrating Geo-Union components. These application systems can be running both in desktop and network environment.

2.2 The Web Application Mode of Geo-union

Actually, Geo-Union component layer is a set of ActiveX controls and correlative programmable objects. ActiveX controls can be embedded into Web pages directly and programmable objects also can be used through ASP, by means of which Web browser and GIS combine with each other. Figure 2 shows the Web application mode.

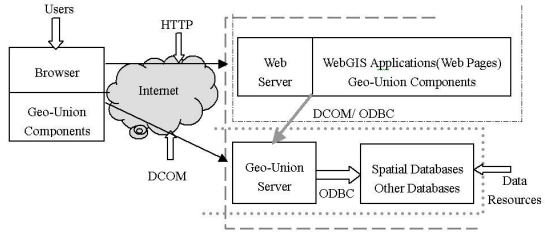


Fig. 2. Web Application of Geo-Union

Users betake WebGIS applications through browser, and browser can attain geographical spatial services by direct mode or indirect mode.

Direct mode means browser use Geo-Union components directly at client side. WebGIS applications are downloaded from Web server and execute at client side with the help of Geo-Union components. Required spatial data is transferred from Geo-Union server to client, and users' requests are processed at client side.

Indirect mode means users' requests are submitted to Web server, and WebGIS applications are execute at Web server side with the help of Geo-Union components to process the requests. Required spatial data is transferred from Geo-Union server to Web server. Finally, a dynamic Web page that reflects the result of the requests is returned to browser.

In both modes, Geo-Union component layer is the indispensable pivot in whole system.

3 Spatial Cache Framework in Geo-union

Cache is an important technique for improving the performance of system. In Geo-Union, there are two aspects affecting the efficiency of data access: one is the access to database, especially when storing spatial data in ORDB; the other is the transmission of spatial data in network. We take different spatial cache modes to solve these two problems in Geo-Union.

3.1 Spatial Cache Framework

Figure 3 shows the spatial cache framework in Geo-Union, which includes three typical cache modes: database cache, network cache and data proxy server.

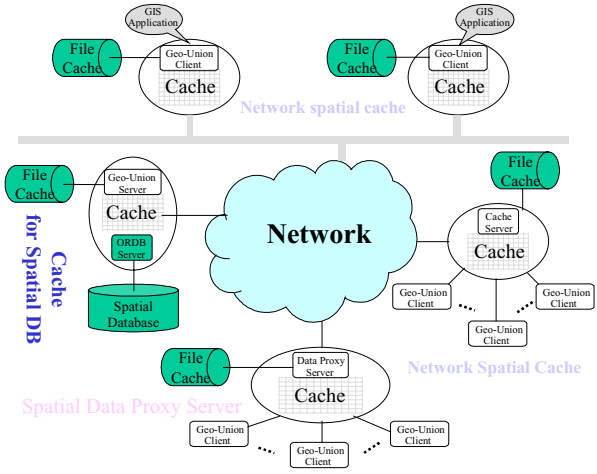


Fig. 3. Spatial Cache Framework

(1) Cache for spatial database

Geo-Union server is the bridge between Geo-Union client and ORDB, so cache for spatial database is established and maintained by Geo-Union server. Cache is stored in both local file and memory. Because Geo-Union server can manage different spatial data resources, and also can reply different spatial data requests from different clients, so cache for spatial database is a global cache.

It is definite that retrieving a group spatial entities from ORDB, especially some neighboring entities in a layer, is much more inefficient than file-structure based GIS. Aiming to this difference, ORDB – memory file cache mechanism is adopted in Geo-Union to speed up spatial entity access. Here memory file is opened in memory and can be random accessed like memory page. If memory in spatial database server is great enough, memory file cache mechanism is more effective than pure file system.

(2) Network spatial cache

The bandwidth of different users in network is different, but public users of Internet always have low bandwidth. So when people use WebGIS applications, it is a lethal delay when spatial data is transferred from Geo-Union server. In Geo-Union client, we adopt a two-level spatial cache mode to relief transmission bottleneck of spatial data in the network:

The first network cache is used to help a single client access remote data: building a spatial cache in local place. This kind of cache is a partial cache, which is also a popular method in today's Web browser.

The second network cache is used to help many clients in a LAN access remote data: building a common spatial cache for a LAN (building a cache server). Once a local client in the LAN accessed some spatial data, other clients can reuse the spatial data in cache server. Cache server is still a partial cache.

Cache server can realize massive spatial data cache by means of the shared resources, which will speed up the hit rate of cache greatly, so as to improve efficiency of all local clients, and save resource of all local clients. Cache server solves the speed

conflict between local disk data access and remote data access, and the speed conflict between the high-speed LAN and WAN with narrow bandwidth.

(3) Spatial data proxy server

Because the distribution of users in Internet is not well proportioned, so different Geo-Union servers are unbalanced. Some Geo-Union server and its communication may overload. Aiming to this problem, we design spatial data proxy server for those Geo-Union servers to improve performance.

Spatial data proxy server is an initiative cache server. Overloaded Geo-Union server selects a suitable Geo-Union client and builds a spatial data proxy server there, cache all or part of spatial data in spatial database, and response a special group of users.

In Geo-Union, spatial data proxy server serves as special server in Internet to provide spatial data access services for public users. The structure and implementation of spatial data proxy server is same as cache server, but they play a different role in Geo-Union.

Cache server is private of a LAN, and clients in the LAN have to get an authorization before accessing to cache server. Spatial data in cache server is changing with different requests of clients in the LAN.

Spatial data proxy server is public to all clients. Spatial data proxy server may serve as a peer of Geo-Union server. Spatial data proxy server can be built anywhere in Internet if needed. If spatial data in a spatial database is unchanged for a long time, spatial data proxy server can cache all spatial data of that spatial database.

Building spatial data proxy servers properly in Internet will make Geo-Union applications more scalable and effective.

3.2 Organization of Spatial Cache

In Geo-Union, spatial cache is organized as three levels: layer, slot and entity.

When creating a layer in spatial database, a GUID (Global Universal Identification) is generated to identify the layer, which is named as *layerID*. When reading a layer into cache, the system will allocate a separate space for the layer according to its *layerID*. Whether a layer is valid in cache is determined by the *layerVersion* of the layer both in cache and in spatial database.

Entities in a layer are always separated into different slots according to a certain rule. When reading a layer into cache, we do not read the whole layer, but read some slots of the layer. Slot brings two benefits: almost all spatial queries do not need a whole layer but only a certain scope in the layer, so when the layer is massive, reading relative slots can satisfy the requirement and will reduce network load greatly; less data will exhaust less computing resource and storage resource. The rules to organize slot are various. We can organize slots in a layer according to a correlativity of geographical location or a neighborhood geographical location. A correlativity of geographical location means we can put entities along a railway into a slot, and a neighborhood geographical location means we will put entities in a certain spatial scope into a slot. Every entity in a layer belongs to a slot. When entities in any slots changed, the *slotVersion* of the layer will change too.

One update operation may modify only one or several entities in a layer, so *layerVersion* and *slotVersion* of the layer cannot reflect the latest modification of entities. We set a *versionNumber* for every entity, and when an entity changes, its *versionNumber* changes too. The *entityVersion* of a layer is the max *versionNumber* in the layer. When the *versionNumber* of a layer in cache is less than that in spatial database, some entities in cache is invalid and those entities that have larger *versionNumber* should be reloaded from spatial database into cache.

3.3 Refresh and Pre-load Spatial Cache

Refreshment of spatial cache can be done online or offline. Online refreshment means updating spatial data in cache at the same time as updating spatial entities in a layer in spatial database. Offline refreshment means updating invalid spatial data of a layer in cache if their versions are invalid when accessing to them.

Information in spatial cache includes *layerVersion*, *slotVersion*, *entityVersion* and the corresponding relations between all entities and slots of a layer. When accessing to entities in a layer, we can determine whether spatial data in cache is valid or not through comparing version information in cache and spatial database.

(1) If *layerVersion* of a layer in cache is smaller than that in spatial database, it means that all entities of the layer are changed, and the whole layer should be refreshed in cache. Otherwise,

(2) If *slotVersion* of a layer in cache is smaller than that in spatial database, it means that all slots of the layer are rearranged, and the corresponding relations between all entities and slots of the layer should be refreshed in cache. Otherwise,

(3) If *entityVersion* of a layer in cache is smaller than that in spatial database, it means that some entities of the layer are changed, and those entities should be refreshed in cache.

(4) If above three conditions are all equal to those in spatial database, it means spatial data of a layer in cache is same as that in spatial database, and no refreshment is required.

When accessing to spatial data, if we can pre-load some spatial data into cache, it will make spatial cache more effective. But how to predict what kind of spatial data will be accessed to?

There are two rules of accessing to memory: there is every probability of accessing to just being accessed memory, and there is every probability of accessing to neighbors of just being accessed memory. There are same rules in accessing to spatial data: there is every probability of accessing to just being accessed spatial data, and there is every probability of accessing to neighbors of just being accessed spatial entities. According to those rules, we can arrange slots by spatial scope, and a slot is an accessing unit of spatial data. When the network is idle, we can pre-load some neighbors of spatial data in cache from spatial database.

3.4 Spatial Cache Object

In Geo-Union, different spatial caches are all implemented by the six spatial cache objects: CGuCacheMgr, CGuCache, CGuSlots, CGuSlot, CGuEntries and CGuOpList. Figure 4 illustrates the hierarchies of them.

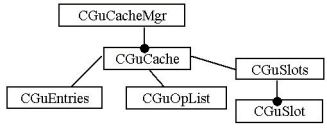


Fig. 4. Hierarchies of Spatial Cache Objects

In every spatial cache, there is only one CGuCacheMgr object, which manages all CGuCache objects. A layer in cache is represented by a CGuCache object. When opening a layer, CGuCacheMgr object will create a CGuCache object or find an existing CGuCache object through its *layerID*.

CGuEntries object is used to manage a mapping list between an *entityID* and entity data of a layer. From the mapping list, we can know the *slotID* where an entity belongs to (if $slotID = 0$, it means an entity is not existing; if $slotID > 0$, it means an entity is in cache; and if $slotID < 0$, it means an entity is not in cache), and the address pointer of the entity data if the entity is in cache.

CGuOpList object is used to record updating operation list of a layer at client side that still doesn't be committed to spatial database. When the updating operation list is committed, it will also update the corresponding spatial data in cache. But when the updating operation list is cancelled, no change will happen in cache.

4 Conclusion

The development trend of WebGIS is the management of massive distributed spatial data and the construction of distributed component-based WebGIS. Component-based WebGIS can be integrated into other development environments seamlessly, therefore can not only decrease the complexity of applications and speed up development progress, but also descend the cost and increase the maintainability. Component-based system makes it more convenient to construct domain-oriented application systems in a distributed environment.

Geo-Union has finished a preliminary component-based model for distributed WebGIS, and has got into use in many fields with sound effects. Although spatial cache and other techniques are adopted in Geo-Union to improve its performance, a lot of works still wait us ahead to make Geo-Union more practicable and effective:

(1) Dynamic load balancing policy. In Geo-Union, most works are completed at client side, and server is only responsible for data access and simple data query. Therefore it is not well balanced between client and server. Especially there will exists a lot of transmission for massive data between client and server. Although, spatial index and spatial cache techniques can improve the performance to a certain extent,

we still want to take full advantage of the computing capability of GIS server, so as to lighten load at client side and decrease the transmission of redundant data in network. This needs a more reasonable component design for the system.

(2) System security. People should pay more and more attention to system security in network when they enjoy shared resources. The security solution in Geo-Union is only based on database system and simple user password mechanism. It is a challenge to provide an effective security mechanism in component system to make users access spatial information conveniently and safely. In addition, some other related problems should be solved, such as controlling different access levels in distributed environment, encrypting important information during transmission, etc.

(3) System concurrency. WebGIS is open to millions of users. How to ensure the correctness, validity, stability and scalability of Geo-Union to satisfy users' requests is another key problem for practicable WebGIS.

(4) System intelligence. Agent technique is a development trend and provides a new model for in distributed software construction. Of course, agent technique will bring a new thought for distributed GIS [8]. How to apply agent technique into distributed GIS, and provide interoperable services, collaborative services and intelligent services with the help of spatial metadata, are most important and significant researches in distributed GIS.

Acknowledgement. This work is supported by the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20020001015; the National Grand Fundamental Research 973 Program of China under Grant No.2002CB312000; the National Science Foundation of China under Grant No.60073016 and No.60203002; the National High Technology Development 863 Program under Grant No. 2002AA135330, No. 2002AA134030 and No. 2001AA113151; the Beijing Science Foundation under Grant No.4012007.

References

1. Li Bin: A Component Perspective on Geographic Information Services, Cartography and Geographic Information Science, 27(1): 75-86(2001).
2. Szyperski C: Component software: beyond object-oriented programming Reading, MA: Addison-Wesley Press, 1998.
3. http://www.esri.com.
4. http://www.mapinfo.com.
5. Dept. of Computer Science and Technology: Peking University, Operation Guide for Geo-Union Enterprise (in Chinese), Technology Material, http://gis.pku.edu.cn.
6. Dept. of Computer Science and Technology: Peking University, Component Guide for Geo-Union Enterprise (in Chinese), Technology Material, http://gis.pku.edu.cn.
7. Li Muhua: Study on Component based WebGIS and Its implementation [Master Dissertation] (in Chinese), Beijing: Peking University, 2000.6.
8. http://map.sdsu.edu/geoagent/.

A Data Structure for Efficient Transmission of Generalised Vector Maps*

Min Zhou and Michela Bertolotto

Department of Computer Science, University College Dublin
Belfield, Dublin 4, Ireland
{min.zhou, michela.bertolotto}@ucd.ie

Abstract. Progressive transmission of a sequence of representations at increasing detail has been identified as a possible solution for the exchange of very large vector datasets across the Internet. In this context an important issue relates to the development of data structures to efficiently store and manipulate multiple representations. In this paper we describe a new data structure used to encode representations obtained by applying Saalfeld's modified RDP algorithm [10]. The data structure includes vertical links between different representations and therefore imposes a hierarchical organisation on the multiresolution sequence.

Keywords: Multiple representations, Progressive vector transmission, Line simplification

1 Introduction

Vector data are stored mathematically as sets of points, lines, and polygons (regions) in Geographic Information Systems (GIS). Vector data associated with location and other attribute information have been widely applied in the fields of digital mapping and location services. Spatial information sharing has become increasingly important with the new developments in communication technologies. However, since vector datasets are typically very large, users wishing to download these datasets from server to client have to suffer a long-time waiting even through faster communication links. Progressive and incremental transmission is proposed in [3-5], where a coarsest version is transmitted and displayed, then progressively substituted by subsequent finer versions until users' needs are satisfied. It ensures that a user receives no more data than desired. This approach relies on the pre-computation of a sequence of representations on the server.

In this paper, we describe a new data structure used to store different representations on the server and to support progressive transmission. The different representations are generalised by using a classical line simplification algorithm (RDP) [7, 9] improved by Saalfeld to guarantee topological consistency [10].

* The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 948–955, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The main objectives we considered in the development of such a data structure are to:

(a) Include all vector objects in the form of points, lines and polygons
(b) Store the entire map (entities and their topological relations) only in the coarsest layer, while all subsequent layers just record newly introduced entities and increments of entities from previous layers
(c) Support hierarchical navigation, browsing and querying capabilities by maintaining vertical links between different representations of the same entities
(d) Be extendible to allow for encoding not only of geometric changes but also of topological changes

The remainder of this paper is organised as follows: In section 2 related work on progressive vector transmission is examined; section 3 describes in detail the proposed data structure for storing multiple representations and vertical links; in section 4 the implementation of the data structure used for progressive transmission in a client server architecture is described; section 5 presents some conclusions and outlines future work.

2 Related Work

Very few models have been developed for progressive vector transmission of geospatial data. So far the only implementations apply to triangular meshes used for digital terrain modelling [4]. For more generic data, two main models have been proposed in the literature. Buttenfield [5] proposes that the pre-computation of multiple representations resulted by line simplification can be performed on a server where each line is iteratively subdivided using Saalfeld's modified RDP algorithm and stored in a hierarchical strip tree. These multiple representations are transmitted progressively to clients. In the implementation, Buttenfield modified Ballard's strip tree data structure [1] to preserve the topological consistency by storing convex hulls of the subdivided lines instead of storing Minimum Bounding Rectangles of the subdivided lines. This data structure is a hierarchical tree structure, which facilitates to reconstruct linear features at a given level of detail and supports efficient progressive transmission. However, as this data structure stores the sub-polylines only, it lacks effective spatial indexing as the complete geometry of the polyline is not maintained. Spatial indexing is essential for optimizing the performance of spatial queries in a large datasets. Drawbacks of this model are: (a) only one line at a time can be simplified; (b) vertical links for efficient navigation between multiple representations are not supported; and (c) topological changes are not explicitly taken into account.

A model proposed by Bertolotto and Egenhofer [3, 4] structures a sequence of map representations at different levels of detail by applying a set of topological operators: line contraction, region contraction, region thinning, line merge, region merge, point abstraction and line abstraction [2]. The model represents a framework for progressive vector transmission in a distributed client-server architecture. This model is applicable not only to linear data but to all vector objects (i.e. points, lines and polygons). Vertical links are added to enhance the sequence of multiple

representations in a hierarchical structure. However, the model performs only topological changes on a vector map and vertical links are represented by topological operators only. Clearly, in real applications, changes in the geometry of objects also occur while generating multiple representations. For example, line simplification is a basic generalisation operation used in cartography.

In this paper we consider multiple representations generalised by line simplification operations. In order to efficiently store and transmit such representations, we have defined an appropriate data structure, which is able to manipulate multiple representations at different levels of detail organised in a hierarchical structure.

Several spatial data structures that provide some limited facilities for multiple levels of detail were described in van Oosterom's review [11]. However, none of the data structures presented in the review combines the geometric capabilities with multiple levels of detail. Furthermore they were not developed with the purpose of supporting progressive data transmission over the Internet.

3 The Proposed Data Structure

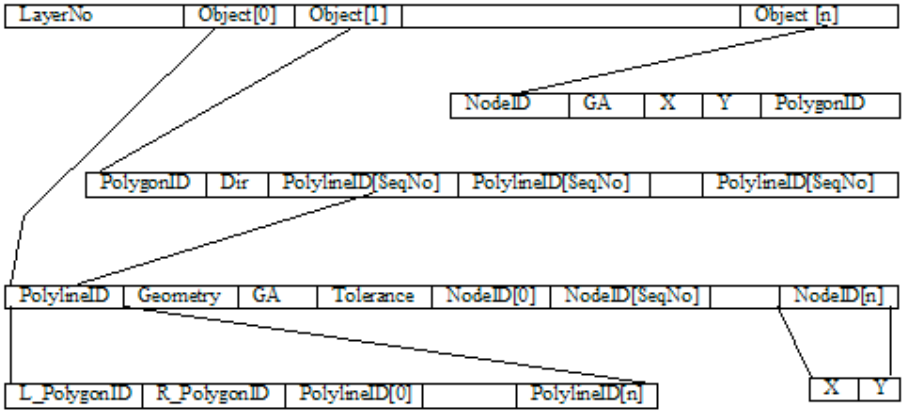
A new data structure, EXDCEL, is developed to implement a mechanism for dynamically interpreting vertical links and compose the maps transmitted incrementally. The increments downloaded progressively are then merged into a complete map.

Our data structure represents a hierarchical extension of the Doubly-Connected Edge List data structure (DCEL), a topological data structure for plane vector maps [8]. The version of DCEL used here is an extension of classical DCEL including information about isolated features entities [6]. Such a data structure stores:

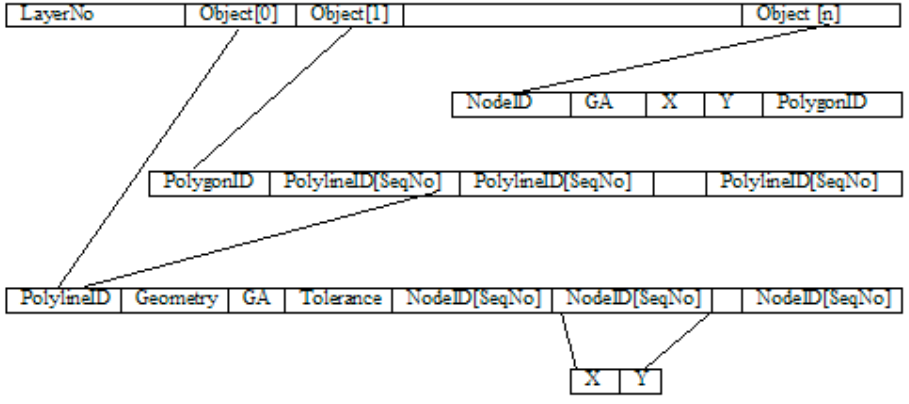
- For each point: its coordinates, a reference to an edge incident to it, and the region containing it if it is a point feature.
- For each line: its geometry, its endpoints, reference to the left and to the right polygon, and 4 lines sharing an endpoint with it.
- For each region: a list of references to edges forming the outer boundary and possible inner boundaries, a list of lines composing linear features, and a list of point features.

The following information is added to the above structure to form the EXDCEL data structure:

- An identifier is attached to each entity. The identifier for a layer is added. The points of a polyline and the polylines of a polygon are ordered by sequence number. Vertical links that consist of these identifiers and sequence number, enhance a sequence of multiple representations with a hierarchical structure.
- A tolerance value is a pre-determined threshold distance, which is used by RDP algorithm to control data reduction. It is added as an attribute of an entity. Multiple tolerance values are applied for sectors of a polyline or polylines to produce a nearly constant number of objects stored on layers.
- Graphical and semantic attributes can be associated with entities. These are used to classify an entity as a geographical feature such as a road, railroad, river, lake etc.



(A) EXDCEL data structure used for storing spatial entities and their spatial relations at the coarsest layer



(B) EXDCEL data structure in the case of storing increments and vertical links only at the subsequent layers.

Where
LayerNo: Layer Identifier. It points to the layer where a spatial object is allocated
Object: an array of objects on a layer. Object can be a point feature, polyline, polygon and line features of polygon.
NodeID: point feature Identifier
PolylineID: Polyline Identifier
PolygonID: Polygon Identifier
GA: Graphic Attributes such as line type, width, colour, etc. to clarify entities as roads , rivers
Tolerance: a parameter used by RDP algorithm to control the data reduction and is applicable to polyline and polygon
SeqNo: the sequence number of a node in its original polyline or the sequence number of a polyline in its original polygon
L_PolygonID: PolygonID on the left side of polyline
R_PolygonID: PolygonID on the right side of polyline
Dir: Direction such as 1 represents an outer polygon and 0 represents an inner polygon
X: X-coordinate of a point
Y: Y-coordinate of a point

Fig. 1. Examples of EXDCEL data structure

In order to avoid redundancies, we follow an approach similar to the one described in [3, 4] for storing a sequence of multiple representations. Only the coarsest layer is stored completely with endpoints of polylines and their topological relations. The subsequent layers store increments (new points of polylines) and vertical links only. Note that spatial relations need not be stored on the subsequent layers (i.e. only the coarsest layer is encoded by means of a DCEL data structure). Only endpoints of a polyline are necessary to completely reconstruct the topology of the polyline itself. Topological relations of polylines are not changed during refinement of lines by merging new points between the endpoints. Fig. 1 (A) shows the data structure used to encode the coarsest layer (entities and their spatial relations), while Fig. 1. (B) shows the data structure in the case of encoding the increments and vertical links only to transmit efficiently.

4 Implementation Details

We have implemented our progressive transmission system as a distributed 4-tier client-server architecture including a client, a map server, an application server and a web server, using the Java programming language. As Java is platform independent, our application can be easily run on any platform without any changes or recompilation required. Different clients are connected to the web server via Internet simultaneously. All communications between a client and the map server are conducted through the web server and the application server.

4.1 Map Server

The map server supports multiple representations of spatial datasets and relies on an Oracle9i Spatial Database Management System. Multiple representations at different levels of detail are pre-computed on the server side and stored in a hierarchical structure. Client applications provide a user interface for transmission requests and data display.

All information is stored in the database based on a topological data structure using relational tables, from which geometric entities are extracted as object views. Most of the spatial functionality offered by Oracle Spatial such as spatial indexing and spatial querying can be exploited using these geometric objects.

Even though the RDP simplification algorithm has become a default simplification routine in many GIS packages, the algorithm treats individual polylines as isolated features. Therefore it can generate self-intersections, intersections between polylines and other topological inconsistencies. Saalfeld indicated conflicts could only occur with vertices of other polylines that lie within the closed convex hull of the original polyline. He proposed a modified RDP algorithm by adding extra checks for the external points in a dynamic convex hull data structure, which can ensure that topological consistency is preserved [10]. In our system, the modified RDP algorithm was implemented in Java to simplify different polylines simultaneously. Multiple representations generalised by the algorithm using different tolerance values preserve

topological consistency. This resolves the critical issue of manipulating multiple representations.

Tolerance is used to obtain a roughly constant number of objects stored on layers. Better output is achieved by using different tolerances in different sections of a polyline or for different polylines. There is no reliable procedure for automatically selecting tolerance values. In our implementation, tolerance-setting can be changed on a case-by-case basis. In doing so, the number of layers and a roughly constant number of objects stored on layers can be optimised by predetermined tolerance values on the server side.

For example, we consider the polyline in Fig. 2 consisting of 28 points. The coarsest level (Layer 0) stores at least two end points of the polyline (point 1 and point 28) and is created by applying the modified RDP algorithm with the largest predetermined tolerance value. The next layer (e.g. Layer 1) is created by storing new points, which are selected by applying the algorithm with a smaller tolerance value. All new points, which lie in between two points (point 1 and 28) of Layer 0, with regard to their original sequence number, are stored. This process is repeated until the tolerance value equals 0. Fig. 2 shows multiple representations of a polyline by performing the modified RDP algorithm with different tolerance values. In such a way, all points of the polyline are grouped into different layers.

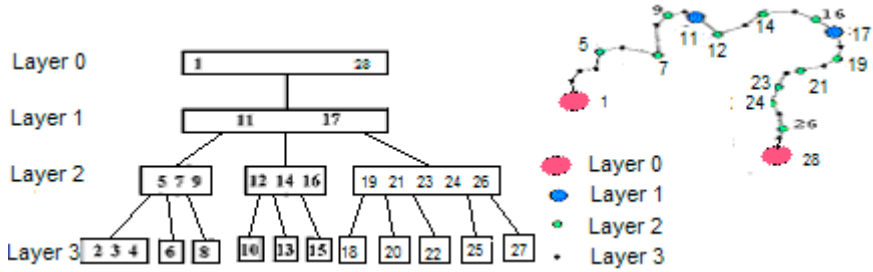


Fig. 2. Points allocated in every layer on server side

4.2 Vertical Links

Every point identified by its NodeID is ordered by its original SeqNo in a polyline allocated into a layer on the server. A pointer is made from the root to the new points consisting of LayerNo, PolylineID (or PolygonID), NodeID and SeqNo. This allows us to know which points are added to the subsequent layers and which ones are preserved. Fig. 3 shows the link from layer 0 to layer 1 consists of LayerNo (1), PolylineID (177267165), NodeID (new point: 32772) and sequenceNo (new point 17). Furthermore, different representations of a polyline at different levels of detail can be linked in a hierarchical structure i.e. layers::polygons::polylines::nodes. Vertical links allow for hierarchical navigation and browsing across levels and facilitate to reconstruct a map at a given level of detail (for more detail about this see section 4.3).

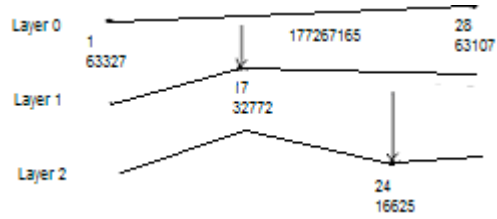


Fig. 3. Vertical links joining different representations of a polyline in a multiple map representation sequence

4.3 EXDCEL on Client Side

On the server, only the coarsest layer is completely stored; increments and vertical links are stored on the subsequent layers. On the client, the dataset corresponding to an intermediate representation must be reconstructed.

Vertical links are transmitted with increments and used to reconstruct the complete topological data structure of intermediate layers. Multiple representations progressively downloaded from server are organised as a hierarchical structure (see fig. 2 as a reference): a polyline with the coarsest level of detail can be created initially and a finer polyline will be created by only merging new points according to its sequence number. A complete representation corresponding to an intermediate level of detail can be obtained by merging the coarsest layer with increments that are progressive transmitted by following vertical links in a hierarchy until the users requests are satisfied.

Reconstructing the topology on the client makes data management locally possible: locally available entities and their spatial relations can be manipulated without having to reconnect to the server. This saves communication across networks and users waiting time.

5 Conclusions

In this paper, we describe the development and implementation of a new data structure for progressive transmission of a sequence of multiple representations generalised by applying a line simplification algorithm. Such a data structure supports hierarchical navigation, topological consistency of multiple representations and spatial indexing.

Our data structure is a hierarchical extension of a topological data structure for plane maps and therefore can be used to store not only linear data but generic vector datasets. Furthermore it can be extended to include the vertical links corresponding to topological changes such as the ones performed by operators described in [2]. We are working towards a complete integration of topological changes and geometric changes within the same data structure. This is essential in real map generalisation applications.

We are also currently testing our system to evaluate the efficiency of the transmission process.

References

1. Ballard. D. 1981. Strip Trees: A Hierarchical Representation for Curves. Communication of the Association for Computing Machinery, vol. 14: 310-321
2. Bertolotto, M.: Geometric Modelling of Spatial Entities at Multiple Levels of Resolution. Ph.D. Thesis, Department of Computer and Information Sciences, University of Genova, Italy (1998).
3. Bertolotto, M., Egenhofer, M.: Progressive Vector Transmission. Proceedings, 7th International Symposium on Advances in Geographic Information Systems, Kansas City, MO: (1999) 152-157.
4. Bertolotto, M., Egenhofer, M.: Progressive Transmission of Vector Map Data over the World Wide Web, GeoInformatica - An International Journal on Advances of Computer Science for Geographic Information Systems, Vol. 5 (4), Kluwer Academic Publishers (2001) 345-373.
5. Buttenfield, B.P.: Transmitting Vector Geospatial Data across the Internet, Proceedings GIScience 2002, Lecture Notes in Computer Science, Vol. 2478. Springer-Verlag, Berlin (2002) 51-64.
6. De Floriani, L., Marzano, P., Puppo, E.: Spatial quieries and data models, in Spatial Information Theory – A theoretical basis for GIS, A.U. Frank, I. Campari (eds.), Lecture Notes in Computer Science 716, Springer-Verlag (1993) 113-138.
7. Douglas, D.H., Peucker, T.K.: Algorithms for the reduction of the number of points required representing a digitised line or its character. The Canadian Cartographer, Vol. 10 (2), (1973) 112-123.
8. Preparata, F.P., Shamos, M.I.: Computational Geometry: an Introduction, Springer-Verlag (1985).
9. Ramer, U.: An Iterative procedure for the polygonal approximation of plane curves, Computer Vision Graphic and Image Processing, Vol. 1, (1972) 244-256.
10. Saalfeld, A.: Topologically consistent line simplification with the Douglas-Peucker algorithm. Cartography and GIS, Vol. 26 (1) (1999).
11. van Oosterom, P:. Reactive Data Structures for Geographic Information Systems. PhD-thesis Department of Computer Science, Leiden University, (1990).

Feasibility Study of Geo-spatial Analysis Using Grid Computing

Yincui Hu[1], Yong Xue[1,2*], Jianqin Wang[1], Xiaosong Sun[1], Guoyin Cai[1], Jiakui Tang[1], Ying Luo[1], Shaobo Zhong[1], Yanguang Wang[1], and Aijun Zhang[1]

[1]Laboratory of Remote Sensing Information Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, P. Box 9718, Beijing 100101, China
[2]Department of Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
{huyincui@163.com, y.xue@londonmet.ac.uk}

Abstract. Spatial applications will gain high complexity as the volume of spatial data increases rapidly. A suitable data processing and computing infrastructure for spatial applications needs to be established. Over the past decade, grid has become a powerful computing environment for data intensive and computing intensive applications. In this paper, we tested and analyzed the feasibility of using Grid platform for spatial analysis functionalities in Geographic Information System (GIS). We found that spatial interpolation, buffers, and spatial query can be easily migrated to Grid platform. Polygon overlay and transformation could achieve better results on Grid platform. To do network analysis and spatial statistical analysis on Grid platform could be no significant improvement of performance. The most un-suitable spatial analysis on Grid platform is the spatial measurement.

1 Introduction

In numerous scientific disciplines, terabyte and petabyte-scale data collections are emerging as critical community resources. A new class of "data grid" infrastructure is required to support management, transport, distributed access to, and analysis of these datasets by potentially thousands of users. Researchers who face this challenge include the climate modeling community, which performs long-duration computations accompanied by frequent output of very large files that must be further analyzed. The number of applications that require parallel and high-performance computing techniques has diminished in recent years due to the continuing increase in power of PC, workstation and mono-processor systems. However, Geographic information systems (GIS) still provide a resource-hungry application domain that can make good use of parallel techniques. GIS applications are often run operationally as part of decision support systems with both a human interactive component as well as large scale batch or server-based components. Parallel computing technology embedded in a distributed system therefore provides an ideal and practical solution for multi-site organisations and especially government agencies who need to extract the best value from bulk geographic data.

* Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 956–963, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Spatial applications will gain high complexity as the volume of spatial data increases rapidly. A suitable data processing and computing infrastructure for spatial applications needs to be established. Over the past decade, grid has become a powerful computing environment for data intensive and computing intensive applications.

Parallel and Distributed Knowledge Discovery (PDKD) is emerging as a possible killer application for clusters and grids of computers. The need to process large volumes of data and the availability of parallel data mining algorithms, makes it possible to exploit the increasing computational power of clusters at low costs. On the other side, grid computing is an emerging "standard" to develop and deploy distributed, high performance applications over geographic networks, in different domains, and in particular for data intensive applications. Cannataro (2000) proposed an approach to integrate cluster of computers within a grid infrastructure to use them, enriched by specific data mining services, as the deployment platform for high performance distributed data mining and knowledge discovery.

Integrating grid computing with spatial data processing technology, Pouchard *et al.* (2003) described the increasing role of ontologies in the, context of Grid Computing for obtaining, comparing and analyzing data. They presented ontology entities and a declarative model that provide the outline for an ontology of scientific information. Relationships between concepts were also given. The implementation of some concepts described in this ontology was discussed within the context of the Earth System Grid II (ESG).

In this paper, we tested and analyzed the feasibility of using Grid platform for spatial analysis functionalities in Geographic Information System (GIS). First, we listed the several basic spatial analysis functions used in GIS systems. Following the definitions of criteria and basic principles for spatial analysis middleware development for Grid platform, we analyzed the feasibilities of above basic spatial analysis functions for Grid platform and give the suggestions on how to develop the middleware for spatial analysis with Grid platform.

2 Spatial Analysis Functionalities Commonly Used in GIS

Spatial Analysis is a set of techniques whose results are dependent on the locations of the objects being analyzed and requiring access both to the locations of objects and also to their attributes (Goodchild 2001). GIS is designed to support a range of different kinds of analysis of geographic information: techniques to examine and explore data from a geographic perspective, to develop and test models, and to present data in ways that lead to greater insight and understanding. All of these techniques fall under the general umbrella of "spatial analysis". In general, it includes Query, Analyses which are simple in nature but difficult to execute manually, such as overlay (topological), map measurement, particularly area, and buffer zone generation, Browsing/plotting independently of map boundaries and with zoom/scale-change such as seamless database, need for automatic generalization and editing, and Complex modeling/analysis (based on the above and extensions). We will focus on the following spatial analysis functionalities: Query and reasoning: the identification

of objects and attributes either by their location or attribute query; Measurement: simple geometric measurements associated with objects; Transformation; Buffers; Spatial overlay; Spatial Interpolation; Network Analysis and Statistical Analysis.

3 Evaluation Criteria and Basic Principles

Two issues will be considered in use of Grid technology. One is the efficiency or high performance, i.e. to get the solution in a very short time period or to solve more complex problem in a same time period. Another is the high throughput computing to reduce the cost, etc. We will only evaluate the performance of spatial analysis algorithms on Grid platform and neglect the limitations of hardware and data storage.

To improve the efficiency of application algorithms is to enhance the parallel processing. Normally, there are two ways to deal with it. One is to parallel the processing algorithms and second is to parallel the data. We use 5 levels of criteria factors: I – worse, II – poor, III – good, IV – better, V – Best. For the parallel of algorithms, how many sub-jobs can a job be divided into? Table 1 shows the explanation of all five levels. For the parallel of data, how many sub-areas can the whole area be divided into? Table 2 explains the meaning of al five levels.

Table 1. The meanings of all five criteria levels of algorithm parallel.

Criteria Level	I	II	III	IV	V
Digital Score (A)	1	3	5	7	9
Description	There is few parallel.	Can be paralleled, but half of sub-jobs are inter-associated.	Can be paralleled, but between 1/2 and 1/3 of sub-jobs are inter-associated	Can be paralleled, less than 1/3 of sub-jobs are inter-associated	Can be paralleled, no sub-jobs are inter-associated

Table 2. The explanation of all five criteria levels for data parallel.

Criteria Level	I	II	III	IV	V
Digital Score (D)	1	3	5	7	9
Description	Data cannot be divided for processing.	Data can be divided with more than half of redundancy.	Data can be divided with half of redundancy.	Data can be divided with less than half of redundancy.	Data can be divided without redundancy.

It is easier to realize the data parallel processing. Because of the higher efficiency for data parallel processing than that for algorithm parallel processing, we define the different weight factors for data and algorithm parallel processing. The overall evaluation criteria score (E) will be

$$E = 0.6 * D + 0.4 * A \quad (1)$$

Where A is the score from Table 1 for algorithm parallel and D is the score from Table 2 for data parallel. The final overall evaluation criteria will be in five levels as shown in Table 3.

Table 3. The overall criteria levels for both data and algorithm parallel

Criteria Level	I	II	III	IV	V
Score (E)	$1 < E \leq 1.8$	$1.8 < E \leq 3.6$	$3.6 < E \leq 5.4$	$5.4 < E \leq 7.2$	$7.2 < E \leq 9$
Description	Worse	Poor	Good	Better	Best

4 Performance of Spatial Analysis on Grid Platform

4.1 Spatial Query

Database query is probably one of the most important and most commonly used application in Geographic Information Systems. Like any database, a GIS allows you to access information held in a data file in a variety of ways. Information can be grouped in categories, sorted, analyzed, printed, etc. The difference, once again, is that GIS deals with *spatially oriented data*. This means that when querying a database you cannot only see the data but its geographic location as well.

Database query simply asks to see already stored information. Basically there are two types of query most general GIS allow: viz., Query by attribute and Query by geometry. Map features can be retrieved on the basis of attributes. The attribute database, in general, is stored in a table (relational database mode.) with a unique code linked to the geometric data. This database can be searched with specific characteristics. However, more complex queries can be made with the help of SQL. GIS can carry out a number of geometric queries. The simplest application, for example, is to show the attributes of displayed objects by identifying them with a graphical cursor. There are five forms of primitive geometric query: viz., Query by point, Query by rectangle, Query by circle, Query by line, and Query by polygon. A more complex query still is one that uses both geometric and attributes search criteria together. Many GIS force the separation of the two different types of query. However, some GIS, using databases to store both geometric and attribute data, allow true hybrid spatial queries.

For spatial query, the database can be divided in to several smaller databases and searched in parallel. The score for criteria level of algorithm parallel is 1 and data parallel is 9.

4.2 Spatial Measurements

GIS makes spatial measurements easy to perform. Spatial measurements can be the distance between two points, the area of a polygon or the length of a line or boundary.

Calculations can be of a simple nature, such as measuring areas on one map, or more complex, such as measuring overlapping areas on two or more maps. Distance measurement refers to measuring straight-line distances between points or between points and their nearest points or lines.

As spatial measurements are always simple and the delay of data communication is much longer that processing, there is no need to use Grid computing technology. The overall criteria score is 1.

4.3 Transformation

The point transformation includes the algebra functions such as addition, subtraction, multiply and division and logical functions such as AND, OR, NOT, NOR, etc; Comparison functions such as GREAT, LESS etc; and other mathematical functions. Those are simple operations and the score for algorithm parallel is 1 and for data parallel is 9.

DEM analysis is one of the most popular transformation operations. Roros and Armstrong (1998) showed that three steps are needed: transect classification, cell classification, and feature topology construction. The score for algorithm parallel is 1, but the score for data parallel is 3 – 7.

4.4 Buffer Analysis

Buffer analysis is used for identifying areas surrounding geographic features. The process involves generating a buffer around existing geographic features and then identifying or selecting features based on whether they fall inside or outside the boundary of the buffer. This kind of analysis is also called proximity analysis. The buffer operation will generate polygon feature types irrespective of geographic features and delineates spatial proximity. The data parallel is scored 9 and algorithm parallel is scored 1.

4.5 Spatial Overlay

One basic way to create or identify spatial relationships is through the process of spatial overlay. Spatial overlay is accomplished by joining and viewing together separate data sets that share all or part of the same area. The result of this combination is a new data set that identifies the spatial relationships. This allows the user to view and analyze those portions of the various layers which cover the same place on the Earth.

Spatial overlay could be done in parallel. The data parallel has a score of 9 of criteria level and the algorithm parallel could be from 1 to 9 depending on the algebra of map overlay.

4.6 Network Analysis

Network analysis is used for identifying the most efficient routes or paths for allocation of services. This involves finding the shortest or least-cost manner in which to visit a location or a set of locations in a network. The "cost" in a network analysis is frequently distance or travel time. Network analysis can also be used to optimize the allocation of resources.

For the case of shortest distance analysis, the algorithm has to be serial and the criteria level's score is 1. Data could be divided into several parts. But the size of each part could be too small as the coordinating between each part takes much long time (Lanthier and Nussbaurm 2003). The data parallel scores 7.

4.7 Spatial Interpolation

A GIS can be used to estimate the characteristics of terrain or ecological conditions from a limited number of field measurements. Spatial interpolation is the procedure of estimating the value of properties at unsampled sites within the area covered by existing observations and in almost all cases the property must be interval or ratio scaled. Spatial interpolation can be thought of as the reverse of the process used to select the few points from a DEM which accurately represent the surface. Rationale behind spatial interpolation is the observation that points close together in space are more likely to have similar values than points far apart (Tobler's Law of Geography). Spatial interpolation is a very important feature of many GISs. Spatial interpolation may be used in GISs:

- to provide contours for displaying data graphically
- to calculate some property of the surface at a given point
- to change the unit of comparison when using different data structures in different layers
- frequently is used as an aid in the spatial decision making process both in physical and human geography and in related disciplines such as mineral prospecting and hydrocarbon exploration

Many of the techniques of spatial interpolation are two- dimensional developments of the one-dimensional methods originally developed for time series analysis. There are several different ways to classify spatial interpolation procedures: Point Interpolation/Areal Interpolation, Global/Local Interpolators, Exact/Approximate Interpolators, Stochastic/Deterministic Interpolators and Gradual/Abrupt Interpolators (Armstrong and Marciano, 1997, Wang and Armstrong 2003). In general, the data parallel criteria level is 9 and algorithm parallel is 5 - 7.

4.8 Spatial Statistical Analysis

All data have a more-or-less precise spatial and temporal label associated with them. Data that are close together in space (and time) are often more alike than those that are far apart. A spatial statistical model incorporates this spatial variation into the stochastic generating mechanism. Temporal information allows this mechanism to be

dynamic. Prediction of unobserveds from observeds and estimation of unknown model parameters are the principal forms of statistical inference. The search for well defined statistical criteria and a quantification of the variability inherent in the (optimal) predictor or estimator are instrinsic to a statistical approach.

It is almost always true that the classical, non-spatial model is a special case of a spatial model, and so the spatial model is more general (spatial-temporal models are even more general). Whether one chooses to model the spatial variation through the non-stochastic mean structure (sometimes called large-scale variation) or the stochastic-dependence structure (sometimes called small-scale variation) depends on the underlying scientific problem, and can be simply a trade-off between model fit and parsimony of the model description.

There are two different categories of spatial statistical analysis: spatial self-correlation and spatial self-regression analysis (Li 1996, Roros and Armstrong 1996). The score for data parallel is 5 and for algorithm is 3.

The summaries of above discussing are illustrated in Table 4. We found that spatial interpolation, buffers, and spatial query can be easily migrated to Grid platform. Polygon overlay and transformation could achieve better results on Grid platform. To do network analysis and spatial statistical analysis on Grid platform could be no significant improvement of performance. The most un-suitable spatial analysis on Grid platform is the spatial measurement.

Table 4. Feasibilities of spatial analysis on Grid platform

Spatial Analysis Functionalities	Algorithm Parallelity	Data Parallelity	Overall Score	Evaluation Level
Spatial Query	1	9	5.8	Better
Spatial Measurement	1	1	1	Worse
Transformation	1	3-9	2.2-5.8	Poor - Better
Buffers	1	9	5.8	Better
Overlay Analysis	1-9	7-9	4.6-9	Good – Best
Network Analysis	1	7	4.6	Good
Spatial Interpolation	5-7	9	7.4-8.2	Best
Spatial statistical Analysis	5	3	3.8	Good

5 Conclusions

Grid computing has emerged as an important new field in the distributed computing arena. It focuses on intensive resource sharing, innovative applications, and, in some cases, high-performance orientation. Grid technology is very effective method for spatial data analysis. It can give strong computing power in grid environment.

As our work was limited on the evaluation of performance of spatial analysis itself, in reality, many other factors such as hardware environment, distribution of data, etc. have to be considered. We are carrying on the research.

Acknowledgement. This publication is an output from the research projects "CAS Hundred Talents Program", "Digital Earth" (KZCX2-312) funded by Chinese Academy of Sciences and “Dynamic Monitoring of Beijing Olympic Environment Using Remote Sensing” (2002BA904B07-2) funded by the Ministry of Science and Technology, China.

References

1. Armstrong, M. P., and Marciano, R. J., 1997, Massively Parallel Strategies for Local Spatial Interpolation. Computers & Geosciences. Vol.23, No.8 , pp.859-867.
2. Cannataro, M., 2000, Clusters and grids for distributed and parallel knowledge discovery. Lecture Notes in Computer Science, Vol. 1823, 708-716, 2000.
3. Goodchild, M. F., 2001, *http://www.csiss.org/learning_resources/content/good_sa*.
4. Lanthier, M., and Nussbaurm, D., 2003, Parallel implementation of geometric shortest path algorithms. Parallel Computing, 29, 1445-1479.
5. Li, B., 1996, Implementing Spatial Statistics on Parallel Computers. In S. I. Arlinghaus, D. A. Griffith, W. C. Arlinghaus, W. D. Drake, & J. D. Nystuen (Eds.), Practical Handbook of Spatial Statistics, (New York:CRC Press) pp.107-148.
6. Pouchard, L; Cinquini, L; Drach, B; Middleton, D; Bernholdt, D; Chanchio, K; Foster, I; Nefedova, V; Brown, D; Fox, P; Garcia, J; Strand, G; Williams, D; Chervenak, A; Kesselman, C; Shoshani, A; Sim, A., 2003, An ontology for scientific information in a grid environment: The Earth system grid. In Proceeding of CCGRID 2003: 3rd IEEE/ACM International Symposium on Cluster Computing and the GRID held in Tokyo, Japan on May 12-15, 2003, pp626-632.
7. Roros, D. -K. D. and Armstrong, M. P., 1996, Using Linda to Compute Spatial Autocorrelation in Parallel. Computers & Geosciences.Vol.22, No.4, pp.425-432.
8. Roros, D. -K. D. and Armstrong, M. P., 1998, Experiments in the Identification of Terrain Features Using a PC-Based Parallel Computer. Photogrammetric Engineering & Remote Sensing.Vol.64, No.2, pp.135-142.
9. Wang, S. and Armstrong, M. P., 2003, A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments. Parallel Computing. 29, 1481-1504.

An Optimum Vehicular Path Solution with Multi-heuristics

Feng Lu[1] and Yanning Guan[2]

[1] State Key Laboratory of Resources and Environmental Information System,
Institute of Geographical Sciences and Natural Resources Research,
Chinese Academy of Sciences,
Beijing 100101, P.R.China,
luf@lreis.ac.cn
[2] Laboratory of Remote Sensing Information Sciences,
Institute of Remote Sensing Applications,
Chinese Academy of Sciences,
Beijing 100101, P.R.China,
guan@irsa.irsa.ac.cn

Abstract. Heuristics have been widely used in artificial intelligence related fields including path finding. In this paper, the author argue that different heuristics can be integrated to solve the path finding problems and set forward a solution integrating greedy heuristic, directional heuristic and hierarchical heuristic. In greedy heuristic, an improved Dijkstra's algorithm based on quad heap is set forward. Then an MBR for ellipse is presented to limit the searching extent. Thirdly, hierarchical spatial reasoning is used to build another heuristic considering the hierarchical structure of road network, which makes the optimum path selection completed in higher hierarchies as much as possible. A case study is carried out with a real road network to verify the efficiency and validation of the solution integrating the above algorithms.

1 Introduction

The single source optimum path with nonnegative arc weights is one of the most natural network optimization problems. A variety of algorithms for single source optimum path has been designed and implemented[1][2]. Many of them use heuristic strategies. Heuristic indicates literally learning by experience, or more generally in the artificial intelligence literature, a heuristic is a 'rule of thumb' and as such is the approach used by almost any human in conducting a searching [3]. In the context of searching algorithms, heuristic implies simply search specific knowledge. The optimum path algorithms based on heuristics include costing algorithm[3], branch-and-bound algorithm[4], hill-climbing algorithms [4], greedy algorithms [5][6][7] and A* algorithm[8]. etc.

Among the known optimum path algorithms, many of them use a kind of heuristic named greedy searching as searching strategies and explore how to design delicate running data structures and searching algorithms, so as to improve the running efficiency of sequential optimum path algorithms under the uniform time complexity.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 964–971, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Literature [1][2][7][9] have made detailed analysis and comparison for the optimum path algorithms. Because the real networks do not concern negative weights, Dijkstra's algorithm, a famous label setting algorithm adopting greedy heuristic, has got wide applications. Dijkstra's algorithm is the most mature optimum path algorithm theoretically up to date and the most robust one in practice[2][6][10]. Further work on fine-tuning Dijkstra's algorithm has been conducted by many researchers and has formed a large family of Dijkstra's algorithms.

There are many kinds of heuristics besides greedy strategy and they can be integrated to solve the optimum path problems. In this paper, the author showed the advantages of integrating the greedy heuristic, directional heuristic and hierarchical heuristic to get the optimum vehicular path in complicated road networks.

The remainder of this paper is organized as follows. Section 2 sets forward an improved Dijkstra's algorithm with quad heap priority queue. Section 3 introduces how to adopt spatial relationship between geographical objects to limit the path searching scale and speedup the searching procedure. Section 4 shows the hierarchical division of road networks and develops a hierarchy selection algorithm. Section 5 illustrates the integration of the above techniques with a case study. Finally Section 6 makes a discussion and draws some conclusions.

2 Quad Heaps and Fine-Tuned Dijkstra's Algorithm

Dijkstra's algorithm uses a heuristic based on greedy strategy. More details about Dijkstra's algorithm can be found in [11][12]. Theoretically, it is difficult to further improve the worst case time bound of Dijkstra's algorithm within the serial algorithm framework, and current research concentrates on how to improve the operational efficiency in practice.

Heap priority queues have been proved excellent data structures for Dijkstra's algorithms [2][6][7][13]. *K*-ary heaps, binomial heaps, Fabonacci heaps and radix heaps were discussed more. It has been argued that *k*-ary heaps are more suitable than binomial heaps and Fabonacci heaps to implement the priority queues for Dijkstra's algorithm in road networks [10]. A *k*-ary heap can be regarded as a complete *k*-ary tree. Every node on the corresponding complete *k*-ary tree corresponds to the element of an array that suffixed the node. The root correspond to the minimum (maximum) element of the *k*-ary heap, and the node elements of a sub tree rooted from one node are greater than (less than) the node element. The heights of *k*-ary heaps are $\lfloor \log n \rfloor$. The operation time bounds of *k*-ary heaps are proportional to the tree heights, i.e., $O(\log n)$.

Dijkstra's algorithm based on *k*-ary heaps concerns three kinds of heap operations, i.e., *heap-insert*, *heap-extract-min* and *heap-decrease-key*. Among these operations, '*heapify*' operation, which is used to keep the characteristics of *k*-ary heaps, plays a major role. Theoretically, the running time of a '*heapify*' operation depends on the *k* value. The optimal *k* value will minimize the running time of concerned heap operations.

The author had solved the optimal integer solution of k is 4. More detail can be found in [12]. The fine-tuned Dijkstra's algorithm implemented with quad heap priority queue has same time bound as that with binary heaps, namely $O((m+n)\log n)$. According to [10], the time bound is only $O((n+\log(1+m)-\log n)\log n$. For the sparse graph such as road networks that hold n<<m<<2n, the bound is $O(n\log n)$.

3 Extent Restriction for Path Searching

Most of the optimum path algorithms ignore the spatial distribution characteristics of networks, which make the searching free. Even if the procedure ends once reaching the destination nodes, there is still a lot of redundant searching. For road networks, Location information of nodes can be used to make it up.

3.1 Ellipse Restriction

In a network, approximate maximum distance or cost M_D can be determined once the source and the destination are defined. Then an extent can be constructed to limit the searching. The nodes outside the extent will not be considered during the searching for they are beyond the maximum distance or cost estimated.

If Euclid distance from a node N to the source S and to the destination T is $|SN|$ and $|NT|$ respectively, the limiting condition can be defined as $|SN|+|NT| \le M_D$. The critical points of N form an ellipse focused at S and T, with the major axis M_D, as shown in figure 1.

Ellipse limiting was firstly introduced in [14] to implement a primitive depth-first tree searching method. A suitable major axis M_D can be calculated statistically.

First, a lot of nodes can be extracted systematically from road networks to construct node sets A, B and their Cartesian product C. Every element in C can be regarded as the source and destination node between which the optimum path needs determined.

Supposing Euclid distance between them is e_{ab}, the optimum path distance is p_{ab}, a set R of ratio $r_{ab} = p_{ab}/e_{ab}$ can be formed for the extracted sample. Then a special value τ can be acquired which makes the elements in R have values not higher than τ under a special confidence. With τ as a product coefficient, the major axis M_D of the ellipse can be calculated using the coordinates of the source and the destination.

Ellipse limiting restricts the searching extent and reduces the scale of optimum path searching. But the procedure judging whether the current nodes are located in the ellipse is time consuming, which offsets the benefit brought through the reduced searching scale. When the destination is far away from the source, the ellipse limiting is sometimes worse than the free searching.

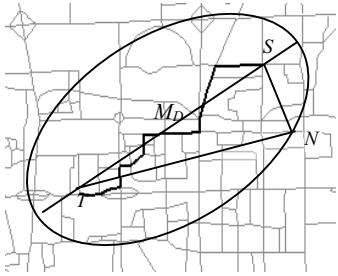


Fig. 1. Ellipse limit searching

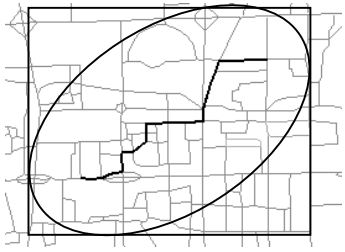


Fig. 2. MBR limit searching

3.2 MBR Restriction

To make up the deficiency of ellipse limiting, we set forward a rectangle limiting method. It avoids the great deal of product and evolution calculation. The major idea is to build the minimum bounding rectangle (MBR) for the ellipse and taking the MBR as a restricted area to reduce the searching scale.

In the ellipse method introduced above, when the product coefficient τ is acquired, the ellipse equation can be expressed as
where
After getting the ellipse equation, we can get the extremums of x and y through the differential equations

The tangents formed by the extremums of x and y compose the MBR for the ellipse, as shown in figure 2.

$$\frac{[\cos\theta(x-a)+\sin\theta(y-b)]^2}{A^2}+\frac{[-\sin\theta(x-a)+\cos\theta(y-b)]^2}{B^2}=1 \tag{1}$$

$$\theta=arctg(\frac{y_T-y_S}{x_T-x_S}) \qquad a=\frac{x_S+x_T}{2} \qquad b=\frac{y_S+y_T}{2}$$
$$A=\frac{\tau}{2}\sqrt{(y_T-y_S)^2+(x_T-x_S)^2} \qquad B=\sqrt{A^2-\frac{(y_T-y_S)^2+(x_T-x_S)^2}{4}} \tag{2}$$

$((x_S,y_S), (x_T,y_T)$: coordinates of the source and destination nodes

$$x_m=a\pm\sqrt{A^2\cos^2\theta+B^2\sin^2\theta} \qquad y_m=b\pm\sqrt{A^2\sin^2\theta+B^2\cos^2\theta} \tag{3}$$

Taking the MBR as the restricted area, judging the nodes processed only need compare whose coordinates with the bounds of the MBR. No complex calculation is needed.

When the azimuth between the source and the destination $\theta = k\pi/2$, the MBR gets the minimum area, i.e., $4AB$. And the area ratio of the MBR to the ellipse is $4/\pi = 1.2732$. When $\theta = (2k+1)\pi/4$, the MBR gets the maximum area, i.e., $2(A^2+B^2)$. And the ratio of the area of MBR to that of the ellipse is

$$\frac{S_R}{S_E} = \frac{4\tau^2 - 2}{\pi\tau\sqrt{\tau^2 - 1}} \tag{4}$$

4 Hierarchical Path Selection

Human cognition to the world has a distinct property of spatial hierarchies, and every hierarchy contains some necessary information to solve specific problems. For example, facing a road traffic map, a person can quickly find an optimum path between two spots chosen arbitrarily on the map. Even if the path is not a shortest path, it is also a good alternate to be referenced. Here the human thinking is none other but a procedure of typical hierarchical spatial reasoning. It does not need to carry out complicated calculation on the completely unwrapped detail, but make judgement on a generalized hierarchy, and decide the optimum path in this way.

4.1 Hierarchical Structure of Road Network

Roads in the network are divided into different classes according to their grades. Hierarchies are composed of classes. Every hierarchy contains the roads within higher hierarchies. Network is divided into sub-regions with boundary formed by the road sections in lower hierarchy. A sub-region is a grid in sub-region with higher hierarchy. Adjacent sub-regions share the boundaries. Adjacent hierarchies share some common nodes, i.e. all the nodes in the higher hierarchy. These nodes are the connector between the hierarchies and lower hierarchies.

The rule of hierarchical division was set forward in [15] and get improved in this paper through keeping the shared nodes in adjacent hierarchies to facilitate backtracking from the lower to the higher hierarchies during the recursion procedure.

4.2 Hierarchy Based Optimum Vehicular Path Selection

Given the source and destination nodes S and T, supposed i and j are the highest hierarchies that contain S and T, S and T are marked as S_i and T_j, then the optimum path algorithm can run on the sub-network level decided by $\min(i,j)$. Although such a lossy algorithm can not guarantee the resulted path is all to nothing the shortest path, it is a preferable path. Furthermore, such a hierarchical spatial reasoning is coincident with the drivers' thinking manner.

Road networks can be divided into several hierarchies, e.g., trunk roads (including superhighways), ordinary road and alley. Adjacent lists are established for the nodes in the three hierarchies respectively. The path searching is performed as follows.

1) Searching the nearest node to S_i and T_j to find out the highest hierarchies that contain S_i and T_j, so as to determine the beginning hierarchy; We have $k = \min(i,j|i \quad 3 \ \& \ j \quad 3)$;

2) Confirming the hierarchical zones that S_k and T_k are located;

3) If Sk and Tk are located in the same zone of hierarchy two, then the optimum path is calculated directly on hierarchy one;

4) Otherwise, if S_k and T_k are located in the same zone of hierarchy three, we can find out the nearest nodes to S_k and T_k on hierarchy two, marked as S_2 and T_2. The optimum paths $S_k \rightarrow S_2$ and $T_2 \rightarrow T_k$ are calculated on hierarchy one. Then the optimum path $S_2 \rightarrow T_2$ is calculated on hierarchy two. $S_k \rightarrow S_2$, $S_2 \rightarrow T_2$, $T_2 \rightarrow T_k$ are linked to get the required paths;

5) Otherwise, if S_k and T_k are not located in the same zone of hierarchy three, we can first find out the nearest nodes S_2 and T_2 to S_k and T_k on hierarchy two according to the method in 4), then calculate the optimum paths $S_k \rightarrow S_2$ and $T_2 \rightarrow T_k$ on hierarchy one. Also we can find out the nearest nodes S_3 and T_3 to S_2 and T_2 on hierarchy three, calculate the optimum paths $S_2 \rightarrow S_3$ and $T_3 \rightarrow T_2$ on hierarchy two. Lastly, the optimum path $S_3 \rightarrow T_3$ is calculated directly on hierarchy three, and the all of the resulted paths, i.e., $S_k \rightarrow S_2$, $S_2 \rightarrow S_3$, $S_3 \rightarrow T_3$, $T_3 \rightarrow T_2$ and $T_2 \rightarrow T_k$ are linked to get the final result.

5 A Case Study

A real urban road network with 12,800 nodes and 17,800 lines, and a PC(with single Pentium 2.4G CPU and 256M memory) are adopted to verify the efficiency of the above algorithms and the integrated solution.

Firstly, a Dijkstra's algorithm is implemented with quad heap priority queue. It uses a greedy heuristic to find the shortest path. Secondly, a directional heuristic with MBR restriction is added onto the greedy heuristic. 400 nodes are extracted from the network systematically to form sets A, B and their Cartesian product set C. Calculating e_{ab}, p_{ab} and r_{ab} for each element in C and statistically analyzing set R formed with all r_{ab},, we get $\tau = 1.379$ under a 95% confidence. It can be resulted from section 3.2 and equation 4 that the MBR limit increases the searching extent to 27.32%~36.29%, compared with the ellipse limit. In spite of the seeming deficiency, the following analysis shows it is still worthy of such an expense.

Nodes are picked up to form five sets from the upper left (1241 nodes), lower left (1155 nodes), upper right (1366 nodes), lower right (1143 nodes) and center (1307 nodes) part of the network randomly to compare the average CPU time of calculating the shortest path between the nodes in the sets with free searching, ellipse limiting and rectangle limiting. The node sets are marked as R_{ul}, R_{ll}, R_{ur}, R_{lr} and R_{cc}, The result is shown in table 1.

It shows that MBR limit always performs better than free searching method, and at most circumstance better than ellipse limit. Moreover, MBR limit increases the confidence of resulted path for its bigger searching area.

Table 1. Average time consumption of the shortest path searching with extent restriction (millisecond)

Searching method	R_{ul}-R_{ur}	R_{ul}-R_{cc}	R_{ul}-R_{ll}	R_{ul}-R_{lr}	R_{cc}-R_{lr}	R_{cc}-R_{cc}	R_{ll}-R_{ur}
free searching	12.49	11.57	17.12	20.05	23.88	8.26	22.82
ellipse limit	6.32	9.10	17.10	20.98	9.41	4.36	25.48
MBR limit	5.86	8.95	14.65	18.51	9.64	4.13	21.87

Thirdly the hierarchical selection heuristic is added onto the greedy and directional heuristics. Figure 3a gets the shortest path only on a single hierarchy. It can be noticed half the path is made of ordinary roads. Figure 3b presents a three-hierarchy division of the network. Figure 3c shows the optimum path with hierarchical path selection. Although the resulted path may not be the shortest path, most of that is located on trunk roads. It decreases the factors difficult to expect or quantify and makes the path more robust.

Fig. 3. Hierarchical optimum path selection

6 Conclusion

Different heuristics have different advantages and can be integrated to solve optimum path problems. Such an idea is experimented in this paper with integrating greedy heuristic, directional heuristic and hierarchical heuristic, with a real urban road network. A Dijkstra's algorithm implemented with quad heap priority queue establishes a perfect basis for the real road network optimum path searching. The spatial distribution characteristic of road networks can be used to form a directional heuristic to limit the searching scale. An MBR for ellipse limit is preferable to free searching and ellipse limit in path searching. The spatially hierarchical structure of road network can be used to build another heuristic to make the optimum path selection completed in higher hierarchies as much as possible, and improves the fault tolerance for the input data and the applicability of the resulted path. On the basis of the presented solution, other kinds of heuristics can also be integrated to contribute to the optimum path problems.

Acknowledgment. This research was supported by the National Natural Science Foundation of China under grant No. 40201043.

References

1. Deo N. and Pang C.Y., Shortest-path algorithms: taxonomy and annotation, Networks, 4(1984) 275-323.
2. Cherkassky B.V., Goldberg A.V. and Radzik T., Shortest paths algorithms: theory and experimental evaluation, Mathematical Programming, 73(1996) 129-174.
3. Fisher P. F., A primer of geographic search using artificial intelligence, Computers & Geosciences, 16(1990) 753-776.
4. Winston, P. H., Artificial Intelligence, 1st edn. Reading: Addison-Wesley, Massachusetts (1984).
5. Pallottino S., Shortest-path methods: complexity, interrelations and new propositions, Networks, 14(1984) 257-267.
6. Ahuja R.K., Mehlhorn K., Orlin J.B. and Tarjan R.E., Faster algorithms for the shortest path problem, Journal of the Association for Computing Machinery, 37(1990) 213-223.
7. Zhan F.B. and Noon C.E., Shortest path algorithms: an evaluation using real road networks, Transportation Science, 32(1998) 65-73.
8. Lester P, A* Pathfinding for Beginners, URL: http://www.policyalmanac.org/games/aStarTutorial.htm (2003).
9. Pallotlino S. and Scutella M. G., Shortest path algorithms in transportation models: classical and innovative aspects, Technical Report, Universita di Pisa (1997).
10. Goldberg A.V. and Tarjan R.E., Expected performance of Dijkstra's shortest path algorithm, Technical Report No. PRINCETONCS//TR-530-96, Princeton University (1996).
11. Tarjan R.E., Data Structures and Network Algorithms, 1st edn. Society for Industrial and Applied Mathematics Press, Philadelphia (1983).
12. Lu F., Zhou C.H. and Wan Q., An improved Dijkstra's shortest path algorithm based on quad heap priority queue and MBR searching method, Proceedings of the 9th Spatial Data Handling Symposium, (2000) 6b:3-13.
13. Fredman M.L. and Tarjan R.E., Fibonacci heaps and their uses in improved network optimization algorithms, Journal of the Association for Computing Machinery, 34(1987) 596-615.
14. Nordbeck, S. and Rystedt, B., Computer cartography --- range map, BIT, 9(1969), 157-166.
15. Car A. and Frank A., General principles of hierarchical spatial reasoning-the case of wayfinding, Proceedings of the 6th International Symposium on Spatial Data Handling, (1994) 646-664.

An Extended Locking Method for Geographical Database with Spatial Rules

Changxiu Cheng[1], Paiwei Shen[1], Mingbo Zhang[1,2], and Feng Lu[1]

[1] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, P.R.China
[2] Shan Dong University of Technology, Zi Bo 255000,P.R China

Abstract. Locking and unlocking operations are crucial to keep data consistency for concurrency control. Compared with traditional pessimistic locking method, the neutral concurrency control scheme can highly improve the concurrency of spatial database management systems. However, in consideration of spatial rules, dead locks brought about by waiting mutually for releasing required resources may frequently arise. To avoid such dead locks, referring to the theories of session management in operation system, we set forward an extended locking method for geographical database with spatial rules to implement the collaboration and competition between concurrent locking and unlocking transactions. The method avoids the expenditure of detecting dead locks and rolling back, and the system becomes more efficient. The test shows the extended locking technique makes the spatial database management system more ordered and efficient in a multi-users environment only with some time consumption of managing logical codes.

1 Introduction

HYPERLINKA heterogeneous, distributed, and loose-coupled working environment is evolving. The cooperative GIS will be popular in the future. As a key technology to realize cooperative GIS, concurrency control is also emerging, i.e. how to ensure that geographical database operations from different users do not interfere with each other (Hanssen 2003).

The early GIS were based on CAD systems where the map base was held as a collection of CAD drawings. The methods of managing such a system where multiple users wish to access and update the map base are based on the manual drawing office approach. Indeed there is a market for Document Management systems(Newell 2003).However, as GIS shift to Relation Database Management Systems form Document Management systems, the concurrency control scheme of GIS has been changed. Most distributed database employ either pessimistic or optimistic concurrency control scheme（Lee and Yang et al 2000）. The former often results in the high communication delay, because the spatial data queried by clients would be hold for a long time. Generally, the latter is employed in distributed GIS application, i.e. version management. The version management in GIS has been put forward for

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 972–980, 2004.
© Springer-Verlag Berlin Heidelberg 2004

decades (Easterfield and Newell et al 1990). Now, it has been widely used, i.e. ArcSDE (Gaskill and Brooks 2000), Oracle Spatial (Arun 2001). But it makes the system complex and perplexes users with undoing and redoing of previous actions on user interface when conflict occurs (Lee and Yang et al 2000, Batty 2002) .In addition, the version management is not reasonable to distribute interactive GIS applications. In this application, the changes are not so complex, the changed entities are not so many, but the results are required to submit in a few minutes. WAVE (Lea and Matsuda 1997) adopts a simple locking mechanism for this interactive application. A user locks objects that would be manipulated and updates them. The updates are propagated to the other users while the lock is released. It was named Neutral Concurrency Control Scheme (NCCS) in this paper.

However, only locking these manipulated spatial entities is not enough considering spatial rule-based processing (Michael 2001, Mayal and Hall 2002, Sha and Bian 2002)in NCCS. So, not only these manipulated objects but also their neighboring objects should be locked in geographical database with spatial rules. Then it will be frequent to encounter dead locks resulted by waiting for releasing required resources by each other. To avoid the emergence of such dead locks, we propose an extended locking method for spatial rule-based processing. It makes the spatial database management system more ordered and efficient in a multi-users environment only with some time consumption of managing logical codes.

The remaining parts of this paper are organized as follows. Section 2 describes the neutral concurrency control scheme and introduces why the dead lock are frequently encountered in spatial rule-based geographical database. Section 3 analyzes the reason why dead locks occur, and presents an extended locking method to avoid the emergence of such dead locks. The implementation of the extended locking method in Oracle Spatial is then introduced in section 4. Section 5 bring forward two tests, which show that the extended locking method makes the spatial database management system more ordered and efficient in a multi-users environment with just some time consumption of managing logical codes. Finally, in Section 6 of this paper, summary as well as forward are made as conclusion.

2 Neutral Concurrency Control Scheme and Deadlocks

In pessimistic concurrency control scheme, there are two types of locks. One is share lock (lock-S); another is exclusive lock (lock-X). Share lock specifies that others may read while you read, but writing objects in the storage group is prevented. Several users can have the share lock at the same time. Exclusive lock prevents other users from getting either of the two kinds of locks. In NCCS, there are also share lock and exclusive lock. But there is something different in compatibility matrix. The share lock allows that one and only one user should write objects while you read. In order to keep consistency, the writing results are propagated to the other users when the exclusive lock is released.

However, only locking these manipulated spatial entities is not enough considering spatial rule-based processing in NCCS. If fig 1 conforms to share edit rule, the shape of entity C, entity D and entity G will be changed on moving the P_1P_2 of entity A.

Therefore, in spatial rule-based geographical database, not only these manipulated objects but also their neighboring objects should be locked. Because the locking granularity is not atomic, a deadlock will occur when two or more users are waiting for data unlocked by each other. If transaction 1 and transaction 2 are executed according to the schedule shown in fig 2, transaction 1 will be choked and wait for transaction 2 to unlock the entity D, and transaction 2 will also be choked and wait for transaction 1 to unlock the entity C at the time point of T5. As a result, the deadlock occurs.

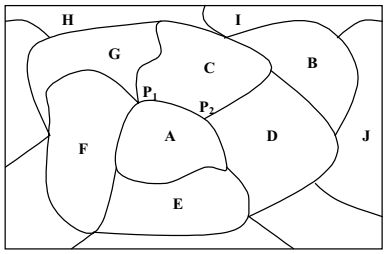


Fig. 1. Map view

	Transaction 1 : Edit A Entity	Transaction 2 : Edit B Entity
T0	Begin	
T1	Lock-X C	
T2	Lock-X G	Begin
T3	Lock-X F	Lock-X J
T4	Lock-X E	Lock-X D
T5	Lock-X D	Lock-X C
…	…	…
T (n-1)	Edit	Edit
Tn	Commit	Commit

Fig. 2. The executive schedule

In geographical database with spatial rule-based, deadlocks occur most often. Because there are many kinds of spatial rule-based in GIS, and some rules maybe lock many entities. For example, when a user is editing the “Changjiang River” of china, he will hold the many spatial data that cover large areas, and then the probability of the occurrence of deadlocks would be high. Some time the deadlocks are no less than the chain reaction specified in the multi-user environment.

3 The Extended Locking Method

To avoid the emergence of such dead locks, referring to the theories of session management in operation system, we set forward an extended locking method to implement the collaboration and competition relationships between locking and unlocking transactions.

3.1 The Collaboration between Locking and Unlocking Transactions

Collaboration is the style or approach used by some processes as they work toward a common goal. Locking, updating, and unlocking are three basic steps in any updating transaction. In the updating transaction, the systems must make them work according to some rules. They are as follows: Rule 1: Firstly lock the data that would be updated. Then update them. At last unlock them. Rule 2: the lock operation must be atomic. Ensures that either all lock operations of a transaction accomplish successfully or all of its effects fail. Rule 3: the lock operations have to wait until others unlock the Data. So, it is a type of collaboration.

The data synchronization is an effective instrument to implement the collaboration. We can set an IsXLocked signal to any spatial entity, which is used for communication. The entity is not locked by transaction, if its IsXLocked equals 0. And it is locked by transaction, if its IsXLocked equals 1. According to the rules of updating transaction, the locking and unlocking logic can be presented as follows shown in fig 3. The UpdateTransaction in fig3 could avoid some deadlocks.

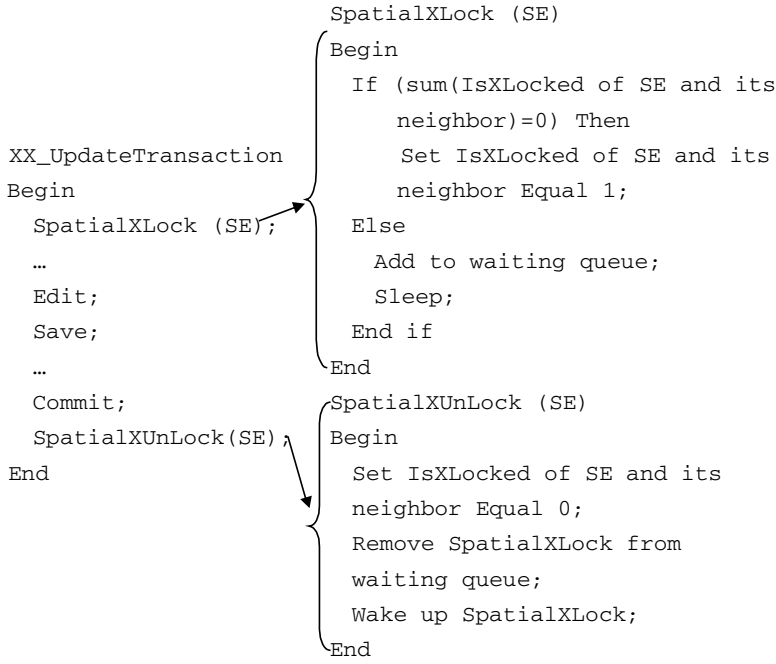


Fig. 3. Logic of locking and unlocking

3.2 The Competition between Locking and Unlocking Transactions

Accessing the shared resources causes the competition between locking and unlocking transactions. In the multi-user environment, many locking transactions exist at the same time. These transactions maybe request the same resource, change the same data, and keep the change for a period. Then the shared data may be inconsistent, and deadlocks will be resulted from. For example, as shown in fig 2, if transaction 1 executes the if clause in fig 3 at time point T0 and transaction 2 executes the if clause at time point T2, the IsXLocked value of the entity C will still be 0 at time point T2, because its value will not change until the transaction is committed. So, transaction 2 will not be chocked, and the deadlocks still occurred at time point T5.

In order to keep consistency of share variables, the locking and unlocking transactions must be mutex. We can define the code section that access IsXLocked as critical section (CS) that only allows one transaction to enter this code section at the same time. The semaphore, SemWait and SemSignal operations can implement the

concept of CS. A semaphore called Mutex, is integer variables that, apart from initialization, can be accessed only through atomic SemWait and SemSignal operations. If insert the SemWait operations after Begin clause of locking and unlocking transaction, and insert the SemSignal operations after Begin clause of them, the locking and unlocking transactions will be mutually exclusive.

4 The Implementation of Extended Locking Method

There are two troublesome in the implementation of extended locking method in Oracle Spatial. One is the implementation of critical section; another is transaction queue mechanism. There isn't the concept of critical section in PL/SQL of Oracle, so it must be simulated with other techniques. And the advanced queue is very complex, so coding it is difficult.

In Oracle, the exclusive table lock is the most restrictive mode of table lock. Only one transaction can obtain an exclusive table lock for a table.So, if we create Table_mutex table as semaphore, then the SemWait operation (EnterCritical Section) can be programmed as follows:

```
Begin
  LOCK TABLE Table_Mutex IN exclusive MODE;  --Exclusive
  Lock
End
```

However, it is needless to program the SemSignal operation, because it operation has been included in commit clause at the end of transaction. The commit will release the exclusively hold of Table_Mutex, and wake up some waiting transactions. So, it is obvious that the queue mechanisms has worked in the lock table and commit clause.

Given the geometry column was name 'shape', we can add IsXLocked column as signal to express the entities whether locked or unlocked. Then, the locking transaction can be programmed as followed:

```
(SpatialEntityID in Number, TableName in Char)
          RETURN Integer IS
BEGIN
 EnterCriticalSection;  -- Enter critical section
 …                          --The define of variables
 SELECT SUM (t1.IsXLocked) INTO LockedCount FROM
 TableName t1, TableName t2 WHERE t1.oid=SpatialEntityID
 AND SDO_RELATE (t1.Shape, t2.Shape, 'MASK= ANYINTERACT
 QUERYTYPE=WINDOW')='TRUE';
 IF LockedCount =0 THEN   --These entities are available
  UPDATE TableName SET IsXLocked =1 WHERE oid IN (SELECT
  t1.oid FROM TableName t1, TableName t2 WHERE
  t1.oid=SpatialEntityID AND SDO_RELATE (t1.Shape,
```

```
    t2.Shape, 'MASK=ANYINTERACT
    QUERYTYPE=WINDOW')='TRUE');
    COMMIT;           --Commit, then exit critical section
    RETURN 1;          --Return success
  ELSE                --Some entities aren't available
    COMMIT;           --Commit, then exit critical section
    RETURN 0;          --Return fail
  END IF;
END Spatial_XLock;
```

The produce is not optimum, because the implicit spatial query of anyinteract is executed twice. We can define some arrays in which the results of the spatial query are stored, which will be reusable. Additionally, the explicit query also can make the produce more effective than implicit query. The unlocking transaction is simpler than the locking transaction; details of which will not be presented here in this paper.

5 Tests and Performance Analysis

The test data is the chinese county boundary scale at 1:4,000,000, in which there are 2525 polygon. The database management systems is Oracle 9i. The database server is Dell4600.

Test1:Set up only one session (single user environment), use the old method described in section 3 to lock and unlock these polygons one by one, and record the elapsed time. Repeat it eight times, and calculate the average elapsed time. Then, use the extended locking method to do the same test.

The test results are shown in Table 1 in the single user enviroment. It is easily calculated that the average elapsed time of locking and unlocking one entity is 0.0209s (52.75/2525) to the old method, and 0.0216s (54.63/2525) to the new method. The 0.007s (0.0216-0.0209) excrescent time of locking and unlocking one spatial entity is few in relative to the elapsed time of its editing produce, which may be lasted for some minutes.

Table 1. The elapsed time using these two method in single user enviroment (Unit:s)

Test times	The elapsed time using the old method	The elapsed time using the new method
1	53	54
2	52	57
3	53	54
4	53	55
5	54	55
6	52	55
7	52	53
8	53	54
Average	52.75	54.63

Test 2: Set up four sessions (multi-user environment), almost simultaneously use the old method to lock and unlock these polygons one by one, and record their elapsed time. Repeat it four times, and calculate the average elapsed time. Then, use the new method to do the same test.

The test results of the old method are shown in Table 2 in the single user enviroment. '∞' stands for the session was enmeshed into deadlocks after one time point. The result show that only session 1 which was start up more ealy would well worked, the others session would enmeshed into deadlocks.

Table 2. The elapsed time using the old method in multi-user enviroment (Unit:s)

Test times	Session 1	Session 2	Session 3	Session 3
1	109	∞	∞	∞
2	108	∞	∞	∞
3	112	∞	∞	∞
4	110	∞	∞	∞

The test results of the new method are shown in Table 3 in single user enviroment. The result show that the four sessions would be worked orderly and finish as soon as possible. The average time elapsed of Table 3 will be quadruple of Table 1, because the four sessions were executed in serial.

Table 3. The elapsed time using the new method in multi-user environment (Unit:s)

Test times	Session 1	Session 2	Session 3	Session 3
1	211	215	214	212
2	210	212	221	219
3	213	216	214	213
4	211	215	214	212

The results of test 2 indicate that the new method is better than the old method as a whole in multi-user environment. The new method keeps the systems work orderly, and the old method makes the most session be in bad deadlock situations.

6 Conclusions

This paper has explored an extended locking method for geographical database with spatial rules. With reference to related theories of process management in operation system, the extended locking method implements the management of collaboration and competition between locking and unlocking transactions, avoids deadlocks, and improves the effectiveness of the geographical database management systems. The two tests show that the extended locking technique makes the systems more ordered and efficient in a multi-users environment at the cost of some time consumption of managing logical codes.

The NCCS is the protocol abided by the extended locking method. Compared with the traditional pessimistic scheme, the NCCS makes the requirements for locking executive less strict and improves the concurrency of the system. While compared with version management, it is easier to be realized and little system expenditure is needed for creating version, recording change, merging two versions, and proposing conflicts.

During the research on geographical database, we should keep in mind the speciality of spatial data and refer theories of computer science to draw conclusions for the use of spatial database.

References

1. Anthony Chiu, Ben Kao and Kam-yiu Lam: An adaptable constrained locking protocol for high data contention environments: correctness and performance, Information and Software Technology. Vol. 42 (2000) 599-608
2. Bharat Bhargava: Concurrency Control in Database Systems, IEEE Transactions on Knowledge and Data Engineering, Vol. 11 (1999) 3-16.
3. Dean Kuo, Volker Gaede, Kerry Taylor: Using Constrains to Manage Long Duration Transaction in Spatial Information System. http://www.wiwi.hu-berlin.de/~gaede/ltm.ps.gz (19997).
4. Dongman Lee, Jeonghwa Yang, Hee Yong Youn, et al: Entity-Centric Scalable Concurrency Control for Distributed Interactive Applications. Proceedings 19th IEEE Int'l Performance, Computing, and Communications Conference, (2000)544 ~ 550.
5. Zeiler Michael: Modeling Our World: The ESRI Guide to Geodatabase Design. New York: ESRI Press, (2001).
6. E. Easterfield, Richard G. Newell , David G. Theriault: Version Management in GIS - Applications and Techniques. Proceedings EGIS '90, (1990) 288-297.
7. Gjermund Hanssen: Concurrency control in distributed geographical database systems. Proceedings 9th Scandinavian Research Conference on Geographical Information Science, (2003)167-180.
8. Gopalan Arun: Oracle Workspace Manager: A Framework for Long Transactions. Proceedings GeoSpatial World 2001 Conference,(2001)
9. Jackie Gaskill, Dale Brooks: Understanding ArcSDE: The Gateway to Your RDBMS. Proceedings 12th Annual ESRI International User Conference, (2000).
10. Kevin Mayal, G. Brent Hall: Generalizing GIS: Development of Spatial Grammars for Landscape Planning. http://www.isprs.org/commission4/proceedings/pdfpapers/360.pdf (2002) HYPERLINK
11. Mainguenaud Michel: Consistency of Spatial Database Query Results. Computer, Environment and Urban Systems, Pergamon Press, Vol. 18 (1994)333-342.
12. Manhoi Choy, Mei-Po Kwan, Hong Va Leong: Proceedings the 27th Hawaii International Conference on System Sciences, Vol.IV(1994)337-346.
13. Oracle Corp: Data Concurrency and Consistency. http://download-west.oracle.com/docs/cd/B10501_01/server.920/a96524/c21cnsis.htm#3178.
14. Peter M. Batty: Version Management Revisited. http://www.gita.org/members_only/downloads/batty.pdf(2002).
15. Prabhudev Konana , Sudha Ram : Transaction management mechanisms for active and real-time databases: A comprehensive protocol and a performance study. The Journal of Systems and Software, Vol. 42, (1998) 205-225

16. Richard G. Newell: The Why and the How of the Long Transaction. http://emea.smallworld.co.uk/support/techpaper/tp9.html.
17. R. Lea, Y. Honda, K. Matsuda: Virtual Society: Collaboration in 3d Spaces on the Internet. The Journal of Collaborative Computing, Vol. 6 (1997)227-250.
18. Sha Zongyao, Bian Fuling: Comprehensive Knowledge Discovery: Theory, Concept and Application.http://www.isprs.org/commission4/proceedings/pdfpapers/191.pdf(2002).HYPERLINK
19. S. Goel, B. Bhargava and S. K. Madria: An adaptable constrained locking protocol for high data contention environments: correctness and performance, Information and Software Technology, Vol. 42(2000) 599-608.
20. Thomas Connolly, Carolyn Begg: Database Systems: A Practical Approach to Design, Implementation, and Management 3/e, Addison Wesley Higher Education(2001).

Preliminary Study on Unsupervised Classification of Remotely Sensed Images on the Grid

Jianqin Wang[1], Xiaosong Sun[1], Yong Xue[1,2*], Yincui Hu[1], Ying Luo[1], Yanguang Wang[1], Shaobo Zhong[1], Aijun Zhang[1], Jiakui Tang[1], and Guoyin Cai[1]

[1]Laboratory of Remote Sensing Information Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, P. Box 9718, Beijing 100101, China
[2]Department of Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
{tian1.wang@163.com,y.xue}@londonmet.ac.uk

Abstract. Grid is a new technology. With corresponding middleware it can give strong computing power. In this paper we mainly discuss the middleware technology and architecture used in remote sensing image classification algorithm. Because unsupervised classification middleware is the key of the classification middleware algorithms, we study the alternant-unsupervised middleware and put forward a non-alternant unsupervised middleware scheme. Based on this scheme, main factors which effect the performance of non-alternant unsupervised classification are analyzed.

1 Introduction

"Grid" computing has emerged as an important new field, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation [1]. They defined "Grid problem", which is defined as flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources –what they referred to as virtual organizations(VO).

The term "the Grid" was coined in the mid 1990s to denote a proposed distributed computing infrastructure for advanced science and engineering [2]. There are several famous grid projects today. Access Grid(ww.fp.mcs.anl.gov/fl/access grid) lunched in 1999 and mainly focused on lecture and meetings-among scientists at facilities around the world. European Data Grid sponsored by European union, mainly in data analysis in high-energy physics, environmental science and bioinformatics. Grid Physics Network (GriDPhyN) [5] lunched in 2000 and sponsored by NSF mainly in data analysis for four physics projects: two particle etectors at CERN's Large Hadron Collider, the Laser Interferometer Gravitational Wave Observatory, and the Sloan Digital Sky Survey. Information Power Grid [6] is the NASA's computational support for aerospace development, planetary science and other NASA research. International Virtual DataGrid Laboratory (iVDGL)[7] sponsored by NSF and counterparts in Europe, Australia, and Japan in 2002. Network for Earthquake Engineering and Simulation labs (NEESgrid) (www.neesgrid.org) intended to integrate

* Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 981–988, 2004.
© Springer-Verlag Berlin Heidelberg 2004

computing environment for 20 earthquake engineering labs. TeraGrid (www.teragrid.org) is the general-purpose infrastructure for U.S. science: will link four sites at 40 gigabits per second and compute at up to 13.6 teraflops. U.K. National Grid [8] sponsored by U.K Office of Science and Technology. Unicore (www.unicore.de) is a seamless interface to high-performance Education and Research computer centers at nine government, industry and academic labs

The famous Grid focused on spatial information includes SpaceGrid, EnvirGrid and EarthObsevation Grid. ESA's SpaceGrid is an ESA funded initiative (http://sci2.esa.int/spacegrid).

EnvirGrid main goals are generalization of Earth Science application infrastructure to become GRID-aware, extend GRID access to European Environmental and Earth Science application to large science communities, to value adding and commercial communities, …, and demonstrate collaborative environment for Earth Science. Dozens of satellites constantly collecting data about our planetary system 24 hours a day and 365 days a year. Large scale of satellite data needed to be processed and stored in real time or almost real time. Real time processing in remote sensing is very significant in some cases, such as fire and flood. So far real time processing in remote sensing confronts much difficulties in one single computer, or even impossibility. Computing grid which is integrated by series of middlewares provide a way to solve this problem [3]. The following is our preliminary study on unsupervised classification middleware on the grid platform.

2 Middleware on Unsupervised Classification on Grid Platform

Classification is a important field in digital remote sensing image processing. Almost all classification algorithms involve processing pixel value one by one. Large amount of remote sensing data classification will take too long time. Luckily, emerging grid technology can provide large computing power in the Internet as the security type. The grid computing nodes resource distribute loosely. They may be supercomputers or common PCs with heterogeneous architecture. The grid computing nodes resource distribute loosely. They may be supercomputers or common PCs with heterogeneous architecture. Classification middlewares on grid platform can solve this problem.

2.1 Testbed of Telegeoprocessing Grid Platform

Telegeoprocessing grid platform in Institute of Remote Sensing Application, Chinese Academy of Science is an advanced High-Throughput Computing grid using condor which copyright is Computer Sciences Department, University of Wisconsin-Madison. Heterogeneous computing nodes including two sets of Linux computers and WINNT 2000 professional computers and one set of WINNT XP computer provide stable computing power. The grid pool uses java universe to screen heterogeneous characters. It is a fundamental bare computing platform and can not run remote sensing image processing program. So remote sensing image processing middlewares running on grid pool are inevitable. Common users can use the heterogeneous grid

and share its strong computing power to process remote sensing images with middlewares as if in one supercomputer.

IP	Name	Arch-OS	Role
192.168.0.5	Manager.linux	Itel/Linux	Manager/client
192.168.0.3	Wjq.linux	Intel/WIN50	Client
192.168.0.102	Tele1.linux	Intel/WIN50	Client
192.168.0.111	Tele2.linux	Intel/WIN51	Client
192.168.0.6	Client2.linux	Intel/Linux	Client

2.2 A Simple Unsupervised Classification Algorithm Description

The aim of the classification middleware on grid is to divide jobs into several assignments and submit them to the computing pool. The rule of dividing jobs is: the more simple the dividing type can be used, the better it is. Difference among all the heterogeneous computing nodes resource determine this rule. With the same algorithm, the amount of the remote sensing image data is the most significant one of all the factors of impacting the classification speed. Naturally, dividing jobs can be transferred into dividing data and sending data to several grid nodes properly

Training areas can be chosen beforehand in the supervised classification algorithm. Then interoperation among the divided areas doesn't exist at all. Dividing assignment is relevantly easy. Contrarily, There exists strong correlation among all the divided data parts in the unsupervised classification. Mean center of classification must change dynamically according to computing result of every time's cycle. So the key technology of classification middlewares is to solve unsupervised classification middleware.

Dynamical clustering method is a typical algorithm of all the unsupervised classification algorithms. Every pixel can be classified into corresponding K class group individually according to least -squares rule. The algorithm iterates over three steps:

1. Compute the mean of each cluster.
2. Compute the distance of each point from each cluster by computing its distance from the corresponding cluster mean. Assign each point to the cluster it is nearest to.
3. Iterate over the above two steps till the sum of squared within group errors cannot be lowered any more. Vector x is the pixel value. The K-means algorithm partitions data into K clusters. The solution is then a set of K centers, each of which is located at the centroid of the data for which it is the closest center. For a partition P of the elements in {1,2…m}, denote by P (i) the cluster assigned to i and C(j) the centroid of cluster j. The intent of the K-means algorithm is the minimization of the objective function

$$Ep(i) = \sum_{i=1}^{m} dist(i, C(P(i))) \tag{1}$$

Where $dist(x,c)$ is the distance between the vectors x and c. In this experiment, we define the distance as the Euclidean distance

$$dist(x,c) = \|x,c\| \tag{2}$$

2.3 Alternant Unsupervised Classification Middlewares Test on the Telegeoprocessing Grid Testbed

Test Data Resource and Ground

We employ IKONOS satellite dada as remote sensing image resource. IKONOS satellite images were acquired on 26 April 2002 with approximately 1 m and 4 m spatial resolution in the visible (panchromatic) and relatively broad bands in the blue, green, red, and near-infrared portions of the spectrum, respectively. We use bands of blue, green, red, and near-infrared for about 4m spatial resolution. Test ground surrounds Olympic park in Beijing.

Fig. 1. The test area showed by IKONOS satellite blue band

Architecture of Alternant Unsupervised Classification Middleware on Grid Testbed

Alternant middleware can provide friendly interface to users. Figure 2 shows the architecture of alternant unsupervised classification middleware development.

User just expresses the number of jobs, which he wants to divide the job into through the user interface. JDMM that denotes job dividing manager module can divide job according to the user's demand. JSMM (Job submission manage module) can submit jobs to grid platform. CJMM is circulating job monitoring manage module the duty of which is monitoring when every job's result can be sent back for each loop. It also computes the new mean center through RRM, which receives result.

CJMM sends message to JSMM when the classification precision is not compatible to the user's demand. DFVM is data fusion and visualization module. It can fuse the divided result and give user visualization of the classification.

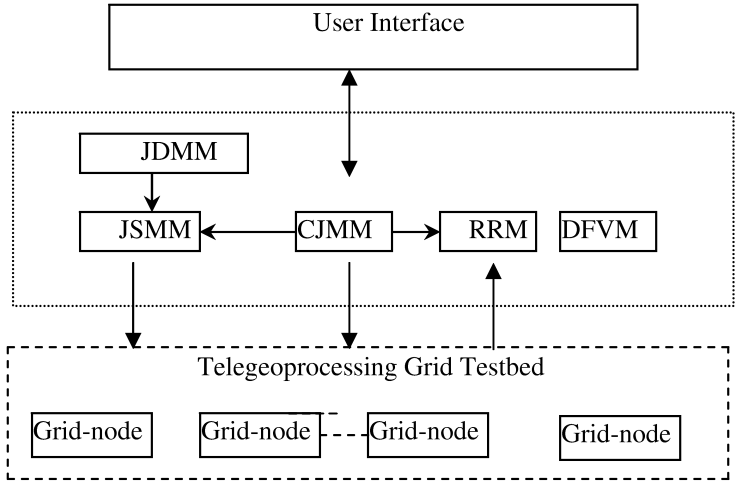


Fig. 2. Architecture of alternant unsupervised classification middleware

2.4 Non-alternant Unsupervised Classification Middleware Study

Alternant unsupervised classification middleware is natural to come to our minds. But alternant algorithm must employs result from every grid node of each loop. Thus job's submission of this time must wait for result of the last time. Frequent communication among grid nodes prolongs computing time. All the computing nodes may be heterogeneous and super-performed nodes must wait for common-performed nodes. Then the total time can be expressed as following:

$$T_{total} = n * \max\{T_{grid_node}\} + n * T_{communication} \tag{3}$$

where:
n: times of loop;
T_{grid_node}: computing time of one grid node for each loop;
$T_{communication}$: communication time of each loop.
It is obvious that alternant unsupervised middleware algorithm is not the ideal one. So non-alternant unsupervised middleware is expected. We image that each job partition data should be computed individually and be submitted only once. After computing they can be fused and displayed. Through deep analysis we find that we can try to solve this problem for several reasons. Object can be distinguished and classified whatever in one whole image or in divided parts according to brightness of every

pixel. The most important issue we care about is which type every object can be classified and the difference among different objects .We can do not care about the brightness of the whole core center. We set the classified number of each divided part. The problem may be that each divided part must be classified into given class number but it may be less the number in just one divided part. Through amalgamating dynamically class number through critical threshold we can adjust the class number in each divided part. Then classifying effect in each part can be almost the same as in one whole image after properly matching colors.

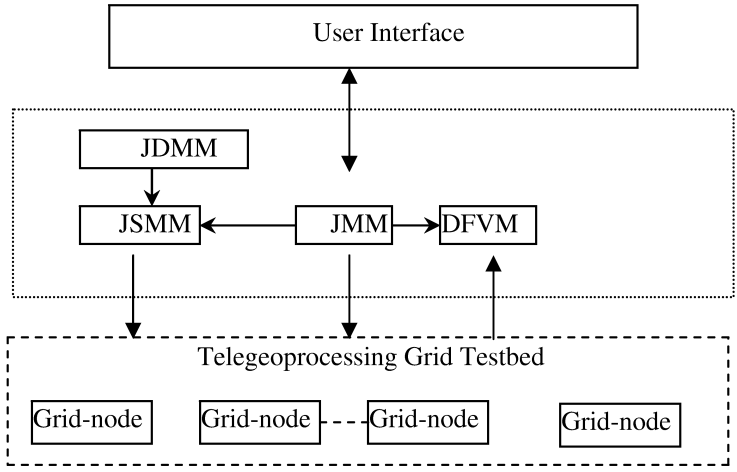


Fig. 3. Architecture of non-alternant unsupervised classification middleware

The assignment can be submitted only once. DFVM (data fusion and visualization module) gives users classification result when JMM (job monitoring module) finding the end result. The following pictures are the classification result comparison between single computer and non-alternant in grid pool. The effect shows almost little difference and approve we can employ this way to exploit our non-alternant middleware in grid environment. To distinguish vegetations from other objects we employ NDVI algorithm (>=0.281) [4].

2.5 Performance Test Using Non-alternant Middleware on Grid Platform

Non-alternant middleware has advanced performance relative to alternant one. Multi-computing nodes in grid environment can give strong computing power. But the idea that the lesser time the assignment can take, the more job will be divided to is not always true. To find out which factor affect the performance most we have several test on this grid platform. Using non-alternant classification middleware we divide assignment into 4 and 16 job units individually. Test result is shown in the following table:

Number of jobs divided	Time (full resource) Second	Time (partial resource) Second
1	35	
4	25	28
16	39	41

When all the grid nodes are free, assignment divided into 4 sub-jobs takes about 25seconds and less than on single computer. But the same assignment divided into 16 sub-jobs takes longer time on the contrary for about 39 seconds. While computers in grid nodes are partially used, the time the same assignment takes will be longer than the free station.

Relationship between the number of the sub-jobs and the number of the computing nodes directly infects the performance of non-alternant classification middleware. When the number of the sub-jobs is almost the same as that of the computing nodes the performance of the middleware will be good. But when the number of the sub-jobs is far more than that of the computing nodes all the sub-jobs will wait in the working queues and the performance will fall.

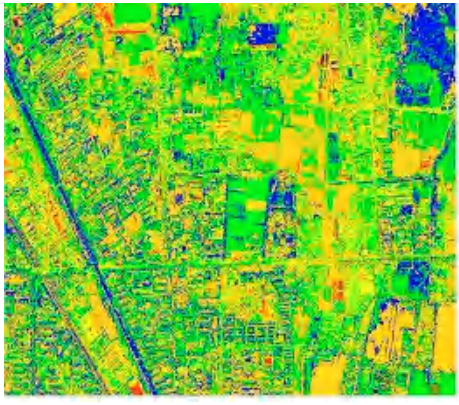


Fig. 4. Classification result on single computer.

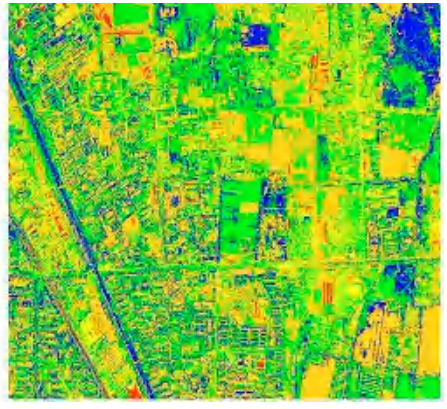


Fig. 5. Classification on Grid platform

3 Conclusions

Grid technology is very effective method for remote sensing image classification. It can give strong computing power in grid environment. Unsupervised classification middleware is the key technology of classification algorithms because of its alternant character. Developing non-alternant middleware of unsupervised classification can successfully use grid to classify remote sensing image objects. It is a good experience for other middleware used in remote sensing image processing. The number of divided sub-jobs must be proper according the actual station of the grid resource. If the number of divided sub-jobs is too more, the performance will be fall on the contrary in the special grid environment.

Acknowledgement. This publication is an output from the research projects "CAS Hundred Talents Program", "Digital Earth" (KZCX2-312) funded by Chinese Academy of Sciences and “Dynamic Monitoring of Beijing Olympic Environment Using Remote Sensing” (2002BA904B07-2) funded by the Ministry of Science and Technology, China.

References

1. I. Foster, C. Kesselman, S. Tuecke The Anatomy of the Grid, Intl J. Supercomputer Applications, 2001.
2. Foster, I. and Kesselman, C. (eds.). The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, 1999.
3. Jianqin Wang, Yong Xue, and Huadong Guo, 2003, A Spatial Information Grid Supported Prototype Telegeoprocessing System. In Proceedings of 2003 IEEE International Geoscience and Remote Sensing Symposium (IGARSS’2003) held in Toulouse, France on 21-25 July 2003.
4. F. Hirota, M. Morisugi, F. Weihua, and H. Imura: “A study on evaluations for vegetation in urban area with GIS and high-resolution satellite data – Reexamination of method to extract vegetation zones with the precedent satellite data -”, Proceedings of 29th Annual meeting of Environmental Systems Research 2001, pp. 163 - 168, 2001.
5. Grid Physics Network. `http://www.griphyn.org/index.php`
6. W. E. Johnston, IPG Chief Architect. Information Power Grid. `http://www.ipg.nasa.gov/`
7. IVDGL Grid. `http://www.ivdgl.org/`
8. UK e-Science Grid Support Center. `http://www.grid-support.ac.uk/`

Experience of Remote Sensing Information Modelling with Grid Computing

Guoyin Cai[1], Yong Xue[1,2*], Jiakui Tang[1], Jianqin Wang[1], Yanguang Wang[1], Ying Luo[1], Yincui Hu[1], Shaobo Zhong[1], and Xiaosong Sun[1]

[1]Laboratory of Remote Sensing Information Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, P. Box 9718, Beijing 100101, China
[2]Department of Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK
{y.xue@londonmet.ac.uk, cgywj@126.com}

Abstract. In this paper, we focused on the remote sensing information modeling and determination using Grid computing platform. We have underdone the experiments using remotely sensed images for thermal inertial modeling in Condor system that is one of the Grid Projects existed nowadays worldwide. We divided remote sensing data into several parts and run them on Condor pool and on one single machine. From these tests, the relationship among the work efficiency of image processing in Condor system and the number of the separated parts of image and the number of machines in this system is presented. Given a certain number of machines, a most efficient image size is existed among varies sized images. Besides, the possible causes of the longer put-off in this process are given, and some possible methods to resolve this problem are also presented. It is feasible to use Grid computing system such as Condor to process remote sensing data. And if the postpone problem can be resolved, the work efficiency of Grid systems will be high. Even with so many problems, it is a good thing that Grid systems do many things for you during all of us are in sleep. Our next major task will concentrate on realizing an arithmetic that can read and divide remote sensing images based on image size and number of machines in Grid system automatically, and transfer results back to the submitted machine as a whole data file.

1 Introduction

Remote Sensing data is characterized by largeness and instantaneousness. Space missions involve the download, from space to ground, of many raw images per day. The analysis and sharing of these huge amounts of data is a big challenge for the remote sensing community. The emerging computational grid technologies are expected to make feasible the creation of a computational environment handling many PetaBytes of distributed data, tens of thousands of heterogeneous computing resources, and thousands of simultaneous users from multiple research institutions (Giovanni *et al.* 2003).

Grid is an integrated computing and resources environment, or a location for computing resources. Grid can absorb all kinds of resources for computing, and

* Corresponding author

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 989–996, 2004.
© Springer-Verlag Berlin Heidelberg 2004

transfer them into available, reliable, standard and economic computing ability. Besides computers, computing resources also include web communication abilities, data, instruments and equipments, and even people (Foster and Kesselman 1998). Grid was firstly proposed from using for reference of electric power web. The final goal of the grid is that users can use grid as so convenient as using electric power. Grid can provide final users with common computing abilities that have no relationship with geographic location and computing facilities. Computing abilities provided by Grid cannot be obtained before. The core of Grid is that Grid breaks through various forced constrains on computing resources. It can make people use this computing resources with a new more free and more convenient way to resolve more complicated problems. Finding answers from gird is grid computing, which is a generalized definition. The narrow sense definition of grid computing means organizing distributed computers to resolve complicated science and engineering computing problems cooperatively.

Within the DataGrid project funded by the European Union, an experiment aiming to demonstrate the use of grid technology for remote sensing applications has been carried out. The first results obtained can be found from the paper by Giovanni *et al.* (2003). Aloisio *et al.* (2000) presented and discussed an architecture that allows transparent access to remote supercomputing facilities from a web gateway. The implementation exploits the Globus toolkit and provides users with fast, secure and reliable access to parallel applications. They showed the usefulness of their approach in the context of Digital Puglia, an active digital library of remote sensing digital data. The paper by Aloisio and Cafaro (2003) presented an overview of SARA/Digital Puglia (Synthetic Aperture Radar Atlas), a remote sensing environment that shows how grid technologies and high performance computing can be efficiently used to build dynamic earth observation systems for the management of huge quantities of data coming from space missions and for their on-demand processing and delivering to final users. SARA/Digital Puglia is a grid-enabled, high performance digital library of remote sensing images, developed in a joint research project with CACR/Caltech, ISI/USC and the Italian Space Agency.

Dobashi *et al.* (2003) proposed an image-based modeling of clouds where realistic clouds were created from satellite images using metaballs. The intention of the paper was for applications to space flight simulators, the visualization of the weather information, and the simulation of surveys of the earth. In the proposed method, the density distribution inside the clouds was defined by a set of metaballs. Parameters of metaballs, such as center positions, radii, and density values, were automatically determined so that a synthesized image of clouds modeled by using metaballs was similar to the original satellite image. They also proposed an animation method for clouds generated by a sequence of satellite images taken at some interval. The usefulness of the proposed method was demonstrated by several examples of clouds generated from satellite images of typhoons passing through Japan.

There are three categories of use of remote sensing data, i.e., processing, analysis, and thematic information modelling and determination. Remote sensing processing involves calibration, transformation, etc. The classification and segmentation etc. from mathematical algorithms are the remote sensing data analysis. It is the thematic information modeling and determination to retrieve in-direct information from remote sensing data via physical models. In this paper, we focused on the remote sensing

information modeling and determination using Grid computing platform. We have underdone the experiments using remotely sensed images for thermal inertial modeling. First, we will explain the Grid testbed used in the experiments. Following the results, we will discuss the possibilities of use of Grid computing for remote sensing information modelling. Finally, the conclusion and further development will be addressed.

2 Grid Computing Testbed

Because of largeness of remote sensing data, it will take a lot of time to process them. It is necessary to have a more efficient processing system to do it. Correspondingly, one of the Grid project-Condor can play a role of sparing time, improving work efficiency (*http://www.cs.wisc.edu/condor/overview*), which will change our image processing to a large extent, so that the remote sensing can fully exerts its important role of rapid and instantaneous monitoring.

Condor is one of the grid projects in the world. It is the product of the Condor Research Project at the University of Wisconsin-Madison. Compared with the High-Performance Computing (HPC), Condor provides a High-Throughput Computing (HTC). What the Condor cares is not the number of computed floats in a second, but how many workloads can be done in a period of time. The core of Condor is efficiently using of available resource (*http://www.cs.wisc.edu/condor/htc.html*). Condor is a specialized workload management system for compute-intensive jobs. Like other full-featured batch systems, Condor provides a job queuing mechanism, scheduling policy, priority scheme, resource monitoring, and resource management. Users submit their serial or parallel jobs to Condor, Condor places them into a queue, chooses when and where to run the jobs based upon a policy, carefully monitors their progress, and ultimately informs the user upon completion (*http://www.cs.wisc.edu/condor/description.html*). Condor has two powerful means, one is that it can make available resources more efficient by putting idle machines to work; the other is that it can expand the resources available to users.

Our Grid computing testbed is built on Windows 2000 operating system. It is composed of five Pentium 4 PCs. Three of them have memories of 512MB with a 2.6GHz frequency of central processing unit (CPU). The other two have memories of 256MB with a CPU frequency of 2.0GHz. They are connected by local area network (LAN) whose velocity is 100.0Mbps. The installation and configuration of release 6.4.7 of Condor software has been performed on all mentioned PCs by following the instructions reported in http://www.cs.wisc.edu/condor/manual/v6.4. Vanilla environment was adopted in our test. Vanilla is best suitable for Windows platform, but it has no checkpoint and remote system call, two of which are two important exceptional features of Condor. Vanilla environment has no shared file systems, but provide a default file transfer mechanism (see Condor manual at http://www.cs.wisc.edu/condor/manual/v6.4), so lots of files should be transferred to the executable machines before running. The machines in our Condor pool are Hu, Zhong, Wang, Jennyjordan, and Irsa-cgy. Hu is the Condor Manager and the submitter is Irsa-cgy. Irsa-cgy is also used as the single machine for data processing.

3 Remote Sensing Information Modelling

Our goal is to test the feasibilities of using Grid Computing (Condor system) for Earth surface geophysical parameters determination from remote sensing data, and to examine the work efficiency.

The concept of thermal inertia was introduced to try to take into account both the specific heat and the thermal conductivity in identifying a quantity that would indicate the reluctance of the surface of the Earth (or whatever) to respond to a given heat input. That is, large thermal inertia gives small amplitude oscillations/variations in the surface temperature over a 24-hour period. Let us consider initially the effects of specific heat and thermal conductivity separately. For a given heat input, a large specific heat, c, leads to small temperature changes and this is indicative of a large thermal inertia. However, for a given heat input, for a large value of the thermal conductivity (i.e. for a good thermal conductor) more heat is taken away to lower layers and so the surface temperature rise is less. That is, a large thermal conductivity, K, leads to more heat conducted below surface, which in turn leads to a smaller surface temperature rise and this is also indicative of a large thermal inertia.

What is actually done, therefore, is to define thermal inertia

$$P = \sqrt{K \rho c} \tag{1}$$

where ρ = density. The ρ is included because what matters here is the specific heat capacity per unit volume whereas c, the specific heat capacity, is usually defined as the heat capacity per unit mass (Cracknell and Xue 1996).

Xue and Cracknell (1995) have developed an operational thermal inertia model. This model will be used in the experiments. The Earth surface albedo and daily maximum surface temperature difference are needed to calculate the thermal inertia values. The Earth surface albedo can be derived from visible band remote sensing data and surface temperature can be determined from thermal band data. Two overpass remote sensing data can be used to calculate the daily maximum surface temperature difference. The Advanced Very High Resolution Radiometer (AVHRR) data of daytime at 14:30pm and nighttime at 2:30am were obtained from NOAA (National Oceanic and Atmospheric Administration) satellite. The location of images (84°8'52.63"E - 96°9'29.82"E, 38°7'21.37"N - 32°1'52.11"N) is at the border between Qinghai and Xinjiang where a heavy earthquake occurred on November 14, 2001. The image size is 3000 x 3000.

4 Results and Analysis

Our major tasks of the experiments are testing the working situations of various divisions of images and various sized images running in a single machine and in Condor pool. Firstly, it will mainly for testing the feasibility of using Condor pool to process images and examine how many separated parts of an image has the best work efficiency for Condor pool to run. An image was processed in a single machine whose name is irsa-cgy and the result was recorded. And then the image was divided into 3, 4, 5, and 10 parts in turn. Those separated images were submitted to Condor pool to

process respectively, each of them was submitted to Condor pool 10 times, and the results were collected. Secondly, based on the above result, the relationship between image size and the work efficiency of Condor pool would be tested and established. Remote sensing images of 1200 x 720, 1200 x 1000, 3000 x 720, 3000 x 810 and 3000 x 1800 were divided into 3 parts respectively, and then submitted them into Condor pool to run. Each of them was also submitted 10 times.

Table 1. Returned and compiled results from Condor system

	Task 1	Task 2	Task 3
Postpone time (second)	16	16	33
Executing time (second)	18	18	14
Total time (from submitting)	34 seconds	34 seconds	47 seconds
Executable machine	Jennyjordan	wang	zhong

Table 1 shows a sample result returned from Condor pool, which records the total information of Condor pool's working including submitted time, executable time, terminal time, machines for executing tasks, and problems occurred during this process. It is feasible to using Condor to process images.

Figure 1 shows the average total time curve and average postpone time curve of tasks running in our Condor pool from 3, 4, 5 and 10 parts of an image with size of 1200 x 1000. Our tests show that it has the best work efficiency to separate an image into 3 parts to run in Condor pool. Figure 1 shows that the average postpone time and the average total running time is in a positive proportion, and the postpone time accounts for a lot in the total time.

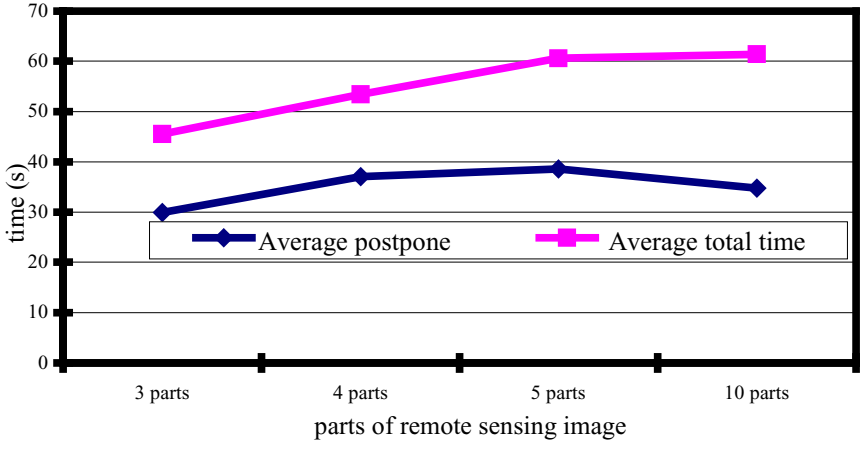


Fig. 1. Average postpone time and average total time curves of the different parts of image with size of 1200 × 1000

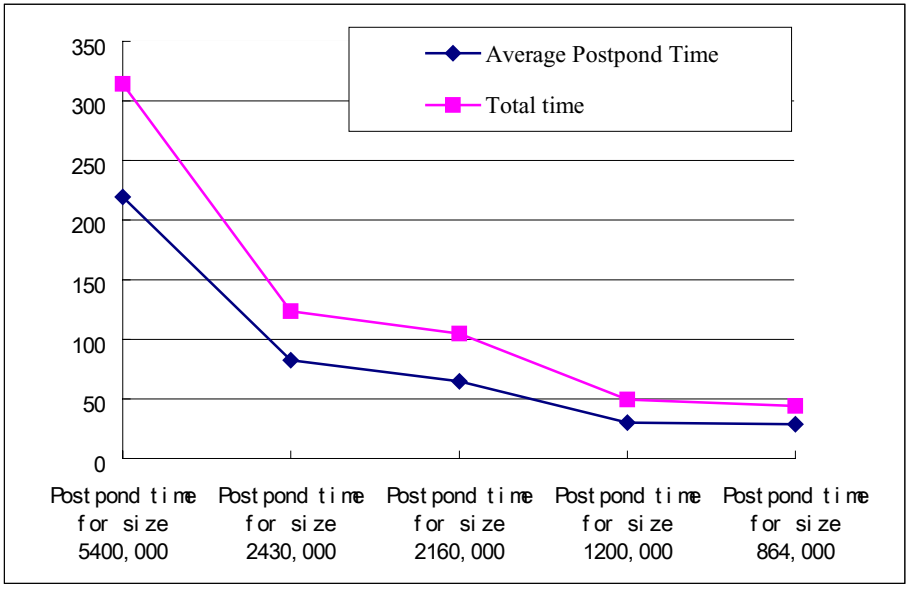


Fig. 2. Returned and compiled results from separated 3 parts of different sized remote sensing images

Table 2. Several parameters to the different sized images

$\times 10^4$ (time unit s)	Time from Condor	Time from S.M[1]	Accelerate index	Postpone/total
A.P[2] for size 540	219.4			
Total time for 540	314.4	181	0.5757	0.69783715
A.P for size 243	82.466667			
Total time for 243	123.5	84	0.680162	0.667746289
A.P for size 216	64.733333			
Total time for 216	104.7	73	0.69723	0.618274435
A.P for size 120	30.1			
Total time for 120	49.4	39	0.789474	0.609311741
A.P for size 86.4	28.833333			
Total time for 86.4	44.1	30	0.680272	0.653817082

[1]: S.M=Single Machine; [2]: A.P.= Average postpone time

Figure 2 shows the returned results of different sized image submitted into Condor pool. It indicates that the most suitable size images for Condor pool exists. To our five machines Condor pool, the best efficiency image size is 1200×1000. But, even the best one, its total running time is longer than the time from a single machine. Table 2 shows the accelerate index which can better indicate system efficiency (Chen and Li 2003). In our system, accelerate index has not exceed 1, which is a very low efficiency. Table 2 also shows the postpone time accounts for more than 60 percent of total time. One important factor to influent Condor's work efficiency is its long postponing time. Its major possible causes are:

- Transfer files are large and web velocity is low. Generally speaking, remote sensing image data are large, so Condor system needs to transfer many different files to the executable machines before running, and then the executable machines will transfer result files back to the submitted machine. During this process, Condor need to compress files at the submitted machine before transmission, and then recompress files at the executable machine before running, thus, the limited web band will limit file transfer, and then influence work efficiency. Besides, Condor needs to move executable file to the remote executable machines. Which also adds workload to the web band.
- Machine numbers in Condor pool. Relatively, the more the machines are in Condor pool, the less files the Condor system needs to transfer to each executable machine. When task number is less then the machine number, Condor will allocate all of tasks to machines for one time, so it is no necessary to check whether there are available machines and transfer files for second time. Which decreases the postpone time and increases the work efficiency. But it is not the case that the more the machines are, the better the condor system is. Under the circumstance of certain web band, more files that need to move also increase put-off time (Chen and Li 2003).
- It takes time that Condor Manager checks whether the remote machines' disk is enough to accommodate files and Condor system's inner message communications.

Condor focus on the HTC, not the HPC, so it is natural that the postpone time is longer in Condor.

5 Conclusion and Further Development

It is feasible to use Grid computing system such as Condor to process remote sensing images. And if the postpone problem can be resolved, the work efficiency of Grid systems will be high. Even with so many problems, it is a good thing that Grid systems do many things for you during all of us are in sleep. Our next major task will concentrate on realizing an arithmetic that can read and divide remote sensing images based on image size and number of machines in Grid system automatically, and transfer results back to the submitted machine as a whole data file.

We can do nothing about web band, but we can increase our machines in Grid system by connecting machines of our two offices together. Then we will do the remote sensing images on Grid system, and the results can be compared with what we have obtained, so that discovery and resolve more problems.

Acknowledgement. This publication is an output from the research projects "CAS Hundred Talents Program", "Digital Earth" (KZCX2-312) funded by Chinese Academy of Sciences and “Dynamic Monitoring of Beijing Olympic Environment Using Remote Sensing” (2002BA904B07-2) funded by the Ministry of Science and Technology, China.

References

1. Aloisio G; Cafaro M., 2003, A dynamic earth observation system. PARALLEL COMPUTING 2003, Vol 29, Iss 10, pp 1357-1362.
2. Aloisio, G; Cafaro, M; Falabella, P; Kesselman, C; Williams, R., 2000, Grid computing on the web using the Globus toolkit. Lecture Notes in Computer Science, Vol. 1823, 32-40.
3. Chen Fulong, and Li Ming dong, 2003, Parallel Processing Based on LAN, *Journal of Sichuan Teachers College (Natural Sciences)*, Vol.24, No.1: 59-62.
4. Cracknell, A. P. and Xue, Y., 1996, Thermal Inertia Determination from Space - A Tutorial Review. *International Journal of Remote Sensing*, 17(3), 431-461.
5. Dobashi, Y., Nishita, T., Yamashita, H., and Okita, T., 2003, Modeling of Clouds from Satellite Images Using Metaballs. www.grid.org
6. Foster, I. and Kesselman, C., (Editors). 1998, The Grid: Blueprint for a new Computing Infrastructure. (*Morgan Kaufmann Publishers*, Inc.)
7. Giovanni, N. A., Luigi, F. B., and Linford, J., 2003, Grid technology for the storage and processing of remote sensing data: description of an application. In Proceedings of the society of photo-optical instrumentation engineers (SPIE), Vol. 4881, 677-685, 2003.
8. Xue, Y. and Cracknell , A. P., 1995, Advanced thermal inertia modeling. *International Journal of Remote Sensing*, Vol.16, No.3, 431-446.
9. An Overview of the Condor System, Condor Team, 2003, available from http://www.cs.wisc.edu/condor/overview
10. High-Throughput Computing, Condor Team, 2003, available from http://www.cs.wisc.edu/condor/htc.html
11. Condor Version 6.4.7 Mannual, Condor Team, 2003, available from http://www.cs.wisc.edu/condor/manual/v6.4
12. What is Condor? Condor Team, 2003, available from http://www.cs.wisc.edu/condor/description.html

Load Analysis and Load Control in Geo-agents

Yingwei Luo, Xiaolin Wang, and Zhuoqun Xu

Dept. of Computer Science and Technology, Peking University, Beijing, P.R.China, 100871
lyw@pku.edu.cn

Abstract. Geo-Agents is an agent-based distributed Geographic Information System (GIS), which is a multi-agent system and consists of four types of GIS agent: GuServer, Facilitator, interface agent and GIS function agent. For a multi-agent system, there contains many distributed agents in different environments, which form a complex distributed system. In order to improve the performance, load control should be a keystone in a multi-agent system. Firstly, the load problems in Geo-Agents are analyzed, and then two load control strategies are discussed in detail.

1 Geo-agents: An Introduction

Geo-Agents is an agent-based distributed GIS (see Figure 1)[1][2]. Geo-Agents consists of four types of GIS agent: Facilitator, Interface agent, GIS function agent and GuServer.

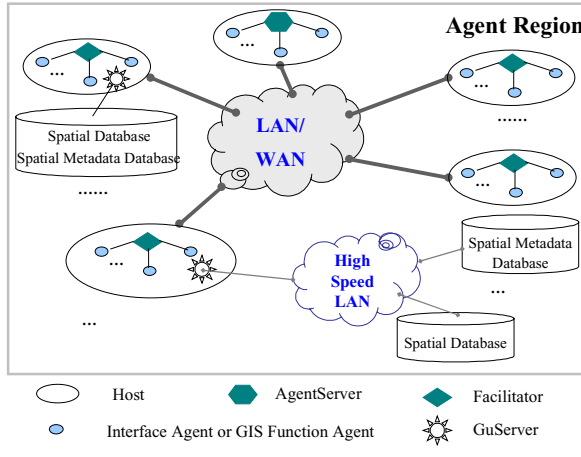


Fig. 1. Geo-Agents System

Facilitator is the manager of Geo-Agents. The functions of Facilitator include registering available GIS agents, searching for practicable GIS agents, managing all active agent instances, coordinating communication and coordinating cooperation.

GIS function agent encapsulates spatial querying, spatial processing or spatial analyzing services. The encapsulated services in GIS function agent may come from dif-

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 997–1003, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ferent existing GIS platforms. Each GIS function agent can complete a same type of problem. According to the features of GIS, GIS function agent is classified into another two types: basic function agent and domain-oriented function agent. Basic function agent completes basic GIS services, such as spatial data search, spatial data access, network analysis, overlay analysis, buffer analysis and so on. Domain-oriented function agent is responsible for application tasks in various domains, and can be constructed by domain-oriented model and used generally in one domain.

Interface agent provides interfaces for users or applications to hand task. Geo-Agents provides GeoScript, an agent manipulating language to describe GIS tasks [2]. When solving a practicable problem, users or applications can simply use GeoScript statements to describe the task, and then hand the statements to interface agent. Interface agent has a GeoScript interpreter and can disassemble the task to subtasks autonomously, recruit GIS function agents to complete the task concurrently.

GuServer is in charge of spatial information accessing services, which manages spatial information and spatial metadata in spatial databases.

In Geo-Agents, GIS agent is reactive agent, and every GIS agent consists of five units: control subsystem, functional subsystem, communication subsystem, human-computer interface and data resource. GIS agent is not only able to carry out its own task independently, but also communicate with other agents, exchange information and cooperate with others.

"Agent Region" mode is adopted to control the distributed scenario for Geo-Agents. An "Agent Region" consists of one or more hosts, which must be installed with a Facilitator (and or other GIS agents). There are many Facilitators in an "Agent Region". Different Facilitators can cooperate to control and coordinate every GIS agent to run correctly, and hold the distributed controls of the whole system.

In an "Agent Region", there is one and only one Facilitator that will be configured as AgentServer. Facilitator is used to manage and coordinate agents inside one "Agent Region". Besides the functions of Facilitator, AgentServer serves as a bridge among different "Agent Regions". That is to say, an agent in one "Agent Region" can only communicate with agents in other "Agent Regions" through AgentServer. Of course, an authorization is required. AgentServer manages a table to register other "Agent Regions" that are authorized with itself each other. Between two directly authorized "Agent Regions", agents between them can communicate via the coordination of their AgentServers. The authorization relation in Geo-Agents can be passed one by on: If a directly authorized AgentServer chain can be found for two "Agent Regions" that have no direct authorization, agents between the two "Agent Regions" can communicate via the coordination of the AgentServer chain.

A multi-agent system contains many distributed agents in different environments, which form a complex distributed system. In order to improve performance, load control should be a keystone in a multi-agent system. The existing researches imply that load control can improve performance of distributed system very well [3-6]. In multi-agent system, agent is the minimum execution unit, and the load control is to dispatch all subtasks of a complex task to different agents in different hosts averagely [7]. In the following two sections, the load problem and load control strategy in Geo-Agents are introduced.

2 The Load Problem in Geo-agents

Generally, the load problem a distributed system faces may involve two aspects: host load and network load [3]. In Geo-Agents, host load indicates the number of agents resided on one host, and network load contains two parts: the communication among agents and the transmission of spatial data. For any distributed GIS system, the transmission of spatial data is the primary network load. Spatial data is not transmitted directly in a message in Geo-Agents. Spatial data is encapsulated as a Geo-spatial object, and the transmitted data in a message is the reference of the Geo-spatial object. Only when spatial data is really needed, the transmission of spatial data takes place. The transmission of spatial data is through the reference of Geo-spatial object, not through the message among agents [2].

In order to achieve a high performance, it is necessary to control the load for a distributed system. Generally, load control can be accomplished from two aspects: reducing load and balancing load [4][5].

In Geo-Agents, the number of agent for completing a special task is fixed, it is impossible to reduce host load. But we can balance the host load through controlling the number of agent in different hosts.

Network load is a serious problem in Geo-Agents, which is because of that: 1) the collaboration among agents will bring a lot of communication cost; 2) massive spatial data is transmitted in network. The communication mechanism in Geo-Agents is point-to-point direct communication, so it is hard to balance network load among different hosts. But we can reduce the network load through some other effective measures.

Reducing network load can be obtained from two aspects: 1) reducing the quantity of each communication, which must be considered when designing system; 2) reducing communication times, especially massive spatial data communication times [5][6]. In Geo-Agents, the distribution of spatial data and the collaboration among GIS agents make communication absolutely necessary. The only way is to let massive spatial data communication take place in a same host.

In those ways, it is easy to find that a conflict happens between reducing network load and balancing host load. In Geo-Agents, because of the hugeness of spatial data, reducing network load is principal. Balancing host load will be considered only after network load is accordant.

3 Load Control in Geo-agents

In Geo-Agents, load is mainly controlled when creating GIS agent. There are two control strategies: agent scheduling strategy and peer host strategy.

3.1 Agent Scheduling Strategy

3.1.1 The Principles to Create an Agent. GIS Agent is designed as reusable GIS agent and ordinary GIS agent [2]. Reusable GIS agent means its result can be used again.

When creating a reusable GIS agent, if its type and task are same with an existing active reusable GIS agent (in Geo-Agents, there are some determinate reusable GIS agent types, and the task of each type is comparable [2]), it is unnecessary to create a new reusable GIS agent. Otherwise, the requirement to create a reusable GIS agent is same with creating an ordinary GIS agent.

When creating an ordinary GIS agent, there are some different requirements because of considering network load:

(1) A GIS agent must be created in an appointed host.

(2) A GIS agent must be created in local host by Facilitator.

(3) A GIS agent will be created in appropriate host (for example, the host where spatial data locates).

(4) A GIS agent can be created in any host.

In the first two cases, the host load they bring is inevitable, and there is no proper method to reduce or balance relative host load. But in the last two cases, it is necessary to control the creation of agent because of the randomicity, otherwise, there will have too many agents in one host.

During the creation of GIS agent, the network speed must be considered. In Geo-Agents, some assumptions are given: accessing an agent in a same "Agent Region" is faster than in other "Agent Region"; accessing data resource in a same "Agent Region" is faster than in other "Agent Region"; the priority of accessing data resource is higher than accessing agent. For example, supposing there are two "Agent Regions", the data resource is in "Agent Region1", and a user in "Agent Region2" wants to create an agent to process the data resource in "Agent Region1". At this situation, the agent should be created in "Agent Region1".

According to above requirements, the following principles must be obeyed when creating an agent:

(1) An agent should be created in the host where data resource it processes locates if possible.

(2) If an agent needs to process several data resources, the agent should be created in one of the hosts where data resources locate if possible.

(3) If an agent couldn't be created in the host where data resource locates, it should be created in one of the "Agent Regions" where data resources locate if possible.

(4) An agent should be created in the same "Agent Region" with its requestor if possible.

(5) In an "Agent Region", an agent should be created in the host where its load is light.

3.1.2 Agent Scheduling Algorithm. Scheduling agents when creating agents is a powerful means to control host load and network load. The goal of scheduling agents is: 1) making the most of CPU of all hosts, in another word, eliminating the situation that some hosts are too busy and some other hosts are idle at a same time; 2) Shorting the average response time of all tasks and reducing the network transmission as greatly as possible.

When designing an agent scheduling algorithm, three problems must be considered firstly [5]:

(1) Determinate or heuristic algorithm: determinate algorithm is adopted only when the actions(such as computing requirements, data requirements, communication requirements and so on) of all agents are foreseeable. But if the load of a system is unpredictable, a heuristic algorithm is needed. Heuristic algorithm means that the host allocation is instructed by experiential rules and heuristic information during the working progress.

(2) Distributed or centralized algorithm: centralized algorithm must gather all global information to give a decision, but distributed algorithm can give a decision only by some partial information. Gathering all global information in centralized algorithm brings heavy burden to the decision-making host, which will become the bottleneck of a system.

(3) Optimized or hypo-optimized algorithm: centralized algorithm can obtain optimized result, but the cost is higher than hypo-optimized result, because more information and complete processing are needed.

According to the agent creating principles and the agent scheduling goals, a heuristic, distributed and hypo-optimized agent scheduling algorithm is designed:

In Geo-Agents, agent scheduling is completed by AgentServer. When creating an agent, a request is sent to AgentServer to ask it find an appropriate agent type. So AgentServer maintains a host-table that records the number of created agents in each host. If an appropriate agent type can be found only in one host, the agent should be created in that host, and the number of created agents in that host increments by 1; if an appropriate agent type can be found in several hosts, after considering the agent creating principles, AgentServer will select the host where the number of created agents is minimum to create the agent, and the number of created agents in that host increments by 1.

We call this scheduling algorithm a heuristic, distributed and hypo-optimized algorithm because that: 1) it is incapable to predict that how many agents will be created in the system and how many agents are executing at a same time; 2) the scheduling criterion is the number of created agents, not the number of executing agents in one host; and 3) the scheduling criterion is obtained at the volley and it doesn't bring additional burden for AgentServer.

3.2 Peer Host Strategy

When it is foreseeable that the host load in one host is too heavy, the peer host strategy is adopted to reduce the host load: connecting the host with some other hosts via high speed LAN, which have same software and hardware configuration with the host

and can complete same tasks with the host. The host is called as peer main host, and other hosts are called as peer secondary hosts. All main and secondary hosts form a peer host group (see Figure 2). When creating an agent in main host, if the host load of main host is too heavy, the agent can be created in anyone of the secondary hosts. Of course, in order to balancing the host load of all peer hosts, the host where the host load is lightest is selected.

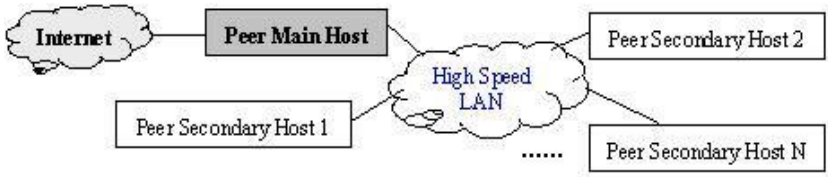


Fig. 2. Peer Host Group

In Geo-Agents, the host crying for "peer host" is the hosts where GuServer locates. GuServer maintains one or more spatial databases and spatial metadata databases [2], and is responsible for managing and accessing spatial data and spatial metadata. Spatial information is centralized in the host where GuServer locates. According to the agent scheduling strategy, all agents who need GuServer to manage and access spatial information will be created in the host where GuServer locates. So the host load of the host where GuServer locates will be too heavy, and the performance of the system will be serious low. Of course, enhancing the software and hardware configuration of the host where GuServer locates can improve performance of the system, but cannot reduce the host load of that host. "Peer host" will play an important role here (see Figure 3).

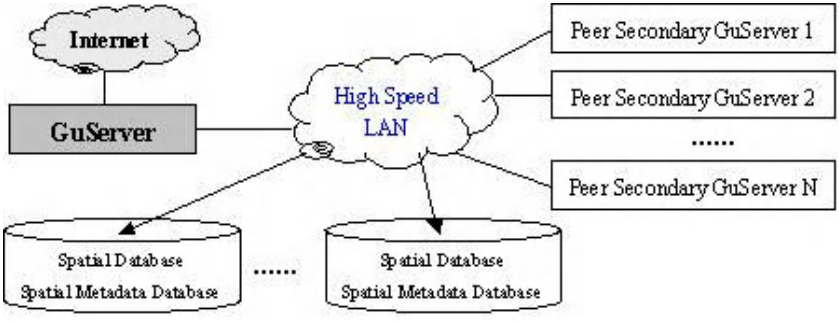


Fig. 3. The Peer Host Group of GuServer

Here, the host installed GuServer is main host, and the hosts installed peer secondary GuServer are secondary hosts. All GuServers manage and access same spatial databases and spatial metadata databases. When creating agents who will use spatial information managed by the GuServer, they can be created averagely in the peer host group of GuServer.

4 Conclusion

The performance is an important factor to evaluate a distributed system, and load control is a powerful means to improve performance. Geo-Agents is an agent-based distributed system, and the granularity of load control is an agent. In this paper, the load problems in Geo-Agents are analyzed and two load control strategies, named as agent scheduling strategy and peer host strategy, are introduced to balance the host load and network load. The performance of Geo-Agents is improved very well by these two strategies [2].

Acknowledgement. This work is supported by the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20020001015; the National Grand Fundamental Research 973 Program of China under Grant No.2002CB312000; the National Science Foundation of China under Grant No.60073016 and No.60203002; the National High Technology Development 863 Program under Grant No. 2002AA135330, No. 2002AA134030 and No. 2001AA113151; the Beijing Science Foundation under Grant No.4012007.

References

1. Luo Yingwei, *et al*: The Model of Distributed GIS-oriented Multi-agent System (in Chinese), Acta Scientiarum Naturalium Universitatis Pekinensis, 38(3): 375-383(2002).
2. Luo Yingwei: The Study on Agent-based Distributed GIS (in Chinese), [Ph.D. Dissertation], Peking University, Beijing (1999).
3. Zhou S.: A Trace-Driven Simulation Study of Dynamic Load Balancing, IEEE Trans. on Software Engineering, 14(9): 1327-1341(1988).
4. Shivararri G, *et al*: Load Distributing for Locally Distributed Systems, IEEE Computer, 25(12): 33-44(1992).
5. Lin Hwa-chu, Raghavendra C S.: A Dynamic Load Balancing Policy with a Central Job Dispatch, IEEE Trans. on Software Engineering, 18(2): 148-157(1992).
6. Luo Zhigang, *et al*: Study on CORBA Load Balancing (in Chinese), Computer Science, 27(10): 31-35(2000).
7. Myung-Hee Jo, *et al*: The Design and Implementation of Dynamic Load Balancing for Web-based GIS Services, The proceedings of 22nd Asian Conference on Remote Sensing, P5-9, Singapore (2001).

Challenges in Transmission Line Modeling at Multi-gigabit Data Rates

Vadim Heyfitch

47-14 Cogswell Avenue, Cambridge, MA 02140, USA
heyfitch@ieee.org

Abstract. A signal trace on a PCB is modeled as a transmission line characterized by four parameters: R, L, C, and G, specified per unit length. A field solver is usually used to determine these frequency dependent parameters for a given trace geometry. This paper describes most recent trends in modeling PCB (Printed Circuit Board) interconnect traces in a context of current design requirements for multi-gigabit data interfaces.

1 Introduction: What Matters and Why?

No matter what signaling scheme is being designed — whether it's a common clock, source synchronous, or serial signaling — there are two main figures of merit: the noise margin and the timing margin.[1] It is the challenge of a signal integrity engineer to design the physical implementation of a bus to ensure that both noise and timing margins meet minimal requirements of the receiver.

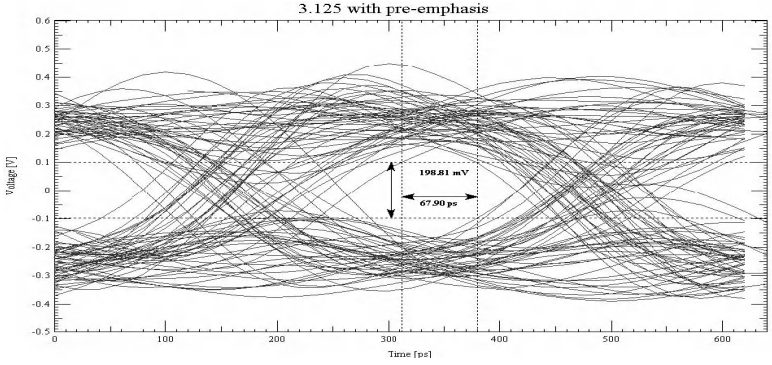


Fig. 1. Figures of merit: timing and noise margin

A receiver data sheet specifies the setup and hold times, as well as the switching threshold and the noise guard band around it. The signal shape at the receiver pad determines these margins in simulations. What is it that affects the signal shape as it arrives to the receiver's end of the trace? After leaving the driver pad, the signal propa-

[1] A system designer can define timing and noise margins in many different ways. Shown in Fig.1 is just one way of doing it.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1004–1011, 2004.
© Springer-Verlag Berlin Heidelberg 2004

gates through the package; then it goes off on to the board where it spends most of the time flying on a trace – whether it's a top or bottom layer microstrip, or a stripline on one of the internal layers. Flying along the copper trace the signal gets attenuated, partly due to the losses in the copper trace and reference plane(s) and partly in the dielectric surrounding the trace. Various frequency components are attenuated to various degrees changing the shape of the arriving signal. Thus modeling the frequency dependent loss in traces becomes of great importance in predicting the signaling scheme performance. That's why so much effort has been spent industry wide on improving the PCB trace models and the algorithms making use of the models.

2 What Is a Trace Model?

From a system level viewpoint, a trace is a segment of an overall interconnect model, which also includes a driver, a receiver, their packages, possibly a connector or two, some vias... What is the trace model as shown in Fig.2? Typically a trace is viewed as a transmission line, and is characterized by the p.u.l. (per unit length) parameters R, L, C, and G.

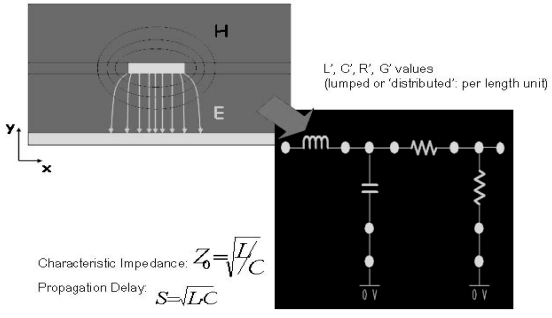


Fig. 2. Transition from physical trace geometry to an equivalent circuit model

Traces on PCBs have relatively low loss.[2] A trace without any loss works as an ideal delay element, and can be represented as a ladder of repeated sections, where each section is a series inductance L connected to a parallel capacitance C to ground. However, when losses become significant, they are accounted for by adding a series resistance R and a parallel conductance G to each section of the ladder model. The former, R, represents conductor loss, i.e. the loss due to the copper resistance, both in the trace and in the reference plane(s). The latter, G, represents the dielectric loss, i.e., loss due to some field energy drained into polarizing the dielectric around the trace, and, sometimes, a direct current leak through the dielectric. Each of these parameters is frequency dependent; it takes on a different value at different frequency. Therefore, one cannot accurately model a trace as a ladder with fixed values of R, L, C, and G. Such a ladder would *not* capture the frequency dependence of these parameters, and

[2] Low loss means R<<Z0.

therefore would *not* exhibit dispersive behavior one observes in reality.[3] Thus Fig.2 should by no means be construed as if a trace is modeled in any particular simulator as a ladder. (Most simulators do better than that.) The ladder representation is used here to merely explain the physical meaning of these parameters. To determine these four parameters at each frequency, a field solver (FS) is used. Below we describe what a FS can and should do.

3 What Are the Important Physical Effects?

There are several physical effects that should be taken into account by a field solver. They are: [4]

- Skin effect
- Proximity, or crowding effect
- Return path resistance
- Frequency dependent dielectric constant and dielectric loss in PCB materials.

Ultimately, we would like a field solver to capture these effects as accurately as possible. Any deviation of the model from the physical reality translates into errors in timing and noise margins as determined in simulations.

3.1 The Skin Effect

When a DC signal propagates through a PCB trace, or through any wire for that matter, the current flows so that it's evenly distributed through the conductor cross-section (see Fig.3). As the signal frequency rises, the magnetic field pushes out of the conductor while the current crowds out toward the wall of the conductor. At even higher frequencies, the current is flowing mainly within the thin layer under the conductor surface, or under the "skin" of the conductor. In the absence of other conductors in proximity, the current distributes itself evenly along the perimeter of the conductor cross-section. This in turn leads to frequency dependence of the inductance and resistance.

3.2 The Crowding Effect

The crowding or proximity effect manifests itself when another conductor is found near the conductor under consideration. In this case, through the distortion of the electromagnetic field, conductors sense each other's presence. This results in the high frequency current bunching up closer toward the side of the neighboring conductor.[5] If this other conductor is a reference plane in the PCB, then there is more current

[3] A more complex ladder model is possible, with many RLC elements in a section. See a paper at NESA, Inc. website. [1].

[4] R and L are affected by 3.1, 3.2, 3.3; G and C are affected by 3.4.

[5] See Eq.5.1 and Fig.5.3 in [2].

flowing along the side of the trace facing the reference plane. If this trace is routed as an asymmetrical stripline, the current tends to lean more toward the closer reference plane.

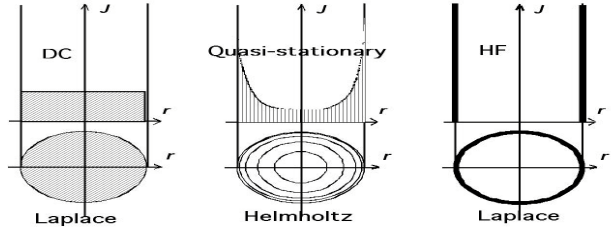


Fig. 3. Current density J as a function of frequency. Current spreads out toward the conductor surface as the frequency rises

3.3 The Return Path Resistance

It's not only the direct current in the trace, but also its return counterpart that faces some material resistance of copper. Previously, trace models in circuit simulators neglected this very real physical effect, treating the reference planes as ideal conductors. The return current density is fairly accurately described by the Lorentzian function. It concentrates more directly under the trace, and tapers off farther away from the trace.

3.4 The Frequency Dependent Dielectric Loss

There are two main mechanisms responsible for the dielectric loss in PCB materials. First, there is some, although very small, amount of direct current leakage through the dielectric even at zero frequency. Second, there is a polarization loss, which can be easily understood if the molecules of FR4, or other PCB materials, are viewed as dipoles. Water, for example, is a very prevalent molecule in FR4; it has a rather high dipole moment, and is responsible for the relatively high loss tangent of FR4 of about 0.02. When put in an alternating electric field, H_20 molecules tend to change their orientation following the electric field, just like a weathercock in a wind constantly changing its direction. It is the signal driver that causes the current to flow in the trace, and therefore causes the high frequency field to move down the trace along with the signal – as one happy cloud. The field, in turn, forces the molecules of the surrounding dielectric to oscillate.[6] While doing so the field loses some energy. Since the signal and the field "around it" are inseparable, the signal loses some of its strength too. The dipole molecules of the dielectric react differently to the external field at dif-

[6] This is exactly how a kitchen microwave works. It generates a narrow band MW field, which causes water molecules in food to oscillate. The kinetic energy of such rotational motion is, in fact, what we feel as "hot."

ferent frequencies. From DC to some high frequencies of about several hundred MHz, these molecules easily follow the external alternating field. However, at even higher frequencies, they find it harder and harder to respond to the external force. As the field frequency increases, the dipole molecules of the dielectric, first, lag behind the field, and then totally ignore it at very high frequencies. Via this mechanism the dielectric molecules drain the energy from the field – but drain it to various degrees at various frequencies. Since a digital signal comprises a bunch of various frequency components, each such signal component gets attenuated differently from others. With various frequency components being attenuated to a different degree, the signal shape changes in a rather convoluted manner. This is why it is very important to have accurate loss tangent data available for a field solver (see more on this issue in section 4 below).

4 How Are These Effects Accounted for in *R, L, C,* and *G* as Extracted by a Field Solver?

At frequencies where the skin depth becomes comparable to the trace's width, the skin effect sets in. The trace resistance bends up from its DC value at about the "knee frequency". For typical PCB dimensions, it happens around 100 kHz to 1 MHz range. Traditionally, SI simulators modeled this skin resistance as the square root of frequency. This is an approximation, however, that is accurate only for a perfectly round wire far away from a reference plane. A typical PCB trace cross section is far from being that. Presently, some circuit simulators take into account this real trace shape as well as the proximity of the reference plane(s) and other (adjacent) traces. This new level of FS sophistication does away with the traditional ω assumption, and results in a truer frequency dependent resistance R. It offers major improvements both around and above the knee frequencies, where, otherwise, estimated resistance may be off by as much as a factor of two. (see Fig.4). Parenthetically, let's mention another effect: the surface roughness of copper traces and planes. It certainly raises the trace resistance by some amount. The RMS roughness of "copper dandruff" can typically be between two and ten micron. In comparison, the skin depth of copper at one gigahertz is about two micron. Therefore, the current's higher frequency components comprising a digital signal do not just smoothly "flow" in the surface layer, but rather hop from peak to peak, thus facing much higher resistance. This issue has been dealt with in microwave design world, but has been mostly overlooked in high-speed digital designs. There is a conspicuous lack of data on this subject. Some engineers use fudge factor to account for this effect.

As far as inductance is concerned, it was typically modeled very accurately at frequencies well above the skin frequency, the higher the better. It was computed at infinite frequency (using BEM – Boundary Element Method), and assumed a constant across the entire frequency range: from DC to infinity. However, at frequencies below the knee, inductance is higher. It can be higher by up to 45% at DC than its value at high frequencies. (See Fig.5)

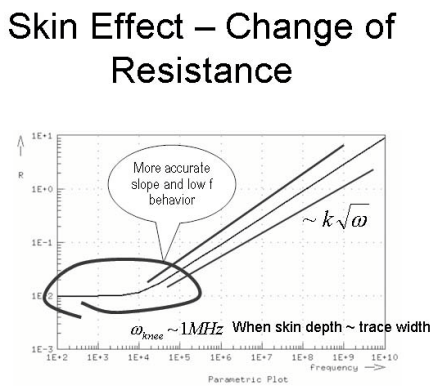


Fig. 4. Accurate field solver does away with traditional assumptions about the resistance's frequency dependence

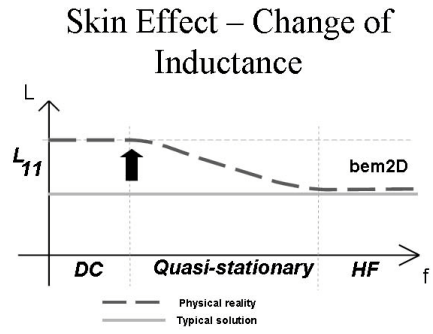


Fig. 5. Due to internal inductance, the DC value can be up to 45% higher than traditionally computed by BEM (at HF limit)

Such is the physical reality. Some circuit simulators successfully take this effect into account. These more accurate inductance values at low frequency result in more accurate modeling of signal shape at the receiving end. As pointed out above, all this translates into more accurate prediction of both the noise and timing margins. Mutual inductance follows the same trend as self-inductance. When modeling capacitance we usually assume capacitance to be independent of frequency. In reality, however, the real part of the dielectric constant is weakly dependent on frequency, and with it the capacitance. Its imaginary part, or the loss tangent, affects the conductance G

Conductance G, the last - but not least of the four p.u.l. parameters - actually becomes the dominant loss mechanism at frequencies above several hundred megahertz. Even "a slow" signal with a rise time of 0.5 ns has a (-3 dB) power density spectrum point at about 700 MHz. At this frequency the dielectric loss is comparable to the loss caused by the skin resistance. Thus including the conductance into the trace model

becomes a must. Conductance is comprised of two parts: the DC and the AC, as detailed in section 3.4 above. The DC component is negligible for all practical purposes, whereas the AC term is to be modeled as accurately as possible. This term is directly proportionate to the dielectric loss tangent. Traditionally, the latter was assumed a constant, and PCB manufacturers provided its value at a single frequency, typically 1 MHz. In that case, conductance rises linearly with frequency, and has no limit at higher frequencies. (top line on Fig.6). However, using this model will create an electrical short on any trace at sufficiently high frequencies.

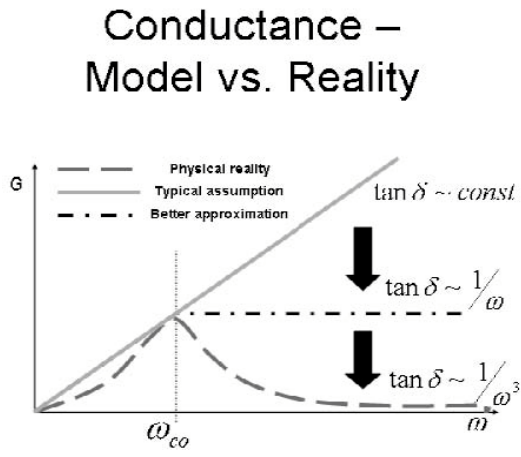


Fig. 6. The ever-growing level of sophistication in conductance modeling: evolutionary view

To counteract this non-physicality, some circuit simulators now clamp G at some frequency, thus again making an assumption about the frequency dependence of the loss tangent (as inverse frequency; see middle line on Fig.6). This is a definite improvement in accuracy. It is known, however, from the material modeling of dielectrics, that asymptotic behavior at very high frequencies should follow the inverse power of three. (Dashed line Fig.6) Signal integrity tools, however, are not in the business of doing molecular modeling of dielectric materials, and are anxiously waiting for the measured dielectric property data to become widely available. In the meantime, some tool vendors are preparing for the day this data becomes readily available by allowing the user to specify arbitrary frequency dependence for the conductance matrix. Thus the onus now is on the user to come up with such data, whether he does measurements in-house or relies on his fab house or any other third party.

5 Summary

In this article we have discussed why accurate modeling of a PCB trace is very important for high-speed design, especially at multi-gigabit data rates. We have identi-

fied what a PCB trace model is. We listed most significant physical phenomena and discussed how they affect the RLCG parameters. Describing each of these parameters, we gave a historical perspective, as well as outlined typical limitations that exist in some commercial Signal Integrity tools. Hopefully, this article will raise the level of awareness of both a high-speed system designer and EDA tool companies, and help them to be mindful of various aspects of modeling PCB and IC package traces.

References

1. Sayre, P.E., Baxter, M.A., Savarino, T.: Development of a New Transmission Line Skin Effect Model for SPICE Evaluations - Simulations and Measurements. Design Superconference (1997); http://www.nesa.com/sknef.html
2. Johnson, H.W., Graham, M.: High-speed Digital Design: a Handbook of Black Magic. Prentice Hall PRT (1993)

MPI-Based Parallelized Model Order Reduction Algorithm

Igor Balk[1] and Sergei Zorin[2]

[1]R3 Logic Inc, 80 Sherman Street, Camridge MA 02140, USA
ibalk@r3logic.com
[2]Parametric Corp., Boston, MA, USA
szorin@ptc.com

Abstract. Progress in integrated circuits and packaging design has made significant changes in the requirements for modeling tools. In this paper described algorithm of parallelization of Krylov subspace model order reduction technique for "full-wave" electromagnetic simulation.

1 Introduction

The topic of model order reduction appears in many disciplines. In control systems, perhaps the most well known approaches are Truncated Balanced realizations and Hankel singular values. Other popular approaches are based on the modal properties of the system, such as the method of Selective Modal Analysis, which is widely used in power systems applications. Unfortunately, most of these methods require $O(n^3)$, where n is the number of unknowns, operations to generate reduced model since they require explicit knowledge of the entire eigenspectrum of the system as in selective modal analysis or the hankel singular values as in truncated balanced realization. In complicated packaging and interconnect problems for which n can be quite large, these methods would require days and gigabytes of memory to compute.

In the area of circuit simulation, asymptotic waveform evaluation has popularized the use of model order reduction. A robust approach for deriving moment matching reduced order models is the Arnoldi process. The Arnoldi process has its origin in eigenvalues computation but was used recently to generate moment matching reduced order models.

In this paper we will give a brief overview of Arnoldi-based passive algorithm for reduction of 'full-wave' electromagnetic models [2, 3] and then introduce MPI-based parallelization strategy.

2 Overview of Krylov Subspace Model Order Reduction

Lets consider a single input – single output linear system, where A is the system matrix

$$A\frac{dx}{dt} = x + bu$$
$$y = c^t x \tag{1.1}$$

A here is m x m matrix and $b, c \in R^m$.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1012–1016, 2004.
© Springer-Verlag Berlin Heidelberg 2004

From (1.1) the relation for the transfer function *H(s)* in Laplace space can be written as

$$H(s) = -c^t (I - sA)^{-1} b \tag{1.2}$$

Transfer function H(s) can be expanded in a Taylor series around zero frequency,

$$H(s) = \sum_{i=0}^{\infty} m_i s^i = \sum_{i=0}^{\infty} -c^t A^{-(i+1)} b s^i \tag{1.3}$$

where m_i, the coefficient of the *i*-th term in the Taylor series, is known as the *i*-th moment of the transfer function. Most commonly used model order reduction algorithms utilize the Arnoldi method of fast generation of orthogonal basis for Krylov subspace defined as

$$K_q(A,b) = span\{b, Ab, A^2 b ... A^{q-1} b\} \tag{1.4}$$

The general idea of this method is to generate *k*-th order orthogonalized Krylov subspace from the *(k-1)*-th order orthogonalized Krylov subspace. The last vector of *(k-1)*-th order from the orthogonalized subspace is multiplied by *A* and then the resulting vector is orthogonalized with respect to the rest of the basis. After *n* steps this algorithm returns a set of *n* orthogonal vectors and we can construct a matrix $V_n \in R^{m \times n}$ and an *n x n* upper Hessenberg matrix H_n such as

$$AV_n = V_n H_n + h_{n+1,n} v_{n+1} e_n^t \tag{1.5}$$

It's easy see that for *k*<*n*

$$A^k b = \|b\|_2 A^k V_n e_1 = \|b\|_2 V_n H_n^k e_1 \tag{1.6}$$

With this relation we can write that moments can be related to H_n by

$$m_k = c^t A^k b = \|b\|_2 c^t V_n H_n^k e_1 = c_n^t A_n^k b_n \tag{1.7}$$

and the n-th order Arnoldi based approximation to H(s) can be written as

$$H(s) \approx \|b\| c^t V_n (I - sH_n)^{-1} e_1 \tag{1.8}$$

corresponding to the steady state representation $A_n = H_n$, $b_n = e_1$ and $c_n = \|b\| V_n c^t$.

And substitution to the representation of original system we will result in following equation:

$$\begin{aligned} V_n^t A V_n \frac{dx}{dt} &= V_n^t x + V_n^t b u \\ \tilde{y} &= c^t V_n^t x \end{aligned} \tag{1.9}$$

And it is easy to see that

$$y \approx \tilde{y} \tag{1.10}$$

3 Guarantee Passive Algorithm for Full Wave Model

Now lets assume that A is frequency dependent and that the transfer function is positive definite for all s with non-negative real part; i.e. *Re(i,H(s)i)* ε for some $\varepsilon > 0$. This kind of problem is quite common in modeling of interconnect circuits or packaging devices at high frequencies when we have to use the 'full-wave' kernel to generate integral equation formulation. The standard Arnoldi based model order reduction algorithm, such as PRIMA, are unable to handle this type of problem, and has been a lot of work done in this field. We will introduce here a passive algorithm for Arnoldi based reduction of such systems.

In order to generate a reduced order model A(s) will be presented in a from:

$$A(s) = \frac{1}{s^p + 1} A_t + \frac{s^p}{s^p + 1} A_t \tag{2.1},$$

where A_t is the power series expansion of A(s) with infinite radii of convergence. And first order infinite dimension matrix will be generated in a way similar to the formulation described in [1].

$$\left[I + \frac{s}{s^p + 1} \begin{vmatrix} \tilde{A}_0 & \tilde{A}_1 & \cdots & \tilde{A}_N \\ -I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & -I & 0 \end{vmatrix} \right] \begin{vmatrix} x_0 \\ x_1 \\ \vdots \\ x_N \end{vmatrix} = \begin{vmatrix} b \\ 0 \\ \vdots \\ 0 \end{vmatrix}$$
$$y = [c^t \quad 0 \quad \cdots \quad 0] \begin{vmatrix} x_0 \\ x_1 \\ \vdots \\ x_N \end{vmatrix} \tag{2.2},$$

where $x_k = s^k x$ and $\tilde{A}_k = A_0^{-1} A_{k+1}$ for $k+1<p$ and $\tilde{A}_k = A_0^{-1}(A_{k+1} + A_{k+1-p})$ for $k+1 \geq p$.

Arnoldi process can be used for the new systems with $\tilde{s} = \frac{s}{s^p + 1}$. And model order reduction algorithm can be formalized as follows:

1. Calculate A_0
2. Set $n=0$ and set some $p>1$
3. Calculate zero order transfer function
4. Increase n by 1
5. Calculate A_n
6. Update p such as $\left| \tilde{s} \max_{i,j} \tilde{A}_{i,j} \right| < 1$ for any A_n and $\tilde{s}$

7. Update A_l for all l from 0 to n
8. Calculate n-th order transfer function
9. Check convergence of transfer function
10. If converge calculate reduced model else back to step 4

Using the properties of power series expansion with infinite radii of convergence and the definition of passivity, it's possible to show that the introduced algorithm will produce a passive reduced model. Proof is based on the fact that we do not have to truncate the expansion of $A(s)$ in order to calculate the n-th order of the reduced model and that expansion of transfer function will converge for any value of $\tilde{s}$.

4 Parallelization Strategy

Is it can clearly can be seen from the matrix structure in Eq. (2.2) the problem has intuitive and easy to observe block structure. First and the easiest way of parallelization would be assigning each block to the separate processor and using one processor to reconstruct the reduced matrix. The disadvantage of this strategy is that it requires explicit computation of submatrixes $A_{i,j}$ and vector matrix products and is quite costly in terms of memory and processor time. The alternative approach does not require explicit computation of all $A_{i,j}$ utilizes the fact that right hand side has only n none zero entries. Each submatrix $A_{i,j}$ is divided in to block and each block is assigned to the processor. Each processor performs computations within it's own block to calculate matrix vector product and then while receiving data from neighbors it computes boundary data for the processors with adjusted blocks. Thus we will get optimal performance as we are doing computations at the same time as waiting for new data.

5 Computational Results and Conclusions

Above-mentioned parallelization was implemented using MPI standard libraries for easy portability. Parallel performance results can be seen on Fig. 1. The test was performed on a cluster of Sun Sparc processors with shared memory. We used transmition line simulation in a frequency range from 1MHz to 7GHz for the test runs.

The results indicate that electromagnetic and signal integrity analysis can be performed faster and it makes simulation of large and complex 3D structures possible. The approach described in this paper requires minimal changes to the serial algorithm. We are currently are working on developing new more effective parallelization strategies.

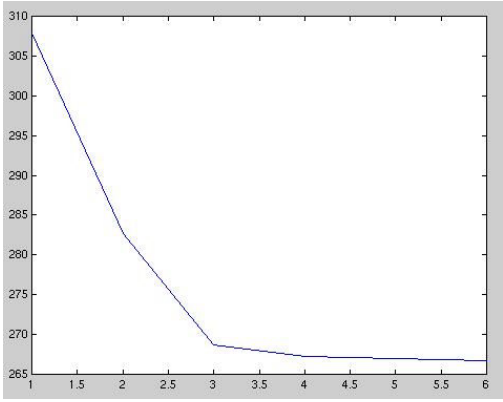


Fig. 1. Computational time vs. number of processors

References

1. Joel R. Phillips, Eli Chiprout and David D. Ling.: "Efficient full-wave electromagnetic analysis via model order reduction of fast integral transforms". 33rd DAC, Las Vegas 1996
2. Igor Balk: "Arnoldy based passive model order reduction algorithm". EPEP 2000, Scottsdale, 2000
3. Igor Balk, Yie He: "Full Wave Electromagnetic Analysis and Model Order Reduction for Complex Three Dimensional Structures". MSM 2001, Hilton Head Island, 2001

3D-VLSI Design Tool

Rajesh Bollapragada

80, Sherman St, Cambridge, MA 02140, rbollapragada@r3logic.com

Abstract. In this paper we present a tool for 3d-integrated circuit technology. The fashion of decreasing feature size and increasing chip dimension has increased the concern of increased interconnect delay and high power-budgets in the design. Adequate design tools are not available which incorporate 3D design capabilities like design rule checking, timing analysis and netlist extraction. We have extended M-Edit, a Layout Editor to incorporate 3D design and verification capabilities. It also features capabilities to perform thermal simulation, layout-vs-schematic checking, and 3D parasitic extraction.

1 Introduction

Today, semiconductor devices are scaling in accordance to Moore's law. The fashion of decreasing feature size and increasing chip dimension has increased the concern of increased interconnect delay and high power-budgets in the design. Hence, new technologies have evolved which can increase the "nearest neighbors", by using three-dimensional integration. Three-dimensional circuit technology has enabled the fabrication of a single chip with multiple device-interconnect layers (wafers) stacked on each other. This technology aids in lowering the interconnect wire length by connecting them vertically.

Not many designers have adapted to this technology as there are not many design tools which incorporate the features available in conventional 2D CAD tools like design-rule checking, netlist extraction, etc. Recently, tools have been introduced, which facilitate drawing of three-dimensional circuit layouts but do not incorporate features like design-rule checking, 3D parasitic extraction etc. There is need for a unified tool, which can seamlessly integrate and facilitate the process of designing and verifying circuits using three-dimensional technology.

1.1 Background

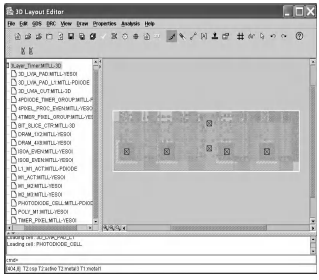
The use of reduced feature size, has helped circuit designers achieve greater frequencies of operation. Smaller geometries are causing a wide variety of deep sub-micron effects like interconnect delays and thermal issues. Three dimensional circuit technologies helps designers to layout a single chip with multiple device-interconnect layers. This technology helps reduce the interconnect delay, but not thermal issues. Sample studies have shown intermediate layers reaching high temperatures in a three-layer process. Three-dimensional integration is an iterative process and there are no definitive rules on the placement or circuit-partition, required to implement the

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1017–1020, 2004.
© Springer-Verlag Berlin Heidelberg 2004

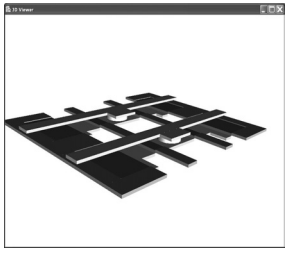
design. There is need for a unified tool, which can help in drawing 3D layouts and which provides tools such as design-rule checks, 3D parasitic extraction, 3D netlist extraction and thermal simulation on the top-level layout. The availability of such a design tool will help in coming up with design methodologies to implement circuits in this new technology.

2 3D Design Tool

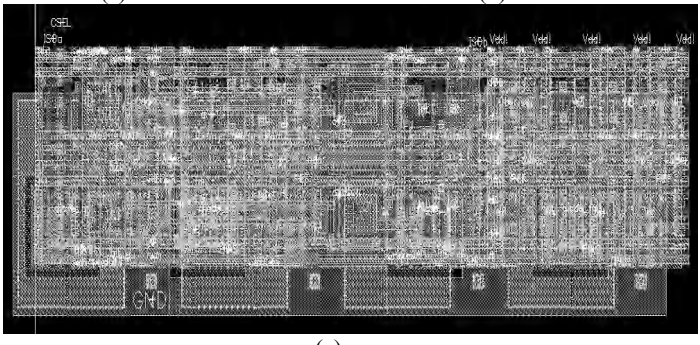
We have extended M-Edit a Layout Editor to incorporate 3D design capabilities. It is being developed in Java and hence is easily portable to various platforms. It is a technology independent layout design tool delivering an extensive set of functions for custom IC design. The user can create custom technology files using the process-design kit.



(a)



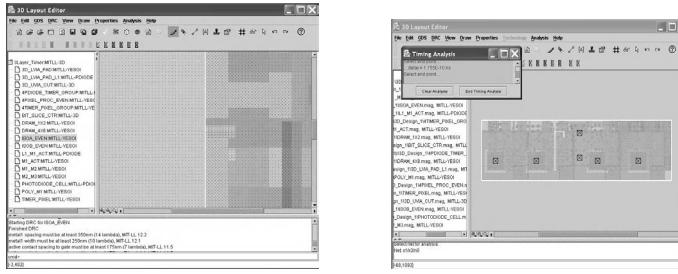
(b)



(c)

Fig. 1. (a) 3D Timer with over 1000 transistors and its (b) 3D view (Partial); (c) Cadence implementation of 3D layout

M-Edit implements a proprietary dynamic design-rule checking algorithm, which reports the complete DRC error on – the –fly for 3D layouts (Fig 2(a)). The tool maintains the electrical connectivity on the top level of the 3D layout; so the designer can extract useful information like delay between components in different layers, and 3D netlist. At any point, the user can access the technology file, to check the effects of the material properties on delay and power (Fig 2(b)).



(a) (b)

Fig. 2. (a) Easy Highlighting for design error correction, (b) Timing analysis on 3D Layouts

3 Design Composition

To implement the circuit using the new technology, the designer needs to create a three-dimensional technology file. This file consists of the individual technology files, which make up the device-interconnect layers and the inter-level vias required to connect layers. The technology file also embeds information regarding material and electrical properties and design rules. A process design kit is provided with the layout tool, which can be used to change the material properties or the design–rules.

The ability of a human to do three-dimensional design is limited by both the depth-free nature of most human interface devices and the inability of a human to keep track of complex three-dimensional interactions. This problem is alleviated as the designer composes the individual layers (wafers) using the corresponding technologies as separate layout files. The choice of the number of layers to partition the design depends on the inter-level via cost factor and other factors such as joule heating of inter-mediate layers, etc. Once the device-interconnect layers are complete, the designer creates an abstract layer in which he can connect the different layers using the inter-level vias. The abstract layer is used as a marker layer to align the different layers and place vias in the right position. Electrical connectivity is maintained across the layers, allowing the designer to examine and extract the nets. The abstract layer acts as the top-level of the design, which can be used to instantiate copies of the design. Upon completion of the 3D Layout, the user can perform tasks such as 3D parasitic extraction and 3D netlist extraction for the complex hierarchical design.

4 M-Edit Applications

Medit has been used to generate layouts for the full-module of a 3D-Imaging Chip (Fig. 3). The layout was done using the MIT Lincoln Labs 3D process. The design involved 3 layers with a total of over 80 million transistors. The first layer contains counters and memory cells, the second layer contains photo-diode cells and the third layer which is the abstract layer combines the above two layers with inter-level vias. We chose to use M-Edit for this application as no other tool had working design rule checking capabilities. The tool was used to extract the netlist for the 3D chip. The

final design was 22 mm x 22 mm in dimension. We were able to successfully tape-out the design and fabricate the chip.

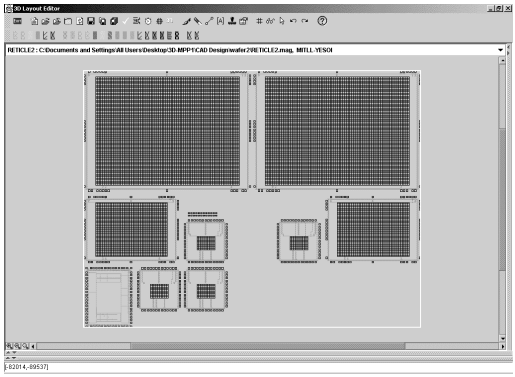


Fig. 3. 22mm x 22mm layout containing over 80 million gates

5 Conclusion

We have presented a design tool, which can be used to take advantage of the three-dimensional circuit technology. It provides the designer with additional functions such as design-rule checking, which is essential to layout the design with minimum errors. The process design kit offers a user-friendly interface to create 3D technology files, so that additional inter-level vias can be added and the necessary design-rules incorporated. M-Edit has been used to design a 3D Imaging chip. Availability of a 3D layout tool will help in the study of other issues like placement and joule heating in the new packaging technology.

Acknowledgements. The paper acknowledges the support from DARPA under the project SBIR Phase II.

References

1. A.C Harter: Three Dimensional Integrated Circuit Layout, *Cambridge University,* Press 1991
2. J. Ousterhout et al. Magic: A VLSI layout system. In *Proceedings of the 21st Design Automation Conference*, pages 152–159, 1984.
3. International technology roadmap for semiconductors, `http://public.itrs.net/`.
4. S.Das, A. Chandrakasan, R. Reif: Three-Dimensional Integrated Circuits: Performance, Design Methodology, and CAD Tools. In *Proc. ISVLSI*, Feb. 2003.

Analytical Solutions of the Diffusive Heat Equation as the Application for Multi-cellular Device Modeling – A Numerical Aspect

Zbigniew Lisik[1], Janusz Wozny[1], Malgorzata Langer[1], and Niccolò Rinaldi[2]

[1] Department of Electronics and Telecommunications Engineering, Technical University Of Lodz, 90-924 Lodz, ul. Wolcza ska 223, Poland, jwozny@p.lodz.pl
[2] Department of Electronics Engineering, University of Naples, „Federico II", Via Claudio 21, 50125 Naples, Italy

Abstract. This work concerns the thermal analysis of solid state devices, namely the solution of the heat transfer problem as part of the coupled electrical-thermal analysis of an IC. We investigate some typical numerical problems which can be found during evaluating of the time dependent solution of the heat diffusion equation for a 3D finite domain with an uniform heat source. Considering a surface heat source, the solutions obtained by means of the Large Time Greens Function and Small Time Greens Function are studied. A simple method to improve the computational efficiency is proposed. A comparison of different approaches is given as well.

1 Introduction

Power dissipation is one of the main problems in the development of ICs, due to the tremendous growth in both integration level and operation speed. Thermal effects can not be neglected in IC design since the phenomenon of the heat conduction markedly affects the electrical behavior of devices. Thermal coupling and self-heating effects should be included. Due to the interdependence of electrical and thermal effects, a coupled electro-thermal analysis is needed. The heat conduction is a 3D phenomenon in nature and any reduction in dimensions may lead to some considerable discrepancies compared to full 3D solution [1].

In any IC, heat dissipating elements comprise single devices or their clusters. Since the thickness of the IC active area is negligibly small in comparison with the substrate, the heat dissipating devices (active cells) can be treated as flat heat sources placed on the top surface of the substrate. In modern ICs, the density of such active cells can reach several millions. The temperature of each active cell can be treated as homogeneous due to very small cell size whereas its magnitude will depend both on the self-heating in the cell and on thermal coupling between adjacent cells. Such an approach to electrothermal simulation allows developing a procedure, in which the electrical and thermal models are split [5]. The electrical part is a combination of temperature-sensitive electrical models representing particular active cells that are treated as points characterized by one temperature, e.g. [5], whereas the thermal model is fully 3D, based on the solution of heat transport equation.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1021–1028, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The devices belonged to one active cell are treated as a point from "electrical" point of view or as the surface heat source from the "thermal" point of view. We will concentrate on its thermal counterpart. It is aimed at the evaluation of 3-D temperature distribution in the whole chip. It can be done by solving the heat transfer differential equation using a numerical or analytical approach. In the case of 3D domain representing the whole IC chip with several heat sources, the numerical approach would require extremely large computational effort giving the temperature distribution in a huge number of nodes whereas its value an a very few points is needed. These values could be obtained much easier from analytical solution if such a solution is available. It is possible in the case of considered thermal model but unfortunately, the analytical expressions have usually the form of infinite series. The number of elements in series must be limited in the computation process if it is to take a reasonable time. It is a crucial problem in the analytical thermal models and therefore it was the subject of consideration in this work. Different forms of the analytical formulas suitable for the 3D heat transport model in IC are investigated with respect to numerical efficiency of the calculation process.

2 The Approach to the Thermal Analysis of IC

Without loosing the generality of the approach, the IC is considered as the chip shown in Fig.1a with active cells represented by N surface heat sources (SHS) with uniform heat generation whose magnitude is determined in the electrical model. Assuming that the system is linear, each source can be treated separately and the temperature at each point can be found as the superposition of temperature rises at the point caused by the different heat sources. The temperature contributions caused by the individual heat sources are given by the solutions obtained for the domain shown in Fig.1b that includes the whole chip and one heat source only.

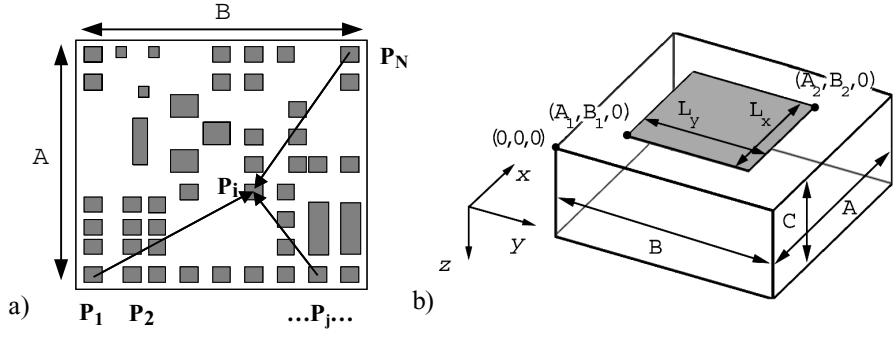


Fig. 1. a) The top surface of a IC with N heat sources b) A domain with a single source

The silicon homogeneous parallelepiped domain (Fig.1a) will be considered for the development of the analytical solutions. These have been obtained for the geometry in Fig.1b under the assumptions that the top and side walls are adiabatic and the bottom one, in contact with an ideal heat sink, is isothermal with a constant temperature T_{hs}

equal to the initial temperature T_0. If we denote the heat generation density g(r,t) as $g(r,t)=g_{0S}$, and the temperature T at a point r as:

$$T(r,t) = T_{hs} + \theta(r,t) \tag{1}$$

The system of equations that has to be solved is as follows:

$$\begin{cases} \nabla^2\theta(r,t) - \dfrac{1}{\alpha}\partial_t\theta(r,t) = -\dfrac{g(r,t)}{k} \\ \partial_x\theta(r,t)\big|_{x=0,A} = \partial_y\theta(r,t)\big|_{y=0,B} = \partial_z\theta(r,t)\big|_{z=0} = 0 \\ \theta(r,t)\big|_{z=C} = 0, \qquad \theta(x,y,z,t=0) = 0 \end{cases} \tag{2}$$

The k and α are the thermal conductivity and diffusivity, respectively. Since the thermal conductivity strongly depends on the temperature, in order to linearize the differential equation, the Kirchhoff transformation over k [3] can be always employed.

The solution of (2) is in a form of continuous function that can determine temperature for each place in the active cell. Since the electrical model requires one temperature for the whole cell only, one considered as the cell temperature that one calculated for the cell center. This way the final i-th cell temperature is determined by the sum:

$$T(r_{c,i},t) = T_{hs} + \theta_1(r_{c,i},t) + \ldots + \theta_i(r_{c,i},t) + \ldots + \theta_N(r_{c,i},t) \tag{3}$$

where $r_{c,i}$ is the i-th cell center coordinate, $\theta_j(r_{c,i},t)$ is the temperature increase in i-th cell caused by the heat dissipation in j-th cell. If the mutual thermal resistance $R_{th\,j,i}(r_{c,i})$ in i-th cell with respect to j-th cells is defined as follows:

$$R_{th\,j,i}(r_{c,i}) = \theta_j(r_{c,i}) / P_j \tag{4}$$

the temperature at the i-th cell can be rewritten in the form, eg. [2,3]:

$$T(r_{c,i}) = T_{hs} + P_1 \cdot R_{th\,1,i}(r_{c,i}) + \ldots + P_i \cdot R_{th\,i,i}(r_{c,i}) + \ldots + P_N \cdot R_{th\,N,i}(r_{c,i}) \tag{5}$$

where P_j power dissipated in j-th cell. Then the temperature for a single active area is characterized by N thermal resistances. Eq. (5) describes the temperature in one active cell whereas the thermal model is to evaluate the temperatures in all the cells determined by N such equations.

If the transient case is considered, the thermal resistance $R_{thj,i}$ is replaced by the thermal impedances $Z_{th\,j,i}(r_{c,i},t)$. The thermal impedance contains all necessary information to describe the thermal properties of a device.

3 Analytical Solutions

Usually, one can solve the same problem using different approaches that leads to different forms of its analytical solution. Let consider two such solutions of the problem described by (2); namely (6) and (7) called the Large Time Greens Function solution

(LTGF) [1,4] and the Small Time Greens Function Solution (STGF) [4,6] (known also as the image method approach), respectively. The both are in a form of complex series but the difference between the formulas are so large that one can expected that their application in the considered thermal model cannot be equivalent from the point of view both computation effort and computation complexity.

$$
\begin{aligned}
Z_{th,LTGF}(r,t) = \frac{2}{kABCL_xL_y}\Bigg\{ & \sum_{p=1}^{\infty}\frac{L_xL_y\cos(\nu_p z)}{\nu_p{}^2}\left(1-e^{-\alpha\nu_p^2 t}\right)+ \\
& 2\sum_{p=1}^{\infty}\sum_{m=1}^{\infty}\frac{L_Y J_m\cos(\beta_m x)\cos(\nu_p z)}{\beta_m\left(\beta_m{}^2+\nu_p{}^2\right)}\left(1-e^{-\alpha\left(\beta_m^2+\nu_p^2\right)t}\right)+ \\
& 2\sum_{p=1}^{\infty}\sum_{n=1}^{\infty}\frac{L_X K_n\cos(\gamma_n y)\cos(\nu_p z)}{\gamma_n\left(\gamma_n{}^2+\nu_p{}^2\right)}\left(1-e^{-\alpha\left(\gamma_n^2+\nu_p^2\right)t}\right)+ \\
& 4\sum_{p=1}^{\infty}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{J_m K_n\cos(\beta_m x)\cos(\gamma_n y)\cos(\nu_p z)}{\beta_m\gamma_n\left(\beta_m{}^2+\gamma_n{}^2+\nu_p{}^2\right)}\left(1-e^{-\alpha\left(\beta_m^2+\gamma_n^2+\nu_p^2\right)t}\right)\Bigg\}
\end{aligned}
\tag{6}
$$

$$
\text{where}: \beta_m=\frac{m\pi}{A},\gamma_n=\frac{n\pi}{B},\nu_p=\frac{(p-0.5)\pi}{C},m,n,p=1,2,3,..
$$

$$
J_m=\sin(\beta_m A_2)-\sin(\beta_m A_1),K_n=\sin(\gamma_n B_2)-\sin(\gamma_n B_1)
$$

$$
\begin{aligned}
Z_{th,STGF}(r,t) = \frac{\alpha}{4kL_xL_x}\int_0^t\Bigg\{ & (4\pi\alpha u)^{-\frac{1}{2}}\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\sum_{p=-\infty}^{\infty}(-1)^p\times 2\,e^{-\frac{(2pC+z)^2}{4\alpha u}} \\
& \left[t_1(-A_1)-t_1(-A_2)+t_1(A_2)-t_1(A_1)\right]\left[t_2(-B_1)-t_2(-B_2)+t_2(B_2)-t_2(B_1)\right]\Bigg\}du.
\end{aligned}
\tag{7}
$$

$$
\text{where } t_1(a)=\operatorname{erf}\left(\frac{2mA+x+a}{\sqrt{4\alpha u}}\right),t_2(b)=\operatorname{erf}\left(\frac{2nB+y+b}{\sqrt{4\alpha u}}\right)
$$

3.1 Efficiency of the LTGF Solution

To check the reliability of the solution, the thermal impedance $Z_{th}(t)$ has been calculated for the center of centered SHS. The domain is described by A=B=175μm, C=100μm and L_x=L_y=5μm but due to the symmetry of the problem, only its quarter is considered to improve the efficiency of the calculation procedure. The different number of terms has been used and four sets of sum limits listed in Table 1 were considered. The number of terms for the single sum over (p) index is Q_1, for the double sums over (m,p) and (n,p) indexes it equals $M_2 \cdot Q_2 + N_3 \cdot Q_3$, and for the triple sum it is $M_4 \cdot N_4 \cdot Q_4$. The total number of terms N_{TOT} is then a sum of these products.

Table 1. Used upper sum limits.

No.	*1*	*2*	*3*	*4*
P_1	225	708	2237	7072
M_2 , N_3	112	239	511	1097
P_2 , P_3	215	469	1013	2158
M_4	83	146	258	457
N_4	83	146	258	457
P_4	160	286	510	908
N_{tot} **[1e+6]**	**1.1**	**6.3**	**35**	**194**

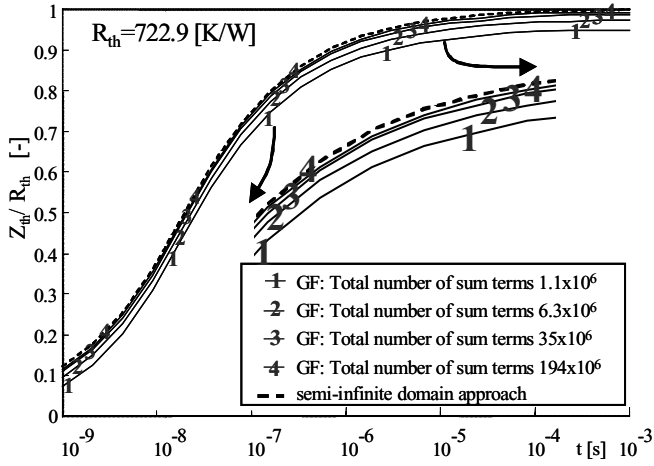


Fig. 2. Normalized thermal impedance affected by the truncation error for different number of terms (in the table) using LTGF approach. The heat source is in the center of the domain

For small times, when the heat diffusion does not reach the side walls, the impedance curves should fit the semi-infinite one but it is not observed. If more terms are included, the thermal impedance tends to fit this curve, but even a large number of terms like 2×10^8 is not sufficient to make the error negligible. To find the most accurate results, 6 hours were needed to compute 20 time points on a 660 MHz PC-computer.

In Fig.2, the curves 1-4 are parallel and the difference between them is constant and does not depend on the time. It means that the error resulting from the series limitation is not a function of time. It is possible in such a case only when the convergence of the time dependent terms is much faster than the convergence of the time independent ones. This remark allows considering the full solution as the sum of the slow converging steady state solution $R_{th}(r)$ and the time dependent part $\Gamma(r,t)$ that converges faster due to covering an exponential function of time [7].

$$Z_{th,LTGF}(r,t) = R_{th}(r) + \Gamma(r,t) \tag{8}$$

3.2 Efficiency of the STGF Solution

The STGF solution (7) has a form of triple series covering integrals with respect to time what means that a numerical integration must be performed. The terms of the series correspond to image sources that are used in infinite domain to set proper boundary conditions. In used formalism each term correspond to 8 image sources. The solution for semi-infinite domain [6] can be also achieved from (7) by taking (m,n,p)=(0,0,0) and neglecting some of the remaining elements.

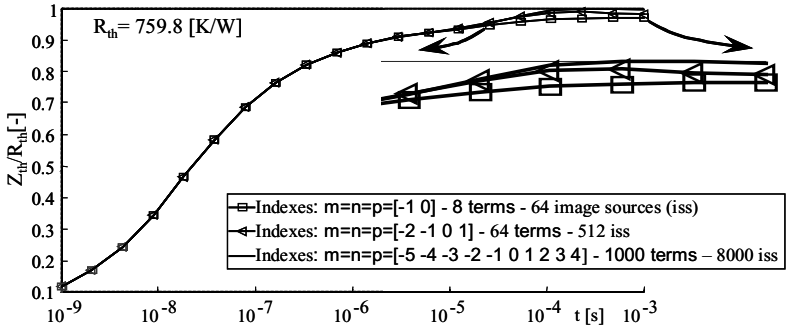


Fig. 3. Normalized thermal impedance for different number of terms. The heat source is in the center of the domain

To investigate the properties of STGF solution, as previous, the thermal impedance in the middle of the centered heat source has been calculated utilizing different number of terms and the obtained curves are depicted in Fig. 3. The dimensions of the domain were fixed as follows: A=B=90μm, C=100μm, L_x= L_y =5μm.

In comparison with the LTGF approach, the truncation error of the STGF solution is not constant with time. The best agreement is found for the small times. When the time increases, the discrepancy appears. It can be explained by considering the terms as related to the image sources. For small times a few terms (i.e. images) are necessary only to set proper boundary conditions. When the time increases more terms should be utilized. If number of terms is too small, the boundary conditions are not well established and the error appears. In the example even 1000 terms (8000 image sources) were not enough to make the error unnoticed for $t=10^{-3}$s.

Compared to LTGF approach, the STGF solution involves an integral over time. It has to be obtained numerically. Thus, although the number of terms is much lower than for the LTGF, the numerical integration is an additional operation that must be performed carefully (due to singularity for u=0) so as no essential additional error related to integration process were introduced. The need of the numerical integration decreases the efficiency of the STGF approach that has better numerical performance for small times.

4 The Time Partitioning Approach

For small times very few terms of the STGF solution can be enough to get the satisfactory accuracy. In some cases even a part of the final expression i.e. the part that describes the semi-finite domain solution can be sufficient. On the other hand, considering the LTGF solution, the larger time is the lower number of terms needs to be included to obtain exact value of the transient part $\Gamma(r,t)$ of the full solution. The time partitioning approach [4] tries to combine these features: for times lower than some value t_p ($t<t_p$) the STGF solution is used and if the solution for $t>t_p$ is needed, the impedance is calculated as:

$$Z_{th}(r, t > t_p) = Z_{th,STGF}(r, t_p) + \Gamma(r, t) - \Gamma(r, t_p) \tag{9}$$

It means that $Z_{th,STGF}(r,t_p)$ is calculated using the STGF approach for the t_p point and the numerical integration limited to the range from 0 to t_p. To complete the thermal impedance to its value at point t, the formula for LTGF approach is used. In this case, however, its transient parts calculated for the while t_p and t are necessary only since the difference $\Gamma(r,t) - \Gamma(r,t_p)$, corresponds to the thermal impedance increase from t_p to t. In the time partitioning solution (9), there is no slow convergent steady state part of the LTGF solution and the STGF method is used only up to t_p time point. The t_p time can be chosen arbitrarily to optimize the speed of the evaluation process. The t_p must be low enough to use as few as possible image sources since numerical integration is necessary. On the other hand the larger t_p is the less terms are used by $\Gamma(r,t)$.

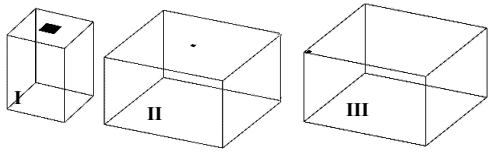


Fig. 4. Domains used to compare the efficiency. C=100μm **I:** L_x=L_y=23μm, A=B=90 μm, **II:** L_x=L_y=5μm, A=B=175 μm, **III:** L_x=L_y=5μm, A=B=175 μm, the center of the heat source (left-down corner of the domain) is at (5,5) μm point. The domains are in scale. The heat sources can be seen as small black rectangles on the top surface

To check the efficiency of the time partitioning approach the thermal impedance has been evaluated at the source centers for 3 test domains (Fig. 4) at 20 time points from 1ns to 1ms in logarithmic scale. Two solutions are considered:

- the standard LTGF solution (Eq. (6))
- the time partitioning method (Eq. (9)).

To make the results comparable, the maximal relative error was made for both methods equal to 6%.

All simulations were done in Matlab6.5 environment using the *quadl.m* procedure for numerical integration, which employs the Gauss-Lobatto quadrature algorithm. The results of calculation are summarized in Table 2 that covers the total time of calculations (ToC), the total number of terms for the LTGF solution (NT) and the time partitioning point t_p. For structures I and II a quarter of the domain for the LTGF solution is used. As the STGF solution the semi-infinite domain solution is used only. The t_p is chosen to be latest available time point where the semi-infinite domain solution is valid for the finite domain.

The results presented in the table show that proper choice of the solution formulae can speed up calculations from 6h to 0.1s that is more than $1 \cdot 10^5$ times. The ToC for the time partitioning method can be optimised by the proper choice of the t_p point.

Table 2. Comparison of efficiency of analyzed methods

Method \ Domain	I	II	III
i	ToC=25min, NT=107×106	ToC= 6 h *(660MHz PC)*, NT=3880×106	-
ii	ToC: 0.11s, NT= 166 t_p=1.4×10^{-}6s	ToC: 0.15s, NT= 8 t_p=4.8×10^{-}4s	ToC: 0.7s, NT=105 t_p=1.8×10^{-}8s

5 Conclusions

The electro-thermal analysis of any device, especially ICs, becomes indispensable and the ICs should be considered in their thermal domain as 3D systems. Describing the electrical behavior of the cell or cluster by simple model as it has been presented, the analytical solutions of the 3D heat transfer phenomenon seems to be well suited in this case. Although different analytical formulas are available, evaluation process of the expressions may involve a lot of computational effort and its efficiency is changes depending on the chosen analytical solution and particular conditions. The LTGF and the STGF approaches have serious drawbacks when considered separately. But both solutions can be combined in time partitioning method. Although this idea has been indicated by Beck *et. al.* in [4], on the best knowledge of authors, it has not been used in any application. Presented analysis shows that the time partitioning approach, as a proper combination of methods, may improve the efficiency by few orders of magnitude.

References

1. d'Alessandro, V., Rinaldi, N.: A critical review of thermal models for electrothermal simulations. Solid-State electronics, Vol. 46, (2002) 487-496
2. d'Alessandro, V., Rinaldi, N., Spirito, P.: Thermal modeling and simulation of multicellular power devices. Proc. MicroTherm 2003, (Łód 2003, Poland).
3. Batty, W., Christoffersen, C. E., Panks, A. J., David, S., Snowden, C. M., Steer, M. B.: Electrothermal CAD of power devices and circuits with fully physical time-dependent compact thermal modeling of complex nonlinear 3-D systems. IEEE Trans. on Components and Packaging Technologies, Vol. 24, (2001) 566-590
4. Beck, J.V., Cole K., Hadji-Sheikh A., Litkouhi B.: Heat conduction using Green's functions. Hemispherev Publishing Corp., Washington DC (1992)
5. Macchiaroli, M., Rinaldi, N., d'Alessandro, V., Breglio, G., Spirito, P.: A new Electrothermal Simulation Tool for the Analysis of Bipolar Devices and Circuits. Proc. 7th Thermal Investigations of ICs and systems (THERMINIC) Workshop, (Paris 2001) 281-286.
6. Rinaldi, N., On the modeling of the transient thermal behavior of semiconductor devices IEEE Trans. on Electron devices, Vol. 48, 2796-2802 (2001)
7. Wo ny, J., Rinaldi, N., Owczarek, M., Nowak, J.: Optimizing The Numerical Procedure of Temperature Evaluating from The Heat Diffusion Equation. Accepted for TCSET conference, Lviv (2004).

Layout Based 3D Thermal Simulations of Integrated Circuits Components

Krzysztof Slusarczyk, Marek Kaminski, and Andrzej Napieralski

Department of Microelectronics and Computer Science
Technical University of Lodz, Poland
{kslusar, kaminski, napier}@dmcs.p.lodz.pl

Abstract. In this paper a three-dimensional (3D) thermal simulations of basic integrated circuit (IC) components are presented. The layout of the real Application Specific Integrated Circuit (ASIC) generated in CAD CADENCE software was loaded into the CFD-ACE environment, where 3D thermal simulations were done. The influence of many heating points was considered in simulations. Comparison among several cooling conditions was also made and presented in the paper. Also the importance of logic circuits thermal simulations in case of their thermal parameters estimation, an approximation of potentially thermally unstable areas, and a redesigning of a logic circuit due to its thermal strength are shown in this paper.

1 Introduction

A constant development in microelectronics causes that new, till now not observed problems have appeared during simulation process. One of the hot topics for IC designers is appropriate analysis of thermal phenomena, which appear in VLSI circuits. Nowadays, more than half of IC fatal failures is caused by overheating a semiconductor structure or not efficient enough heat transfer in a device itself. This is only one of many reasons why thermal analysis has become one of the most important research directions in modern microelectronics. A miniaturization and an increase of the circuit integration degree influences strongly on an increase of the power density generated inside the IC and also on a deterioration of cooling conditions (Fig. 1).

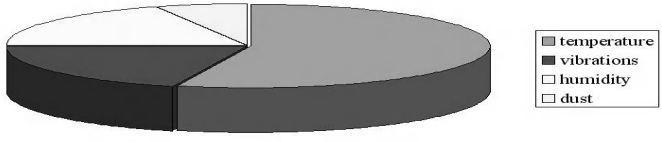


Fig. 1. The most common reasons of IC failures [6]

All kinds of problems accompanying heat generation and abstraction became multiplied in multilayer integrated circuits. In such ICs, layouts must be designed much more precisely than in a single layer IC. Although in this paper a single layer ASIC components were tested, the presented thermal optimization path is valid for all ICs.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1029–1036, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Thermal simulator is becoming an important CAD tool, which allows designer to validate the structure, estimate the time life of the structure and also implement real (changed by the increase of the temperature) electrical parameters. The complexity of modern layouts makes impossible importing data for the thermal analysis manually, so one of the most required elements of such simulator is layout converter, which task is to extract from the 2D layout the 3D structure that is loaded into the simulator. The conversion of this methodology is described in first part of this paper.

The differential equation of thermal conduction cannot be solved analytically for such a complex structure as IC. Only numerical methods (e.g. finite differences or finite elements) can be used for solving this equation. Problems connected with proper mesh generation are the most important issue, which has strong influence on accuracy, correctness and usefulness of thermal simulation results. On the one hand too sparse mesh is the cause of incorrectness and possibility of an omission points of high temperature, on the other, too dense mesh causes numerical problems (convergence and accuracy problems), and also strongly extends simulation time. Due to complexity of the analyzed structures and also determinable compute power of a computer, the simulation time cannot be too long, if a practical application of such simulation is considered.

Another problem is an accurate boundary conditions definition, which is strongly connected with IC cooling (most often convection is the most important phenomenon, but sometimes also radiation must be considered), and also precise modeling heat generation inside the structure. Simulator must be a specialized tool, which allows designer for generating the most optimal mesh, defining boundary conditions equal to the real cooling conditions in the designed IC, and also proper definition of the heat source. These problems are considered in second part of this paper.

Nowadays, nearly all IC are designed to decrease the usage of semiconductor area. Such approach is the reason of designing projects in which heat sources are placed very close to each other. These wrongly routed heat sources become the hottest areas of a chip (so called hot spots), and are responsible for the thermal strength of the whole circuits. The thermal simulation allows to estimate, where such spots are placed in the logic circuit layout, and what is the temperature of these spots in a function of a power release. Redesigning of the layout by spreading hot spots effects in larger, but far more thermally resistant circuit [5]. These problems are considered and discussed in third part of the paper, on the example of a NAND gate and a operational amplifier.

2 Exporting Layout to CFD-ACE

A layout generated in CADENCE environment is usually saved in CIF or GDSII format. Data written in GDSII file are used in a silicon foundry to prepare masks used in production process. The format of data contains list of geometrical figures (most often rectangles), their coordinations and layer type. Such GDSII file cannot be directly imported into CFD-Micromesh software, due to lack of information about typical thickness and placement of each layer [1,2,3]. These supplementary information must be saved in technology file, which contains names of all layers used in the project (it

should be remembered, that GDSII file does not contain any information about passivation layers), and also conversion table allowing for ascription GDSII layer number to CFD-Micromesh layer name. Exemplary conversion table used during simulation is presented in Tab. 1.

Table 1. Conversion table used during loading layout

Layer name	GDSII symbol	Material	Depth [nm]	GDSII number
Passivation	PROT	SiO_2	750	—
Metal 3	MET3	Al	940	39
Via 2	VIA2	Al	—	38
Oxy_m2_m3	IMD2	SiO_2	650	—
Metal 2	MET2	Al	650	37
Via 1	VIA1	Al	—	36
Oxy_m1_m2	IMD1	SiO_2	650	—
Metal 1	MET1	Al	720	35
Contact	CONT	Al	—	34
Oxy_m1_poly	ILD	SiO_2	650	—
Poly 2	POLY2	Si	250	30
Poly 1	POLY1	Si	250	20
Field oxide	FOX	SiO_2	400	10
Diffusion1	DIFF	Si	100	10
Wafer	—	Si	50000	—

3 Mesh Generation, Boundary Conditions, and Heat Generation

A proper mesh selection is a complicated process. There is an optimal number of mesh nodes, which is a compromise among simulation time, convergence of solution and also accuracy of results. The mesh generation is the important process during model creation, although it is also the most time and effort consuming procedure[3,7].

Authors ran series of simulations for the same heating and cooling conditions, but for models of different mesh densities. Also additional surfaces in Z-direction were added. The total number of mesh nodes is presented in Tab. 2

Table 2. Total number of mesh nodes

Number of surfaces in Z	Number of cells in XY surface			
	20 (voxel = 1μm)	200 (voxel = 0.1μm)	2000 (voxel = 0.1μm)	10000 (voxel = 0.1μm)
19 (voxel = 0.1μm)	269	7174	17867	87269
47 (voxel = 0.04μm)	587	27320	39949	194842
78 (voxel = 0.04μm)	903	48733	62896	—

The most crucial element of every thermal simulation is the proper selection of cooling conditions. Heat is being removed from the IC in three different ways:
- heat conduction in the semiconductor structure and packaging contact area
- heat conduction in metal wires connecting semiconductor structure and pins
- convection from upper IC surfaces

Complex IC thermal simulation should include heat distribution in both semiconductor structure and packaging, but due to time consumption of such estimation, model dimensions and also mesh density of the structure, usually the whole simulation is reduced to semiconductor area.

If only a small part of the IC is considered, also side walls boundary conditions should be set. The best solution in that case is to define adiabatic boundary conditions, which is similar to a situation, when simulated structure is surrounded by comparable structures of the same temperature. This is the worst condition case, so the behavior of the logic gate in severe situation can be predicted.

Authors had chosen cooling conditions, that allowed for decreasing simulation time, and also increasing the computations convergence, an isothermal or adiabatic condition at the semiconductor structure – package contact, and convectional (linear) at the passivation – air surface (Fig. 2).

An appropriate heat generation model is the most fundamental part of a properly run thermal simulation. Unfortunately, it is impossible to build an electro-thermal simulator for ICs. Although size and efficiency of heat sources inside the IC can be estimated easily during designing process, it is impossible to define every heat source separately; some global rules must be accepted. It is assumed that inside the IC, the basic heat generation area is MOS transistor channel. In CFD-ACE environment, heat source can be easily defined as a boundary condition for one of walls or interfaces of the model. In that case every MOS channel in the IC layout can be easily defined as a heat source. An interface between POLY1 and diffusion always matches transistor area, so these interfaces can be automatically defined as heat sources. The drawback of such solution is a uniform distribution of the heat in every MOS transistor. Although some transistors switch more frequently than others, every device is treated uniformly, so every transistor is a source of the same power density.

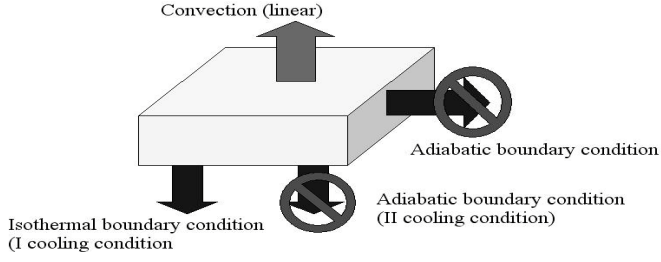


Fig. 2. Boundary conditions for the examined structure

The thermal analysis of basic IC elements demands high level of attention to detail in modeling heat sources allocation. The most unsuitable method of simplifying the model and increasing calculations is averaging generated power on larger area in order to attain omit some structure details (e.g. transistors' channels). As an example, a

NAND gate was simulated with three different heat sources. In the first case the heat source was defined properly – only in transistors' channels, in the second case (shown in Fig. 3) heat source was spread in whole diffusion area, in the third case the whole gate surface was defined as a heat source.

Fig. 3. Diffusion areas defined as a heat source

Further averaging is the source of loosing information about the temperature distribution and also the computed maximum temperature is much reduced. It is impossible to estimate the resistance of the circuit to the thermal breakdown.

In Tab. 3 comparison of results obtained during NAND gate simulations for three different heat emission areas is presented. Enlarging the heat emission area (to decrease the complexity of the model) is the reason of decreasing the maximum temperature in the structure.

Table 3. Simulations results for different heat emission areas

Heat emission area	S [μm^2]	H_S [Wm^{-2}]	T_{max} [K]
Transistor channel	17	10^8	303.008
Diffusion area	85.74	1.92E+7	302.463
Total logic gate surface	263.58	6.45E+6	301.342

4 NAND Gate and Operational Amplifier

As it was mentioned in "Mesh Generation" section, three different mesh were used during simulations (Tab. 4 and 5). In Fig. 4 (right) the exemplary mesh is presented.

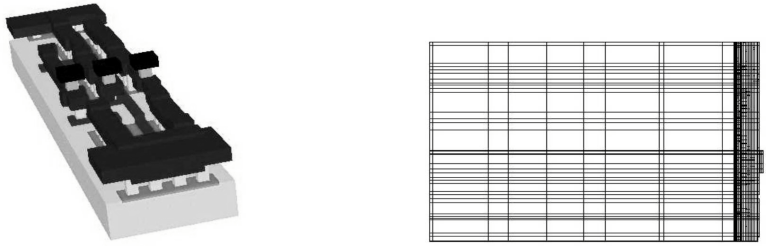


Fig. 4. The 3D view of the NAND gate structure (left) and an exemplary mesh (right)

During simulations two kinds of cooling conditions were considered (isothermal condition for lower surface, and convection k = 40 $Wm^{-2}K^{-1}$ and $H_s = 10^8$ $Wm^{-}2$ for upper surface or adiabatic condition for lower surface, and convection k = 4000 $Wm^{-2}K^{-1}$ and $H_s = 10^5$ Wm^{-2} for upper surface. A graphical representations of simulations results for the first cooling conditions are presented in Fig. 5.

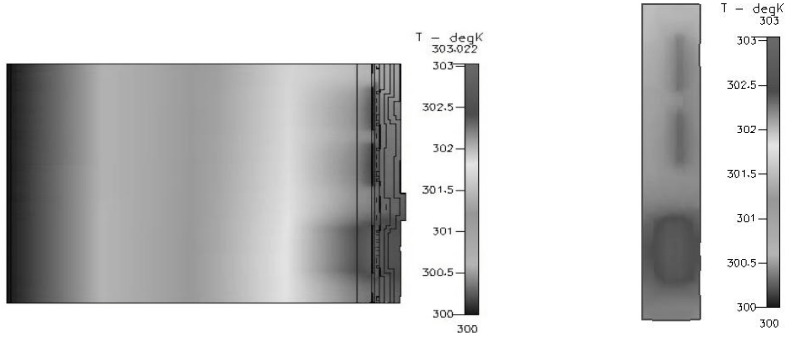


Fig. 5. NAND gate simulation results for the first cooling condition (Z – cut and upper surface)

Table 4. Simulations results for the spare mesh

Number of surfaces in Z	Number of cells in XY surface			
	20 (voxel = 1μm)	200 (voxel = 0.1μm)	2000 (voxel = 0.1μm)	10000 (voxel = 0.1μm)
19 (voxel = 0.1μm)	1588.45 K	303.007 K	303.02 K	303.004 K
47 (voxel = 0.04μm)	309.812 K	303.038 K	303.061 K	303.022 K
78 (voxel = 0.04μm)	309.654 K	302.998 K	302.998 K	—

Table 5. Simulations results for the dense mesh

Number of surfaces in Z	Number of cells in XY surface			
	20 (voxel = 1μm)	200 (voxel = 0.1μm)	2000 (voxel = 0.1μm)	10000 (voxel = 0.1μm)
19 (voxel = 0.1μm)	305.965 K	301.583 K	301.59 K	301.59 K
47 (voxel = 0.04μm)	304.348 K	301.568 K	301.593 K	—
78 (voxel = 0.04μm)	304.77 K	301.546 K	301.512 K	—

Comparing results presented in Tab. 4 and 5, it can be found, that sparse mesh was a source of serious numerical errors. When the density of the mesh achieved some crucial value, on the one hand simulations results became nearly unchangeable, but on the other, if the mesh density was still being increased, simulation time was increasing drastically. It was obvious, that implementing very dense meshes was unbeneficial. The most optimal mesh for examined NAND gate model was a structure containing 19 surfaces and 200 cells for each surface.

The opamp device was simulated for the same cooling conditions as the NAND gate. Simulation results are presented in the Fig. 6. In the successful simulation opamp model consisted of 39392 nodes and 60717 cells.

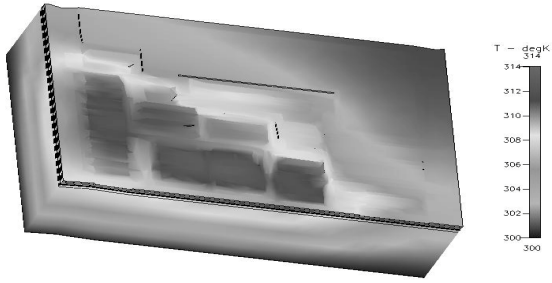


Fig. 6. Opamp simulation results (3D view)

5 Logic Gates Designing Due to Their Thermal Strength

As could be expected, transistors' channels are the hottest areas on the surface of a NAND gate. Due to the fact that NMOS channels are placed close to each other, this area can be treated as the hot spot for this gate. In case of larger heat emission, and bad cooling conditions, this spot is potentially thermally unstable and can be easily overheated (although the temperature difference between NMOS and PMOS channels is only 0.5K, more important than the temperature itself is the power density, that is emitted from the NMOS area).

For some applications, in which the minimum chip area is not a crucial parameter, it is much more profitable to redesign the basic logic gate to achieve more uniform heating of all chip's area. The comparison between maximum temperatures obtained for standard and redesigned NAND gates simulations is presented in Table 6.

Table 6. Maximum temperatures obtained for standard and redesigned NAND gates

Logic gate	Surface area [μm^2]	T_{max} [K]
Standard	263.58	303.1
Thermally optimized	301.78	302.6

As can be seen, although the redesigned hate is slightly larger, the maximum temperature achieved in simulations is 0.5K less than in standard gate (Fig. 6). The temperature distribution on the surface of the redesign gate is presented in Fig. 9.

The very similar situation can be also found for the opamp structure. In the structure optimized for minimal layout area, heat is generated unequally in the structure. All transistors are grouped in one area and become the hot spot for this device (Fig. 7). Redesigning the structure (e.g. placing transistors on both sides of the capacitor) will

make the structure a little bit larger, but the thermal strength of the structure will be much more increased.
Logic gates that are thermally optimized should be used in all areas of the IC structure, which are potentially thermally unstable. Electrical simulations should be an integral part of a layout design, according to estimated power emission in different basic components, it is possible to select logic gates, which should be optimized also thermally.

6 Conclusions

In this paper a full 3D thermal simulations of basic elements of the real ASIC IC were presented. Such simulations allow for very detailed analysis of thermal phenomena inside of the IC structure. The method for automatic transformation of the ASIC layout from the CADENCE environment into the CFD-ACE environment 3D model was also presented in this paper. The obtained model was extremely accurate comparing to the real silicon structure The implementation of the heat sources into the ASIC modeling quite simple and no advanced knowledge is necessary.

Although the heat generation model that was used during simulations was much simplified (heat was generated evenly in the transistor channel), results obtained in simulations can be used as a basis for redesigning the IC. It can be assumed, that such full 3D thermal simulation can be necessary not only during designing process of the simple ASICs (such as sigma-delta ADC), but far more complicated IC structures as microprocessors or multilayer chips (where not only many heat sources, but also heat transfer problems have to be taken into account) or even power devices.

Due to thermal simulations of basic layout components it is possible to optimize thermally basic logic gates in order to decrease the maximum heat release in the ASIC structure. Redesigned elements, which thermal strength properties are enlarged, can be used in all potentially thermally unstable or insufficiently cooled devices.

References

1. A. Raman, M. Turowski, and M. Mar “Layout Based Full Chip Thermal Simulations of Staked 3Dintegrated Circuits,” *2003 ASME International Mechanical Engineering Congress and Exposition*, Washington D.C., Nov. 16 – 21 2003
2. CFD-ACE Tutorials, CFDRC, 2003
3. CFD-Micromesh Tutorials, CFDRC, 2003
4. M. Szermer, A. Napieralski “Modeling of the Sigma-Delta Analogue to Digital Converters with Application of VHDL-AMS,” to be published at *TCSET’2004,* Slavsko, Feb. 2004
5. A. Napieralski, M. Janicki “Zjawiska cieplne w ukladach elektronicznych,” Technical University of Lodz Scripts, (in Polish), Lodz, 2001
6. G. De Mey “Thermal Problems in Electronics,” Advanced Training Course *Mixed Design of VLSI Circuits*, Debe, Poland, 1994, pp. 25 – 31
7. M. Furmanczyk “Elektroniczna symulacja ukladow VLSI ze szczegolnym uwzglednieniem integracji w srodowisku projektowym,” Technical University of Lodz, Ph.D. thesis (in Polish), Lodz 1999

Simulation of Electrical and Optical Interconnections for Future VLSI ICs

Grzegorz Tosik[1,2], Zbigniew Lisik[1], Małgorzata Langer[1], Frederic Gaffiot[2], and Ian O'Conor[2]

[1] Institute of Electronics, Technical University of Łódź, 91-924 Łódź, ul. Wólczańska 230
[2] LEOM Ecole Centrale de Lyon

Abstract. At present, metallic interconnections become the "bottleneck" of the further progress in VLSI technology. The optical solution is considered as an alternative that could allow overcoming the limitations but its advantage should be confirmed at the level of modeling approach. The clock distribution network (CDN) that is the most representative component of the modern VLSI circuits has been used as the test circuit and its numerical models for H-tree architecture have been worked out and used. The model of the electrical CDN as well as some results of simulations presenting its expected features, also with the comparison to the results obtained for its optical counterpart are presented.

1 Introduction

The advances of semiconductor fabrication process made possible design and fabricate chips with several millions of transistors operating at a very high speed. These advances together with innovative hardware organisations of modern integrated circuits (IC's) give high performance ICs at low cost. However it becomes evident that most of the known technological capabilities will approach or reach their fundamental limits and one will require substantial changes in device technologies and structures that will become more and more 3D ones.

The progress of the VLSI systems has been driven by downsizing of their components and increasing the operating speed. In contrast to transistor scaling the interconnect scaling improves the integration density but at the cost of the degraded propagation delay and the power consumption. In the new technology, the interconnect delay dominates over logic delay even in spite when new metallization technologies such as copper or new low-k dielectrics are applied [1]. The downsizing process leads to smaller transistors characterized by smaller power consumption but the number of transistors and its density increases in such a degree that the resulting density of power dissipation becomes larger still. In spite of the multi-level (3D) design, the simultaneous growth in circuit complexity leads to larger and larger chip dimensions and the total length of interconnection lines. They are manufactured as multilevel structures, like the one shown in Fig.1, characterized by the total length reaching a few kilometers at the width below 0.3μm on the lowest level, giving a considerable contribution to the total power dissipation. As the result, in modern VLSI systems, the power dissipation increases rapidly especially in its interconnection part and in a VLSI circuit with the power dissipation of 100W [2] only the clock tree uses at least 30-50% of this power [3]. Due to natural limits in thermal management, this is a real barrier to further progress of modern VLSI systems

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1037–1044, 2004.
© Springer-Verlag Berlin Heidelberg 2004

and just now, it has happened that manufactures must reduce their ratings due to the thermal reason only.

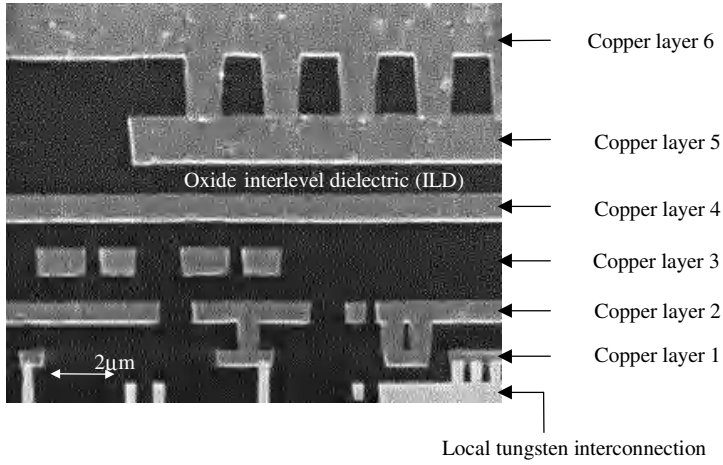


Fig. 1. Scanning electron micrographs of interconnect architecture with six levels of copper wires/vias and tungsten contacts/local interconnects

As a possible alternative that can overcome the limitation of metallic wires an optical interconnection are proposed. Applying optical interconnections to ICs has been the subject of many researches [4, 5] but no real remedy has been presented till now. It is obvious that this alternative can be acceptable only if it demonstrates significantly improved performance over the electrical solution. Unfortunately, due to the obvious reasons, the electrical metallic interconnections foreseeing by ITRS do not exist at present as well. In such a case, an approach basing on their models is the only way to evaluate and to compare the prospective features of the both solutions. This approach has been also applied in the project [6] aiming at the new optoelectronic VLSI solution with a layout covering silicon waveguides formed on the silicon chip. The clock distribution network (CDN) that is the most representative component of the modern VLSI circuits has been used as the test circuit and its model for the H-tree architecture has been worked out and used. The model of the electrical CDN and some results of simulations presenting its expected features, also with the comparison to the results obtained for its optical counterpart, are presented in the paper.

2 Clock Distribution Network (CDN)

2.1 General Description

The semiconductor technologies operate at increasingly higher speeds, and the system performance has become limited not by the delays of the individual logic elements but by the necessity to synchronize the flow of the data signals. A clock network distributes the clock signal from the clock generator, to the clock inputs of the synchronizing components. This must be done while maintaining the integrity of the

signal and minimizing such clock parameters like: the clock skew, the clock slew rate or the clock phase delay. Additionally, these objectives must be attained while minimizing the use of system resources such as the power and the area. The CDN of a modern microprocessor uses a significant fraction of the total chip power and has a substantial impact on the overall performance of the system.

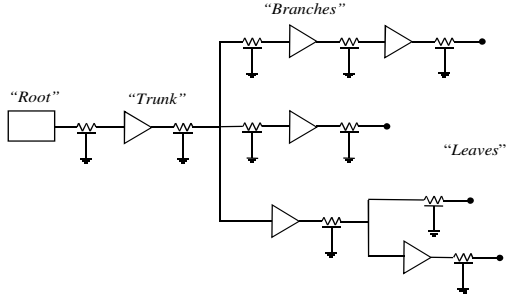


Fig. 2. The tree-like structure of the clock network

The proper generation and the distribution of the clock signal are critical for the high performance systems operating. The most common strategy for this distribution is using a tree-like structure shown in Fig.2. The clock input is connected to the "*trunk*" and the signal is subsequently split and distributed to the"*branches*" connected hierarchically that finally feed logic elements ("*leaves*"). In order to keep the quality of clock signals, the tree is completed by repeaters (buffers) built-in into the branches.

2.2 H-Tree Architecture

Fully symmetrical structures of CDN are desirable to eliminate the clock skew. The most popular solution that one considers here is the H-tree one, shown in Fig.3. That makes all the paths from the clock signal source to the clocked register being identical and therefore the clock skew is minimized. The distribution scheme repeats the H-shaped structure recursively and the lowest level of these structures covers points used to drive the local registers. Some design rules must be fulfilled to ensure the proper work of the H-tree CDN. The impedance of the branch leaving each branch point must be twice the impedance of the one providing the signal to the branch point to avoid the reflection effect at the branch point [7], and appropriate buffers must be built-in into the H-tree branches. The both demands have been met in the worked out model.

3 Model of Electrical H-Tree CDN

3.1 General Description

The most recent microprocessors have as many as seven metal layers [2], while the ITRS predicts the use of up to 10 levels of wiring for the 45nm technology node in

2010 at up to $2x10^9$ transistors per chip, and clock frequencies up to 10 GHz. Coping with the analyze of electrical interconnections under the above conditions is a task that requires an adequate tool allowing to investigate the signal propagation in the considered system, and dedicated numerical models seem to meet this demand. Such a model for H-tree CDN has been worked out in a form of ICAL software package. It allows designing any H-tree structure for any technology level predicted by ITRS, extracting the electric parameters from technology data both for transmission lines and necessary buffers, and creating an equivalent circuit model presented in a form of SPICE netlist useful in further circuit simulations.

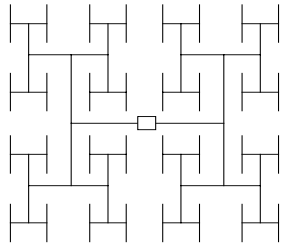


Fig. 3. The H-tree topology

3.2 Transmission Line Model

The interconnection system in a VLSI chip covers two power lines and one signal line as shown in Fig.4. The technology used to its realization allows considering the power lines as a meshed network that plays the role of distributed current sources supplying all electronic elements equally. The signal line must, however, be considered as a transmission line coupling buffers sketched as CMOS inverters, and characterized by its resistivity R, inductivity L and capacity C that can be used to model it as an RC or RLC transmission line shown in Fig.4a. When such models are created for CDN transmission lines, the specificity in realizing them consists in shielding the signal wires at the sides by power, and ground lines as shown in Fig.4b.

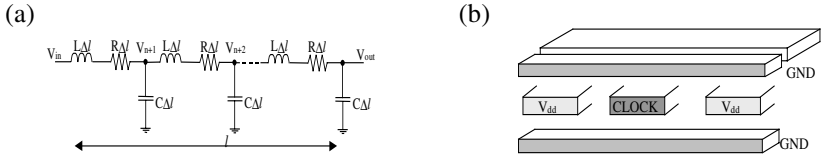


Fig. 4. Distributed RLC interconnection model (a) and clock wire shielding (b)

Interconnection resistance. The interconnection resistance per unit length, R_o, is generally determined by (1) where ρ - resistivity of the metal layer, W - width of the wire and T - its thickness. At large signal frequencies, the resistance can by changed by the skin effect and such a possibility is introduced into ICAL. It takes place when the skin effect depth δ is smaller than the wire dimension, what can, however, happen

very seldom in considered CDN systems since δ=0.7μm for the maximal ITRS frequency, f=10GHz.

$$R_o = \frac{\rho}{W*T} \tag{1}$$

Interconnection capacitance. Parasitic capacitances associated with the interconnection lines are shown schematically in Fig.5. They have become the primary factors in the evolution of the very high speed integrated circuit technology, their evaluation is not a trivial task and has been the subject of many investigations [8-10]. To get an accurate interconnection capacitance, 2D or 3D electric field simulation in the whole interconnection system should be applied. It is, however, such a huge task that in real applications approximate approaches have been used only. They treat the total interconnection capacitance as a sum of a few different capacitances resulting from the particular design and differ the way of the components capacitances identification.

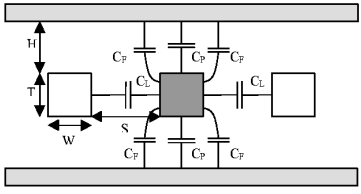


Fig. 5. Interconnect capacitance components: C_P - parallel plate capacitance component, C_F - fringing field component, C_L - lateral coupling capacitance component

The most popular approaches were proposed by Saravat [11], Sakurai [12], Chern [13] and Wong [14] and their application in ICAL has been considered. Since each of them gives another evaluation of the total capacitance C_{tot} for designs predicted by ITRS and their experimental verification is impossible, the numerical simulation has been used to recognize which one of them could be used. In this goal the software package OPERA [15] using finite element techniques to solve the Maxwell equations has been used to analyze electric field in 2D domain shown in Fig.4b. The electric field distributions, like the one in Fig.6, were used to calculate the real values of C_{tot} that were compared to the evaluations obtained by means of tested approaches The comparison showed that only for the Chern approach the discrepancy was lower than 7% over a wide range of ITRS parameters whereas it was much larger for the others. The Chern approach has been chosen for ICAL package.

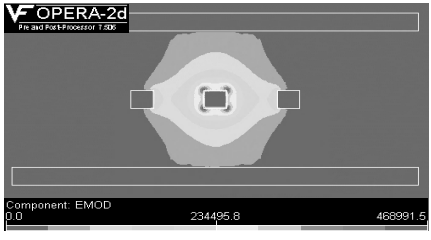


Fig. 6. Interconnect structure used in numerical capacitance calculation

Interconnection inductance. In today's VLSI circuits, the inductance starts to become significant due to longer metal interconnections, the reduction in wire resistances and the higher operation frequency. The analytical expressions (2) and (3) have been obtained [16] for partial self and mutual inductances, respectively, under the assumption of uniform current distribution and rectangular shape of wires shown in Fig.7.

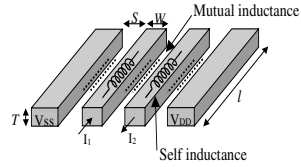


Fig. 7. On-chip interconnection inductance

$$L = \frac{\mu_o l}{2\pi}\left[\ln\left(\frac{2l}{W+T}\right) + 0.5 + \frac{0.2235(W+T)}{l}\right] \quad (2)$$

$$M = \frac{\mu_o l}{2\pi}\left[\ln\left(\frac{2l}{S}\right) - 1 + \frac{S}{l}\right] \quad (3))$$

3.3 Buffer Model

The accurate characterization of the MOSFET transistor is crucial for any high speed digital integrated circuit analysis and design. In order to calculate the clock path delay, buffers incorporated into ICAL program are modeled by an equivalent circuit shown in Fig.8a (where D_{Inv} is the buffer internal delay, R_{Inv} is the buffer output resistance, C_{in} is the buffer input capacitance and V is the logic swing on the clock lines) and by the transistor level model shown in Fig.8b. Taking into account the demands resulting from the technologies predicted by ITRS, the Berkeley BSIM3v3 and BSIM4 MOS models were used in the transistor level model.

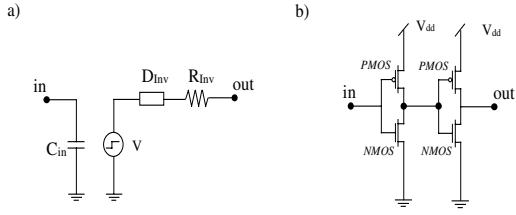


Fig. 8. Models of buffer devices. a) Equivalent circuits. b) Transistor level model

4 Results of Simulation

The majority of investigations have been performed for 128 and 256 symmetrical global H-trees manufactured on 300-450 mm^2 chips depending on the ITRS demands for the assumed technology. Some of these results are collected in Fig 9-12.

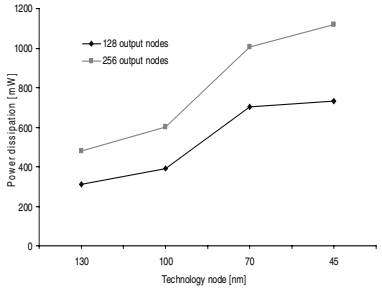


Fig. 9. Power dissipated in electrical H-trees vs. technology

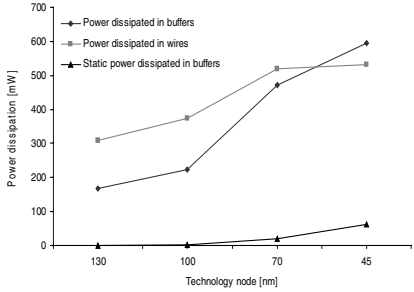


Fig. 10. Power budget in the electrical H-tree networks at 256 output nodes

Fig.9 shows the electrical power consumption in the global H-trees versus technology node, and the operating frequencies for considered systems are (in GHz/technology): 1.6/130nm, 3/100nm, (5.6 and 11.5)/(70nm and 45nm). It is obvious that the total power consumption in electrical clock distribution network tends to grow despite technology improvements. Fig.10 shows the power dissipated by buffers and wires in for various technology nodes. It shows that along with technology scaling the contribution of the power consumed by buffers to the total power consumption tends to grow, and for the 45nm technology it will be even bigger than the power consumed by wires due to increasingly large number of repeaters used in clock distribution systems. Fig.11 shows the dependence of the power consumption for both electrical and optical CDN's on the frequency for 70nm technology node. Whereas the power consumption in the electrical system increases rapidly with the increase of the clock frequency, in the optical system it remains almost on the same level. In Fig.12 the power consumption is plotted as a function of the chip width for f=5.6GHz and 70nm technology node rules. In opposite to the optical CDN, in the electrical one, the power consumption increases drastically when the die size increases.

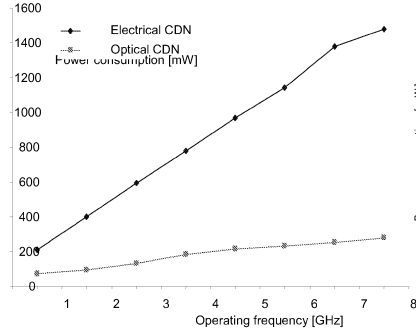


Fig. 11. Power consumption in optical and electrical CDN's versus operating frequency at 256 output nodes

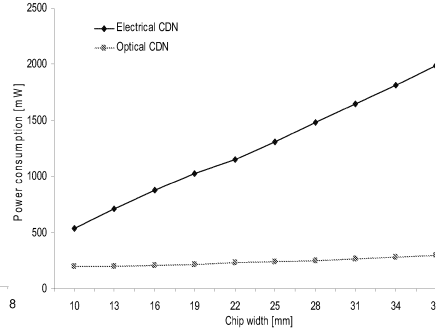


Fig. 12. Power consumption in optical and electrical CDN's versus chip width. at 256 output nodes

References

1. SIA The National Technology Roadmap for Semiconductors San Jose CA, USA 1994
2. N.,Barkatullah et al IEEE J. Solid-St. Circuit. 36 (2001) 1647-1652
3. S.,Rusu et al IEEE J.Solid-St. Circuit., 35, (2000),1545-1552
4. J.W.Goodman et al Proc. IEEE,72 (1984) 850–866.
5. D.A.Miller et al Proc. IEEE 88 (2000) 728-749.
6. Integration of optical and optoelectronics technologies for the realization of interconnections in microelectronics systems, Polonium Project No 4337.I/2002
7. D.C.Keezer et al IEEE Int. Conf. Wafer-Scale Integration (1992) 168-175.
8. A.Ruehli et al J. Solid-St. Circuits 10 (1975) 530-536.
9. E.Barke IEEE Trans.CAD 7 (1988) 295-298
10. M.Lee J.Solid-St. Circuits.33 (1998) 657-661
11. K.C.Saraswat J. Solid-St. Circuits 17 (1982) 275-280.
12. T.Sakurai IEEE J.Solid-St. Circuits 18 (1983) 418-425.
13. J.H.Chern et al IEEE EDL 13 (1992) 32-34.
14. S.C.Wong et al IEEE Trans. Semicon. Manuf. 13 (2000) 219-223.
15. Vector Fields (http://www.vectorfields.com/op2d).
16. F.Grover Instrum. Soc. of America (1945).

Balanced Binary Search Trees Based Approach for Sparse Matrix Representation

Igor Balk[1], Igor Pavlovsky[2], Andrey Ushakov[3], and Irina Landman[3]

[1] R3 Logic Inc, 80 Sherman Street, Camridge MA 02140, USA
[2] Terra Discount Club LLC, 110 Strathmore Rd 302, Brighton MA 02135 USA
[3] Corning Ltd, 4 Birzhevaya Linia, 199034, St. Petersburg, Russia

Abstract. In this paper novel method of memory allocation for sparse matrixes is presented. Sparse matrixes are widely used in computational electro-dynamics. This paper show that use of balanced binary search trees for memory allocation of sparse matrixes guarantees O(ln(n)) access and insertion time. Comparison with traditional hash map memory allocation method was also made and presented in the paper.

1 Introduction

Many aspects of modern life require sophisticated computations to be performed. The majority of these computations and simulations are based on operations with sparse matrixes. A sparse matrix is a matrix with number of zero elements much larger than number of non-zero elements. These matrices are used in many computational disciplines such as image processing, computational electro and hydrodynamics and many others. Utilizing this characteristics, we can significantly reduce memory space required to store these matrixes and use special algorithms with this type of matrixes much faster than with dense matrixes. Good examples of such computations utilizing sparse matrices are image processing and electromagnetic simulation. Performance of mathematical libraries used for these simulations becomes critically important as complexity level growth. Direct algorithms will give $O(n^3)$, where n is the size of the system to be solved, increase in computational time and memory required, and it is challenging problem even for modern supercomputers. To solve this problem we developed fast numerical linear algebra library for sparse matrixes, which is based on data structures with O(ln(n)) access time, where n is dimension of the matrix.

Several data structures with O(ln(n)) access time are known in graph theory. In our work we are utilizing different kinds of binary trees in order to accelerate performance of the library. The most natural assumption is to use red-black trees and AVL trees. These trees are well known in computer science theory and widely used for solution of network routing problem. We decided to use these trees to create a data structure in which sparse matrix will be stored.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1045–1048, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Theory

There are many publications available describing the theory of binary trees. These trees are popular since they allow the search of the element to be performed in c*ln(n) operations, where c is some constant and n is the number of elements in the tree. The major and still unanswered question is how to reduce the constant c. There are several techniques, which will give improvements in c value depending on the problem. Two most well known techniques are AVL trees and red-black trees. Use of this trees gives performance of a search and insertion operation in a tree close to optimal. [1, 2, 3]

A binary search tree (BST) is an AVL tree if the difference in height between the subtrees of each of its nodes is between -1 and +1. Said another way, a BST is an AVL tree if it is an empty tree or if its subtrees are AVL trees and the difference in height between its left and right subtree is between -1 and +1. In order to demonstrate performance we will mention some theorem [2]: an AVL tree with n nodes has height between log2 (n + 1) and 1.44 * log2 (n + 2) - 0.328. An AVL tree with height h has between pow (2, (h + .328) / 1.44) - 2 and pow (2, h) - 1 nodes.

For comparison, an optimally balanced BST with *n* nodes has height *ceil* (*log2* (*n* + 1)). An optimally balanced BST with height *h* has between *pow* (2, *h* - 1) and *pow* (2, *h*) - 1 nodes.The average speed of a search in a binary tree depends on the tree's height, so the results above are quite encouraging: an AVL tree will never be more than about 50% taller than the corresponding optimally balanced tree. Thus, we have a guarantee of good performance even in the worst case, and optimal performance in the best case [2]. And for the red-black theorem, similar to the one we introduced for AVL trees [1, 3], can be formulated: a red-black tree with n nodes has height at least log2 (n + 1) but no more than 2 * log2 (n + 1). A red-black tree with height h has at least pow (2, h / 2) - 1 nodes but no more than pow (2, h) - 1.

And for comparison, an optimally balanced BST with *n* nodes has height *ceil* (*log2* (*n* + 1)). An optimally balanced BST with height *h* has between *pow* (2, *h* - 1) and *pow* (2, *h*) - 1 nodes. Thus we can see that if we will use AVL and red-black trees to store data in sparse matrixes we will get performance close to optimal.

3 Implementation

Based of on the theory described above we have developed memory allocation algorithm for sparse matrixes. We used AVL trees to represent rows and colons of space matrixes. As it was shown above similar results could be obtained using red-black trees. Algorithm was implemented in ANSII C for maximum performance and portability. Important to mention, that described approach allows easy parallelization and further speed up on multiprocessor architecture.

4 Computational Results

We performed some preliminary tests comparing our new library with the data hash map memory allocation model, which is widely used in sparse matrix manipulation

for example in such popular package as Matlab. Both memory models were implemented in ANSII C with same compiler optimization. We performed two tests to demonstrate performance of new memory model. The results of the tests are shown on Figures 1and 2.

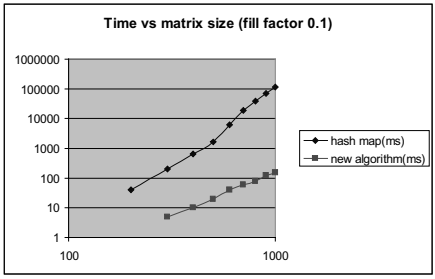


Fig. 1. Filling time vs. matrix size for fixed filling factor

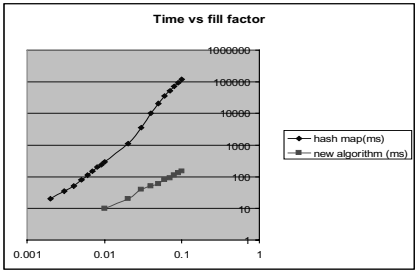


Fig. 2. Filling time vs. filling factor for fixed matrix size

On these figures dependence of the time of random filling, for the matrix of given size, from zero up to given filling factor is demonstrated. The Figure 1 demonstrates this dependence for the fixed filling factor = 0.1 and matrix size varying from the $n=10^2$ to the $n=10^3$. On vertical axis filling time is shown and on horizontal - size of the matrix. As it's clear from the graph even with relatively small matrixes new algorithm of memory allocation gives order of magnitude saving in time. Figure 2 shows filling time (vertical axis) for the sparse matrix with fixed size and various filling factors (horizontal axis). The size of the matrix n is set to be 10^3 and filling factor is varying from 0.002 to 0.1.

Table 1. Speed of algorithm depending of filling factor for fixed matrix size (n=1000)

Fill factor	Hash map (ms)	New algorithm (ms)
0.003	35	
0.004	50	
0.005	80	
0.006	110	
0.007	150	
0.008	200	
0.009	240	
0.01	291	10
0.02	1122	20
0.03	3646	40
0.04	10344	50
0.05	20900	60
0.06	34840	80
0.07	51143	90
0.08	71342	110
0.09	93955	130
0.1	117979	150

Table 2. Speed of algorithm depending of matrix size for fixed filling factor (f=0.1)

Matrix size	Hash map (ms)	New algorithm (ms)
200	40	
300	200	5
400	630	10
500	1623	20
600	6329	40
700	18307	60
800	39156	80
900	71403	120
1000	117549	150

5 Conclusion

We believe that the described library will provide engineers and researchers with a fast, reliable, and cost effective tool to satisfy the growing need for computational tools. A simple, intuitive interface, fast data access algorithms, and low cost will make this product a “must have” tool for people using numerical linear algebra in their work.

References

1. Cormen, C. H., C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. McGraw-Hill, 1990.
2. Knuth, D. E., *The Art of Computer Programming, Volume 3: Sorting and Searching*, 2nd ed. Addison-Wesley,
3. Sedgewick, R., *Algorithms in C, Parts 1-4*, 3rd ed. Addison-Wesley, 1998
4. Trefethen, Lloyd N.(Lloyd Nicholas)., David Bau, III, Numerical linear algebra SIAM, Philadelphia, 1997
5. Gene H. Golub, Charles F. Van Loan, Matrix Computations, 2 nd ed.The John Hopkins University Press, Baltimore, 1998
6. Igor Balk, Andrey Ushakov, Irina Landman “FNLAL: AVL-Tree based fast numerical linear algebra”, Corning Technical Report, Corning, NY, 2001

Principles of Rectangular Mesh Generation in Computational Physics

Vladislav Ermolaev[1], Evgeniy Odintsov[2], Alexander Sobachkin[2], Alexey Kharitonovich[3], Marina Bevzushenko[4], and Sergei Zorin[5].

[1] LTX Corp., Boston, MA, USA
vermolaev@mail.ru
[2, 3] NIKA GmbH., Moscow, Russia
[3] alkharit@nika.sokol.ru
[4] Boston University, Boston, MA, USA
marinavb@math.bu.edu
[5] Parametric Corp., Boston, MA, USA
szorin@ptc.com

Abstract. Many real world problems may be represented with mathematical models. These models often set complex mathematical, namely partial differential equation (PDE) problems hard to solve analytically and will often require computational approach. Definition of such computational problems will usually imply having a geometry model and initial conditions set on, in or around this model. Computational techniques have to deal with discrete space and time in order to approximate large and complex PDEs with ready to calculate simple arithmetical equations. Discontinuous or discrete space is called mesh. The scope of this article is problems of mesh generation and ways of their solution.

1 Introduction

Whether we have to deal with bioscience, fluid dynamics, electromagnetism or any other field where real world problems require computational approach it will often be a PDE problem. Its definition will usually imply having a geometry model, computational domain around it and boundary or initial conditions set in this domain. In order to apply computational approach for PDE solving it is necessary to convert continuous space into discontinuous or so-called mesh. This conversion results in partial loss of initial data or accuracy as geometry model and initial conditions in discontinuous space become a finite element model (FEM) that is merely an approximation of its prototype in continuous space. The finer the mesh the closer the approximation is. On the other hand fine mesh means great number of finite elements or cells every of which is participating in calculation thus consuming computational power and increasing overall calculation time.

Process of optimal mesh generation is very complex and will often consider specifics of the scientific field in general as well as specifics of every particular problem (adaptive mesh refinement) and sometimes even workflow of a problem solution (solution adaptive mesh refinement). The scope of this article is principles of rectangular mesh generation by the example model in which one of many 3D modelling CAD products is used as a basis for problem definition tool.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1049–1055, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Mesh Generation

2.1 Basic Mesh

We will be discussing rectangular mesh. It is formed by intersection of three sets of planes that are parallel to each other within every set and sets being orthogonal to each other thus forming a Cartesian frame. Spans between adjacent parallel planes may vary to better resolve geometry model and flow features. This mesh is called *basic mesh* and is usually too coarse to yield any reasonable results.

Every cell of basic mesh (*basic cell*) may be split into eight equal rectangular cells. These smaller cells are *children* with respect to basic cell, their *parent*. This process is called *refinement*. Similarly, every group of eight children may be merged back to one cell. This is *unrefinement* process. Every child in turn may be split further to eight new cells and so on. *Level of refinement* or simply *level* of cell is non-negative number that shows how many times parents of the cell were refined to build this cell. Thus, level of any basic cell is zero, level of their children is one and further refinement yields cells of level two and so on. There is a rule requiring level difference of adjacent cells be not higher than one. This requirement is held for smoother and more accurate solution.

2.2 Initial Mesh

Mesh built before calculation is started has special name *initial.* This is because it may change further during calculation. Initial mesh is constructed in several stages:

- constructing basic mesh for a specified number of cells and stretching or contracting it locally to better resolve geometry model and flow features
- splitting basic mesh cells either to capture small solid features or to resolve curvature of interfaces between other materials with different physical properties
- narrow channel refinement - refinement of obtained mesh to better resolve narrow channels and conductors
- irregular cells refinement - if irregular cells appear, they are split to maximum allowed level or until become regular

First step to obtain optimal mesh is to build basic mesh wisely. We may contract it at regions of most interest or stretch somewhere to sacrifice accuracy for calculation time. This is done by means of *control planes*. They are set by engineer who defines problem and are part of basic mesh. Every such plane has associated number that tells mesh generation algorithm how much should mesh be stretched or contracted around this plane. This number N is dimension multiplier of basic cell of a uniform mesh. In other words, cells around this plane will be N times the size of cells of simple uniform mesh built in the same region and amount of cells per its dimension. Size of neighbour cells changes linearly from control plane to computational domain boundaries or to another control plane forming gradually contracted or stretched mesh.

Control planes with multipliers may be rendered automatically given some specifics of the problem (field, geometry model, initial conditions, etc.). For instance, given a fluid dynamics problem of a projectile moving through atmosphere with relatively high Mach number there may be expected flow compression in the vicinity of projec-

tile's geometry and high turbulence trace behind it. Thus, problem definition contains enough information to automatically detect regions where physical parameters will most probably have large gradients and build more dense mesh in those regions.

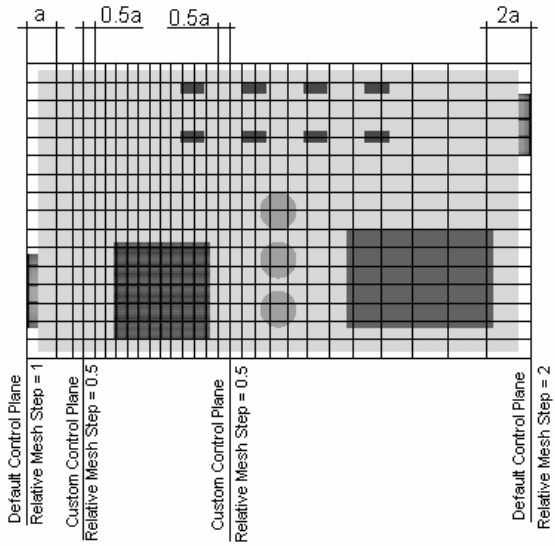


Fig. 1. Using control planes to vary mesh density

2.3 Geometry Capturing

For algorithm that computationally solves defined problem to run properly it has to have information about geometry in every cell of the mesh as well as initial conditions, if any, applied to that cell. Hence, next step after basic mesh is generated is to capture model geometry from CAD and refine basic mesh well enough not to lose important geometry features while keeping total number of cells low to save computational time.

Most of modern CAD products publish variety of tools in form of application programming interface (API) for third-party software to gain access to geometry model and operations on it. They may export different operations to determine relations between given point and the model, functions to determine all intersection points of given ray with the model and so on. However, the fastest way to work with geometry information proved to be by using parametric representation of surfaces that build up the model. This representation allows finding relations between every given cell of the mesh and geometry model.

Cell types. Depending on geometrical relation between a cell and model all cells are classified as *fluid*, *solid* or *partial* cells. Fluid cells have no common points with geometry model. Solid ones are those that completely belong to model, in other words fully occupied with solid body. Partial cells are those in between. They have intersections with model surfaces and situated partly inside and partly outside the model.

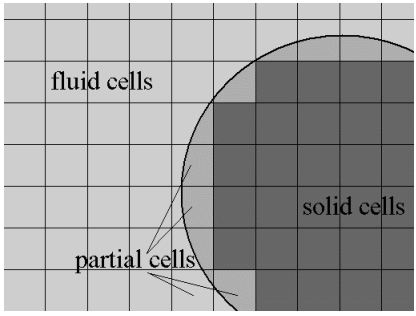


Fig. 2. Cell types

As solver considers every cell during calculation it has to know whether matter can flow in the cell and what valid directions of the flow are. Gas, liquid, heat, electrical current or other form of energy may flow through the cell. In case of pure fluid and solid cells this information is given by cell type itself as solver knows what matter can flow through every cell type. For example, gas and liquid flux is not limited by geometry in fluid cells and may flow in any direction through all cell facets. Same with heat flux in full solid cells, electricity in conductors, etc. Special case is partial cells however. Solver requires information about geometry in these cells to know what cell facets or parts of facets are valid for fluxes to pass. Since geometry inside partial cell may be very complex and solver refers to it very often during calculation it is virtually impossible to keep this information in original parametric form and have solver recalculate everything it needs every time. Instead, model surfaces inside every given cell are linearized, i.e. replaced with polygons based on points where cell edges intersect model surfaces. Not only such approximation reduces solver load, but it requires little memory to store geometry information inside cell also. It may be as little as coordinates of one point and normal vector to represent geometry model inside given cell.

Although such geometry linearization has its advantages it is a source of certain difficulties as well. If cell intersects some fine geometry feature such as thin solid plate, or on the contrary, a thin crack in solid model, or some spire, it becomes very hard to approximate that geometry with one polygon. Such approximation may change physics of processes dramatically, e.g. allow fluids to flow where they should not. This problem may partly be solved by building base mesh wisely. By preliminary analysis of geometry model basic mesh may be projected so that opposite faces of thin solid plane occur inside two adjacent cells instead of one. Needle or spire may appear coaxial with cell edge occupying four adjacent cells and therefore be formed with four polygons. There are situations however when little can help but refining cell into smaller ones. Such cell where geometry can not be approximated with one polygon without principal changes to physics in the cell is called *irregular*. Irregular cells are usually subject to refinement to maximum allowed level or until become regular solid, fluid or partial cells.

2.4 Mesh Refinement

After base mesh is rendered every its cell undergoes a number of tests to decide should it be split into smaller ones. There are several reasons why it may be necessary to refine a cell. All these reasons are represented with some mathematical criterion thus forming a list of *refinement criteria*. Every cell is tested according to this list.

Curvature criterion. All geometry model surfaces are approximated with polygons inside mesh cells. If radius of surface curvature is comparable to characteristic size of cell then approximation of this surface with polygons inside cells will be very coarse. The greater the radius with respect to cell size the smoother approximation will be.

To control geometry approximation accuracy *curvature criterion* rule is applied to every partial cell. Normal to surface polygon of partial cell is compared by pairs with normals to surface polygons of immediate neighbour cells. These normal differences $/n_1\text{-}n_2/$ should be less than certain number c, curvature criterion. Otherwise the cell is split into next level cells to improve accuracy of geometry surface approximation.

Passing areas ratio criterion. Polygon approximating geometry surface divides some cell faces into two areas which belong to different media, e.g. solid and fluid, different solids, porous media, etc. Solver uses these areas during calculation when it needs information what part of every cell's face is passable for one or another flux, e.g. fluid flow, or heat flow in solids or electrical current in conductors.

Every cell face is common to two adjacent cells with the exception of computational region boundaries. Actual surface and polygons that approximate it generally intersect cell edges in different points. It is likely that polygons of two adjacent cells intersect common cell face differently. Thus, same cell face may have different ratio of two areas mentioned above depending on what cell is considered. For conservation laws to hold solver always uses same ratio for the same cell face regardless of cell it considers. What solver uses is some average of area ratios of two adjacent cells.

Even though conservation laws are not violated areas solver uses are not exact. They are different from what would real geometry surface produce. To control error related to this issue there is another refinement rule and criterion. Difference of area ratios $/s_1\text{-}s_2/$ given by considering adjacent cells should be less than some number s.

Narrow channel resolution criterion. Main focus of rules and criteria above is model geometry and its approximation in the mesh. This however is not the only concern in initial mesh generation. Physical behaviour in certain areas of computational region is sometimes predictable even before calculations. Thus, by initial conditions and geometry model base mesh is stretched and contracted in some places in anticipation of certain physical behaviour there. Another example when anticipated physical phenomena affect mesh generation is refinement of narrow channels or conductors.

Narrow channel is geometry feature with one linear dimension several times greater than any other. Pipes of various cross sections, cracks and narrow gaps between solid bodies, and so on, are common examples of narrow channels. If fluid can flow there it is likely to flow with high speed. The narrower the channel the higher flow speed can be expected. For high speed flows in narrow channels physical phenomena in boundary layers may have significant influence. Boundary layer is region of fluid close to solid body where flow speed is considerably lower than in the middle of the channel due to friction. The narrower the channel and higher the flow speed the more effect of

boundary layers is. Sometimes disregard of this effect may crucially change overall solution.

Similarly, high frequency currents will flow mostly through boundary layer of conductor due to skin effect. Energy flow occurs mostly on conductor surface exponentially diminishing with depth.

To take boundary layer phenomena into account there should be enough number of cells across every narrow channel. There is separate criterion for this purpose, average number of cells across any channel in the model. Regions inside all channels are refined until there are enough cells. This refinement is not linear and generates more cells close to channel walls and less in the middle of assumed stream. This pattern helps resolve future boundary layers while keeping overall number of cells minimal.

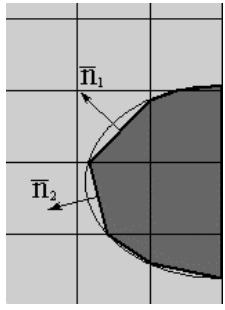


Fig. 3. Linear approximation of geometry

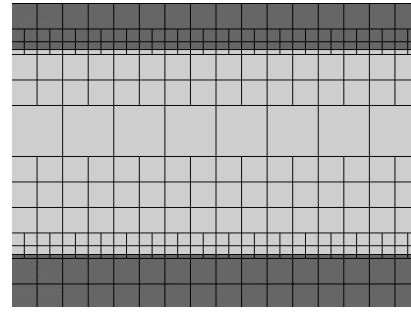


Fig. 4. Channel with eight cells fit across it

2.5 Local Mesh Settings

Great variety and complexity of real life problems makes it very difficult to automate all stages of mesh generation process. It is possible to recognize certain patterns of initial conditions and geometry in problems and build mesh accordingly. However, an engineer will sometimes know problem specifics better than any general purpose algorithm may detect. To give engineer some flexibility in these cases and therefore improve overall accuracy and timing performance *local mesh settings* are implemented.

Local mesh settings are non-overlapping regions of computational domain where all settings and criteria mentioned above have their local values. This is a tool to select specific region manually and demand higher geometry resolution in it or refine some solid body to certain level to better resolve heat flow in it or set different narrow channel resolution criteria for different channels, etc. All problem specifics known beforehand may be reflected with local mesh settings to improve calculation performance.

3 Refining Mesh during Calculations

Above principles help generate appropriate mesh to achieve good calculation accuracy and keep number of cells low to save calculation time. Also it is proved to be very efficient tactics to adjust mesh as problem solution develops.

These mesh adjustments or *solution adaptive refinements* may take place several times during calculation as solution picture changes. The idea of solution adaptive refinement is to locate regions with extreme gradients of physical parameters, either very low or conversely very high. Cells in regions with high gradients are subject to extra refinement to better resolve physical peculiarities such as pressure drops. At the same time cells of regions with almost constant parameters are merged into lower level cells.

Parameters to take into account may vary depending on cell medium and problem specifics. Pressure and velocity vector are important for gases and liquids; medium density for compressible flows; temperature is typical parameter for solids; wave magnitude and phase angle or field energy may be considered for electromagnetic fields.

In order to decide whether given cell should be split, merged or left untouched physical parameters in it and its immediate neighbours are compared. Differences of these values are used to calculate some weighted average that shows should this cell be split or merged if it should at all.

4 Summary

In this article we revealed basic principles of rectangular mesh generation. Problem of building optimal mesh to maximize ratio of achieved accuracy to computational time is very complex. One of the fundamentals is to use maximum information about every particular task to design mesh for its solution. This may help dramatically increase performance by concentrating on details of crucial importance while saving on others that may only contribute little error.

Different aspects of a problem may affect mesh generation. We have seen how model geometry features may influence this process as well as initial conditions and physical peculiarities anticipated in future solution. There are several steps in mesh design to take into account every such aspect of the problem.

Principles discussed here are not limited for use with PDE solvers and may be generalized for larger class of problems including integral equations.

References

1. Joe F. Thompson, Bharat Soni and Nigel P. Weatherrill: Handbook of Grid Generation. CRC Press Pub. (1998)
2. Vladimir D. Liseikin: Grid Generation Methods. Springer-Verlag (1999)
3. Graham F. Carey: Computational Grids : Generation, Adaptation, and Solution Strategies: Series in Computational and Physical Processes in Mechanics and Thermal Science. Taylor & Francis Pub. (1997)
4. Joe F. Thompson, Z. U. A. Warsi and C. W. Mastin: Numerical Grid Generation : Foundations and Applications. North Holland Pub. (1985)
   http://www.erc.msstate.edu/publications/gridbook/

Inter-finger Connection Matrices*

Vladimir M. Zatsiorsky[1], Mark L. Latash[1], Fred Danion[2], Fan Gao[1], Zong-Ming Li[3], Robert W. Gregory[4], and Sheng Li[5]

[1] Department of Kinesiology, The Pennsylvania State University, USA
[2] Faculté des Sciences du Sport, Université de la Mediterranée, France
[3] Musculoskeletal Research Center, University of Pittsburgh, USA
[4] School of Education, University of Kansas, USA
[5] Rehabilitation Institute of Chicago, USA

Abstract. Fingers of the hand are interdependent: when a person moves one finger or produces a force with a fingertip, other fingers of the hand also move or show force production. Hence, no direct correspondence exists between the neural commands to individual fingers and finger forces. The relations among fingers are described with *inter-finger connection matrices*, IFM. The IFMs depend on the number of fingers involved in the task. This presentation addresses three aspects of the IFMs: (1) computation of the IFMs, (2) role of finger interdependence during manipulation of hand-held objects and (3) inter-individual differences in the IFMs.

When a person moves one finger or produces a force with a fingertip, other fingers of the hand also move or show force production[12,6,4]. This phenomenon has been termed enslaving[18,19]. The finger interdependence is due to three sources/mechanisms: (1) peripheral connections, both tendinous[8] and intermuscular myofascial[5], (2) multi-digit motor units in the extrinsic flexor and extensor muscles[6], and (3) central neural connections[13]. Due to the enslaving, there is no direct correspondence between neural commands to individual fingers and finger forces.

The relations among fingers can be described with inter-finger connection matrices, IFM [18,11]. The IFMs depend on the number of fingers involved in the task. The reason behind this dependence is a so called force deficit: a maximal force exerted by a finger in a multi-finger task is smaller than a maximal force produced by this finger in a single-finger test. The deficit increases with the number of fingers involved in the task[9,10]. Existence of the force deficit makes determination of the IFMs in static tasks nontrivial: recording of finger forces while the subject tries to press with only one finger does not account for the force deficit and, hence, is not sufficient to determine an IFM. This presentation addresses three aspects of the IFMs: (1) computation of the IFMs, (2) role of finger interdependence during manipulation of hand-held objects and (3) inter-individual differences in the IFMs.

* This study was partly supported by NIH grants AR 048563, NS-35032 and AG-18751. The support from the Whittaker Foundation to Dr. Z.M. Li is also acknowledged.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1056–1064, 2004.
© Springer-Verlag Berlin Heidelberg 2004

1 Computation of the IFMs

So far, two techniques have been used to compute the IFMs: (a) neural networking and (b) algebraic approximation.

1.1 Neural Networks

The three-layer network model is shown in Fig. 1. The model consists of three layers: the input layer that models a central neural drive; the hidden layer modeling finger flexors serving several fingers simultaneously, and the output layer representing finger force output. Note the existence of direct input-output connections that model muscular components that serve individual fingers. The networks incorporate the following ideas/hypotheses:

(a) Existence of two groups of muscle components. Each muscle/compartment of the first group serves an individual finger (unidigit muscles; intrinsic muscles of the hand) and each muscle/compartment of the second group serves several fingers (multi-digit muscles; extrinsic muscles of the hand). The first group of muscles is represented in the neural networks by a direct one-to-one connection from the input to the output layer. The second group is represented by the middle layer and its multiple connections.
(b) The force deficit phenomenon is modeled by specific transfer characteristics of the middle layer neurons: the output of the middle layer was set as inversely proportional to the number of fingers involved. Note that in the model, the force deficit effects are only assigned to the multi-digit muscles of the hand.
(c) The enslaving effects are modeled by the connection weights from the middle to the output layer.

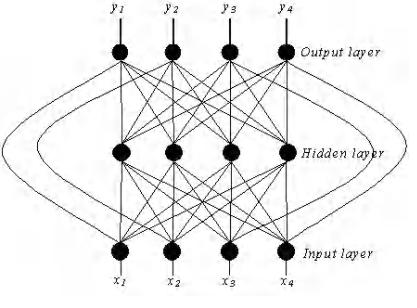


Fig. 1. Basic network. The hidden layer models the extrinsic hand muscles (those that are located in the forearm) having multiple connections to all four fingers while direct input-output connections represent the intrinsic hand muscles (those that are located in the hand) that serve individual fingers. The index, middle, ring and little finger correspond to 1, 2, 3, and 4, respectively.

The net input to the j^{th} unit of the hidden layer from the input layer is

$$s_j^{(1)} = \sum_{i=1}^{4} w_{ij}^{(1)} x_i \qquad j = 1, 2, 3, 4 \tag{1}$$

where $w_{ij}^{(1)}$ are connection weights from the i^{th} unit in the input layer to the j^{th} unit in the hidden layer. The characteristic of input/output in the hidden layer is described as

$$z_j = f_1(s_j^{(1)}) = \frac{w_{jj}^{(1)} x_j}{s_j^{(1)}} \qquad j = 1, 2, 3, 4 \tag{2}$$

where z_j is the output from the hidden layer. The net input to the k^{th} unit in the layer ($s_k^{(2)}$) from the hidden layer is expressed as

$$s_k^{(2)} = \sum_{j=1}^{4} w_{jk}^{(2)} z_j + v_k x_k \qquad k = 1, 2, 3, 4 \tag{3}$$

where $w_{jk}^{(2)}$ are connection weights from the j^{th} unit in the hidden layer to the k^{th} unit in the output layer. v_k are the connetion weights directly from the k^{th} unit in the input layer to the k^{th} unit in the output layer. An identity input/output transfer relationship was defined at the output layer, i.e.

$$y_k = f_2(s_k^{(2)}) = s_k^{(2)} \qquad k = 1, 2, 3, 4 \tag{4}$$

The inputs to the network were set at $x_i = 1$, if finger i was involved in the task, or $x_i = 0$ otherwise. The weights from the input layer to the hidden layer were set as a unit constant ($w_{ij}^{(1)}$). The network was trained using a backpropagation algorithm[1].

The developed network yielded a relation between the neural commands and the finger forces:

$$[F] = k[w][c] + [v][c] \tag{5}$$

where $[F]$ is a (4×1) vector of the finger forces, $[w]$ is a (4×4) matrix of inter-finger weight coefficients, $[c]$ is a (4×1) vector of dimensionless neural commands (command to a finger ranges from 1.0, when a finger is intended to produce maximal force, to 0.0, if the finger is not intended to produce force), $[v]$ is a (4×4) diagonal matrix with gain coefficients that model the input-output relations for single-digit muscles, and k is a coefficient that depends on the number of fingers in the task ($0 \leq k \leq 1$). The value of k was set either at $1/n$, where n is the number of intended fingers in the force production task, or computed by the network; the two approaches yielded similar results[18]. From (5) it follows that a command c_j sent to a finger j ($j = 1, 2, 3, 4$) activates all other fingers to a certain extent (*enslaving effects*). For a given n, in particular for $n = 4$, (5) can be reduced to

$$[F] = [W][c] \tag{6}$$

where $[W]$ is the (4×4) IFM accounting for both force enslaving and force deficit[17,18,11].

1.2 Algebraic Approximation

Recently, Danion et al.[2] suggested a different equation in which the force deficit is represented by a coefficient related to the number of explicitly involved fingers and the IFM accounts only for enslaving:

$$[F] = \frac{1}{n^{0.66}}[W'][c] \tag{7}$$

where $1/n^{0.66}$ is an empirical force deficit coefficient and $[W']$ is a 'pure' enslaving matrix. The elements of $[W']$ can be easily estimated without the neural network computations from single-finger maximal force contraction (MVC) tasks.

The above approach inspired a mode-control hypothesis of finger coordination. According to the hypothesis, for each single-finger task, the CNS controls a unique variable (a *Mode*) leading to force production by the master finger, as well as by the enslaved fingers. For instance, when a subject produces force voluntarily with the index finger (I), Mode-I is recruited by the CNS. Due to the enslaving phenomenon, Mode-I also leads to force production by the middle, ring, and little fingers. Similarly, voluntary force production by the middle (M), ring (R) or little finger (L), is assumed to involve corresponding Modes (Mode-M, Mode-R, and Mode-L, respectively). Therefore, a Mode can be viewed as a collective variable, which leads to activation of many hand muscles bringing about a specific pattern of force production by several fingers.

2 The Role of Finger Interdependence during Manipulation of the Hand-Held Objects

An interest to the IFMs greatly increased when it was shown that the enslaving occurs during natural grasping[17]. Knowledge of the IFMs allowed to reconstruct the intensity of neural commands sent to individual fingers and to estimate the magnitude of the enslaving effects, the force exerted by finger i due to the command sent to finger j.

If the vector of finger forces $[F]$ and IFM matrix $[W]$ are known the vector of the neural commands can be determined by inverting (6). The vector of neural commands is then

$$[c] = [W]^{-1}[F] \tag{8}$$

To test whether the various observed force-sharing patterns were optimal, optimization methods have been employed. The norms of the following vectors were employed as cost functions:

(G1) Finger forces.
(G2) Finger forces normalized with respect to the maximal forces measured in single-finger tasks.
(G3) Finger forces normalized with respect to the maximal forces measured in a four-finger (IMRL) task.

(G4) Finger forces normalized with respect to the maximal moments that can be generated by the fingers while grasping an object with five digits.
(G5) Neural commands.

The main distinction between the first four cost functions and the fifth one lies in the way of finger interdependence being accounted for: the cost functions based on the finger forces neglect the finger interdependence while the optimization of neural commands accounts for it.

In experiments with static holding of a handle with an attached load, when the subjects were required to produce different combinations of force and torque, some ('agonist') fingers generated moments in the direction required by the task while other ('antagonist') fingers produced moments in the opposite direction. Optimization of neural commands was able to model such 'antagonist' finger force production and resulted in a better correspondence between the actual and predicted finger forces than the optimization of various norms of the finger forces (Fig. 2, next page). Hence, during grasping strong commands to particular fingers activate also fingers that generate moments of force in the direction opposite to the direction required by the task.

3 Inter-individual Differences in the IFMs

Individuals forming the general population may be expected to show both similarities and differences in the way their central nervous system organizes finger interaction. These differences may be related to such gross characteristics as the amount of total maximal finger force a person can produce[15] or to more subtle individual factors such as experience with particular tasks that require precise finger coordination[16]. An IFM for a hand contains 16 numbers. Hence, in general, individual differences in finger interaction, as reflected in IFMs, may require 16 variables to be fully described. We hypothesize, however, that a significantly smaller number of variables may be sufficient to describe differences among individuals without special skills. Hence, we aimed to discover such variables and relate them to indices of finger interaction introduced in earlier studies.

To this end, we recorded maximal finger forces in single-finger and multi-finger tasks in 20 right-handed university male subjects. The data were analyzed in the following way:

(A) The individual IMFs were computed by the artificial neural network shown in Fig. 1. We will call these matrices the *non-normalized* IFMs. The sum of the elements of a non-normalized IFM equals the total force of all fingers in the four-finger task.
(B) *Normalized* IFMs were computed by dividing the elements of a non-normalized IFM by the sum of its elements, i.e. by the total force of the four fingers in the IMRL task. The sum of the elements of a normalized IFM equals one. Both normalized and non-normalized IFMs were used for the further analysis.

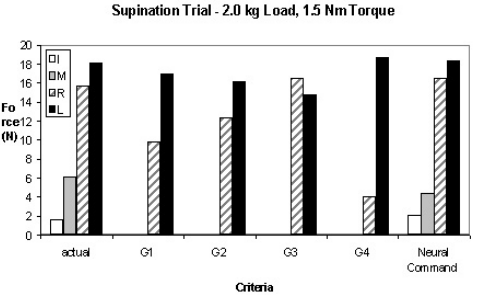


Fig. 2. Comparison of actual force data with force patterns predicted by different optimization criteria. Criteria G1-G4 do not predict the antagonist moments while the optimization of neural commands does.

(C) The similarities/dissimilarities (proximities) between the individual matrices have been quantified. Suppose we have two IFMs computed on subjects $s1$ and $s2$. The corresponding matrices are $\mathbf{A}$ and $\mathbf{B}$ and their difference is $(\mathbf{A}-\mathbf{B})$. The dissimilarity of the matrices was computed as a square root of the trace of a matrix, $(\mathbf{A}-\mathbf{B})^{\mathbf{T}}(\mathbf{A}-\mathbf{B})$, that is

$$\delta_{ij} = \{trace[(\mathbf{A}-\mathbf{B})^{\mathbf{T}}(\mathbf{A}-\mathbf{B})]\}^{\mathbf{0.5}} \tag{9}$$

where the superscript T denotes the transpose. In order to avoid confusion with the 'distances' determined in the multi-dimensional scaling (MDS) method (see below) and following the accepted terminology, we called δ_{ij} a *proximity* of matrices $\mathbf{A}$ and $\mathbf{B}$, i.e., the proximity of subjects $s1$ and $s2$. The 190 computed values of δ_{ij} were arranged in a 20×20 proximity matrix Δ.

(D) The MDS was performed on the proximity matrix. The proximity matrices Δ for both normalized and non-normalized IFM were input into MDS function module to perform MDS analyses (Statistica, Statsoft Inc, OK, USA). The MDS program transforms the proximity data into Euclidean distances in a low-dimensional space. The Euclidean distances and the monotonic function that transforms the proximities into distances[7] were estimated by minimizing the following stress function:

$$stress = \sqrt{\sum_i \sum_j [f(\delta_{ij} - d_{ij}]^2 / \sum_i \sum_j d_{ij}^2} \tag{10}$$

where d_{ij} are the distances and $f(\delta_{ij})$ indicates a monotonic transformation of the elements δ_{ij} of the input proximity matrix Δ. To estimate the goodness-of-fit the following values of stress are usually recommended: 0.2 - poor, 0.1 - fair, 0.05 - good [7].

(E) To identify the meaning of the dimensions in the 2- and 3-dimensional spaces yielded by the MDS, individual coordinates of the subjects along the

two/three dimensions (using both non-normalized and normalized IMFs) were correlated with the following variables: (1) sum of finger forces during four finger (IMRL) task; (2) - (5) individual finger forces during single finger (I, M, R, L) tasks, N; (6)-(9) shares of finger contribution into the total force during a four-finger task, % (I/IMRL, M/IMRL, R/IMRL, L/IMRL); (10) location of the resultant force of the four fingers along the medio-lateral axis, mm (Location) [10]; (11) and (12) traces of the IFMs, both normalized and non-normalized, - they represent the total sum of the 'direct' finger forces, in N and percents, respectively; (13) and (14) sums of the off-diagonal elements of IFMs, both normalized and non-normalized,- they represent the total amount of the enslaved finger forces, in N and percents, respectively; (15) and (16) the size of the hand, its length and width, respectively.

Number of dimensions. The stress as a function of the number of dimensions (the so-called 'scree plot') is presented in Fig. 3. For non-normalized IFMs the stress can be represented sufficiently well in two dimensions (although the three-dimensional representation is more accurate) while at least three dimensions are necessary to account for the individual differences in the normalized IFMs. The decrease in dimensionality achieved by the MDS was substantial: from the 20×20 proximity matrix to 2- or 3-dimensional spaces.

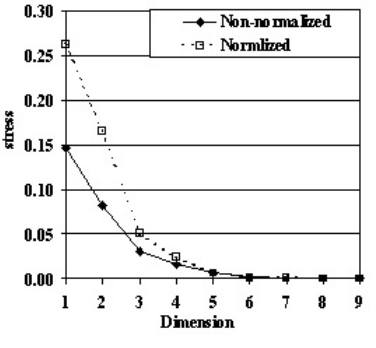


Fig. 3. Scree plots (stresses vs. number of dimensions). For non-normalized IFMs the stress was 0.098 at two dimensions and it was 0.027 at three dimensions. The stress for the normalized IFMs equaled 0.102 at two dimensions and it was 0.045 at three dimensions.

Interpretation of the dimensions. To interpret the meaning of the computed dimensions, they were regressed on a set of finger force parameters described in the text. For the non-normalized IFMs an interpretable dimension was the strength of the subjects. For the normalized IFM's, two dimensions were interpreted: (a) the location of the point of resultant force application along the medio-lateral axis that is defined by the pattern of force sharing among the fin-

gers and (b) the total contribution of the enslaved forces into the total finger force.

In summary, the employed set of methods: experimental recording of finger→ forces artificial network modeling and determining of individual IFMs→ computing differences (proximities) between the individual IFMs→MDS offers promise for future research.

References

1. Bose, N.K., Liang, P.: Neural network fundamentals with graphs, algorithms and applications. McGraw-Hill, New York (1996)
2. Danion, F., Schoner, G., Latash, M.L., Li, S., Scholz, J.P., Zatsiorsky, V.M.: A mode hypothesis for finger interaction during multi-finger force production tasks. Biol Cybern **88** (2003) 91–98
3. Gao, F., Li, S., Li, Z.M., Latash, M.L., Zatsiorsky, V.M.: Matrix analyses of interaction among fingers in static force production tasks. Biol Cybern **89** (2003) 407–414
4. Hager-Ross, C.K., Schieber, M.H.: Quantifying the independence of human finger movements: comparisons of digits, hands, and movement frequencies. J Neurosci **20** (2000) 8542–8550
5. Huijing, P.: Muscular force transmission: a unified, dual or multiple system? A review and some explorative experimental results. Arch Physiol Biochem **107** (1999) 292–311
6. Kilbreath, S.L., Gandevia, S.C.: Limited independent flexion of the thumb and fingers in human subjects. J Physiol **479** (1994) 487–497
7. Kruskal, J.B., Wish, M.: Multidimensional Scaling. Sage Pub., London (1978)
8. Leijnse, J.N.: Measuring force transfers in the deep flexors of the musicians' hand :theoretical analysis, clinical examples. J Biomech **30** (1997) 873–882
9. Li, Z.M., Latash, M.L., Newell, K.M., Zatsiorsky, V.M.: Motor redundancy during maximal voluntary contraction in four-finger tasks. Exp Brain Res **122** (1998) 71–78
10. Li, Z.M., Latash, M.L., Zatsiorsky, V.M.: Force sharing among fingers as a model of the redundancy problem. Exp Brain Res **119** (1998) 276–286
11. Li, Z.M., Zatsiorsky, V.M., Latash, M.L., Bose, N.K.: Anatomically and experimentally based neural networks modeling force coordination in static multi-finger tasks. Neurocomputing **47** (2002)259–275
12. Schieber, M.H.: Individuated finger movements of rhesus monkeys: a means of quantifying the independence of the digits. J Neurophysiol **65** (1991) 1381–1391
13. Schieber, M.H., Hibbard, L.S.: How somatotopic is the motor cortex hand area. Science **261** (1993) 489–492
14. Schieber, M.H., Poliakov, A.V.: Partial inactivation of the primary motor cortex hand area: effects on individuated finger movements. J Neurosci **18** (1998) 9038–9054
15. Shinohara, M., Li, S., Kang, N., Zatsiorsky, V., Latash, M.: Effects of age and gender on finger coordination in maximal contractions and submaximal force matching tasks. J Appl Physiol **94** (2003) 259-270
16. Slobounov, S., Chiang, H., Johnston, J., Ray, W.: Modulated cortical control of individual fingers in experienced musicians: an EEG study. Electroencephalographic study. Clin Neurophysiol **113** (2002) 2013–2024

17. Zatsiorsky, V.M., Gregory, R.W., Latash, M.L.: Force and torque production in static multifinger prehension: biomechanics and control. I. Biomechanics. Biol Cybern **87** (2002) 50–57
18. Zatsiorsky, V.M., Li, Z.M., Latash, M.L.: Coordinated force production in multi-finger tasks: finger interaction and neural network modeling. Biol Cybern **79** (1998) 139–150
19. Zatsiorsky, V.M., Li, Z.M., Latash, M.L.: Enslaving effects in multi-finger force production. Exp Brain Res **131** (2000) 187-195

Biomechanics of Bone Cement Augmentation with Compression Hip Screw System for the Treatment of Intertrochanteric Fractures

S.J. Lee[1], B.J. Kim[1], S.Y. Kwon[2], and G.R. Tack[3]

[1]Department of Biomedical Engineering, Inje University, 607 Obang-Dong, Kyung-nam, Korea
{sjl, kbj93}@bme.inje.ac.kr
[2]Department of Orthopaedic Surgery, Youido St' Mary's Hospital, Catholic University, Seoul, Korea
sykwon@cmc.cuk.ac.kr
[3]Dept. of Biomedical Engineering, Konkuk University, 322 Danwol-Dong, Chungju, Chungbuk, Korea
grtack@kku.ac.kr

Abstract. Recently, cement augmentation technique was introduced to enhance screw fixation within the femoral head. This study investigates biomechanical efficacies of cement augmentation technique. Finite element models of the femur with sliding hip screw assemblies were constructed with and without bone cement augmentation. Appropriate contact conditions with varying friction coefficients were assigned to simulate the fracture planes and other interfacial regions. With cement augmentation, 80% reduction in stresses was found in the cancellous bone, suggesting reduced possibility of cancellous bone fracture and screw cut-outs. The peak von Mises stress within the cement mantle was about 1/3 of its fatigue strength, which suggested the longevity of the cement mantle and less likelihood of osteolysis due to cement debris. Micromotions at the hip screw interfaces were also dwon from 0.275mm to 0.008mm, an indication for strong fixation after the surgery.

1 Introduction

Recent findings estimate that intertrochanteric femoral fractures (IFF) occur in more than 200,000 patients each year in the United States alone, with reported mortality rates ranging from 15% to 20% [1]. Most IFFs occur in patients over 70 years of age and they are likely to increase each year as the population of these age group increases. Traffic accidents contribute significant increase in patient numbers as well.

One of the most commonly used surgical treatments for IFF is using internal fixation devices such as sliding compression hip screws with side plate assemblies. These devices are considered to be safe with minimum amount of drilling of the cancellous bone in the femoral head and neck region. They allow early weight-bearing and limit interfacial movement on the fracture plane by providing strong compressive forces.

Clinical studies show, however, that superior cutting-out through the femoral head by the sliding hip screw is one of the major complications [2,3]. Improper screw placement has been cited as one of the major contributing factors. Frequent non-union

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1065–1072, 2004.
© Springer-Verlag Berlin Heidelberg 2004

on the fracture plane often requires revision surgery that can compound more surgical difficulties [2,4]. For patients with severe osteoporosis in which the primary compression trabecular structures have diminished in considerable proportions can be more prone to this type of complication and may add further surgical challenges. Therefore, it is very important to maintain strong fixation of the sliding hip screws within the femur with minimum amount of sliding motion for optimal bony-healing.

Biomechanical studies have been done to elucidate the effects of biomechanical factors such as the placement angle of the hip screws, positioning within the femoral head and neck, and the effect of anatomical reduction [5,6]. Recently, augmenting of hip screws with polymethylmethacrylate (PMMA) has been suggested to provide more secure fixation of the screws within the femoral region [3,4]. Clinical study by Cheng et al (1989) found that cement augmentation may provide initial stability but can lead to late complications if not properly used [7]. Bartucci et al (1985) advocated limiting the use of PMMA on the proximal side to prevent intrusion of the cement into the fracture plane [4]. Using cadaveric femurs Choueka et al (1995) investigated biomechanical effects of using various types of sliding hip screws and/or dome plunger in terms of load-bearing and fixation strength [2].

This study was designed to investigate the biomechanical feasibility of the bone cement augmentation technique. The changes in stress distributions due to cement augmentation were studied. For this purpose, finite element models of the operated femur with sliding hip screw assemblies were constructed with and without bone cement augmentation. Close attentions were given to the peak von Mises stress within the cement mantle region to assess the likelihood of cement failure that might lead to osteolysis due to cement debris. In addition, changes in micromotion at the interfacial junctions such as screw-cement-surrounding cancenllous bone and fracture plane were assessed to study the fixation strength that is critical for post-op bony healing.

2 Material Methods

A three-dimensional finite element model of the femur was constructed using geometric data acquired from computed tomography (CT) scans (Fig. 1). Intertrochanteric fracture was simulated by assigning a fracture plane from the greater trochanter to the lesser trochanter at 30 degrees with respect to the long axis of the femur (Fig. 2). Based on this hip fracture model, two types of surgical model were generated. One is the non-cemented model (NC) in which only the hip screw assembly was added to the femur and the other is the cement-augmented model (CA) where the cement mantle was added to surround the hip screw (Fig. 3).

The dimension of the hip screw assembly as well as screw profiles(screw threads, pitches, inner and outer diameters) were based on the Richard compression hip screw (Osteo, Swiss). The distance from the apex of the femoral head to the tip of the screw known as TAD (Tip-Apex Distance) was set at 20-24mm near the center of the femoral head according to suggestion by Baumagaertner ea al [8]. The average thickness of the cement mantle was determined as 12.5mm distributed uniformly over the hip screw based on the clinical experience of one of the authors (SYK).

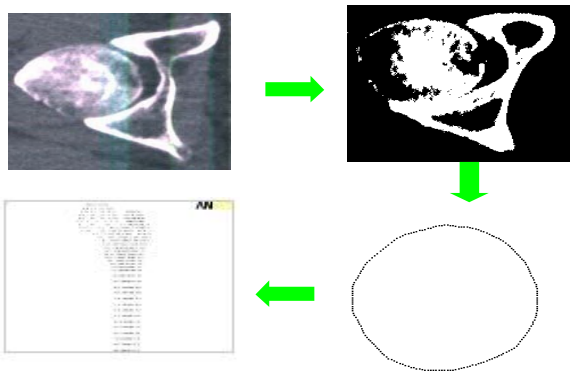


Fig. 1. Geometric data acquired from CT scans

The material properties for the anatomical and implants were based on the findings from literatures (Table I, [9]). To simulate the interfacial conditions, appropriate friction coefficients (□) were assigned at the fracture plane (□=0.5), the interface between screw and surrounding cancellous bone (□=0.5), and between the screw and the cement (□=0.3). The cement-bone interface was assumed to be rigidly and the side plate assembly was rigidly fixed to the lateral aspect of the femur. The sources for the friction coefficient are also listed Table I. 2014N of compressive loading in a cubic cosine distribution was simulated on the top of the femoral head at angles of 12 and 26 degrees in sagittal and transverse planes, respectively (Fig. 4). This loading condition is designed to simulate the after heel strike phase in a gait cycle in which the highest compressive load is applied to the femur. It was assumed that the distal part of the model was fixed in all directions. Eight-noded 3-D brick elements were used for the entire model except for the interfacial regions where the 3-D contact elements were used.

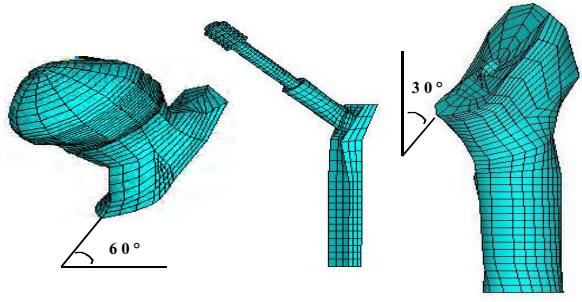


Fig. 2. Construction of a finite element model

3 Results

Strain data from three different locations on the cortical surface of the femur were compared for the validity check (Table II). Here, two sets of results were in a very close agreement with each other, thereby confirming the validity of our model.

Results showed that cement augmentation had resulted in decrease in stresses in the hip screw and in the bone, both cortical and cancellous (Table III), an indication of favorable stress transfer due to the addition of bone cement. Highest decrease in stresses was noted with cement augmentation at the cancellous bone region (reduction of 80%), which makes the further fracture of the femur far less likely in cement-augmented (CA) case than in non-cemented (NC).

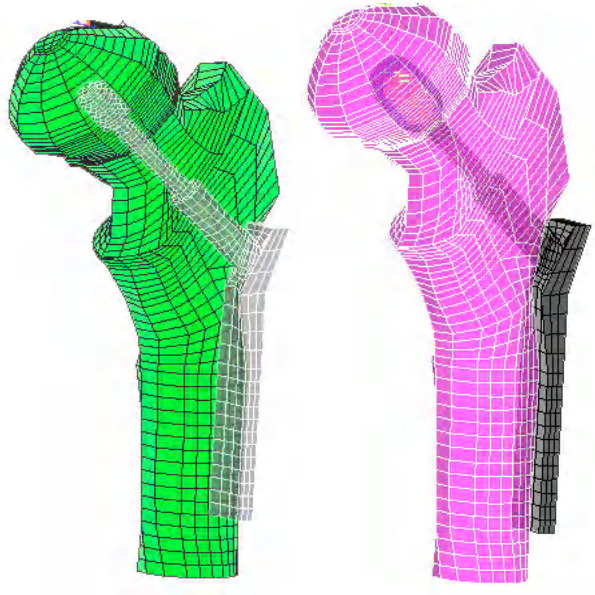


Fig. 3. FE models, (A) non-cemented(NC) (B) cement-augmented(CA)

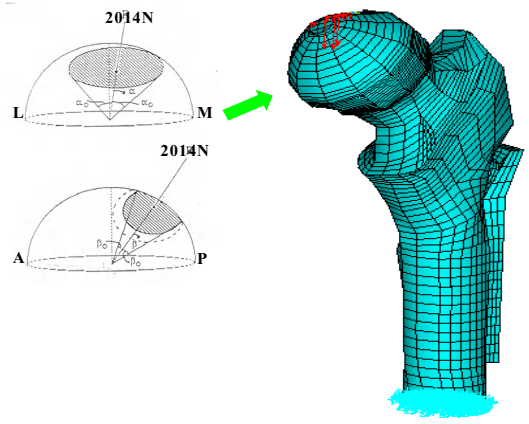


Fig. 4. Loading & boundary conditions of a finite element model

Table 1. Mechanical properties of the proximal femur for finite element modeling

		Elastic modulus (MPa)	Possion's ratio (ν)	Coefficient of friction
Cortical bone	Subchondral bone	2,000	0.32	
	Pure cortical bone	14,000	0.32	
Cancellous bone	Femoral head	550	0.32	
	Proximal region	411	0.32	
	Distal region	345	0.32	
Bone cement		2,200	0.23	
Compression hip screw		200,000	0.30	
Cement-screw				0.3 [13]
Cancellous bone – screw				0.5 [14]
Fracture plane				0.5 [15]

In fact, the peak von Mises stresses (PVMS) of 27MPa was aspsessed at the screw-bone junction in the cancellous bone with NC. This exceeded its yield strength of 22MPa suggesting the likelihood of micro-fracture in this region. On the other hand, with CA the corresponding PVMS decreased to 5MPa, making it far less susceptible to fracture. The PVMS at the screw was considerably lower than its yield strength. For example, the PVMS at the hip screw were only 322MPa, slightly above one third of its yield strength of 860MPa (stainless steel). The peak von Mises stress at the cement region was about 9MPa that was far less than 1/3 of its reported shear strength of 30MPa and endurance limit of 28MPa at one million cycles [11].

The micromotions at the screw-bone (NC) and the screw-cement (CA) interfaces were 0.275mm and 0.008mm, respectively. The high level of micromotion in NC coupled with high PVMS that exceeds the yield strength may all contribute the eventual cut-out of the screw because the microscale failure of the cancellous bone and increased micromotion may be initiated

Table 2. Validation results

Method	Experimental results from Oh et al [8]	Results from the present study using FEM
Location	Mean ± SD (μ ϵ)	
Medial 1	1827 ± 601	1273.6
Medial 2	1419 ± 628	988.4
Lateral	1019 ± 404	623.2
Remarks	Results from strain gauge measurement	Predicted values from matching nodes

Table 3. Comparison of peak von Mises stress (PVMS) for the non-cemented (NC) and the cement-augemented (CA) cases

Model location	Non-cemented (NC)	Cement- augmented (CA)	relative %change in PVMS	Yield Strength (MPa)
	PVMS (MPa)	PVMS (MPa)		
Hip screw region	322.61	174.97	-45.76	860.00
Cement region		9.16		30.00
Cortical region	48.81	54.63	11.92	560.00
Cancellous region	27.74	4.88	-82.41	22.00

4 Discussion

This study investigated the biomechanical advantages of adding bone cement to reinforce the hip screw fixation during the surgical treatment of intertrocanteric fracture of the femur by using clinically relevant finite element (FE) models. Although the intertrochanteric fractures often occur in elderly patients with osteoporosis, the material properties of the bone in this study used those of the normal person in order to elucidate its feasibility in a simplified way. It would be very interesting to conduct similar studies for varying degrees of osteoporosis to assess its feasibility by introducing appropriate conversion values for the elastic stiffness in accordance with the BMD data of a given patient.

Our results clearly suggested the biomechanical advantages of bone cement augmentation. In particular, the most drastic reduction in stresses was seen at the cencellous bone. Almost 80% of reduction was noted. This makes the cut-out of the screw that has been advocated as one of the major complications of the hip screw systems far less likely to take place. In fact, the PVMS with non-cemented case was higher than its yield strength (27MPa vs. 22 MPa), a sign of impending loosening of the hip screw at the screw-bone interface that may progress to cut-out of the screw.

No difference in micromotion at the fracture plane was assessed between two cases (CA and NC). At the cement-bone interface, we assumed that interdigitization of cement took place with the irregularities of the bone (i.e., microinterlock). Therefore, we did not assign any friction coefficient. Rather we took it as continuity (i.e., the nodes are shared) and no micromotion was assessed here. It should be noted that at the interface between screw and cement, we adopted the friction coefficient value of 0.3 as suggested by Mann et al [13], which was lower than the that of screw-bone interface (μ = 0.5, suggested by Shrazi-Adl et al [14]). The initial micromotion assessed immediately after surgery has been cited as one of the indicators whether successful bony healing can take place. Studies have shown that micromotion should be less than (0.15mm) for bony fusion at fracture sites [12]. Based on the magnitudes of the micromotion predicted from this study (0.008mm with CA), cement augmentation could reduce the risk of non-union or delayed union of the fracture that might have had progressed otherwise (0.275mm in NC). Our results also demonstrated that despite the lower friction coefficient assigned for the screw-cement interface (CA) its micromotion was substantially less than that of non-cemented (NC). This may due to the fact that the deformation is inversely proportional to the elastic

modulus. It appears higher modulus of bone cement (E= 22000MPa) as opposed to the cancellous bone (E= 550MPa) overcame the relative lack of friction and resulted in less micromotion.

The peak stress found in cement mantle was about less than 1/3 of its fatigue strength. This suggests that the bone cement mantle may theoretically withstand the repeated load indefinitely under the loading condition provided in this study. Although much in-depth study should be done, this result may relieve clinicians from worries about the likelihood of debris-causing fatigue failure that may initiate the vicious cycles of osteolysis and implant loosening

5 Conclusion

Our finite element analysis was able to show the following advantages of bone cement augmentation technique in the treatment of IFF. They can be summarized as:

- A significant reduction in stresses in the cancellous bone: This indicates that further fractures of the cancellous bone or the screw cut-outs are far less likely to take place.
- Cement mantle could withstand against the physiological load and further against the repeated fatigue load.
- Reduced micromotions at the interfacial regions suggest stronger chance of bony healing.

This work was supported by a grant from Korea Science & Engineering Foundation (R01-2001-0500).

References

1. Sledge, C.B.: *Master technique; in orthopaedic surgery, The Hip*, Lippincott-Raven, Philadelphia (1998) 106
2. Choueka, J., Koval, K.J., Kummer, F.J., Crawford, G.: Biomechanical comparison of the sliding hip screw and the dome plunger. The Journal of Bone and Joint Surgery (Br), 77(2) (1995) 277-283
3. Choueka, J., Koval, K.J., Kummer, F.J., Zuckerman, J.D.: Cement augmentation of intertrochanteric fracture fixation. Acta Orthop Scand 67(2) (1996) 153-157
4. Bartucci, E.J., Gonzalez, M.H., Cooperman, D.R., Freedberg, H.I., Barmada, R., Laros, G.S.: The effect of adjunctive methylmethacrylate on failures of fixation and function in patients with intertrochanteric fractures and osteoporosis. The Journal of Bone and Joint Surgery (Am) 60(7) (1985) 1094-1107
5. Hartog, B.D., Bartal, E., Cooke, F., Kansas, W.: Treatment of the unstable intertrochanteric fracture. The Journal of Bone and Joint Surgery (Am), 73-A(5) (1991) 726-733
6. Kyle, R.F., Wright, T.M., Burstein, A.H.: Biomedical analysis of the sliding characteristics of compression hip screws. The Journal of Bone and Joint Surgery (Am), 62-A(8) (1980) 1308-1314

7. Cheng, C.L., Chow, S.P., Pun, W.K., Leong, J.C.: Long-term results and complications of cement augmentation in the treatment of unstable trochanteric fractures. Injury 20(3) (1989) 134-138
8. Baumgaertner, M.R., Curtin, S.L., Lindskog, D.M., Keggi, J.M.: The value of the tip-apex distance in predicting failure of fixation of peritrochanteric fractures of the hip. The Journal of Bone and Joint Surgery (Am), 77-A(7) (1995) 1058-1064
9. Brown, T.D., Ferguson, J.A.: Mechanical property distributions in the cancellous bone of the human proximal femur. Acta Orthop Scand 13 (1980) 687-699
10. Oh, I.D., Harris, W.H.: Proximal strain distribution in the loaded femur. The Journal of Bone and Joint Surgery (Am) 60(1) (1978) 75-85
11. Kuhn, K.D.: *Bone cement, Springer*; Berlin Heidelberg New York (2000) 19-20; 142-147
12. Pilliar, R.M., Lee, J.M., Maniatopoulos, C.: Observations on the effect of movement on bone ingrowth into porous-surfaced implants. Clinical Orthopaedics and Related Research 208 (1986) 108-113
13. Mann, K.A., Bartel, D.L., Wright, T.M., Burstein, A.H.: Coulomb frictional interfaces in modeling cemented total hip replacements: a more realistic model. J. Biomechanics 28(9) (1995) 1067-1078
14. Shirazi-Adl, A., Dammak, M., Paiement, G.: Experimental determination of friction characteristics at the trabecular bone/porous-coated metal interface in cementless implants. J. Biomed. Mater. Res. 27 (1993) 167-175
15. Natarajan, R.N., Chen, B.H., An, H.S., Andersson, B.J.: Anterior Cervical Fusion; A finite element model study on motion segment stability including the effect of osteoporosis Spine 25(8) (2000) 955-961

Comparison of Knee Cruciate Ligaments Models Using Kinematics from a Living Subject during Chair Rising-Sitting

Rita Stagni, Silvia Fantozzi, Mario Davinelli, and Maurizio Lannocca

Department of Electronics, Computer Science and Systems, University of Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
{rstagni, sfantozzi, mdavinelli}@deis.unibo.it
mlannocca@libero.it
http://www-bio.deis.unibo.it/

Abstract. The knee joint is a key structure of the human locomotor system. Any lesion or pathology compromising its mobility and stability alters its function. As direct measurements of the contribution of each anatomical structure to the joint function are not viable, modelling techniques must be applied. The present study is aimed at comparing cruciate ligaments models of different complexity using accurate parameters from RMN and 3D-fluoroscopy of a single selected subject during chair rising-sitting motor task.. The complexity of the model was not relevant for the calculation of the strain range of the cruciate ligaments fibres. On the other hand, three-dimensionality and anatomical twist of the modelled fibres resulted to be fundamental for the geometrical strain distribution over the ligament section.

1 Introduction

The knee plays a fundamental role in determining the human locomotor ability. Any alteration of its anatomical structures can compromise its function. The development of effective methods for surgical reconstruction and rehabilitation is of great clinical interest, regarding both joint replacement and surgical reconstruction of the main anatomical structures. This interest is demonstrated by the 259000 total knee replacements, 25000 ligaments reconstructions and 15000 other repairs of the knee performed in the USA in 1997 as reported by the American Association of Orthopaedic Surgeons (AAOS). For the development of these procedures, an accurate knowledge of the mobility and stability of the whole articular structure, as well as of its different anatomical sub-units, is necessary. The need for this deeper knowledge led to a bulk of in-vitro and in-vivo studies, which allowed to clarify several aspects of the physiological behaviour of this complex joint. In-vitro testing allows to directly observe and measure different aspects of joint mechanics, but not in physiological conditions. During its normal function, the knee lets the shank move with respect to the thigh, maintaining the stability of the structure under articular load and torque. These are the result of several contributions: the inter-segmental contact load, ligament tensioning, loads applied by the muscles, the inertia of body segments. All these contributions are strongly dependent on the analysed motor task, as well as on

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1073–1080, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the physical characteristics of the subject. Thus, if we want to quantify the contribution of each anatomical structure in determining the physiological function of the knee, modelling is the only possible solution, as direct measurements cannot be performed.

The problem of knee modelling has been approached at different levels of complexity. Two-dimensional models were designed in order to investigate the role of the cruciate ligaments in simple conditions, such as isometric quadriceps contraction [1,2]. Three-dimensional models, including articular surfaces and ligaments, were also proposed. Even these more complex models were applied in conditions far away from those of the physiological knee [3,4,5,6]. The natural evolution of this approach is inserting the model into a context which allows to evaluate the boundary conditions of the knee-structure during the performance of a simple task of daily living [7]. Even if the model is designed properly for the application devised, its potentials can be nullified by the effect of errors within the definition of subject parameters and during the acquisition of experimental inputs. In previous modelling attempts, these errors were due to discrepancies in the origin of parameters and inputs, which were often obtained from different and non-homogeneous subjects.

In order to avoid this possible source of error, in this paper, different cruciate ligament models were compared using parameters from a single selected subject analysed as accurately as possible. The specific geometry of articular surfaces and ligaments insertions were reconstructed using the three-dimensional reconstruction of segmented bone and soft tissues, obtained from Nuclear Magnetic Resonance (NMR). The specific accurate kinematics was obtained from cine-fluoroscopic images of a chair rising-sitting motor task. Cruciate ligaments models of different complexity: from the simple bi-dimensional untwisted one to the more realistic three-dimensional twisted with circular insertion were compared. The aim was to select the best compromise between accurate anatomical description and model simplicity for the investigation of knee biomechanics.

2 Material and Methods

Overview. A subject-specific model of the right knee of a young male living subject (height 168 cm, weight 62 kg, and age 30 years) was developed from a high resolution NMR data set. Three-dimensional outer surfaces of the biological structures of interest were generated.

The subject performed chair rising-sitting with the knee under analysis inside the fluoroscopic field of view. The accurate 3D pose of the bones was reconstructed by means of single-plane lateral 2D fluoroscopic projections and relevant models previously obtained.

The cruciate ligaments fibres were modelled with six geometrical equivalents and relative fibres strain compared: 2D, 3D with rectangular insertions, 3D with circular insertion, each twisted and untwisted.

The NMR data set. A data set of high resolution NMR images was collected with a 1.5T Gemsow scanner (GE Medical Systems, Milwaukee, Wisconsin). Details of the scanning parameters are shown in Table 1.

Table 1. The NMR scanning procedure parameters

Scanning sequence	Spin Echo (T1 weighted)
Number of slices	54
Pixel spacing	0.037x0.037 (cm·cm)
Scanned region length (across the knee)	15.9 (cm)
Slice thickness	2.5 (mm)
Slice spacing	3 (mm)

The segmentation procedure. A 3D tiled surface geometrical representation was generated using the software Amira (Indeed - Visual Concepts GmbH, Berlin, Germany), for the distal femur, the proximal tibia, and the insertion areas of the anterior (ACL) and posterior cruciate ligaments (PCL).

A segmentation of the NMR data set was performed with an entirely manual 2D segmentation technique. For each slice, the outer contour of the structures of interest was detected and outlined, as shown in Fig. 1. The resulting stacks of contours were interpolated to generate polygonal surfaces which represent the outer boundary of the objects to be modelled. The model used for the kinematic analysis is shown in Fig. 2.

Fig. 1. Outlined contours of femur and ligaments in a slice of the NMR data set

Kinematics. Series of lateral images were acquired at the frequency of 6 samples per second with a standard fluoroscope (SBS 1600, Philips Medical System Nederland B.V.). Images of a 3D cage of Plexiglas with 18 tantalum balls in known positions and of a rectangular grid of tin-leaded alloy balls 5 mm apart were collected in order to calculate respectively the position of the camera focus and the parameters necessary for image distortion correction. This was obtained using a global spatial warping technique[8]. An established technique for 3D kinematics analysis of a known object from a single view was implemented [9] (Fig. 3). Bone poses in space were obtained from each fluoroscopic image by an iterative procedure using a technique based on tangent condition between projection lines and model surface.

Previous validation work on prosthesis components [9] showed that relative pose can be estimated with an accuracy better than 1.5 degrees and 1.5 mm.

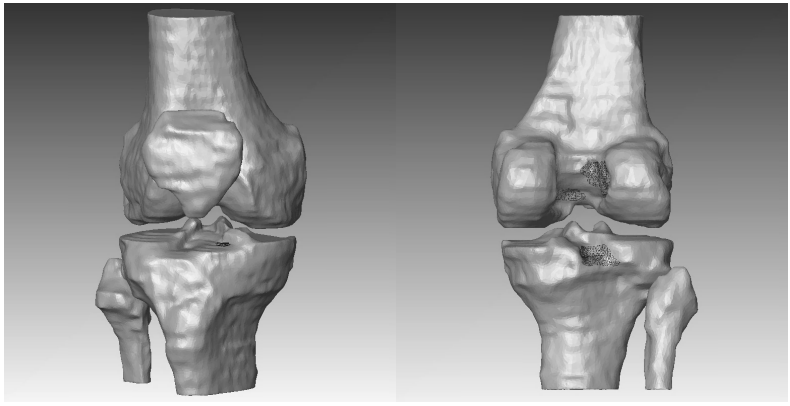


Fig. 2. Anterior and posterior view of the complete knee model. The areas of insertion of ligaments are the dotted regions on the femur and the tibia

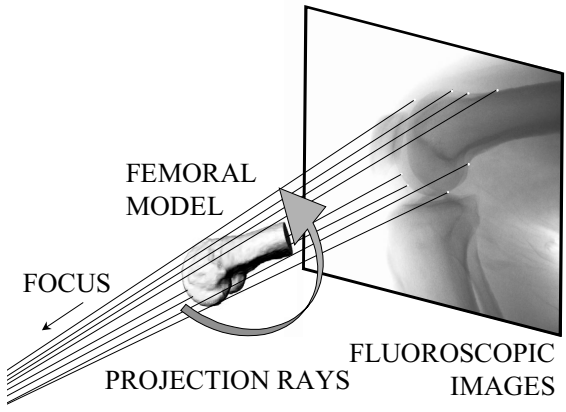


Fig. 3. Sketch of the model for fluoroscopic image generation process

Cruciate ligament geometrical models. The geometrical models of the cruciate ligaments differ for dimension, shape of insertion and twist:

1. Bi-dimensional - untwisted. The insertions were modelled as follows: a) the line that minimizes in a least squares sense the points of the insertion area was calculated, b) insertion segment was identified on this line between the anterior and posterior limits of the insertion surface, and c) 25 uniformly distributed points were identified on this segment. Thus, 25 fibres were modelled for both ACL and PCL. In both ligaments the fibres connected these points of the insertions from the most

posterior to the most anterior on both femur and tibia, i.e. the most posterior point of femur insertion was connected to the most posterior point of the tibia insertion.

2. Bi-dimensional - twisted. The insertions and the fibre were modelled as in model 1 except for the fact that the points from the most posterior to the most anterior of the femur were connected to those from the most anterior to the most posterior of the tibia.
3. Three-dimensional - rectangular insertions - untwisted. The insertions were modelled as follows: a) in the plane approximating the insertion points in a least square sense a rectangle including 80% of these points was estimated, and b) a 5x5 uniform grid of points was identified on the rectangle. In both ACL and PCL the 25 fibres connected points of the insertions with no twisting.
4. Three-dimensional - rectangular insertions - twisted. The insertions and the fibre were modelled as in model 3 except for the fact that in both ligaments a twist angle of 90° was introduced.
5. Three-dimensional - circular insertions - untwisted. The insertions were modelled as follows: a) in the plane approximating the insertion points in a least square sense a circle including 80% of these points was estimated, and b) a 25 uniformly distributed points were identified on the circle. In both ACL and PCL the 25 fibres connected points of the insertions with no twisting.
6. Three-dimensional - circular insertions - twisted. The insertions and the fibre were modelled as in model 5 except for the fact that in both ligaments a twist angle of 90° was introduced.

For each model, for each single fibre, the strain, ε, was calculated as follows:

$$\varepsilon(t) = \frac{L(t) - L_0}{L_0} \tag{1}$$

were L is the length of the fibre at time sample t, and L_0 is the maximum length the fibre reached during the motor task.

3 Results

The modelled PCL always showed a larger elongation, with an average strain of about 33% versus 19% of the ACL. The strain calculated for the fibre approximately connecting the mean point of the insertions of the ACL and PCL was equivalent for all ligament models. The range of the strains calculated for the ACL [-16%;-20%] and for the PCL fibres [-24%;-38%] was similar for the different models. The geometrical distribution of the strain over the ligament section resulted model-dependent.

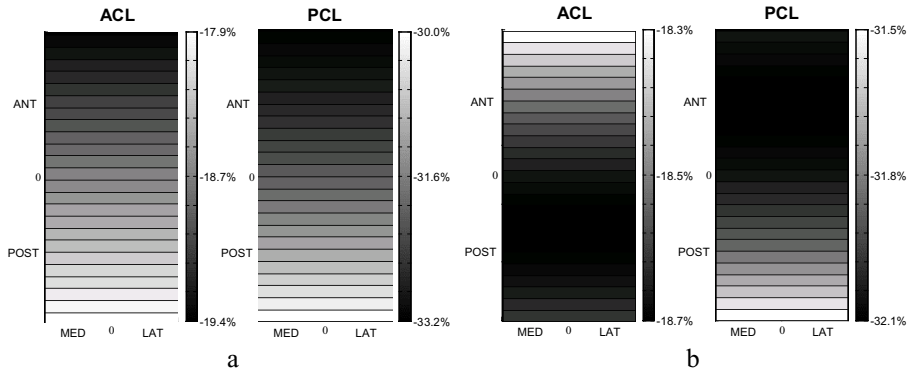


Fig. 4. The maximum value of strain over the section of modelled ligament during the execution of the motor task for each of the 25 fibres is plotted for model 1(a) and model 2(b)

The strain calculated for the other fibres resulted also model dependent, in particular the bi-dimensional models produced different results with respect to the three-dimensional ones.

For the bi-dimensional models (Fig.4) the PCL showed the largest strain at the anterior fibres independently from the twist. The strain of the fibres of the ACL was larger for the anterior ones when untwisted and for the posterior ones when twisted.

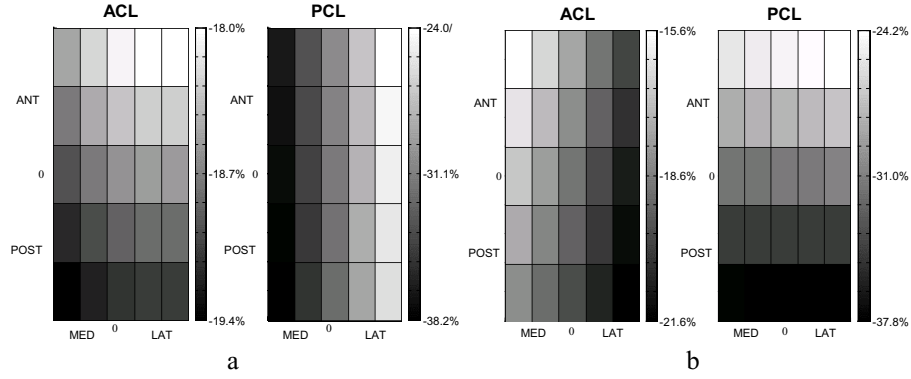


Fig. 5. The maximum value of strain over the section of modelled ligament during the execution of the motor task for each of the 25 fibres is plotted for model 3 (a) and model 4 (b)

The strain behaviour of the fibres was similar for the two three-dimensional models (Fig.5 and Fig.6). The largest strain resulted for the ACL at the postero-medial fibres when untwisted, and for the postero-lateral ones when twisted, for the PCL the medial fibres when untwisted and for the posterior fibres when twisted.

4 Discussion

Six different cruciate ligament models were compared using parameters from a single selected subject analysed as accurately as possible. Plane, rectangular and circular

sections were considered, and the mechanical effect of the anatomical twisting of the ligament fibres was also investigated.

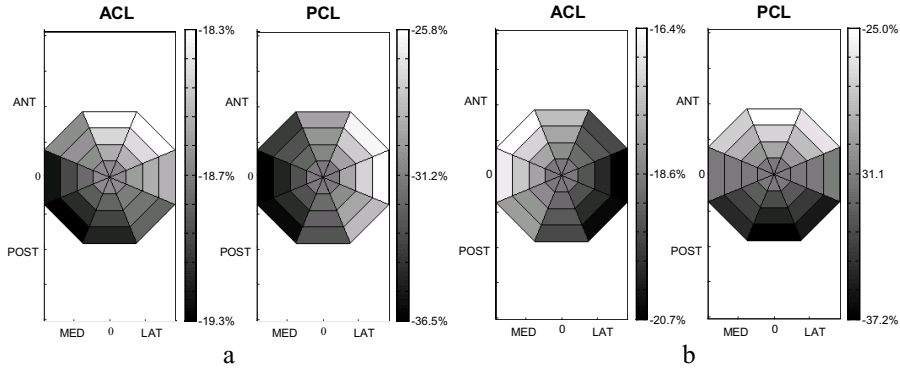


Fig. 6. The maximum value of strain over the section of modelled ligament during the execution of the motor task for each of the 25 fibres is plotted for model 5 (a) and model 6 (b)

The strain range of the modelled fibres was not relevantly influenced by the model adopted, which resulted to influence the geometrical distribution of the strain over the fibres in the section of the ligament.

The more conventional bi-dimensional model [10] showed the largest differences from the two three-dimensional ones. No significant difference could be highlighted between the rectangular insertion and the circular insertion three-dimensional models.

The twist showed significant influence in the strain distribution for each model.

In conclusion, when only the magnitude of the fibres elongation is to be calculated the selected model does not considerably affect the results. Instead, the model should be accurately selected when the geometrical distribution of the strain over the section of the ligament is required, i.e. when the strain is used for the calculation of the load applied to the joint by the ligament [11]. In this case, a three-dimensional model is suggested, independently from the selected insertion shape, and the anatomical twist of the fibres has to be taken into account, as it strongly influences the strain distribution over the section.

References

1. Gill, H.S., O'Connor, J.J.: Biarticulating two-dimensional computer model of the human patellofemoral joint. Clin Biomech 11 (1996) 81-89
2. Lu, T.W., O'Connor, J.J.: Lines of action and moment arms of the major force-bearing structures crossing the human knee joint: comparison between theory and experiment. J Anat 189 (Pt 3) (1996) 575-585
3. Huss, R.A., Holstein, H., O'Connor, J.J.: The effect of cartilage deformation on the laxity of the knee joint. Proc Inst.Mech.Eng [H.] 213 (1999) 19-32
4. Kwak, S.D., Blankevoort, L., Ateshian, G.A.: A Mathematical Formulation for 3D Quasi-Static Multibody Models of Diarthrodial Joints. Comput Methods Biomech Biomed Engin. 3 (2000) 41-64

5. Mommersteeg, T. J., Blankevoort, L., Huiskes, R., Kooloos, J. G., Kauer, J. M.: Characterization of the mechanical behavior of human knee ligaments: a numerical-experimental approach. J Biomech 29 (1996) 151-160
6. Mommersteeg, T.J., Huiskes, R., Blankevoort, L., Kooloos, J.G., Kauer, J.M., Maathuis, P.G.: A global verification study of a quasi-static knee model with multi- bundle ligaments. J Biomech 29 (1996) 1659-1664
7. Piazza, S.J., Delp, S.L.: Three-dimensional dynamic simulation of total knee replacement motion during a step-up task. J Biomech Eng 123 (2001) 599-606
8. Gronenschild, E.: The accuracy and reproducibility of a global method to correct for geometric image distortion in the x-ray imaging chain. Med.Phys. 24 (1997) 1875-1888
9. Zuffi, S., Leardini, A., Catani, F., Fantozzi, S., Cappello, A.: A model-based method for the reconstruction of total knee replacement kinematics. IEEE Trans.Med.Imaging 18 (1999) 981-991
10. Zavatsky, A.B., O'Connor, J.J.: A model of human knee ligaments in the sagittal plane. Part 1: Response to passive flexion. Proc Inst.Mech.Eng [H.] 206 (1992) 125-134
11. Zavatsky, A.B., O'Connor, J.J.: A model of human knee ligaments in the sagittal plane. Part 2: Fibre recruitment under load. Proc Inst.Mech.Eng [H.] 206 (1992) 135-145

Computer and Robotic Model of External Fixation System for Fracture Treatment

Yoon Hyuk Kim and Soon-Geul Lee

School of Advanced Technology, Kyung Hee University, 1 Sochon-ni, Kihung-eop,Yongin-shi, Kyongki-do, 449-701, Korea
yoonhkim@khu.ac.kr

Abstract. A computer graphic model combined with a robotic model of a unilateral external fixation system for a fractured tibia was developed to simulate the pre-operative planning of bone fracture and deformity correction by adjustment of fixator joints. An inverse kinematics analysis algorithm was developed to quantify the necessary rotations and translations at each joint of the fixator. The graphic model was developed using commercial CAD software, and the robotic model was developed to have same configurations. The accuracy of each motor in the robotic model was successfully tested. For a given rotational deformity with a fracture gap, the simulation results of the fracture reduction process were fully matched with the robotic execution. The presented models and techniques can be used for knowledge-based fracture treatment and bone deformity correction using external fixators.

1 Introduction

External fixation has been widely used in bone fracture fixation and deformity correction [1]. External fixation has a distinct advantage in that it allows adjustment of the bone deformity intra-operatively in one setting or post-operatively through gradual distraction of the bone segments at the osteotomy site [2,3]. Furthermore, it allows ongoing monitoring and correction of bone segment alignment to avoid residual deformity in rotation and translation. In addition to stabilization and adjustment, external fixation also provides mechanical stimulation at the fracture site through the elastic properties of the pins and the joints.

However, it is often difficult to accurately correct a given deformity because optimal adjustment is determined by the trained eye of a clinician, and not by biomechanical data, which is the necessary for precise adjustment of the fixator joints [4]. In addition, the length of radiographic exposure to the clinicians during the procedure can be prolonged since bone segment alignment is evaluated by empirical methods and trial-and-error. It would be very useful to determine the precise fixator joint adjustments needed to accurately execute the correction plan. The plan could also be validated pre-operatively using computer simulation techniques [5].

* This study was supported in part by Res. Inst. of Med. Instr. & Reh. Eng., Kosef, Korea.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1081–1087, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Robotic systems are used in various industries to provide stable and accurate manipulation of objects. Such robotic systems may also be applied to the accurate and efficient execution of fracture reduction and deformity correction. To the author's knowledge, there have been no studies thus far attempting to apply robotic systems towards fracture treatment execution.

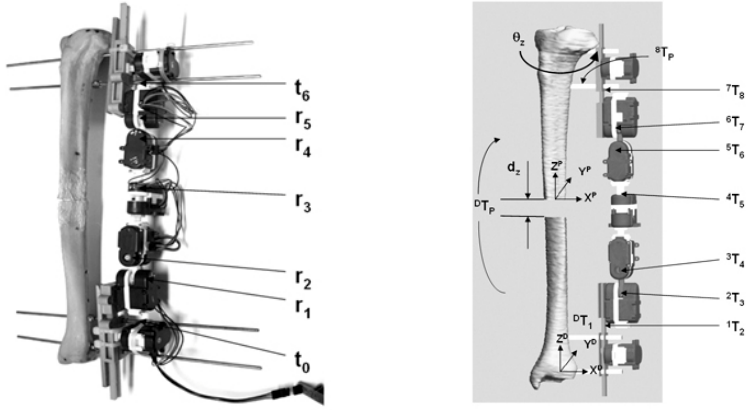
The objective of this study was to develop an open-link type robotic external fixator model combined with a corresponding computer model of a unilateral external fixator to simulate the required adjustments necessary for optimal bone fracture reduction and deformity correction. The inverse kinematics analysis algorithm was developed to determine the necessary fixator joint adjustments under given bone deformities and fixator application configurations. A computer graphic model of both the fixator and fractured tibia was developed to visualize and validate the analysis result. A robotic model with the same configuration as the computer graphic model was developed and tested. These models and simulation techniques assure precise execution of the desired deformity correction.

2 Materials and Methods

2.1 Development of Robot and Computer Model of External Fixation System

A serial link type robot model based on a clinically popular unilateral external fixator for fracture fixation and bone deformity correction (Dynafix®, EBImedical, USA) was developed in this study. The fixator robot model is composed of four pins inserted into the bones, two pin clamps, four revolute joints, and two sliding joints, respectively (Fig. 1a). In the robot model, AI-motors system® (Megarobotics, Korea) was used to construct the revolute joints, and rack-and-pinion gears were added to the revolute joints to represent the prismatic joints. The AI-motor system is a module-type motor system that has both a control and link unit. It is relatively light, small in size, and very inexpensive, making it ideal for construction of the preliminary robot model. The robot system is controlled by Matlab® (Mathworks, USA) with the accuracy of 1.3^{o} in rotation and 0.154 mm in translation, according to the manufacturer's report. Geometric dimensions of the robot model were measured and the joint types were defined to facilitate the inverse kinematic analysis.

The computer graphic model of a transversely fractured tibia at the diaphyseal area was modeled from CT data (Visible Human data, NIH, USA). The graphic model of the external fixator was developed and the bone-fixator system assembled using commercial CAD software, SolidWorks™ (Solidworks, MA, USA). Then, computer simulation of the deformity correction process was performed using commercial animation software (3D Studion Max®, Autodesk Inc., USA) in order to visualize the computed results of joint rotation and translation required to achieve the desired fracture reduction process and evaluate the performance of the robot model. In the computer model, the geometric dimensions and configurations of the fixator, including the pin insertions, are the same as those of the robot model.



(a) (b)

Fig. 1. The robot, (a), and the computer graphic, (b), models of the Dynafix® external fixator system (EBImedical, USA)

2.2 Inverse Kinematics Analysis Algorithm

The mathematical model of the motion of each link of the bone-fixator system can be represented as a open-link serial manipulator system interconnected by five revolute and two prismatic joints, thus having 7 degrees-of-freedom (DOF).

A 4 x 4 homogeneous transformation matrix was utilized to express the kinematic loop equations of the fixator-tibia system in order to define six bone deformity parameters at the fracture site [5]. The rotational sequences at the fracture site follows the x-y'-z" Euler-angle system, and the global coordinate system was fixed to the distal segment of the tibia. All local joint coordinate systems were defined identically to the global coordinate system. In the mathematical model, ${}^{D}\mathbf{T}_{P}$ is the transformation matrix from the proximal tibial segment to the distal segment, was expressed by the matrix or chain equation shown below:

$$ {}^{D}\mathbf{T}_{p} = {}^{D}\mathbf{T}_{1}\cdot{}^{1}\mathbf{T}_{2}\cdot{}^{2}\mathbf{T}_{3}\cdot{}^{3}\mathbf{T}_{4}\cdot{}^{4}\mathbf{T}_{5}\cdot{}^{5}\mathbf{T}_{6}\cdot{}^{6}\mathbf{T}_{7}\cdot{}^{7}\mathbf{T}_{8}\cdot{}^{8}\mathbf{T}_{P}. \quad (1) $$

${}^{D}\mathbf{T}_{1}$ and ${}^{8}\mathbf{T}_{P}$ represent the rigid body translations of the local coordinate systems between the bone segments and the pin clamps (Fig. 1b). The matrices ${}^{2}\mathbf{T}_{3}$, ${}^{3}\mathbf{T}_{4}$, ${}^{5}\mathbf{T}_{6}$ and ${}^{6}\mathbf{T}_{7}$ are pure rotations at the revolute joints. ${}^{4}\mathbf{T}_{5}$ represents the axial rotation at the central body. ${}^{1}\mathbf{T}_{2}$ and ${}^{7}\mathbf{T}_{8}$ represent translation of the prismatic joints at the pin clamps. After substituting the unknown fracture (${}^{D}\mathbf{T}_{P}$) and fixator geometric parameters into the transformation matrices, the seven unknown joint variables, $\mathbf{t_0}$, $\mathbf{r_1}$, $\mathbf{r_2}$, $\mathbf{r_3}$, $\mathbf{r_5}$, $\mathbf{r_5}$ and $\mathbf{t_6}$, can be determined by solving Eq. (1) (Fig. 1a). The resulting systems of non-linear equations were solved using the nonlinear least square optimization method (MATLAB™, Mathworks, MA, USA) [5].

2.3 Validation of the Models and Simulation of Fracture Treatment

To validate the developed computer and robot models of bone-fixator system, in the first, the accuracy of each revolute and prismatic joint was tested. The accuracy of a joint was defined by the difference between the arbitrarily applied joint values in the control program and the encoder values of the joint motor after joint movement.

A correction of 45° of rotational deformity combined with 20 mm of fracture gap was then simulated using the computer graphic animation and the robotic execution. Different adjustment options, such as sequential adjustment with small increments and simultaneous adjustment, were tested in order to fully illustrate the 3-D fixator adjustability and the ability of the simulation to achieve accurate alignment correction during treatment. Based on analysis results and the corresponding graphic animation, the deformity correction planning process can be clearly visualized for evaluation of the analysis results.

3 Results

3.1 Accuracy Test

Table 1 shows the error values between the input joint values and the measured encoder values from the motors for the robot model. In the table, the maximum error of the all five revolute joints was 1.3^{o}, and these values coincided with the manufacturer specification. The maximum error of the prismatic joints at each pin clamps was 0.14 mm, since the rack-and-pinion gear system used in this model translated 1.3^{o} of rotational error to 0.14 mm.

3.2 Simulation of Fracture Reduction Process

In the case of the correction of a 45° rotational deformity with 20 mm of fracture gap, the solution of the fixator joint variables was obtained for the developed fixator system, with the following result: ($\mathbf{r_1} = 17^{o}$, $\mathbf{r_2} = -7^{o}$, $\mathbf{r_3} = -43^{o}$, $\mathbf{r_4} = -7^{o}$, $\mathbf{r_5} = -17^{o}$, $\mathbf{t_0}$ = 17mm, $\mathbf{t_6}$ = 12mm). Upper and lower bounds imposed on the rotational and prismatic fixator joints based on the fixator design constraints improve the solution convergence.

Based on the joint values calculated from the inverse kinematics analysis, the robotic execution of the fracture treatment to correct the given deformity was performed (Fig. 2a). The adjustment sequence of the robotic execution was from the distal revolute joint, $\mathbf{r_1}$, to the proximal joints, $\mathbf{r_5}$, then the two prismatic joints, $\mathbf{t_0}$ and $\mathbf{t_6}$.

Upon implementing the resulting inverse kinematics analysis values into the fixator joint variables, the computer graphic simulation indicated achievement of perfect reduction the given deformity (Fig. 2b and 2c). Two types of fracture reduction processes were simulated for testing clinical options. In Fig. 2b, we experimented with new adjustment techniques, by first inserting the pins into the appropriate positions of the bone segments and assembling the fixator frames according to the analysis results,

then readjusting all the joints to their neutral positions. In Fig. 2c, the conventional approach executed by clinicians was simulated, by first constructing the fixator frames in their neutral positions, then changing the fixator joints according to the analysis results. The resulting actions on the bone segments in both cases should perfectly reduce the given deformity.

In order to examine the accuracy of the robotic execution, the reverse adjustment procedure was performed from the neutral positions to generate the same deformity, which was a 45° rotational deformity with 20 mm of gap. As shown in Fig. 3, the final rotations and the gap generated after the robotic execution was 43° and 20 mm, respectively. This error came from the low stiffness of the AI motors used in this study.

Table 1. The rotational errors of the revolute joints and the translational errors of prismatic joints in the robot model of the external fixation system

Input (deg)	Rotational Error (deg)					Input (mm)	Trans. Error (mm)	
	R1	R2	R3	R4	R5		T0	T6
-52.00	0.00	1.30	0.00	1.30	1.30	0.00	0.00	0.15
-31.20	0.00	0.00	0.00	1.30	1.30	4.63	0.00	0.00
-10.40	0.00	0.00	0.00	0.00	0.00	9.25	0.00	0.15
0.00	1.30	1.30	0.00	0.00	0.00	13.88	0.15	0.15
10.40	1.30	1.30	0.00	0.00	0.00	18.50	0.00	0.00
31.20	1.30	1.30	0.00	0.00	0.00	23.13	0.15	0.00
52.00	1.30	1.30	0.00	0.00	1.30	27.75	0.15	0.15

4 Discussion and Conclusions

External fixation is commonly used to stabilize long bone segments following fracture or for bone lengthening [1,2,3]. This surgical treatment has several advantages such as adjustment capability, elastic fixation, easy removal from the bones, and mechanical stimulation. External fixation methods also provide different surgical options such as gradual correction of the residual deformity at the fracture site can be accomplished by rotating and translating the fixator joints gradually on an adjustment plan. Despite the many advantages of external fixation, it has not been favored as the treatment of choice even when clinical indications are favorable for such treatment because of improper pre-operative planning and inaccurate execution from the surgical planning [4]. In order to obtain long-term good clinical results, computer-assisted 3-D preoperative planning and precise execution of external fixation is necessary [4,6,7].

Some limitations of the current robot model have been found, such as oscillation of the fixator frames during adjustment and correction errors of the fracture sites. These limitations come from the low torque and low stiffness of the AI-motors used in this study, since this model was developed primarily to demonstrate the feasibility of robotic execution techniques for external fixation. These problems could be eliminated easily by adopting high-torque and high stiffness low velocity motors for the robot system.

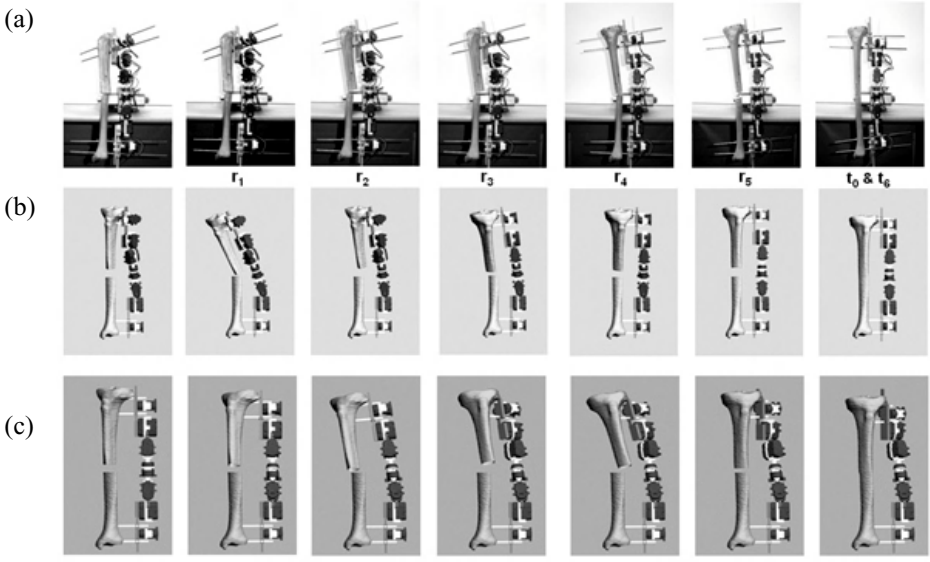


Fig. 2. The computer simulation and robotic execution of 45° of external rotation with a 20 mm gap by sequential adjustment from the proximal to distal revolute joints following the two end prismatic joints. (a) Robotic executions of initial fixator system with adjusted positions of the joints. (b) Computer simulation of fixator system with joints initially in adjusted positions. (c) Computer simulation of fixator system with joints initially in neutral positions

Fig. 3. The results of rotation and gap after execution of robot model of the external fixation system for a 45° of external rotation with a 20 mm

In summary, this paper presents the development of a computer graphic model with the robotic model of an external fixation system for fracture treatment and deformity correction. Using the developed inverse kinematics analysis, the required amount of the rotation and translation of the fixator joints are determined to reduce given fracture deformities. Based on these calculated joint values, the developed robot model of the bone-fixator system demonstrates the execution of the adjustments for a given fracture deformity. In addition, the computer graphic model developed in this study can accurately simulate the adjustment process to visualize and evaluate the pre-operative

planning procedure. In the future, the combined tools could be used for executing knowledge-based computer-aided fracture treatment, enhancing clinical performance and facilitating changes in the design configuration for the external fixator.

References

1. Chao E.Y., Hein T.J.: Mechanical performance of the standard Orthofix® external fixator. Orthopedics 11 (1988) 1057-1069.
2. Paley, D., Tetsworth, K.: Mechanical axis deviation of the lower limbs; Preoperative planning of uniapical deformities of the tibia and femur. Clin. Orthop. Related Res. 280 (1992) 48-64
3. Choi, I.H., Chung, C.Y., Cho ,T.J., Park, S.S.: Correction of genu recurvatum by the Ilizarov method. J. Bone Joint Surg. 81-B (1999) 769-774
4. Hsu, R.W.W., Himeno, S., Coventry, M.B., Chao, E.Y.S.: Normal axial alignment of the lower extremities and load-bearing distribution at the Knee. Clin. Orthop. Related Res. 255 (1990) 215-227
5. Kim, Y.H., Inoue, N., Chao, E.Y.S.: Kinematic simulation of fracture reduction and bone deformity correction under unilateral external fixation. J. Biomech. (2002) 1047-1058
6. Ellis, R.E., Tso, C.Y., Rudan, J.F., Harrison, M.M.: A surgical planning and guidance system for high tibial osteotomy. Comput. Aided Surg. 4 (1999) 264-274
7. Chao, E.Y.S., Sim, F.H.: Computer-aided preoperative planning in Knee osteotomy. The Iowa Orthop. J. 15 (1995) 4-18
8. Larsen, S.T., Magyar, G., Onsten, I., Ryd, L., Lindstrand, A.: Upper tibial valgus osteotomy using a dynamic external fixator. J. Bone Joint Surg. 80-B (1998) 295-297
9. Chao E.Y.S., Rim K., Smidt G.I., Johnston RC.: The application of 4X4 matrix method to the correction of the measurements of hip joint rotations. J. of Biomech. 3 (1970) 459-471

Robust Path Design of Biomechanical Systems Using the Concept of Allowable Load Set

Jin Ho Chang[1], Jae Hyun Kim[2], and Byung Man Kwak[3]

[1] Graduate Student, Department of Mechanical Engineering, KAIST,
305-701 Daejeon, Korea
jhchang@khp.kaist.ac.kr
[2] Researcher, Digital Appliance Research Laboratory, LG Electronics Ltd.,
153-802 Seoul, Korea
jhkim@lge.com
[3] Professor, Department of Mechanical Engineering, KAIST,
305-701 Daejeon, Korea
bmkwak@khp.kaist.ac.kr

Abstract. A new formulation for robust design of mechanical systems is presented using the concept of allowable load set. A new measure of safety of a multi-body system is suggested. By maximizing this measure with respect to design variables one can attain the most robust system. In this approach detailed probability information is not required. Applications to multi-body systems show new solutions to complex problems. As a specific category of problems, trajectories of a human motion of lifting a weight with or without low back pain are considered.

1 Introduction

Uncertainties in a structure have long been considered in the literature. Several methods have been studied to deal with uncertainties in structural design. Reliability based design is one important direction and its methodologies are relatively well developed. Detailed statistical data are required for reliability-based design, but they are not readily available in most problems. For that reason, non-probability approach is studied as an alternative. Ben-Haim [1] suggested such a method using convex models. The measure that he developed is the amount of uncertainty the system can tolerate before failure. This measure is related to robustness with respect to uncertainty, and he called it a robust reliability.

Among the uncertain factors, loading is the most dominant property and not controllable in many cases. Kwak and Kim [2] introduced a new concept called "Allowable Load Set," ALS in short, to deal mainly with the problems with loading uncertainty. ALS can be utilized for understanding characteristics of a structure and defining safety of a multi-body system by finding the weakest configuration. A new meas-

* This research was supported by the Center for Concurrent Engineering Design, a National Research Laboratory of the Ministry of Science and Technology and also from Samsung endowment fund.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1088–1094, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ure of safety, called a relative safety index, is suggested, which is defined as the distance from a given nominal load to the boundary of ALS.

Robust optimal design maximizing a relative safety index of a system is presented. The methodology is applied to the design of a multi-body system composed of several subsystems. Because there may be several local minima in calculating the relative safety index of multi-body systems, a global optimization method is utilized.

2 Definition of the Measure of Safety

The reliability based on the probability of failure has been used as the most popular measure to represent the safety of a structure. However, in most problems there is no reliable information on distributions of uncertain variables. To consider robustness of the structure against a constraint with no probabilistic data, a new safety index is introduced in this paper, denoted by γ. It is simply defined as the smallest distance from the mean value to the boundary of ALS in a load space. It can be obtained without knowledge of probability data of the random load. It is a relative measure useful to compare each other. It is called here, a relative safety index. When it is used in optimal design, the result will lead to the most reliable design in a relative sense.

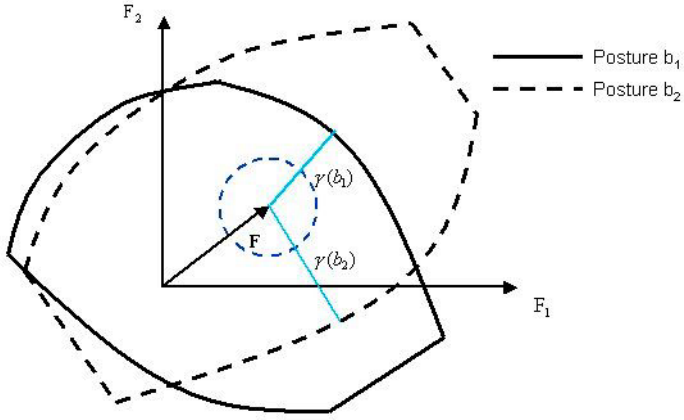


Fig. 1. ALS of two postures of multi-body system

When the structure has n constraints, the corresponding relative safety index γ_i can be calculated. The smallest index among them can be taken as a global measure of safety of the structure. The structure is safer as this index is bigger.

Furthermore, the relative safety index can be extended to a multi-body system. Since all constraints are dependent upon its configuration, the index is a function of configurations, $\mathbf{b}$, that is, $\gamma = \gamma(\mathbf{b})$ in the multi-body system. Figure 1 shows that the

relative safety index between two configurations is different. Thus, the relative safety index of the multi-body system can be defined as follows:

$$\gamma = \min_{\mathbf{b}} \gamma(\mathbf{b}) = \min_{\mathbf{b}} \min_{i} \gamma_i(\mathbf{b}). \tag{1}$$

The configuration $\mathbf{b}$ corresponding to γ is critical. It is quite possible that $\gamma(\mathbf{b})$ has multiple local minima and so a global optimization method may be necessary.

3 Robust Path Design of Multi-body Systems

3.1 Formulation

The proposed robust design is to maximize the relative safety index of a multi-body system.

$$\begin{aligned} &\max \quad b_{n+1}. \\ &\textit{subject to} \quad b_{n+1} \le \gamma_i,\ i = 1, 2, 3, \cdots, N_1. \\ &\qquad\qquad g_{N_1+j} \le 0,\ j = 1, 2, 3, \cdots, N_2. \end{aligned} \tag{2}$$

where b_{n+1} and g_{N_1+j} are an artificial design variable and geometric constraints, respectively. In case the value of γ_i becomes large, the numerical optimization hardly converges. To overcome numerical difficulties the above formulation can be transformed as follows:

$$\begin{aligned} &\max \quad b_{n+1}. \\ &\textit{subject to} \quad G_i \le 0,\ i = 1, 2, 3, \cdots, N_1. \\ &\qquad\qquad g_{N_1+j} \le 0,\ j = 1, 2, 3, \cdots, N_2. \end{aligned} \tag{3}$$

where G_i can be determined by a global optimization method.

$$G_i = \max_{\mathbf{q}} \max_{j} \max_{\mathbf{F} \in D} g_j(\mathbf{F}, \mathbf{q}_k, \mathbf{b}),\ j = 1, 2, 3, \cdots, m. \tag{4}$$

In Formulation 1 load uncertainties are considered into the path design by Eq. (3). Also the results are compared with those obtained by a conventional approach. Formulation 2 is a deterministic formulation that maximizes the minimum value of constraints, which is considered as a conventional approach.

$$\begin{aligned} &\min \quad b_{n+1}. \\ &\textit{subject to} \quad b_{n+1} \ge g_i,\ i = 1, 2, 3, \cdots, N_1. \\ &\qquad\qquad g_{N_1+j} \le 0,\ j = 1, 2, 3, \cdots, N_2. \end{aligned} \tag{5}$$

3.2 A Biomechanical Model

A biomechanical model was developed consisting of body-segments namely; lower leg (LL), upper leg (UL), trunk (T), upper arm (UA) and lower arm (LA). The spine is the one of the most important part, because many people have been troubled by back-ache. So modeling of spine needs to be more careful. It has been reported that about 85% and 95% of all disc disorders occur at L5/S1 and L4/L5 levels, respectively [3]. In this model, the compressive load at L5/S1 is considered. To evaluate the constraint functions, moments and inter-segment forces are calculated at the ankle, knee, hip, shoulder and elbow.

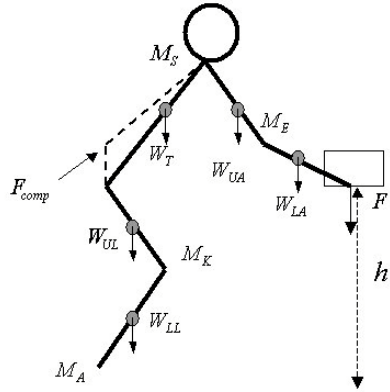


Fig. 2. A biomechanical model

Hwang suggested a mechanical model to calculate the compressive load at L5/S1 for lifting work [3]. The load at L5/S1 is calculated with the abdominal force, the erector spine muscle force and the resultant force at the shoulder joint. Figure 2 shows the biomechanical model used in this study. The spine is modeled with one beam element instead of two beam elements as Hwang's model. But the length of the beam changes like a cylinder element. With the length of the beam, we can obtain equivalent spine model with Hwang's. The model has six independent variables: the angles of bodies and the length of trunk.

3.3 Robust Path Design

Assuming that a man lifts a 20kg object from the ground to 1.2m high, two different optimizations are performed. Formulation 1 maximizes the relative safety index to reduce the effect of load uncertainties. It is a legitimate conjecture that the natural way of force exertion by a human, although not yet well known, should be in such a way to keep away as far as possible from the boundary of constraints, that is, hazardous conditions. Formulation 1 matches this principle.

Table 1. Geometric constraints and loading constraints

Geometric constraints	Allowable load constraints
$g_1 = -M_A / M_A^* - 1 \le 0.$	$G_1 = (\theta_{UL} - \theta_{LL}) / \pi - 1 \le 0.$
$g_2 = M_A / M_A^* - 1 \le 0.$	$G_2 = (\theta_{LL} - \theta_{UL}) / \pi \le 0.$
$g_3 = -M_K / M_K^* - 1 \le 0.$	$G_3 = (\theta_T - \theta_{UL} - c_1) / \pi \le 0.$
$g_4 = M_K / M_K^* - 1 \le 0.$	$G_4 = (\theta_{UL} - \theta_T + c_2) / \pi - 1 \le 0.$
$g_5 = -M_S / M_S^* - 1 \le 0.$	$G_5 = (\theta_T - \theta_{UA} - c_3) / \pi - 1 \le 0.$
$g_6 = M_S / M_S^* - 1 \le 0.$	$G_6 = (\theta_{UA} - \theta_T) / \pi \le 0.$
$g_7 = -M_E / M_E^* - 1 \le 0.$	$G_7 = (\theta_{LA} - \theta_{UA} + c_4) / \pi - 1 \le 0.$
$g_8 = M_E / M_E^* - 1 \le 0.$	$G_8 = (\theta_{UA} - \theta_{LA}) / \pi \le 0.$
$g_9 = F_C / F_C^* - 1 \le 0.$	$G_9 = (h - h^*) / H \le 0.$
	$G_{10} = (h^* - h) / H \le 0.$

It is difficult to construct loading constraints of human body due to its complexity and non-linearity. Moreover, several constraints related to postures cannot be expressed by numerical equations without simplification. The constraints are defined as shown in Table 1. There are nineteen constraint equations. Nine constraints are related to ALS and the others are geometric constraints. Allowable loads are set to half of the maximum loads a human can take.

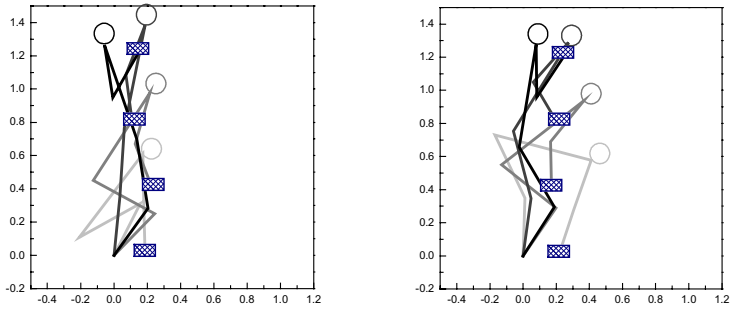


Fig. 3. Optimal lifting posture (a) Formulation 1 (b) Formulation 2

The optimum paths are obtained by raising the hands from the ground to a height of 1.2m with an increment of 0.4m. The set of design variables, corresponding to the worst postures is found by the global optimization method. Among several local solutions, the local solution that seems kinematically most natural is taken as the optimal path. The two paths in Fig. 3 obtained by the two formulations show differences as expected.

To investigate the results in detail, the ALS are drawn with a scale factor 4.e-4 when h* is equal to 0.0m and 0.4m (Fig. 4). The head of the load arrow denotes the location of the mean load. Figure 4(b) reveals that the postures obtained by Formulation 2 are near the borders of constraints 5 and 9, which are related to shoulder and spine. That is, the man with this posture is more liable to hurt than those by Formulation 1. Researchers in ergonomics recommend workers to take the postures shown in Fig. 4(a). Considering instability and narrow safety margin that might occur when lifting a heavy load, the postures that adopt Formulation 1 seem more natural and safer than those by Formulation 2. It is thus our conclusion that one of the guiding principles suitable for predicting a human body motion is to use the relative safety index, comparing with those previously used in the literature.

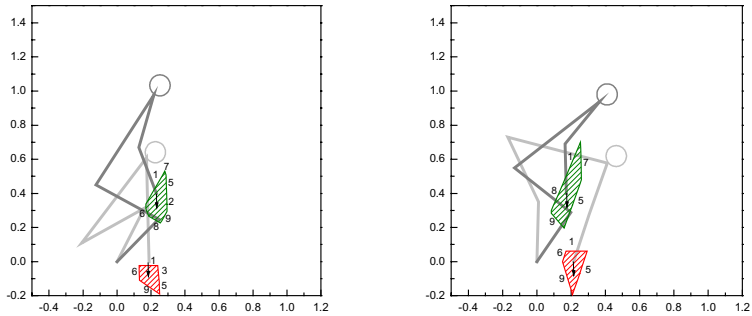


Fig. 4. The comparison of ALSes of optimal postures by two formulations (a) Formulation 1 (b) Formulation 2

Figure 5 shows the optimized postures of a man that has a backache. It is assumed that he can endure only 2400N compressive force at L5/S1 instead of 3400N in the normal condition. As against normal condition, the man with a backache tends to get the trunk straighter and keep the object closer to his body. To investigate how the variation of the spine load capacity affects other parts of human body, the rates of change of relative safe indices and constraint value are calculated. Most parts are affected, even though there are some variations by the postures, but the knee gets more severe load regardless of postures.

4 Conclusions

A methodology using the concept of ALS is introduced and shown very effective in path design of biomechanical systems. A new measure of safety of multi-body systems, which is called a relative safety index, is defined. To calculate this index does not require any probability data, but it is a good indicator of structural integrity.

By maximizing the relative safety index, a robust design can be obtained. This criterion is adopted to obtain a new formulation of obtaining stable working postures and results are compared with those by previously used force minimization approach. It is also shown that how low back disorders affect working postures and human body reactions. For the multi-body system a global optimization algorithm is usually necessary to locate several possible local minima.

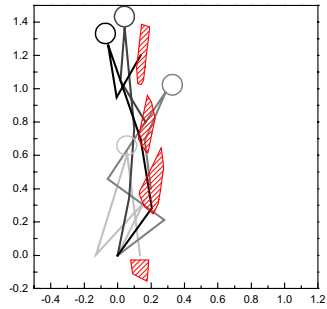


Fig. 5. Optimal lifting postures of a man who has low back disorders (Formulation 1)

In summary, it is shown that the concept of ALS is useful not only to show the characteristics of the structural integrity of a biomechanical system, but also to do a robust design without knowing detailed probability data.

References

1. Ben-Haim, Y.: Robust Reliability of Structures. Adv. Appl. Mech. 33 (1997) 1-41
2. Kwak, B. M., Kim, J. H.: Concept of Allowable Load Set and Its Application for Evaluation of Structural Integrity. Mechan. Struct. Mach. 30(2) (2002) 213-247
3. Hwang, D. Y.: Effects of Working Postures on the L5/S1 Compressive Forces. MS thesis. KAIST (1995)

A New Modeling Method for Objects with Branching Problem Using Non-uniform B-Spline*

Hyo Shin Kim[1], Yoon Hyuk Kim[2], Yeon Hyeon Choe[3], Sung-Min Kim[4], Taeg-Sang Cho[5], and Joung Hwan Mun[1]

[1] Dept. of Bio-Mechatronic Engineering, College of Life Science & Technology, Sungkyunkwan University, Korea
drrose@skku.edu, jmun@skku.ac.kr
[2] School of Advanced Technology, Kyunghee University, Korea
yoonhkim@khu.ac.kr
[3] Sungkyunkwan University School of Medicine Korea
yhchoe@smc.samsumg.co.kr
[4] Dept. of Biomedical Engineering, School of Medicine, Konkuk University, Korea
smkim@kku.ac.kr
[5] Dept. of Electrical Engineering and Computer Science, KAIST, Korea
taegsang@kaist.ac.kr

Abstract. In many applications, objects are reconstructed from cross-sections for visualization, finite element and dynamic analysis. Although cross-section of an object may contain multiple contours, a few papers have dealt with branching problem. Moreover ends of branches are described flatly. In this paper, as a basic study for dynamic analysis of a human knee joint, we present a new modeling method which proposes a data-set for solving branching problem and handling convex-end-condition of branches. We select an initial standard point from lowest slice and decide a nearest standard point of the next slice and the next, in turns. Based on standard points, we complete the data-set by applying contour alignment. For 3D reconstruction, the surface is approximated by bicubic non-uniform B-spline surface fitting. This method provides the smooth surface model with C^2 continuity and describes the convexity of ends of branches.

1 Introduction

In many applications, objects are often reconstructed from 2D cross-sections for visualization of 3D structure, finite element and multibody dynamic analysis. For 3D reconstruction, several methods have been proposed, ranging from polyhedron-based approaches[1, 3, 4, 5], intensively studied during the past decade, to tiling algorithms[1, 3, 4, 5], skinning methods[6, 7, 8] and surface fittings[2, 9, 10, 13].

* This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea.(02-PJ3-PG6-EV06-0002)
Corresponding author: Joung H. Mun, Ph.D., Dept. of Bio-Mechatronic Engineering, College of Life Science and Technology, Sungkyunkwan University, Suwon, Korea. (Tel) +82-31-290-7827 (e-mail) jmun@skku.ac.kr

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1095–1102, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In general, each cross-section may contain multiple contours. Some studies have nevertheless proposed branching algorithms to link one or more contours in one cross-section to multiple contours in an adjacent cross-section with triangular facets[1, 4, 5]. And only a few papers have approximated a set of 2D contours with branching problems. Park and Kim[10] proposed a method for B-spline surface approximation to a set of contour with branching problem but triangular facets were constructed over each branching region and built triangular surface patches over these facets. Jeong et al[2] proposed B-spline surface approximation to cross-sections using distance maps. This method provided a smooth surface model, yet realized efficient data reduction and described ends of branches as a flat surface by using the end tangent condition[2, 11]. However this flat surface affects seriously force recovery if this model is applied to dynamic analysis of the joint, because contact points occurring at the joint are usually placed on the flat surface.

In this paper, we present a new modeling method for objects with branching problems which uses a data-set for solving branching problem and handling convex-end-condition of branches simultaneously. For 3D reconstruction, the surface is approximated by tensor product non-uniform B-spline surface fitting. This method provides bicubic non-uniform B-spline surface with C^2 continuity and describes the convexity of ends of branches.

2 Surface Fitting

If an object has complex shape, double branching situation occurs frequently[5]. So in this paper, we consider objects with double branching. Double branching is a case in which a contour C_r^{k+1} at level z_{k+1} must be linked to two contours, C_i^k and C_j^k, at its lower level z_k shown in the Fig. 1. Here, data points have a set of heights $z_0 < z_1 < ... < z_k < ... < z_{max}$ and points of the kth contour are saved counterclockwisely.

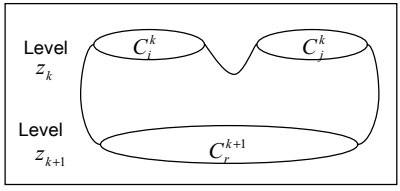


Fig. 1. Definition of double branching problem

The proposed method consists of three main phases, data-set, contour alignment, surface fitting. First, we briefly review definitions of B-spline curve and surface fitting. A detailed discussion of B-spline theory can be found in the literature[11].

2.1 Non-uniform B-Spline Curve Fitting

As a foundation for the surface fitting algorithm, we first apply open non-uniform B-spline curve fitting to contours. Here, a cubic B-spline is employed as the basis

function (p=3) and satisfies Q_0=C(0) and Q_m=C(1). In order to fit the open B-spline curve to given m+1 contour points Q_k, we use the least squares formulation below to find control points P_i which is unknown as n+1 and assume $n \geq p$.

$$f = \sum_{k=1}^{m-1} \left| Q_k - C(\bar{u}_k) \right|^2 \tag{1}$$

is a minimum with respect to n+1 variables, P_i. f is a scalar-valued function of n-1 variables, $P_1,\ldots, P_{n-1}$. To minimize f, we set derivatives of f with respect to n-1 points, P_l, equal to zero. Finally one linear equation remains below and we can calculate control points $P_1,\ldots, P_{n-1}$.

$$(N^T N)P = R \tag{2}$$

The $\bar{u}_k$ and U affect the shape and parameterization. The $\bar{u}_k$ is the parameter value precomputed by the chord length method[11] which is most widely used and generally adequate.

The placement of knots should reflect the distribution of the $\bar{u}_k$. We need a total of n+p+2 knots and there are n-p internal knots, and n-p+1 internal knot spans. Let

$$t = \frac{m+1}{n-p+1} \tag{3}$$

then define internal knots by

$$\begin{aligned} i &= \text{int}(jt), \qquad \alpha = jt - i, \\ u_{p+j} &= (1-\alpha)\bar{u}_{i-1} + \alpha\bar{u}_i, \qquad j = 1,\ldots,n-p. \end{aligned} \tag{4}$$

This guarantees that every knot span contains at least one $\bar{u}_k$, and under this condition the matrix ($N^T N$) is positive definite and well-conditioned. It can by solved by Gaussian elimination without pivoting.

2.2 Tensor Product Non-uniform B-Spline Surface Fitting

The final step is to generate a bicubic open non-uniform B-spline surface and the resulting B-spline surface is expressed as

$$S(u,v) = \sum_{i=0}^{n} \sum_{j=0}^{m} N_{i,p}(u) N_{j,q}(v) P_{i,j} \tag{5}$$

where P and S have x, y and z components. The control points $P_{i,j}$ are (n+1) by (m+1) network in u and v directions and we also calculate these by least squares method defined by

$$\sum_{k=1}^{r-1} \sum_{l=1}^{s-1} \left| Q_{k,l} - S(\bar{u}_k, \bar{v}_l) \right|^2 . \tag{6}$$

Again, the first order of business is to compute reasonable values for the $(\bar{u}_k, \bar{v}_l)$ and knot vectors U and V. A common method like the curve fitting is used to compute

parameters $\bar{u}_0^l, \ldots, \bar{u}_r^l$ for each l, and then each $\bar{u}_k$ is obtained by averaging across all $\bar{u}_k^l$, $l = 0,\ldots,s$, and $\bar{v}_l$ is also calculated. Once $(\bar{u}_k, \bar{v}_l)$ are computed, knot vectors U and V can be obtained like the method at Sect. 2.1.

2.3 Branching Problem and Convex-End-Condition Handling; Data-Set

It is simple to construct a data-set for reconstructing a geometric model of an object with double branching problem. The number of points of left and right branch extracted from branches must be same and the number of row and column of points at one side (left side in this paper) of branches are followed;

$$\text{Number of row at one side of branch} = x + 2\times(n-1), \tag{7}$$

$$\text{Number of column at one side of branch} = 2\times(2\times n) \tag{8}$$

where x and n(=1, 2, …) mean a quarter of the total number of points of slice at the z_0 and the slice number, respectively. And the total number of points demanded at the first base and the number of row and column of points at the base are followed;

$$\text{Total number of points at the first base} = 2\times\{x+2\times(n_{\max}-1)+2\times m_1\}+2\times(4\times n_{\max}+2\times m_1)-4, \tag{9}$$

$$\text{Number of row at the base} = x+2\times(n_{\max}-1)+2\times m, \tag{10}$$

$$\text{Number of column at the base} = 4\times n_{\max}+2\times m \tag{11}$$

where n_{max} is the number of slice(=k+1) at the z_k, m_1(=1)and m(=1, 2, …) mean the slice at the z_{k+1} and the number of slice counted only at the base, respectively. Thus the size of data-set is $(x+2\times(n_{\max}-1)+2\times m, 4\times n_{\max}+2\times m)$. Here each value of column includes three components, x, y and z.

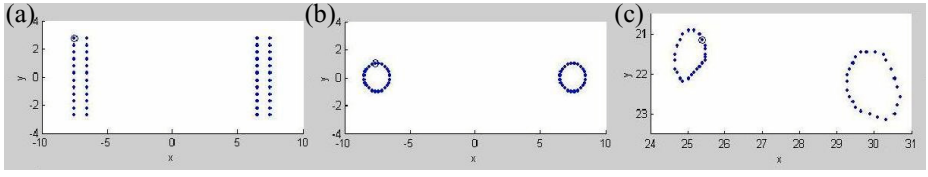


Fig. 2. Initial standard point (*surrounded by small circle*), (a) rectangle-based object(*RBO*), (b) circle-based object(*CBO*) and (c) femur

For constructing the data-set, we select an initial standard point from a slice at the z_0. As shown in Fig. 2, it is efficient to select a maximum(or minimum) point placed at the perpendicular direction to the line passing through approximate centers of two branches at the left branch or a nearby point of a mentioned point. Based on this point, the others are rearranged counterclockwisely. Points from branches and bases are located at the left and right side inside of a data-set and outside of a data-set to surround points from branches, respectively

2.4 Contour Alignment

As the control net of a B-spline surface is composed of a set of control polygons, a number of different control nets can be defined by variously aligning the given control polygons. A different alignment of control polygons gives rise to a different control net and results in the different shape of the surface. Since points of *u*- and *v*-direction pass through different slices, neighboring points have to be arrayed as closely as possible. After an initial standard point is selected by Sect. 2.3 we select a shortest one of the next slice. Then this point becomes a standard point of the next slice. Like this way, standard points are selected from all of slices. Standard points are located at the closest place from a previous standard point, just above. By the contour alignment, we can prevent unwanted twists or wiggles.

2.5 Approximation to within a Specified Accuracy

The bicubic B-spline surface can be obtained by (5) and the sum-of-squares error can be calculated by comparing surface points at assigned data parameter values to original data points[12], as shown by

$$E(P) = \sum_{k=0}^{r}\sum_{l=0}^{s}\left(\left(\sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(\overline{u}_k)N_{j,q}(\overline{v}_l)P_{i,j}\right) - Q_{k,l}\right)^2 . \tag{12}$$

The average error in the surface fit is given by

$$E_{ave}(P) = \frac{1}{(r+1)(s+1)} E(P) . \tag{13}$$

To obtain the approximating data to within some specified error bound, E, we use iterative method which have three steps:

1. Starting with (p+1, q+1) control point (the minimum) and fit a surface.
2. Checking if the surface deviation is less than E. A record is maintained for each knot span, indicating whether or not it has converged.
3. If the surface deviation does not satisfy E, a knot is added at the mid point of each nonconvergent span, thus adding control points. And if the deviation is less than E, stopping the reiteration.

3 Experimental Results

The proposed data-set has been applied to three different sets of contour data. We use the synthetic data of rectangle-based object(RBO) and circle-based object(CBO) to verify the method and the CT data of inferior region of a femur as a general case with double branching problem.

The information of two synthetic data and CT data is denoted in Table 1 and their cross-sections are shown in Fig. 3. In Table 1, the size of data-set, (row×column),

means total number of points which have x, y, and z components. In Fig. 3, the contour in each cross-section has been segmented and compressed into a closed polygon. And small circles on a red line passing through contours are standard points of each contour by contour alignment, starting at an initial standard point of lowest slice.

Table 1. Information of two synthetic data and CT data of a femur

	RBO	CBO	Femur
Number of slices	14	14	18
Distance between neighboring slices	0.5	0.2	0.125(mm)
Size of data-set	38×42	38×42	46×48

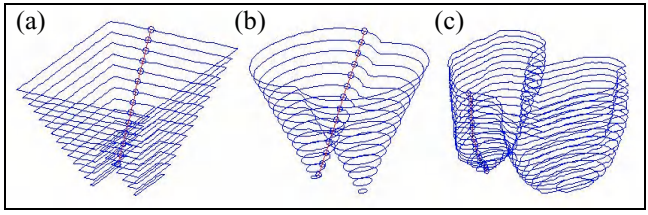


Fig. 3. Cross-sections of (a) RBO, (b) CBO and (c) femur

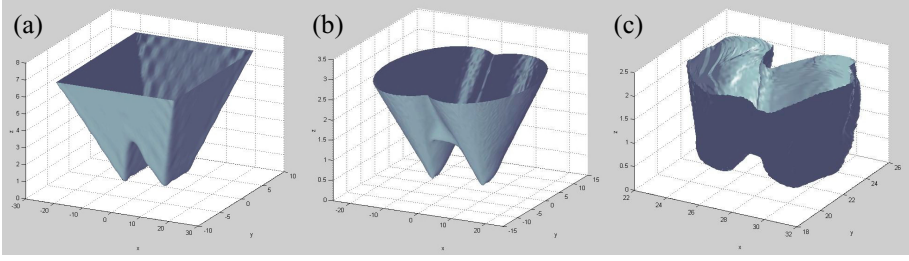


Fig. 4. B-spline surface of (a) RBO, (b) CBO and (c) femur

Table 2. Number of control points, total error and average error

	RBO	CBO	Femur
Number of control points (u direction, v direction)	(20, 21)	(19, 21)	(16, 19)
E(P)	827.0186	796.4465	88.2719
E_{ave}(P)	1.9691	1.9961	0.2904

Figure 4 shows shaded surfaces. The resulting surface is approximated by bicubic non-uniform B-spline surface. Table 2 presents the number of control points fitted in

given accuracy, a total error and an average error occurring at data points with respect to the surface. In the number of control points, (*u* direction, *v* direction), *u* direction and *v* direction mean the number of points needed to approximate the vertical and horizontal data of a data-set, respectively.

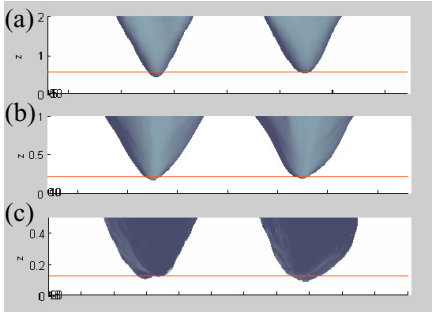


Fig. 5. Ends of branches, (a) RBO, (b) CBO and (c) femur

Figure 5 shows states of ends of branches and each example has double branches. A straight red line means a lowest slice, n_1, and z values are 0.5, 0.2 and 0.125 for (a) RBO, (b) CBO and (c) femur, respectively. If there is just a previous slice of a lowest slice, its z value is 0 for all examples. As shown in figure, ends of branches are reconstructed convexly and its surface is smooth for all direction. We can also verify lowest points for all ends of branches are within z value, 0.

4 Conclusions and Discussions

As a basic study for more advanced dynamic analysis of knee joint, we have presented a new modeling method for objects with branching problem. For constructing a data-set, we select an initial standard point and complete the data-set based on standard points by applying contour alignment. For 3D reconstruction, the surface is approximated by tensor product non-uniform B-spline surface fitting. This method has provided a smooth bicubic B-spline surface model with C^2 continuity, as shown in Fig. 4.

In addition to a surface approximation, another important fact for surface fitting is the error between data points and points on the surface. Control points are calculated by least squares sense so that the surface is closer to the data. The more the number of control points increases, the more approximation improves. However, as the number of control points approaches the number of data points, noises or unwanted wiggles can occur according to [11]. So the tolerance for an average error allowable was 2.0 for RBO and CBO and 0.3 for a femur which are determined empirically. As shown in Table 2 and Fig. 4, average errors are satisfied with the given tolerance and resulting surface models have no noises or unwanted wiggles.

Moreover, as shown in Fig. 5, ends of branches of all examples are convex and within the distance between neighboring slices. Sun et al[13] used ‘doom’ feature in SolidWorks to describe the terminal region convexly for finite element analysis or multibody dynamic analysis of human middle ear. Thus convex-end-condition is very

important and we think these results will play an important role in conducting more advanced dynamic analysis of knee joint hereafter.

The number of points extracted from images is restricted within small number if the resolution of images is low. This can be limitation but we can overcome this problem by improvements in data acquisition and imaging techniques such as computed tomography, magnetic resonance imaging and ultrasound imaging. Since the location of points affects significantly to the shape of a model, the data-set must not have sharp turn. When there are sharp turns, it need to use the centripetal method for parameter values or rise the degree of B-spline, as described in the literature[11].

We concluded this new method provided C^2 continuous bicubic B-spline surface model of objects including branching problem and it described the convex surface, required absolutely for dynamic analysis of joints, occurring at ends of branches smoothly within distance between neighboring slices. In the future study, we have a plan to execute more advanced dynamic analysis of a human knee joint.

References

1. Park, H., Kim, K.: 3D shape reconstruction from 2D cross sections. Journal of Design and Manufacturing. 5 (1995) 171-185
2. Jeong, J., Kim, K., Park, H., Cho, H., Jung, M.: B-Spline Surface Approximation to Cross-Sections Using Distance Maps. Advanced Manufacturing Technology. 15 (1999) 876-885
3. Zyda, M., Allan, R., Hogan, P.: Surface construction from planar contours. Computers and Graphics. 11(4) (1987) 393-408
4. Ekoule, A., Peyrin, F., Odet, C.: A triangulation algorithm from arbitrary based multiple planar contours. ACM Transactions on Graphics. 10(2) (1992) 182-199
5. Choi, Y., Park, K.: A heuristic triangulation algorithm for multiple planar contours using an extended double branching procedure. Visual Computer. 10 (1994) 372-387
6. Woodward, C.: Skinning techniques for interactive B-spline surface interpolation. Computer Aided Design. 20(8) (1988) 441-451
7. Odesanya, O., Waggenspack, W., Thompson, D.: Construction of biological surface models from cross-sections. IEEE Transactions of Biomedical Engineering. 40(4) (1993) 329-334
8. Kaklis, P., Ginnis, A.: Sectional-curvature preserving skinning surfaces. Computer Aided Geometric Design. 13(7) (1996) 601-619
9. Rogers, D., Fog, N.: Constrained B-spline curve and surface fitting. Computer-Aided Design. 21(10) (1989) 641-648
10. Park, H., Kim, K.: Smooth surface approximation to serial cross-sections. Computer-Aided Design. 28(12) (1996) 995-1005
11. Piegel, L. and Tiller, W.: The NURBS Book. Springer Berlin (1995)
12. Klingensmith, J., Vince, D.: B-spline methods for interactive segmentation and modeling of lumen and vessel surfaces in three-dimensional intravascular ultrasound. Computerized Medical Imaging and Graphics. 26 (2002) 429-438
13. Sun, Q., Chang, K., Dormer, K., Dyer, R. Jr., Gan, R.: An advanced computer-aided geometric modeling and fabrication method for human middle ear. Medical Engineering and Physics. 24 (2002) 595-606

Motion Design of Two-Legged Locomotion Process of a Man

Svetlana Novikava[1], Kastus Miatliuk[1,2], and Krzysztof Jaworek[2]

[1]Hierarchical Multilevel Systems Laboratory, Minsk, Belarus,
220064, P.O. Box 48
hmsl@buk.edu.by
[2]Technical University of Bialystok, WM, Poland
kmiat@pb.bialystok.pl

Abstract. The geometric construction of bio-kinetic apparatus of man's locomotion representation in cybernetic form is actual task. It is very important for motion design and locomotion coordination (control) tasks solving. For this aim the hierarchical multilevel systems construction was chosen and applied in this paper for bio-kinetic apparatus of man description. The hierarchical multilevel systems coordination method realisation in geometric design of man's locomotion is also presented in the paper.

1 Geometric Representation of Bio-kinetic Apparatus Construction

Geometrical construction of men in our case is regarded as a set of bio-kinetic chains and limbs. It can be presented as hierarchical multilevel system by Mesarovi and Takahara [1] in form of *aed* [3-6] (*aed* is an ancient Greek word):

$$_{g}S^{\ell} = \{\omega^{\ell}, S_0^{\ell}, \sigma^{\ell}\} \tag{1}$$

where:

ω^{ℓ} - aggregated dynamic realisation (model), which represents the construction (body and bio-kinetic chain) of men as the unit in its environment;

σ^{ℓ} - geometric construction of a bio-kinetic apparatus;

S_0^{ℓ} - coordinator, which connects σ with ω and in this way, coordinates the locomotion process.

ℓ - index of level.

ω^{ℓ} is described with the help of dynamic systems $(\overline{\rho}, \overline{\varphi})$ [2]:

$$\begin{aligned} \overline{\rho}^{\ell} &= \{\rho_t : C_t \times X_t \to Y_t \;\&\; t \in T\}^{\ell}, \\ \overline{\varphi}^{\ell} &= \{\varphi_{tt'} : C_t \times X_{tt'} \to C_{t'} \;\&\; t, t' \in T \;\&\; t' > t\}^{\ell}, \end{aligned} \tag{2}$$

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1103–1109, 2004.
© Springer-Verlag Berlin Heidelberg 2004

where C – states, X –inputs, Y – outputs, ρ - reaction, φ - states transition function, T –time of level ℓ.

ω^{ℓ} in general case includes the aggregated dynamical realisations of object ${}_{o}S^{\ell}$, its environment ${}_{\varepsilon}S^{\ell}$, actions ${}_{o}P^{\ell}$ of ${}_{o}S^{\ell}$ in ${}_{\omega}S^{\ell}$ and actions ${}_{\omega}P^{\ell}$ of ${}_{\omega}S^{\ell}$ with ${}_{o}S^{\ell}$. Structure σ^{ℓ} is defined as:

$$\sigma^{\ell} = \{S_0^{\ell}, \{\overline{\omega}^{\ell-1}, {}_{\sigma}U^{\ell}\}\}, \tag{3}$$

and contains aggregated representations $\overline{\omega}^{\ell-1}$ of lower level elements (body, thigh, shank, foot, fingers) and their connections ${}_{\sigma}U^{\ell}$ – the joints.

For example, the common part of the thigh $\omega_1^{\ell-1}$ and the shank $\omega_2^{\ell-1}$ is their structural connection – the knee ${}_{\sigma}U_{1,2}^{\ell}$.

Coordinator is also presented in this case in form of *aed*:

$$S_0^{\ell} = \{\omega_0^{\ell}, S_{00}^{\ell}, \sigma_0^{\ell}\}, \tag{4}$$

and coordinates the locomotion process on its layers of choice, learning and self-organisation. All metrical characteristics of bio-kinetic apparatus (relational angles, length of limbs etc.) are described by numeric positional systems – numerical characteristics.

2 Coordination Method Realization in Design of Men's Locomotion

Geometric design process (motion design) of two-legged locomotion of man contains two stages and executed by coordinator [11]:

- Synthesis (creation) of geometric construction of bio-kinetic men's apparatus;
- Motion design and evaluation of kinds (states) of men locomotion (walk, run and others).

On the first stage coordinator S_0^{ℓ} realise (on its choice layer) the synthesis of geometric elements $\overline{\omega}^{\ell-1}$ of bio-kinetic apparatus and after that connects the elements by their structural common parts ${}_{\sigma}U^{\ell}$ – joints. In this way the synthesis process of structure of whole bio-kinetic apparatus of man is realised. This process may be described as follows:

$$S_0^{\ell} : \{\overline{\omega}^{\ell-1}, {}_{\sigma}U^{\ell}\} \to \sigma^{\ell}. \tag{5}$$

The structure $\overline{\sigma}^{\ell-1}$ of elements $\overline{\omega}^{\ell-1}$ (for ins. shank) is constructed in computer from geometrical objects $\overline{\omega}^{\ell-2}$ (conditional points) – cubes (Fig.1).

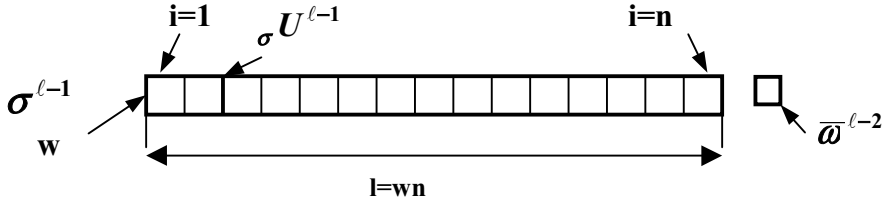


Fig. 1. The example of geometric structures $\overline{\sigma}^{\ell-1}$ synthesis of elements $\overline{\omega}^{\ell-1}$ (for ins. shank)

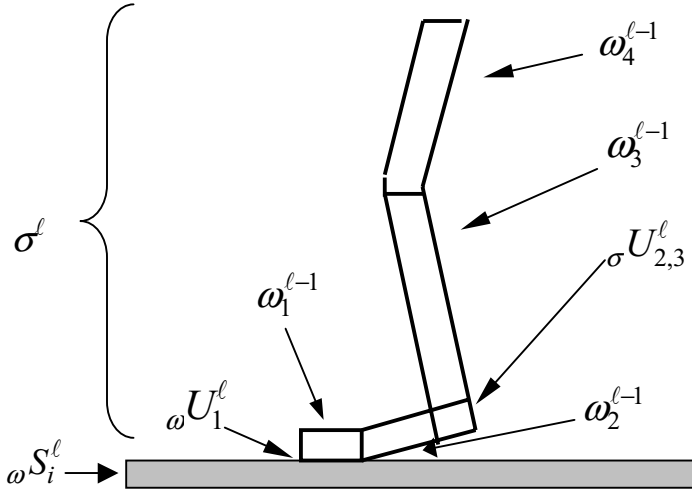


Fig. 2. The structure of the one part (one leg) of bio-kinetic apparatus of a man

Conditional points $\overline{\omega}^{\ell-2}$ connected by its sides ${}_{\sigma}U^{\ell-1}$. The number of elements ***n*** and its sizes ***w*** defines the width ***w*** and length ***l*** of whole geometrical structure of the object. The higher level structure σ^{ℓ} is constructed from $\overline{\omega}^{\ell-1}$

elements with taking into account the interactions ${}_{\omega}U_1^{\ell}$ with the environment ${}_{\omega}S_i^{\ell}$ (Fig.2).

To solve the task of *second stage* (motion design), man is presented in accordance with hierarchical multilevel systems representation as the object ${}_{o}S^{\ell}$. The locomotion process of man ${}_{o}P^{\ell}$ in its environment ${}_{\omega}S^{\ell}$, the environment, which actively affects on the men through its process ${}_{\omega}P^{\ell}$ and this process are also regarded.

Men's environment can be heterogeneous in general case, for instance it is possible to have an obstacles on the trajectory of motion. That is why, coordinator realises its several coordination strategies while walking – from the strategy of *full determination* (the trajectory is known, the external state (state of the environment) is also known) to the coordination strategy of *full information uncertainty* – the trajectory is uncertain (motion in unknown environment); the absence of the possibility to receive the information about the state (structure) of both the path and environment. For instance, in case of man locomotion in conditions of full darkness.

The motion ${}_{o}P^{\ell}$ of man ${}_{o}S^{\ell}$ in its environment ${}_{\omega}S^{\ell}$ is called the change of man's interconnections ${}_{\omega}U^{\ell}$ with elements S_i^{ℓ} ($i \in I^{\ell+1}$) of environment system ${}_{\omega}S^{\ell}$ (Table 1). The contact state ${}_{\omega}U_1^{\ell}$ of man's stand (base) leg with path surface and the state ${}_{\omega}U_2^{\ell}$ of geometric form of bio-kinetic apparatus (Fig.3) (both base and swing legs) we will regard as man's interconnections with its environment while walking:

$$ {}_{\omega}U^{\ell} = \{ {}_{\omega}U_1^{\ell}, {}_{\omega}U_2^{\ell} \} \tag{6} $$

Table 1. Various states of ${}_{\omega}S^{\ell}$, ${}_{\omega}U_1^{\ell}$ and ${}_{o}S^{\ell}$

Systems	C (states)			
${}_{\omega}S^{\ell}$ (way)	Hard	Snow	Hard + Water	Send
${}_{\omega}U_1^{\ell}$ (foot)	Heel	Foot	Toe	Foot + toe
${}_{o}S^{\ell}$ (men)	Different geometrical parameters of bio-kinetic apparatus of men			

Interactions ${}_{\omega}U^{\ell}$ are changeable in time and differ for various kinds (types) of motions: for instance for walking faze the ${}_{\omega}U_1^{\ell}$ can't be equal to zero on whole period of walking time, but for running faze there are moments of time when ${}_{\omega}U_1^{\ell}$**=0**.

There are deformations in bio-kinetic apparatus which cause (during locomotion process) the coordinated movements of its parts – hands and legs: flexion, straightening, inclination (bending), swing. The change of the connections ${}_{\sigma}U^{\ell}$ of above mentioned elements, which are the parts of bio-kinetic apparatus structure σ^{ℓ} we will recognise as deformations. The coordinator task in motion design is to provide the cohered motion of bio-kinetic apparatus and deformations of its elements, i.e. ${}_{\sigma}U^{\ell}$.

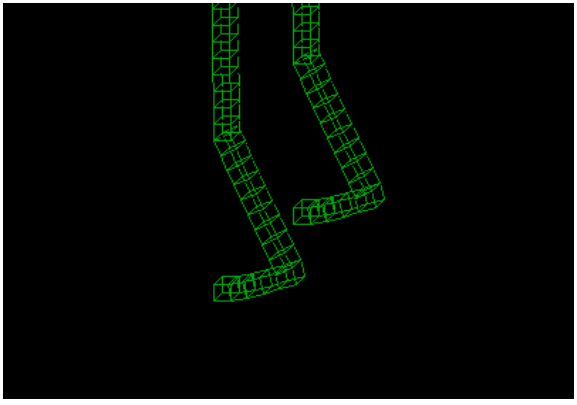


Fig. 3. One state of bio-kinetic apparatus geometric construction in locomotion process (obtained from computer monitor)

Structural connections in geometrical sense are the angles between parts of bio-kinetic apparatus (fingers, foot, led). So, coordinator task is to change continuously these angles to provide the locomotion process. Formally, the task solving is presented in following form:

- for coordinator (as its reaction): $\rho_0^{\ell} : C_0^{\ell} \times \sigma_t^{\ell} \to \sigma_{t'}^{\ell} \,\&\, \sigma_{t'}^{\ell} = Y_0, \sigma_t^{\ell} = X_0$, (7)

- for system (as state transition): $\varphi_{tt'}^{\ell} : \sigma_t^{\ell} \times_{\sigma} U_{tt'}^{\ell} \to \sigma_{t'}^{\ell}$,

where σ_t^{ℓ} and $\sigma_{t'}^{\ell}$ are the state of bio-kinetic apparatus structure at moments of time t and t' correspondingly, $t'>t$; ${}_{\sigma}U_{tt'}^{\ell}$ is the change of angles during the period of time tt', is the state of bio-kinetic apparatus structure at moment of time t' and $t'>t$. The dependence of the angle changing in locomotion process is reflected in [7].

In this case we deal with the biomechanical type of motion for which the next dependence is true:

$$ {}_{\tau\omega}U^{\ell} \left| \overline{T}_{tt'}^{\ell} \right. = const \quad . \tag{8}$$

In other words, the external connections ${}_{\omega}U^{\ell}$ of regarded apparatus depend only on the change of its structural connections ${}_{\sigma}U^{\ell}\left|\overline{T}_{tt'}^{\ell}\right.$ during the period of time $\overline{T}_{tt'}^{\ell}$.

It is possible to present the motion measures (velocity and others) as the change measures of dynamic systems states in elementary time ${}_{\tau}\eta_{T}^{\ell}$ in metrical M^{ℓ}, structural Ξ^{ℓ} and boundary Δ^{ℓ} (geometric forms) spaces. The motion measures have their directions in dependence of the element address in the environment, in accordance with which the present process is regarded.

3 Conclusion

In comparison with known methods [8-10] the proposed coordination one describes the bio-kinetic apparatus of man as hierarchical multilevel system, that allows to solve the design task of man locomotion as interconnections coordination of cohered structural and dynamical representation of apparatus.

In known methods the general rotation angles of limbs (as kinematical mechanism) in 3-D space are calculated. In proposed one we regard the relative angles between limbs as their interconnections only. It allows reduce the number of calculations.

During the motion design processes the dependence of functional parameters (kinds of motions) from structural ones (sizes, angles etc.) of bio-kinetic apparatus elements while walking was established (determined).

Computer program is written on Pascal and implemented on the simplest PC what allows to solve the task fast with minimal computer memory and speed.

References

1. Mesarovic, M.D., D.Macko, D., Takahara, Y.: Theory of Hierarchical Multilevel Systems *A*cademic Press, New York and London (1970).
2. Mesarovic, M.D., Takahara, Y.: General Systems Theory: Mathematical Foundations, Academic Press, New York and London(1975).
3. Novikava, S., Miatliuk, K., et al.: Aed Technology for Ecological, Social and Engineering Systems Coordination. Proc. of 8th Int. Symposium on Modular Information Computer Systems and Networks, Dubna, Russia (1991) 145-152.
4. Novikava, S., Miatliuk, K., et al.: Hierarchical Multilevel Systems in Aed realization. Proc. of 9th Int. Conf. on Mathematical and Computer Modeling, Berkeley, USA (1993) 71.
5. Novikava, S., Mialtiuk, K., Gancharova, S., Kaliada, W.: Aed Construction and Technology in Design. Proceedings of 7th IFAC/IFORS/IMACS Symposium on Large Scale Systems, LSS'95, Pergamon Press, London (1995) 379-381.
6. Novikava, S., Miatluk, K., et al.: Aed Theory in Hierarchical Knowledge Networks. In: Journal of Studies in Informatics and Control 6(1) (1997) 75-85.
7. Jaworek, K.: Indices method of assessment human gait and run. Pub. IBIB-PAN, Warsaw, Poland, (1992). (in Polish)

8. Jaworek, K.: About the index of movement of terrestial mammals. Proc. of the Biomechanics'2003. Int. Conference. Ed. By Acta of Bioengineering and Biomechanics 5 S1 (2003) 199-205.
9. Pennycuick, C.J.: Newton rules biology. A physical approach to biological problems. Pergamon Press, Glasgow (1992).
10. Vaughan, C.L.: Theories of bipedal walking: an Odyssey. Journal of Biomechanics 36 (2003) 513-523.
11. Jaworek, K., Miatliuk, K.: Coordination method in geometric design of two-legged human gait. Proc. of the Men's Locomotion Conf., Gdansk, Poland (2003) 20. (In Polish)

Adaptive Microcalcification Detection in Computer Aided Diagnosis

Ho-Kyung Kang[1], Sung-Min Kim[2], Nguyen N. Thanh[1], Yong Man Ro[1], and Won-Ha Kim[3]

[1] Multimedia group, Information and Communication University, Korea
{kyoung, nnthanh, yro}@icu.ac.kr
[2] Department of Biomedical Engineering, School of Medicine, KonKuk University, Korea
smkim@kku.ac.kr,
[3] Kyung Hee University, Korea
wonha@khu.ac.kr

Abstract. Microcalcification detection is an important part of early breast cancer detection. In this paper, we propose a microcalcification detection method in mammography CAD (computer-aided diagnosis) system. The proposed microcalcification detection includes two parts. One is adaptive mammogram enhancement algorithm using homomorphic filtering in wavelet. The filter parameters are determined by background characteristics. The other is multi-stage microcalcification detection method. To verify our algorithm, we performed experiments and measured free-response operation characteristics (FROC) curve. The results show that the proposed microcalcification detection method is more robust for fluctuating noisy environments.

1 Introduction

From recent medical report, breast cancer is a leading cause of death for middle-aged women in USA and Europe. Thus, early detection and improvement in cancer treatments can reduce the mortality of breast cancer. Especially, the detection of microcalcification is a major part of diagnosis in early stage breast cancer. However, the microcalcification is too small to detect by palpable breast diagnosis. It is known that mammogram is the best modality to detect microcalcification [1].

Various types, sizes, and distributions of microcalcification make it difficult to detect actual microcalcification. In NMD (ACR National Mammography Database), 11 types of microcalcification and 5 types of distribution are categorized. These types of calcification are classified as benign calcification that is usually larger than the calcification associated with malignancy. Regional distributed calcifications scattered in large volume are not malignant. But the clustered calcifications are very dangerous type of microcalcification [2].

Image resolution and film contrast of mammogram are major problems in mammography. Numerous technologies and equipments have been developed during last several

* Images were provided by University of South Florida DDSM (Digital Database for Screening Mammography). This paper is supported by the development of digital CAD system (02-PJ3-PG6-EV06-0002) of Ministry of Health and welfare in Republic of Korea.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1110–1117, 2004.
© Springer-Verlag Berlin Heidelberg 2004

years. However, mammography has remained as one of the diagnostic modalities which are being enhanced. In mammogram, high attenuation properties and small size of microcalcification are major reasons of poor visualization and small lesions. To improve the visibility of breast cancer, image enhancement methods have been performed [3, 4, 5, 6, 7, 8]. The aim of these studies was to increase the contrast of microcalcification.

These previous algorithms may effectively detect microcalcifications. But some mammograms are taken from different environment such as different noise condition, X-ray intensity and concentration of sensitizer of mammogram films. In most previous works, they took the same parameters of image enhancement, denoising, and feature extraction. Some of DDSM (Digital Database for Screening Mammography) images include various kinds of noise in background and the noise is diffused through the breast area. Therefore, we must eliminate these kinds of noise to improve the performance of microcalcification detection.

In this paper, we propose an adaptive microcalcification detection method, which uses an adaptive image enhancement and noise reduction by using noise characteristics in background region of each mammogram. This paper consists of several sections. In Sect. 2, we describe our microcalcification detection system. Section 3 presents the performance of our method in DDSM database. We make conclusions in Sect. 4.

2 Adaptive Microcalcification Detection

The entire scheme of the proposed method is shown in Fig.1. Total system is composed into major two parts. One is image enhancement and the other is detection using ANN (Artificial Neural Network).

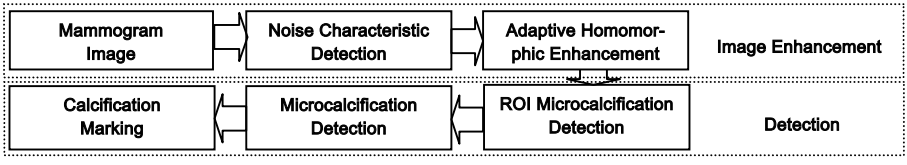


Fig. 1. Total system of our proposed method

In image enhancement part, we take background areas to obtain noise characteristics. Next, we use adaptive homomorphic enhancement and noise reduction in breast areas. The enhancement parameters are determined by noise characteristics of background. In detection part, first we get pixel information to find a potential microcalcification in breast areas that is to find ROI (Region of Interest). In microcalcification detection, entire image is decomposed into sub images corresponding to ROIs and microcalcifications are detected within these sub images.

2.1 Adaptive Mammogram Enhancement Using Homomorphic Filtering

The homomorphic filter function decreases the energy of low frequencies and increases those of high frequencies in the image. The homomorphic filter is used to find the gain K_m [4]. With the mammography images, the homomorphic filter gives contrast stretching for lower gray level, thereby enhancing the contrast. Based on the characteristics of homomorphic filter function, we determined the gain of mapping function, *i.e.*, weighting wavelet coefficients of channels corresponding to homomorphic filter function. Figure 2 represents the gain K_m that is determined according to the discrete homomorphic filtering.

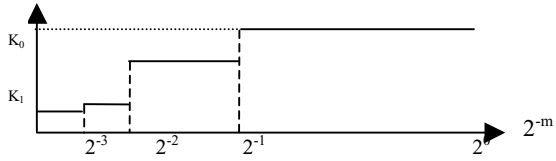


Fig. 2. Filter function for applying to wavelet coefficients

Figure 3 is a diagram of homomorphic filtering approach in this framework. Here, we first take logarithmic function for input signal. It also inverts the exponential operation caused by the radioactive absorption, which is generated in the process of obtaining mammography image.

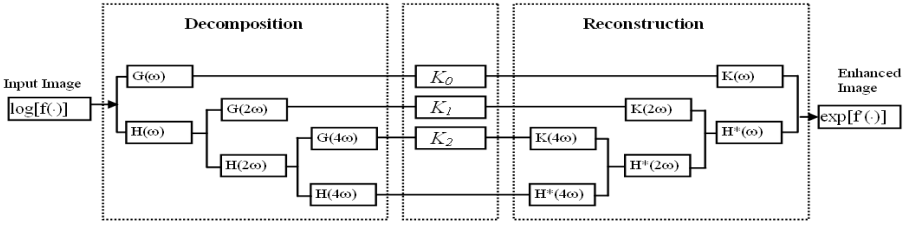
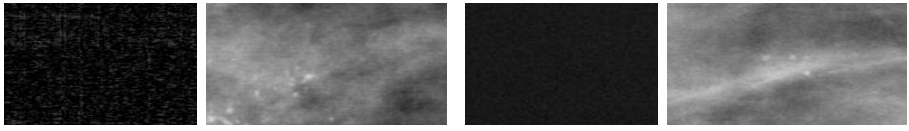


Fig. 3. Homomorphic filtering approach for image enhancement

Noise reduction is a considerable issue in image enhancement. One method of denoising is wavelet shrinkage that was presented in [3]. We can see that each mammogram contains its own noise characteristics because mammograms are taken from different environment. So applying the same parameters in wavelet shrinkage for every mammogram is not efficient. Taking into account noise properties of each mammogram, we propose an adaptive method for mammogram enhancement.



(a) (b) (c) (d)

Fig. 4. Examples of background noises and microcalcifications. (a),(b) are high noise (var.>100) in background and breast, (c),(d) are low noise (var.<50) in background and breast

Figure 4 (a), (b) images have some colored noise with high variance value. One can notice noise effect in breast area. To obtain noise characteristics of the image, the background of mammogram is segmented by thresholding the value combining the gray-level, mean and variance of pixels. The background areas are supposed to contain the noise of image. So we can take noise characteristic in this area. The noise characteristic is measured by background noise variance, which can be written as

$$\mathrm{var}_{background} = \frac{1}{N_{background}} \sum_{(x,y) \in background} \{(I(x,y) - mean(x,y))^2\}, \tag{1}$$

where $N_{background}$ is the number of pixel in background area. If high variance of background noise exists, we need to reduce the gain of homomorphic filter in high frequency domain as

$$K_m' = K_m \times \frac{A}{\mathrm{var}_{background} + A}, \quad if\ m = 0\ and\ m = 1 \tag{2}$$

where A is constant value to normalize noise variance. The m is level of wavelet, K_m means the gains of each wavelet level. In (2), m=0 means highest frequency level in wavelet and m=1 means second highest wavelet level.

Further, an adaptive method of denoising is included in the enhancement framework in wavelet domain. To achieve the edge-preserved denoising, nonlinear wavelet shrinkage method is proposed. A shrinking operator ($S(u)$) for denoising can be written as

$$S(x) = \begin{cases} sign(u) \times \left(|u| - \mathrm{var}_{backgrount}{}^2 / \sigma\right) & if\ |u| > \mathrm{var}_{backgrount}{}^2 / \sigma \\ 0 & otherwise, \end{cases} \tag{3}$$

where σ is variance of reconstructed image using wavelet coefficients in a sub-band. *sign(u)* means positive or negative sign of wavelet coefficient u. Threshold in this wavelet shrinkage is called a nearly optimal threshold [9].

Taking adaptive gains of high frequency area of wavelet domain and optimal denoising operators, microcalcification can be enhanced and also noise can be enhanced in breast area.

2.2 Microcalcification Detection

After adaptive image enhancement, we extract features to find ROIs which contain potential microcalcifications. Further, an artificial neural network with two hidden layers is employed to detect the pixels that may belong to microcalcifications. The features are obtained using statistical characteristics of gray level. We used two effective features: median contrast at position and contrast to noise ratio same as in [8].

In the neural network, inputs are two features mentioned above. And outputs are two categories of pixels, e.g., the pixel belonging to the ROI or not belonging to. In this paper, the first layer of ANN has 10 units and second layer has 4 units.

To train the neural network, fifty mammogram images containing cancer are selected. Pixels in calcification area of these images are obtained to make positive training samples. Negative training samples are obtained randomly in the mammogram

images of normal cases. Using the trained ANN, potential microcalcifications are found in a mammogram image to be tested. Then, the close detected pixels are grouped into one ROI as potential microcalcification.

From each potential microcalcification, which is found in the ROI detection, we find a sub-image whose size are triples of the ROI and its center is located at the center point of the ROI. In the sub-image, some features are obtained using local edge histogram, high-pass mask filter and Number of neighboring ROI.

In the local edge histogram, we measured two features with edge direction using 2x2 edge mask shown in Fig. 5. One is the maximum value of one direction and the other is non-directional edge. Using the local edge histogram, we can obtain high non-directional edge histogram value in microcalcification region while, in tissue or the mammary gland case, directional edges are emphasized. In the ROI detection, we found suspicious pixels such as microcalcification and abrupt change area. Using this edge information, above ambiguous pixels in breast area could be differentiated.

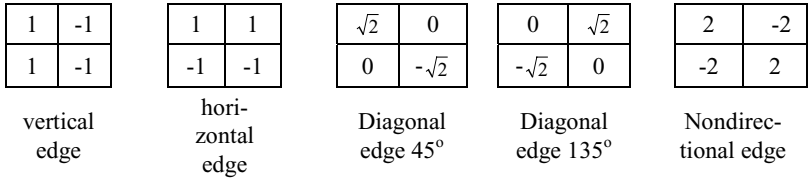


Fig. 5. Directional filter for edge detection

The two features in local edge histogram can be written as

$$Feature_Edge_{directional} = \max\left(E_{vertical\,edge}, E_{horizontal\,edge}, E_{diagonal\,edge45^o}, E_{diagonal\,edge135^o}\right), \tag{4}$$

$$Feature_Edge_{non-directional} = E_{non\,directional\,edge}\,, \tag{5}$$

where $E_{vertical\ edge}$, $E_{horizontal\ edge}$, $E_{45diagonal\ edge}$, and $E_{135\ horizontal\ edge}$ are average value after filtering shown in Fig.5, which represent four kinds of directional edge information.

To obtain the feature using high pass mask filter, the feature for the high pass mask can be obtained by

$$Feature_Highpass = 9\times \sum_{(i,j)\in ROI} p(i,j) - \sum_{(i,j)\in W} p(i,j), \tag{6}$$

where, W represents total window of high pass mask and ROI means potential microcalcification. For this feature, microcalcification is supposed to has higher value compared with adjacent area.

Finally the feature using number of neighboring ROI can be obtained by

$$Feature_N_{ROI1}(x_1,y_1) = \sum ROI(x_i,y_i) \quad if\ (x_1-x_i)^2+(y_1-y_i)^2 < R^2\,, \tag{7}$$

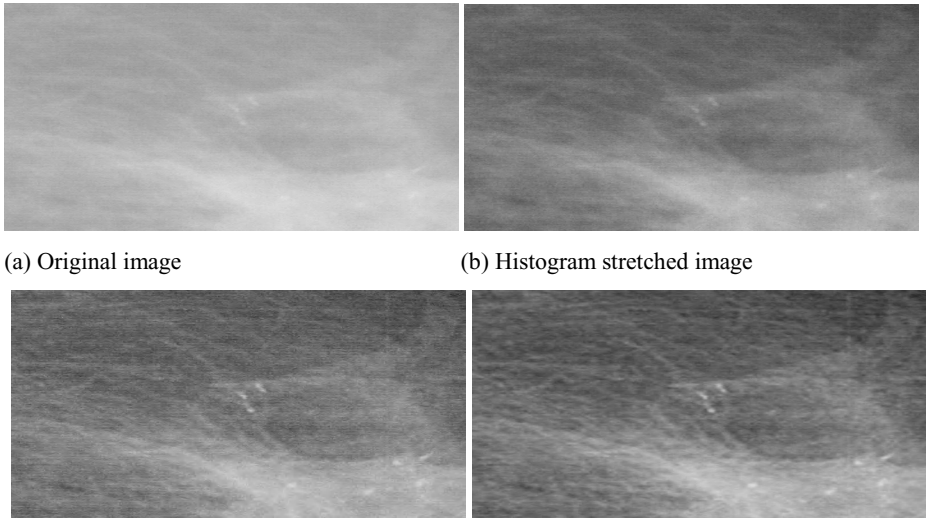
where, $ROI(x,y)$ means center position of an ROI. R is threshold which is related with microcalcification distribution. Equation (7) represents how many potential microcalcifications are located in adjacent area. To detect malignant microcalcification, the

distribution type of microcalcification is important. Real microcalcification is distributed in a way of highly clustering with neighboring potential microcalcification. Therefore, we believe high value of the feature means real microcalcification.

A back-propagation neural network is used to detect the microcalcification with features mentioned in above. The marked calcification areas of mammogram, which are confirmed by medical doctors, are used as positive training data. ROI in the normal areas are used as negative training data.

3 Experiment

To verify the effectiveness of the proposed method, experiments are performed with DDSM mammogram database. In DDSM database, the pixel size of microcalcifications ranges from 1x1 to 10x10, resolution of mammogram image is 50 μm/pixel, and gray level depths are 12bit and 16bit. The ANN training data and test data are obtained from DDSM database. The training database contains 56 cases of microcalcification and 56 cases of normal. And test data contains 118 cases of microcalcification and 118 cases of normal. Figure 6 (a) shows an original image, (b) is a local histogram enhanced image, (c) is a homomorphic enhancement image and (d) is adaptive enhancement image using the proposed algorithm. In Fig. 6 (b), (c) and (d), one can see more pattern of microcalcification than original image. Histogram stretched image is known to give good detail information of microcalcifications [3].



(a) Original image (b) Histogram stretched image

(c) Homomorphic filtering image (d) Proposed method

Fig. 6. Examples of enhancement images

A quantitative measure of contrast improvement is calculated using contrast improvement index (CII) [3]. It is defined by the ratio of enhanced contrast and original

contrast. Table 1 shows CII for original image and homomorphic enhanced image and adaptive enhanced in this paper. In Table 1, C means contrast value of an object, and standard deviation (STD.) of noise is obtained by background noise. As is seen, the proposed method gives higher CII value while smaller noise. Therefore proposed algorithm reduces noise compare with homomorphic filter.

Table 1. One example of Contrast Improvement Index and noise (Low noise case)

	Original Image	Homomorphic filtering	Proposed Method
C	0.0183	0.0600	0.0710
CII	-	3.2860	3.8858
STD. of noise	23.1	11.5	10.3

Because of the low contrast, original image has relatively low performance in microcalcification detection. We compared microcalcification detection using the proposed method with other methods such as histogram stretching and homomorphic enhancement method.

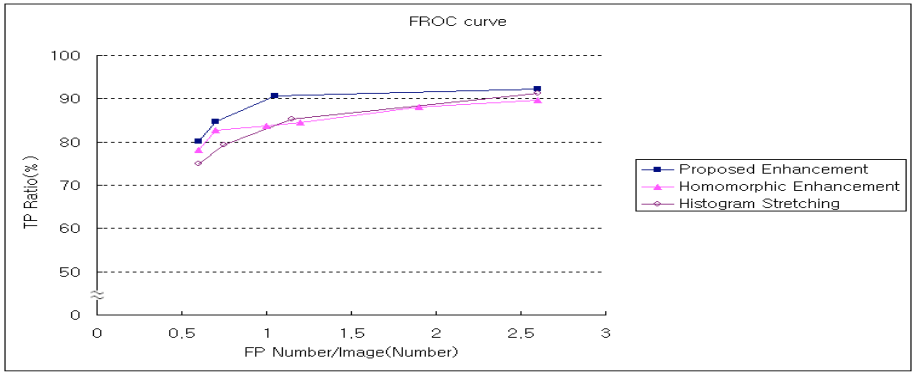


Fig. 7. ROC curve of microcalcification detection

To evaluate the performance of microcalcification detection, we measured FROC (free-response operation characteristics) curve shown in Fig.7. In FROC curve, we test 3 kinds of image enhancement methods. One is proposed method. Others are histogram stretching method and homomorphic enhancement method. TP ratio means true positive detection ratio, which shows how many microcalcifications are detected. And FP number/image means false positive number.

In the FROC curve, the proposed method shows higher TP ratio value than others methods at same FP number. The result of histogram stretching TP ratio is 79.3% for 0.75FP number/image. In homomorphic case, TP ratio is 82.8% for 0.7 FP number/image. In proposed method, TP ratio is 84.6% for 0.7 FP number/image. In Fig. 7, the proposed method show better result for every FP value. Homomorphic enhancement case is almost same as with proposed method for low FP number. But, this method could amplify noises which could be dealt with similar feature as microcalcification. Therefore it could have poor performance for high FP number/image.

4 Conclusion

Mammogram contains fluctuating noise characteristics because mammograms are taken from different environment, such as different noise condition, X-ray intensities and concentration of sensitizer of mammogram films. Therefore, the same parameters in enhancement gain and denoising for every mammogram are not efficient.

In this paper, we proposed an algorithm where noise characteristic is estimated in background area and eliminated in breast area. Therefore, we can obtain more enhanced mammogram using modified homomorphic filtering and adaptive denoising methods. Experimental result shows that the proposed algorithm gives better result compared with conventional image enhancement methods.

References

1. Johns, P. C., Yaffe, M. J.: X-ray characterization of normal and neoplastic breast tissues. Physics in Medicine and Biology. 32(6) (1987) 675-695
2. ACR National Mammography Database(NMD) http://www.acr.org/
3. Laine, A., Fan, J., Yang, W.: Wavelets for contrast enhancement of digital mammography. IEEE Engineering in Medicine and Biology Magazine 145 (1995) 536-550
4. Yoon, J. H., Ro, Y. M.: Enhancement of the contrast in mammographic images using the homomorphic filter method. IEICE Transactions on Information and Systems 85-D(1) (2002) 291-297
5. Karssemeijer, N.: A stochastic model for automated detection of calcifications in digital mammograms. Proc. 12th Int. Conf. Information Processing Medical Imaging (1991) 227–238
6. Strickland, R. N., Hahn, H. I.: Wavelet transform for detecting microcalcifications in mammograms. IEEE Trans. Med. Imag. 15 (1996) 218–229
7. Yu, S., Guan, L.: A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram film. IEEE Trans. Med. Imag. 19(2) (1998) 115–126
8. Sakellaropoulos, P., Costaridou, L., Panayiotakis, G.: A wavelet-based spatially adaptive method for mammographic contrast enhancement. Physics in Medicine and Biology 48 (2003) 787-803
9. Chang, S. G., Yu, B., Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Processing 9(9) (2000) 1532-1546
10. Zheng, B., Qian, W., Clarke, L. P.: Digital mammography: Mixed feature neural net-work with spectral entropy decision for detection of microcalcifications. IEEE Trans. Med. Imag. 15 (1996) 589–597
11. Yoshida, H., Zhang, W., Cai, W., Doi, K., Nishikawa, R. M., Giger, M. L.: Optimizing wavelet transform based on supervised learning for detection of microcalcifications in digi-tal mammograms. Proc. IEEE ICIP 3 (1995) 152–155

The Impact of Information Technology on Quality of Healthcare Services

Mariusz Duplaga

Jagiellonian University Medical College, Skawinska Str.8, 31-066 Krakow, Poland
mmduplag@cyf-kr.edu.pl

Abstract. The incorporation of information technology tools in medical practice results not only in opportunities related to improved quality and availability of services but also in transformation of the model of care delivery. Spreading use of the computer applications in healthcare is propelled by the growing expectations of citizens about quality of medical care. Furthermore, modern healthcare system relies on the concepts of patient's empowerment, shared care and continuity of services. The fulfilment of all these requirements may be possible through establishment of the e□health environment allowing for safe "electronic" contacts between physicians and their patients. In turn, the use of telemedical applications improves the process of information exchange between professionals representing different referential levels and areas of medicine. The awareness of the fact that medical errors themselves may be meaningful cause of morbidity and mortality, accerelated searches for effective countermeasures. The use of computer-based solutions is perceived currently as one of the effective remedies. The paper explores the opportunities of improvement of healthcare performance and quality with solutions brought by information technology.

1 The Role of Information Technology in the Vision of Modern Healthcare

The perception of healthcare has been evolving from the moment when humans started to perform conscious actions for preserving health and treating ailments. The approaches based on the irrational beliefs or individual experience of practitioners were substituted by evidence-based clinical practice in 20th century. Simultaneously, radical changes in lifestyle and nutrition had profound impact on life duration and quality of life. The expenditures for healthcare remain important part of national products in developed countries. Increased awareness of health-related problems results in high expectations about healthcare performance among the citizens. The increased life duration, the emphasis put on maintaining high quality of healthcare services, the sophisticated medical technology and high incidence of chronic conditions in modern societies add to growing trends in healthcare expenditures.

The progress in telecommunication and computers science leads to essential transformations in other domains. The great expectations about opportunities for increased cost-effectiveness and quality of services exist also in healthcare. Information technology opens the window for assuring appropriate healthcare quality obtained within reasonable costs. Particularly, rapid growth of Internet community is

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1118–1125, 2004.
© Springer-Verlag Berlin Heidelberg 2004

perceived as the factor which will radically change the model of health services delivery. The traditional approach to medical care requires substantial changes, if vital health-related needs of developed societies are to be fulfilled. Traditional codes of practice in medicine put stress on direct contacts between the physician and the patient. The form of this contacts is frequently defined by appropriate legal regulations. However, there are many forms of healthcare services in which traditional approach could be modified with the use of modern tools of communication. The best examples of such services include follow-up visits of chronic patients, monitoring of symptoms severity or educating patients about their diseases. In this context, it is understandable that the patient with established diagnosis, remaining under long-term control of his or her physician, returning for the consecutive visits does not require detailed history taking and physical examination on every occasion. Instead, the physician should focus on issues raised by the patient and problems that are priority in the context of the individual patient and specific disease.

The communication between the patients and health professionals may be considerably enhanced as questions raised by the patients are answered without the visit to physician's office. The doubts about disease course may appear quite often in patients anxious about their health status, and the digital platform of communication offers more flexible approach to the provision of explanations by health professionals. Simultaneously, the change in the forms of patient-physician communication brings new challenges to formal and legal regulations of standards of care.[1] The use of electronic communication for routine contacts between physicians and the patients is associated with some risks, but generally, it may be perceived, assuring technical feasibility and security of the specific technical solution, as convenient substitute for repeated visits of the patient in the physician's office performed for checking the patient's status and renewing prescriptions.[25] Internet and e-mail as new modes of communication between patients and physicians allow for greater flexibility and optimum use of health professionals' time.[24]

The Internet-based applications have also the potential for improvement of clinical practice services through better timing, adjustment according to patient's needs and integration of evidence-based guidelines in the information systems, e.g. in the form of decision support tools. After first clinical trials, it became obvious that even severe conditions may be efficiently treated at home providing appropriate communication with the patient as well as monitoring of vital signs and symptoms via telemedical infrastructure is assured. Teleconferencing system may be installed in patient's home and used by the patient with assistance of family members or social workers.

The inclusion of patients in the process of diagnostics and therapy is perceived nowadays as one of the key conditions of successful care.[30] Their active role in monitoring of disease severity became the obvious aim of medical management in chronic conditions, such as diabetes, bronchial asthma or arterial hypertension. However, self-observations and self-measurements conducted by the patients should be supported by interaction with health professionals having the access to patients' data. The Internet and telecommunication links shape perfect environment for such activities. Extensive use of web technologies for chronic care is possible in those societies which have appropriate access to Internet. The rapid growth of the mobile telephony penetration brings another tool suitable for telemonitoring applications.

The Internet health-related platforms are devoted to various aspects of chronic care. The website focused on specific disease offers many opportunities for improvement of care effectiveness. Patients are able to enter their self-observations and self-measurements results on-line but they can also receive answers to questions on every day activities related to their medical conditions without going to physician's office.[13] Simultaneously, health professionals can monitor health status of their patients through monitoring centre.[11] The process of monitoring may be enhanced with the feature of alert triggering in case of deterioration of parameters monitored by the patient prompting for appropriate action from the physician.

The Internet users surfing on the Web frequently use it as the source of health-related information.[16] The consumers of health services search for information about specific disorders, providers or health programmes as well as for contacts with support groups. The WWW-based medical record gives the patients access to his or her data from home environment.[31]

Self-education is important opportunity for chronic patients related to the widespread use of Web-based applications. Internet technologies enable individualised education and training process. More advanced solutions emerge the patient in the environment supporting self-management of the medical problem but also imposing the regularity of activities. In such environment the patient gains self-assurance that in case of deterioration the appropriate information will be forwarded to healthcare provider.

Full exploration of advantages related to the use of information technologies yields many other fields of application in healthcare such as the implementation of preventive measures, the efficient use of individual genetic information in early stages of health-supporting interventions (e.g. lifestyle modification, diet, avoiding specific factors), the transformation of the traditional contacts between health professional and patients and development of virtual support environments for people with special needs.The clinical care incorporates new modes of services delivery through the use of systems supporting electronic healthcare record, telemedical applications or electronic prescribing. Telemedicine offers broad spectrum of clinical scenarios enabling the e-visits for patients, teleconsultations between health professionals representing different medical centres or specialities, telepresence and teletraining in interventional specialities or even virtual surgery. The availability of information system in medical facility enables the implementation of reminding and alerting applications or decision support tools enhancing the quality of the process of care.

2 Improvement of Patient Safety

The awareness of the incidence of medical errors exploded with the report of the Institute of Medicine from USA published in 1999.[23] The main issue emphasized in the report was the great magnitude of the harm resulting from medical errors, the importance of appropriate organization of healthcare services in avoiding medical errors as well as the need for mandatory reporting programmes. The conclusions included in the Report were based on the publications in pear review journals issued in preceding 10 years. The data from this publications allowed for the estimations

about the consequences of medical errors on national scale in USA. The most shocking information was realization that even 98000 deaths annually may be related to medical errors. In 2001, the next report of the Institute of Medicine, titled "Crossing the Quality Chasm" was published.[21] The six aims for improvement indicated in this report include safety, effectiveness, patient-centeredness, timeliness, efficiency and equity. These aims may be approached with the application of information technology tools.[4]

The positive role of technology in reduction of the incidence of medical errors should result in broader use of such solutions, like bar-coding of medication, wireless PDA devices for access to patients record and drug information as well as systems for drug order entry. Among the most frequent adverse effects occurring in healthcare, first three places are hold by adverse drug events, nosocomial infections and procedural complications.

Adverse drug event are the largest, single category of medical errors. Medication errors are responsible for considerable number of physician's office visit and hospitalizations. The broader use of computer applications should be particularly effective in diminishing the number of adverse drug events related to medication errors. Three type of solutions are proposed in this context: computerized physician order entry (CPOE), automated dispensing machines (ADMs) and computerized medication administration records (CMARs).[22] The use of information systems, especially in the areas where automation process is available, is perceived as the way of errors containment in medical management. CPOEs offered in the IT market are able to detect negative drug interaction, unexpected shifts of dosage or conflicts between medication and medical conditions.[5], [26] Another important solution influencing three most frequent types of adverse effects in healthcare is the use of bar codes. [27]

The use of e-prescribing systems in outpatient practices should be encouraged as their implementation is potentially beneficial for patient safety and efficiency of management.[8], [9] The main benefits include the generation of legible and accurate prescriptions checked for harmful interactions, adherence to guidelines and health programmes as well as the option of access to patient information and availability of decision support functions at the point of care.

It is also obvious that great global impact on incidence of medical errors would have the widespread use of electronic medical record systems with functions of order entry, medication administration records and decision support than implementation of many single-purpose applications.

3 The Improvement of Access to Health Care Services with Telemedicine

Telemedicine means delivery of healthcare services conducted with the application of teleinformatic infrastructure. Advanced telemedical system or network enables diversified forms of medical services and information transfer among institutions and health professionals when direct contact is impossible. Telemedical system should fulfill such criteria as geographic distance between centres which exchange medical

information and the use of teleinformatic technologies as substitute of direct contacts. This type of services is also called telehealth, cybermedicine, or more recently, e-health. The appearance of new definitions of telemedicine is related to continuous development of this domain and finding new areas of applications for teleinformatic technologies.[7]

First telemedicine implementations were developed in 50-ies of 20th century. Telemedicine underwent considerable evolution from this time.[29] Today, telemedicine is perceived not only as the way of delivery of healthcare services in some extreme circumstances (as geographic isolation or emergency states) but more as the use of modern telematic tools for improvement of availability and quality of medical services. The use of telemedical application is not even over the world. Paradoxically, developed countries seem to be more devoted to establish telemedical applications in their healthcare environments as they see potential substantial benefits from their implementation. These benefits became more obvious due to rapid growth of the population of Internet users as it offers less expensive, more homogeneous and easily available technology to population. Even if benefits from the growing use of Internet in healthcare are very convincing, there are some issues which require careful recognition e.g. problems related to security and confidentiality of the e-health communications. The assurance of the quality of healthcare services delivered by Internet is also important issue.[10], [28]

Telemedicine employs various types of technical solutions. The systems available today include store-and-forward applications, videoteleconferencing systems and applications based on the WWW technologies. Store-and-forward applications are used in asynchronous mode as some time elapses between the moment of data transfer and their interpretations. Such system may be used for transmission of patients data to consultant who, in convenient moment, will read them and respond to questions.[3], [20] Store-and-forward systems are usually used in situations, in which videoteleconferencing contact is not necessary for efficient telemedical interaction or available telecommunication have limited band of transmission. This may occur in rural, peripheral locations without advanced telecommunication infrastructure.

Real-time interactions are possible in videteleconferencing systems.[18] Multi-point teleconferencing infrastructure enables numerous participants involvement in the teleconsultation sessions. Tumor boards appointed for making decision on oncologic problems are good example of such telemedical services.[6] Videoteleconferencing telemedical scenarios require broadband connections to enable the transmission of high volumes of data. They usually rely on the use of ISDN lines or fiberoptic links.[14] Specialistic peripheral equipment is used for registration of medical images and other types of data (stethoscopes, otoscopes, dermoscopes or whole body examination cameras). Document cameras enable the transfer of data which are not available in electronic form.

There are no single technology which could satisfies all requirements of specific implementation. The successful realization of telemedical scenario depends on choice of technology which satisfies users' requirement related to the volumes of transmitted data and band of the connection. Appropriate choice of teleinformatic infrastructure is also dictated by the scale of available financial support in healthcare system.[12], [19]

4 Chronic Care – Evolving Models of Care Based on Telemonitoring

Chronic diseases became the major health burden in developed countries. The estimations made all over the Europe and in the USA indicate that even 70% of health care expenditures is spent for management of chronic diseases in these areas. The number of patients suffering from chronic conditions is growing, but simultaneously a great number of them want to conduct active professional and social life. This results in the need for the improved efficiency of health care systems. The patient's role in the therapy and symptoms control, the importance of close patient follow-up, community services and prevention are not emphasized properly in healthcare systems as they served for years mainly for the provision of care to patients with acute illnesses.[32] The use of Internet technologies in health services provision in chronic care brings new opportunities and potentially cost-effective models of care.[15] E-health environment offers additional characteristics to modern healthcare through shaping new types of relation between patients and physicians, enabling automatic screening for exacerbations of disease, and creating the network of support for the patient.

The increase of the population of patients requiring chronic care and monitoring became the factor triggering the search of cost-effective but safe, health services delivery models. The rapid adoption of Internet in many domains brought also the promise for improvement of efficiency of healthcare delivered to patients with chronic diseases. The WWW technologies seem to be particularly feasible to obtain the appropriate level of interactivity for chronic patients.[2], [17] Patient empowerment means his or her involvement in prolonged observations of disease course, conducting self-measurement of parameters essential for specific disease. The main trace of self-management in chronic diseases is based on registration of self-observations and self-measurements in long-term perspective. Prolonged tracing of physiologic parameters enables the description of the course of disease and forecasting of exacerbations. Self-observations and self-measurements may be the basis for automatic triggering of alerts sent to health professionals and patients. Trend observed in the disease course are also the basis for modification of medication by the patient (self-adjusted treatment).

5 Summary

The traditional relation between physician and patient underwent considerable changes in last decades. The interactions between patient and healthcare system evolved to various forms. There is also the increasing pressure on healthcare providers to deliver services of high quality in cost-effective manner. The wider use of information technology in healthcare brings new opportunities for healthcare services delivery observing evidence-based approach.

The achievement of desired features of modern healthcare, such as seamless and shared care as well as continuity of care may be supported with web-based and mobile applications. Internet adds considerably to the concept of pervasive services,

especially in the area of chronic care. Patients with chronic conditions became the main consumers of healthcare services and telemonitoring services is new efficient way of care delivery in this population demanding frequent interactions with healthcare providers. Computer-based applications are also perceived as important solution enabling the reduction of medical errors. The access to limited specialized medical resources may be enhanced due to telemedical application. Their use increase also the collaboration within clinical team taking care of individual patient and formed by health professionals based in various institutions.

References

1. Balas A.E., Jaffrey F., Kuperman G.J., Boren S.A., Brown G.D., Pinciroli F., Mitchell J.A.: Electronic Communication With Patients: Evaluation of Distance Medicine Technology. JAMA 1997; 278(2): 152-159
2. Balas E.A., Iakovidis I.: Distance Technologies of Patient Monitoring. BMJ 1999; 319: 1309
3. Bangert D., Doktor R.: Implementing store-and-forward telemedicine: Organizational issues. Telemed. J. e-Health 2000; 6(3): 355-360
4. Bates D.W., Cohen M., Leape L.L., Overhage J.M., Shabot M.M., Sheridan T.: Reducing the frequency of errors in medicine using information technology. JAMA 2001; :299-308
5. Bates D.W., Leape L.L., Cullen D.J., Laird N., Petersen L.A., Teich J.M., Burdick E., Hickey M., Kleefield S., Shea B., Vander M., Seger D.L.: Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 1998;280: 1311-1316,
6. Davison A.G., Eraut C.D., Khan N., Haque S., Tanqueray A., Trask C.W., Lamont A., Uppal R., Sharma A.: Telemedicine for lung cancer: a new multidisciplinary approach. Thorax 1999; 54(3): A37
7. DeBakey M.E.: Telemedicine has now come of age. Telemed. J. 1995; 1(1): 3-4.
8. Elison B.: Electronic Prescribing in Ambulatory Care. J Manag Care Pharm 2001;7(2)
9. E-Prescribing. California Healthcare Foundation. November 2001
10. Ferguson T.: Digital Doctoring – Opportunities and Challenges in Electronic Patient-Physician Communication. JAMA 1998; 280(5): 1361-1362.
11. Field M.J., Grigsby J.: Telemedicine and Remote Patient Monitoring. JAMA 2002; 4: 423-425
12. Field M.J.: Telemedicine: a guide to assessing telecommunications in health care. National Academy Pr. Washington, 1996.
13. Forkner-Dunn J.: Internet-based Patient Self-care: The Next Generation of Health Care Delivery. J Med Internet Res 2003; 5(2): e8 [url:http://www.jmir.org/2003/2/e8]
14. Galvez P., Newman H.: Networking, videoconferencing and collaborative environments. Comp. Physics Communications 1998; 110: 43-50
15. Gillespie G.: Deploying an I.T. cure for chronic disease. Health Data Manag 2000; 8(7): 68-74
16. Glick T.H., Moore G.T.: Time to learn: the outlook for renewal of patient-centred education in the digital age. Med Educ 2001; 35: 505-509
17. Gomaa W.H., Morrow T., Muntendam P., Smith G.: Technology-based disease management: A low-cost, high-value solution for the management of chronic disease. Dis Manag Health Outcomes 2001; 9(10): 577-588
18. Gustke S., Balch D.C., Rogers L.O., West V.L.: Profile of users of real-time interactive teleconference clinical consultations. Arch. Family Med. 2000; 9(10): 1036-1040

19. Harrison R., Clayton W., Wallace P.: Can telemedicine be used to improve communication between primary and secondary care? BMJ 1997; 313(7069): 1377-1381.
20. Houston M.S., Myers J.D., Levens S.P., McEvoy M.T., Smith S.A., Khandheria B.K., Shen W.K., Torchia M.E., Berry D.J.: Clinical Consultations Using Store-and-Forward Telemedicine Technology. Mayo Clinic Proceed. 1999; 74(8): 764-769
21. Institute of Medicine: Crossing the Quality Chasm. A New Health System for the 21st Century.National Academy Press, Washington, DC,2001
22. Kilbridge P., Welebob E., Classen D.: Overview of the Leapfrog Group Evaluation Tool for Computerized Physician Order Entry. Leapfrog Group and First Consulting Group, 2001
23. Kohn L.T., Corrigan J.M., Donaldson M.S. (eds): To Err is Human: Building a Safer Health System: a report from the Committee on Quality of Healthcare in America. Institute of Medicine, National Academy of Sciences. National Academy Press, Washington, DC, 1999
24. Leslie S.: Online consulting. The experience of a commercial service. J Telemed Telecare 2001; 7(Suppl. 2): S2:78-82
25. Mandl K., Kohane I., Brandt A.: Electronic Patient-Physician Communication: Problems and Promise. Ann Intern Med 1998; 129(6): 495-500
26. Mekhjian H.S., Kumar R.R., Kuehn L: Immediate benefits realized following implementation of physician order entry at an academic medical center. J Am Med Inform Assoc, 2002; 9: 529-539
27. Patterson C., Cook R., Render M.: Improving Patient Safety by Identifying Side Effects from Introducing Bar Coding in Medication Administration. J Am Med Inform Assoc 2002; 9: 540-553]
28. Stanberry B. The legal and ethical aspects of telemedicine. 2: Data protection, security and European law. J. Telemed. Telecare 1998; 4: 18-24.
29. Strode S.W., Gustke S., Allen A.: Technical and Clinical Progress in Telemedicine. JAMA; 281(12): 1066-1068
30. Tattersall R.: The expert patient: a new approach to chronic disease management for the twenty-first century. Clin Med 2002; 2(3): 227-229
31. Tsai C., Starren J.: Patient Participation in Electronic Medical Record. J Am Med Ass 2001; 285(13): 1765
32. World Health Organization “Innovative Care for Chronic Conditions: Building Blocks for Action. Global Report.” 2002

Computer Generated Patient Plans Based on Patterns of Care

Ole Martin Winnem

NTNU, SINTEF, S.P. Andersensvei 15b, Trondheim, Norway
Ole.M.Winnem@sintef.no

Abstract. This article presents a method to create patient plans from Patterns of Care (PoC). PoC's are either developed based on Clinical Guidelines or as suggested in this article, based on generalized patient plans. The PoC's are then the basis for generating patient plans. The patient plan generation is a two step process where the first step is based on knowledge intensive case based reasoning and identifies the PoC to be used. The second step adapts the PoC to a patient plan. The way this approach uses PoC's and measured patient information to generate plans is a promising way of provide best practice and to reduce the work required by the medical experts.

1 Introduction

The medical community has so far been concentrating on developing Clinical Guidelines (CG) for manual use. Since these CG's has been developed for human interpretation and adaptation, they are not machine-interpretable. In order to create human supported automated processes, the CG's must be instantiated and structured. In our approach we introduce some context in the adaptation of the CG's in order to fit a specific guideline to a given care model. The care model we are using as example is based on a WHO care model [1, 2] for treatment and care of chronic conditions, where patient plans are becoming more and more important.

In general patient plans are developed based on the experience of the medical expert and with no interaction with other professions. Reasons for this are the lack of tools to support a collaborative approach and the organisational model used. This article identifies components that are needed to improve the creation of patient plans and suggest a model that link the identified components together.

2 The Clinical Guidelines – Structure and Content

Clinical guidelines are normally presented in a text document with minimal structure. They have been developed in order to help physicians in performing their work in a unified way – following best practise, but as generic that adaptation to situation and care model is needed. Below is a figure of a typical Clinical Guideline. The figure shows a Clinical Guideline with chapters. These chapters are the only structure that is introduced and the content is based on descriptive text that is easy to understand for

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1126–1132, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the expert, but is difficult to understand without the context presented and to other possible caregivers.

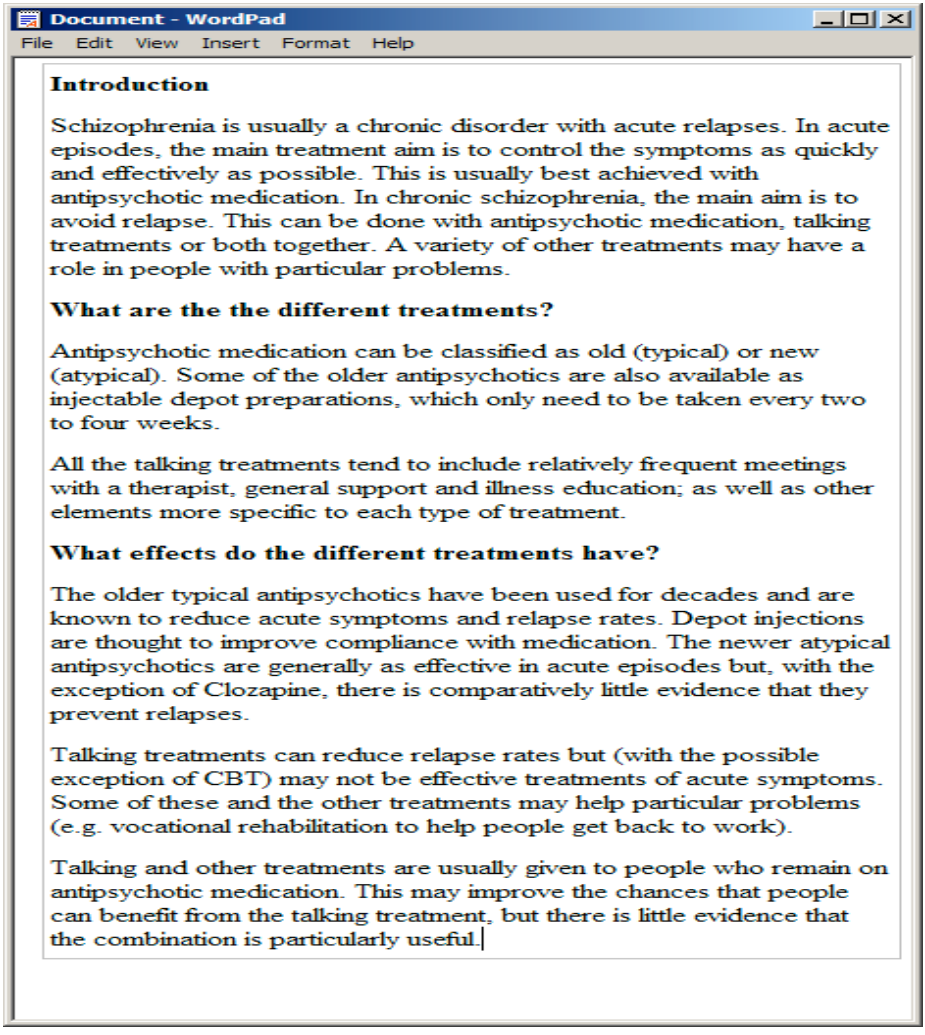


Fig. 1. Example of Clinical Guideline[1]

Several approaches have been tried out in order to make Clinical Guidelines computer interpretable [3]. One of the approaches is Asbru [4], which is a framework for representing time-oriented Clinical Guidelines based on XML. Transforming a textual Clinical Guideline into a Asbru syntax can be performed by using a tool called Degel [5]. Degel is a two step method to transfer Clinical Guidelines into Asbru syntax. The first step creates a semi-structured representation and is performed by

[1] The CG is copied from the web-page: www.nelmh.org

medical experts. This semi-structured representation is not possible for a computer to understand, but the second step that is performed by a knowledge engineer with medical domain knowledge which adapts it into a structured guideline represented in XML. Asbru is developed with focus on representing time constraints. This makes Asbru an alternative to the Creek [6] representation we are adapting. Creek is a semantic network representation with focus on representing general knowledge in combination with case specific information. We have chosen to use this because of its descriptive power and easy access to adapt it to our needs.

Common to these solutions are that they are only in limited use and that it costs a lot to build up them and to update and support them. Anyway we believe that it is an advantage to use case specific information and case based methods to identify the PoC to use, due to a simple and easy to understand way of updating the systems knowledge.

3 Structure of PoC

The PoC structure is defined in order to make it easy to include the information into a knowledge model and at the same time be expressive enough to capture all the vital information from a CG. In order to find this simplified structure we have studied CG's in order to identify generic structures. Generic structures alone are not enough and we have identified structures specific to diagnose and care model. Below is a part of a generic structure of the PoC. The structure shows how the PoC is built up, using XML. The most important part of this structure is the plan part that is used to create a specific plan for the patient. It contains information about all activities that should be created for the selected PoC. The activities have commonalities, but also differences. Typically, medication activities have information describing the ATC-Code and volume, which are parameters that are not useful for a nurse visit activity. In a nurse visit the human tasks describing what is needed to be performed and how this should be done following best practice is much more relevant.

The second important part is the structure that is matched to the patient diagnose and measurements. This part is used in order to select the PoC to be used for creating the patient plan. It is the major diagnostic parameters set by the medical expert. These are used in combination with the measurements that have been performed. Examples of diagnostic parameters are the ICPC or ICD10 code and parameters describing how sever the situation is. Examples of measured parameters are PEV or blood pressure used in the diagnostic process for a COPD or asthma patient. Further the structure contains a pointer to which CG it represents and eventually what level of the CG it is related to. It also includes general information about the purpose and care model it is related to.

The XML structure is used to build up a computer model that makes it possible to evaluate which PoC to use and how to adapt it into patient plans. The XML structure should be easy to understand both for human and computer, which makes it difficult to develop. The suggested structure will be further developed in order to fit new needs that are experienced during the test and evaluation period.

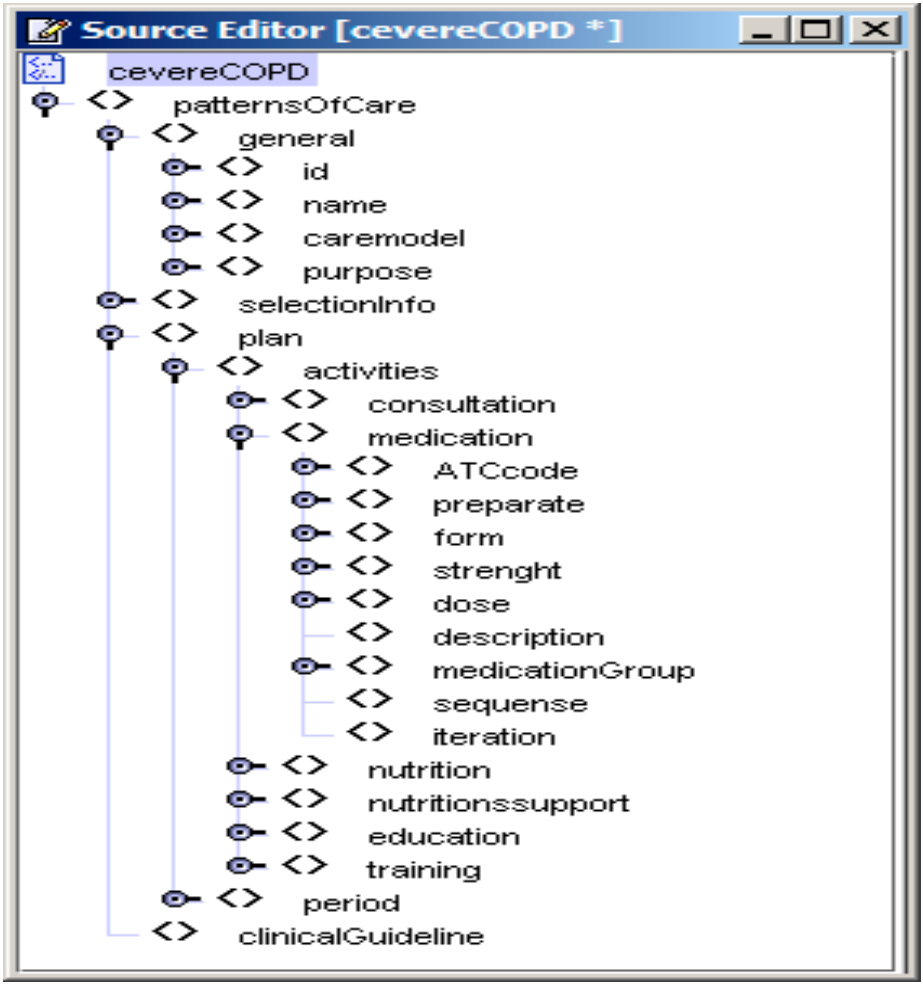


Fig. 2. An XML diagram of PoC

4 Creating PoC

At the moment the process of creating PoC from CG is based on a manual process with technology support. Our initial thoughts have been to design a semi-automated tool which gives knowledge support to the medical expert creating PoC based on CG. But because of the structure of CG's (created with textual structure to be easy to read for humans – not machines), we find it promising to start in the other end.

The similarity between PoC and a detailed patient plan makes it promising to study the possibility to create PoC from patient plans. This requires that the patient plans have been evaluated and controlled that they follow related clinical guideline. So far

we have not gained experience with this method, but the results using this approach will be published as soon as the solution has been tested. The process suggested is a tree step process as follow:

1. Identify representative patient plans
2. Generate PoC from the identified patient plans
3. Evaluate and verify the resulting PoC according to related CG.

In the first step a team of experts identifies detailed plans that represent the diversity of the treatment for the selected CG. These plans are then categorized and made available for the induction algorithm that creates a generic plan or Patterns of Care. Before using this newly created PoC, it should be evaluated up against related CG. This is a manual process where the expert panel can edit the PoC in order to fit it to the CG.

5 Knowledge Model and Resources

The knowledge model in the system is based on a semantic network model describing all terms in the system and their relations to each other. This semantic network is combined with a case based retrieval system that uses the semantic model to explain the similarity between previous used plans and the new current problem to be solved. The knowledge representation system is an adapted and extended version of CREEK [6, 7]. The explanation power is based on the product of all relations needed to explain how one term relates to another.

Other resources in the system are database resources related to plans for the care givers and patient constraints. The database resources are relational databases available to select consultation time with needed care providers. The most important care provider within advanced homecare is the home care unit [1]. This unit is the centre of the care provision for advanced homecare and it is the unit with the largest number of constraints. The needs of the patient are also most related to this resource unit. The patient constraints are related to medication, measurements, travel information, nutrition, patient knowledge level and available local caregivers (patient relatives), where the medication constraints are related to CAVE and mismatch between medications. Measurement constraints are related to problems regarding time of measurement and measurement dependability of medication.

6 Generating Patient Plans

The creation of patient plans from PoC contains two sequential steps. The first step is related to selection of PoC and the second is related to adapting the selected PoC to the available resources.

The selection of PoC is the most knowledge intensive and difficult task. Diagnose and available measurements give the input information which is mapped to available PoCs in order to select the most promising PoC. The selection of PoC is based on the CREEK [6, 7] retrieval engine which is able to match both quantitative and explain

qualitative information and in this way it is much suited to match input information about the patient with the information represented in the PoCs.

Adapting the selected PoC to a patient plan is an optimisation problem which includes among others the constraints described in chapter 5. At the moment the optimisation in the system is only based on selecting the nearest resource and adding activity to the plan based on a sequential approach where the PoC is evaluated starting on the first activity.

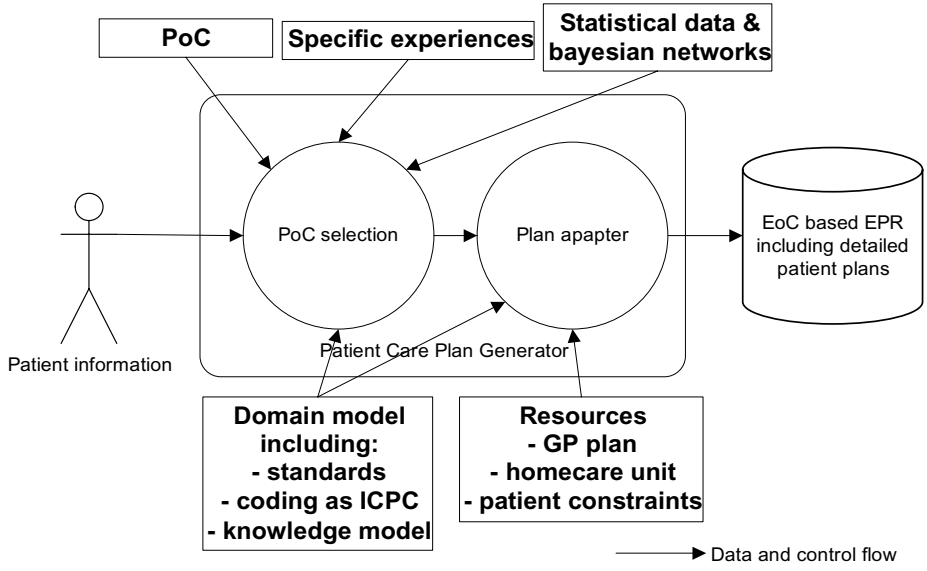


Fig. 3. The figure shows the two processes related to the Patient Care Plan Generator. The two processes uses different kind of information and knowledge sources in order to create patient plans

7 Conclusions and Further Work

This article has described a method for generating patient plans based on predefined Patterns of Care. This is an approach that will shorten the time used to develop plans for the treatment and care of chronic patients. Since the approach is based on structured clinical guidelines, it is reason to believe that the method will help giving the patient unified care according to best practice. In order to make this approach viable it is needed to have access to patient data and care giver resources. Access to these data sources are under development in Norway and the first projects where patient plans have been shared between caregivers at different level and profession have given promising results in 2003.

The design has not been tested in a laboratory or real world environment. This should be done in order to make further conclusions. The system will be further developed in order to make it possible to test the real use of PoC as standard for patient plan generation. At the end a large scale trial should be performed in order to make the final conclusions.

References

1. Winnem, O.M., Walderhaug, S. *Distributed, role based, guideline based decision support*. in *eHealth in Common Europe*. 2002. Krakow: Springer.
2. WHO, *Innovative care for chronic conditions*. 2002, World Health Organization: Geneva. p. 61.
3. van Bemmel, J.H., Musen, M.A., *Handbook of Medical Informatics*. 1997: Springer. 621.
4. Shahar, Y., Miksch, S. & Johnson, P., *The Asgaard Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines.* Artificial Intelligence in Medicine, 1998. 14: p. 29-51.
5. Shahar, Y., Young, O., Shalom, E., Moskovitch, D., Boaz, D. and Galperin, M. *A Hybrid Framework for Representation and use of Clnical Guidelines*. in *AMIA*. 2002.
6. Winnem, O.M., *Integrating knowledge- and symbol level modelling - THE CREEST WORKBENCH*, in *IDI*. 1996, NTNU: Trondheim. p. 134.
7. Aamodt, A.P., E, *Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches*. AICom - Artificial Intelligence Communications. Vol. 7:1. 1994: IOS Press. 39-59.

On Direct Comparing of Medical Guidelines with Electronic Health Record*

Jana Zvárová, Arnošt Veselý, Petr Hanzliček, Josef Špidlen, and David Buchtela

European Center for Medical Informatics, Statistics and Epidemiology, Institute of Computer Science AS CR, Pod Vodárenskou věží 2, Praha 8, The Czech Republic
zvarova@euromise.cz

Abstract. Electronic Health Record (EHR) systems are now being developed in many places. The EuroMISE Center has developed a system of the Multimedia Distributed Electronic Health Record (MUDR) that provides universal structure for collecting data and multimedia objects. Moreover, a method to compare data items on the input of EHR system with medical guidelines has been developed. Therefore it is possible to reveal if the input data item, that represents patient diagnosis or proposed patient treatment, contradicts with medical guidelines or not. The method on direct comparing of medical guidelines with EHR has been tested in hospital information system settings and in settings of the remote GP office.

1 Introduction

The motivation to use electronic data storage in medicine is based on the obvious drawback of the paper medical record including illegible handwriting, poor organization of documents, missing data and ambiguous data. Current developments in electronic health records focus on advanced features, including strategies for data entry and retrieval, multiple views and transparent integration of diverse information sources of different platforms [1], [2]. Moreover, the availability of an EHR (Electronic Health Record) is a must for using advanced decision support in medicine and health care, including medical guideline support. The European project I4C of the 4th Framework Programme (1996–1998) was carried out for the further advancement of cardiac care [3]. It focused on clinical applications and its main goals were to integrate access to data wherever stored, to support evidence-based care by remote electronic consultation and peer review and to record more comprehensively and more consistently patient data, images, videos and bio-signals. With the support of the I4C project the multimedia Open Record for Care (ORCA) was developed. It integrates structured patient data entry including history, medication, symptoms and more with multimedia objects as ECG, angiography or laboratory data.

* The research was partially supported with the project LN00B107 Ministry of Education CR.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1133–1139, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Based on the experience gathered in the EU projects, especially TripleC [4] and MGT [5], [6] the EuroMISE Centre has been focused on specific tasks of EHR, Electronic Formalized Medical Guidelines (EFMG), data mining, decision support and knowledge based systems and on the practical application of new developments in the Czech healthcare environment. There is running development of EHR named MUDR (MUltimedia Distributed electronic health Record), see e.g. [7], [8]. The universal architecture of MUDR EHR provides a flexible platform for storage of medical data of any type and MUDR EHR takes into account obligatory standards for Czech health care providers. Moreover, since the EHR serves as a tool for data collection, decision support programs can use its data for automatic generation of alerts, reminders and suggestions when standards of care (e.g. based on medical guidelines) are not achieved. The development of EFMG and the task on making them available for physicians using Internet, especially with the simultaneous use of EHR is desirable.

2 MUDR Architecture

The MUDR EHR is based on a three-layer architecture with a data layer, an application layer and a user interface. This decomposition enables separating different system modules to small functional parts, which makes the system more flexible. In our system we define the global architecture with communication interfaces based predominately on XML and HTTP. We also define an interface for connecting medical guidelines formalized in the form of dynamic libraries. These libraries extend the capabilities of the MUDR EHR.

MUDR^{II} architecture extends the basic three-layer architecture. The communication between the MUDR DB Server and the MUDR Application Layer Service uses the MUDR DB Connection Module, which enables implementing the data layer on various database platforms. The application layer chooses the right module for the database currently used. Using this scheme, the application layer communicates transparently with the data layer using always the identical interface.

The main difference lies in the application layer interface, used to provide the functionality of the application layer. The application layer integrates communication modules to communicate with different client types. The main communication module is called MUDR WS. This module provides objects using the MUDR .NET Remoting API (MUDRNRAPI). Using .NET Remoting, the remote call of methods of the shared objects is possible. This functionality is utilized by another application layer component called MUDR Web Service. This web service makes the MUDR Web Service Application Interface (MUDRWS-API) accessible to common clients. A typical general practitioner (GP) uses a PC on his table to work with the MUDR EHR remotely. The communication is enabled by the HTTP Server installed on the application server. The commands and parameters are encoded using the SOAP standard.

For the eventual usage of MUDR thin clients in the form of HTML and WAP browsers we use a MUDR WS Proxy Service. This service is implemented as a

Common Gateway Interface (CGI) program; it provides classical HTML or WAP pages to thin clients and appears in the role of a classical MUDR Web Service client at the other side.

3 Electronic Formalized Medical Guidelines

Many medical guidelines were elaborated to improve quality of medical care and to achieve standardization of treatment. Usually, the first version of medical guidelines is worked out by a group of medical experts in a free text. For computer implementation and processing it is necessary to have medical guidelines explicitly structured and developed in an electronic form. A lot of modeling means were suggested for this purpose. The most important and nowadays mostly used is the GLIF (Guideline Interchange Format) model. The GLIF model is a result of collaboration among Columbia University, Harvard University, McGill University and Stanford University. Version 2.0 of GLIF (GLIF2) was published in 1998 [9].

The main goal of GLIF was to enable sharing of electronic formalized medical guidelines (EFMG) among health institutions and across computer applications. GLIF specifies an object-oriented model for guidelines representation and syntax for guidelines utilization in software systems as well as for their transport. GLIF guidelines are mostly given as a flowchart representing a temporarily ordered sequence of steps. Different types of steps in the flowchart represent clinical actions and decisions [9], [10]. To enhance usability of GLIF, new version GLIF3 was worked out. GLIF3 builds upon the framework set by GLIF2, but augments it by introducing several new constructs and extending GLIF2 constructs to allow a more formal definition of decision criteria and patient data [11]. The GLIF model gives process-oriented view of guidelines. The model consists of a set of classes for guideline entities, attributes of these classes and data types for attribute values. Instances of classes (objects) have only attributes and no methods.

A particular guidelines encoded in GLIF are an instance of a general guideline model. It could be represented in a form of oriented graph. The nodes of the graph are guideline *steps* and edges represents continuation from one step to the other one. *Guideline steps* are action step, decision step, branch and synchronization steps and patient state step. *Action steps* specify clinical actions that are to be performed. It can be an application of some therapy, carrying out some examination or measurement etc. Action step also may name subguidelines, which provide greater detail for the action. *Decision steps* are used for conditional branching. There are two kinds of decision steps: case step and choice step. *Case step* is used, when branching is determined by evaluation of defined logical criteria based on data items. *Choice step* is used when the decision cannot be precisely specified in guidelines themselves and decision should be made by the user. *Branch* and *synchronization steps* enable concurrency in the model. Guideline steps that follow branch step could be performed concurrently. Branches with root in branch step eventually converge in synchronization step.

In this step all branches are synchronized. It means, that actions, that follow the synchronization step, could not be performed, unless all actions, following branch step and preceding the synchronization step, are finished. *Patient state step* characterizes patient's clinical state.

Guideline model can be equivalently expressed also in a language form. Syntax for guideline describing language is a part of guideline model specification. In language form encoded guidelines consist of a sequence of guideline steps. Some attributes of a guideline step contain next guideline steps. It enables sequential representation of graph structure in the guideline describing language.

Action steps of medical guidelines specify different kinds of physician actions:

- An action could mean to carry out some examination or some laboratory measurement. The gained results we will call *medical guideline examination data*.
- An action could promote some treatment or application of some therapy. Such actions we will call *medical guideline actions*.

Patient state steps of medical guidelines specify patient states, for example diagnosis or risk group. We will call them *medical guideline states*. Suppose that it is possible to establish relation between medical guidelines examination data, medical guideline actions and medical guideline states on the one side and input data of EHR on the other side. Suppose further that a *translation table* could express this relationship. EHR input data correspond to the *medical guideline examination data* or *medical guideline actions*. EHR input data, which correspond to the medical guideline states we call *patient states*.

What means, that some patient state or physician action at the input of EHR contradicts the medical guidelines? We suggest defining this contradiction in the following way. If some patient state or physician action at the input of EHR has its counterpart in medical guidelines and all occurrences of its counterpart in the medical guideline model are according to the content of EHR database unreachable, then this patient state or physician action contradicts medical guidelines.

4 General Views on Comparing of Medical Guidelines with EHR

EHR systems are now being developed in many ways in many places. EHR system enables to store all relevant data about patient in a structured way and facilitates their further computer processing [13]. The main part of each EHR are medical observations, laboratory results, stated diagnoses, medical treatments etc. The main purpose of EHR system is to store and retrieve data in user-friendly dialogue with physician. Advanced EHR systems should also utilize medical knowledge and provide following facilities.

Checking facility. We check if values of input data are in accordance with medical knowledge. Usually the system verifies if the input value is inside defined interval.

Reminder facility. If a stated diagnosis or a chosen treatment is not in agreement with common medical knowledge, the system warns the user and suggests more probable diagnosis or more appropriate action.
Decision support facility. Based on EHR data the system supports decision-making activities of a physician.

Automatic checking of input data should be an integrated part of each electronic health record (EHR) system. EHR system should analyze input data and warn the user if he inputs data item which is out of interval of possible values or if he inputs data item, which suggests some decision or medical action, that contradicts common medical knowledge. For implementing this task, EHR system usually uses set of logical if - then rules. These rules could be extracted from text form of medical guidelines. The other possibility is to compare input data and medical guidelines directly using electronic medical guidelines. A method for direct comparing of EHR and medical guidelines was proposed in [12].

We will concentrate on the reminder function of the EHR system. The usual way how to design reminder system is to design a set of rules, based on medical knowledge. These rules are then during user-EHR interaction checked against input data. The main problem is how to find out these rules. Usually they are to be built up by a team of experts and knowledge engineers. The methods of their elaboration can be similar to those used in building expert systems. When medical guidelines are at our disposal, the rules can be extracted from them. But is it necessary to follow this way? Perhaps it could be possible to compare data at the input of EHR system with medical guidelines directly.

EHR system can realize the reminder function itself or reminder function can be realized by a Stand-Alone Reminder System (SARS). The collaboration between EHR system and SARS could be described as follows:

- When user puts into EHR system a patient state or a physician action, EHR system sends an activation message to SARS. The part of the message must be the patient state or the physician action.
- SARS translates the patient state or the physical action into a medical guideline state or medical guideline activity according to the translation table.
- SARS, according to the guideline model, determines supports of all occurrences of medical guideline states or medical guidelines actions.
- SARS translates all EHR input data into medical guideline examinations according to the translation table.

SARS finds out if user follows electronic medical guideline. If not, it sends a warning message.

5 Conclusion

The advantages of proposed method of comparing medical guidelines and EHR using EFMG are the following.
General applicability. System based on proposed method can work with arbitrary guidelines. Only at first the guidelines must be transformed into GLIF

graph model and then the model must be encoded into guideline describing language. The transformation of free text guidelines into GLIF model or some similar structured and precisely defined formal model should be accomplish anyway, because only in this way one can be certain, that guidelines are unambiguous and non-contradictory. The encoding of GLIF graph model into guideline describing language is then straightforward as each algorithm coding.

Facilitation of changes in the system. When some part of guidelines is changed, it is not necessary to correct the set of rules used for checking of input data. What is sufficient to do is making corresponding change in guideline model.

Support of distributed computing. The EHR system can be running at one site and the system for comparing input data with guidelines at another site. The system based on the proposed method could be even more distributed. EFMG could be accessible for example on a server at department with medical experts, who worked out guidelines and are responsible for their maintenance and updating. The system for comparing EHR data and a medical guideline can be at another site and it could download EFMG from authorized site whenever necessary. In this case arises problem with translation table. Who should build and maintain it? The simplest solution could be sharing the same vocabulary. EHR system site could at the beginning inform comparing site which standard medical vocabulary, as SNOMED for example, is used. The first experience with comparing EHR data and EFMG on hypertension treatment has been gathered in both information system settings and in settings of the remote GP office.

References

1. Zvárová J., Hanzlíček P., Špidlen J.: Electronic Health Record in Cardiology: Pilot Application in the Czech Republic. In: MIST2002 Proceedings, Taiwan, John Wiley & Sons (2002) 10–13
2. Iakovidis I.: Towards Personal Health Record: Current Situation, Obstacles and Trends in Implementation of Electronic Healthcare Record in Europe. Int. J. Med. Inform. 52 (1998) 105–115
3. Pierik F. H., Ginneken A. M., Timmers T., Stam H., Weber R. F.: Restructuring Routinely Collected Patient Data: ORCA Applied to Andrology. In: Bemmel J. H., McCray A. T. (eds.): Yearbook of Medical Informatics 98, Health Informatics and the Internet, Schattauer, Stuttgart (1998) 257–263
4. Přibík, V., Grünfeldová, H., Hanzlíček P., Peleška, J. and Zvárová J.:, Czech National Data Standards Implementations in ORCA Electronic Patient Record in Cardiology. In: Hasman A., Blobel B., Dudeck J., Engelbrecht R., Gell G., Prokosch H.U. (eds.) Medical Infobahn for Europe, IOS Press, Amsterdam (2000) 652–655
5. Svátek, V., Říha, A., Zíka, T., Zvárová J., Jiroušek, R., Zdráhal, Z.: Informal, Formal and Operational Modelling of Medical Guidelines. In: Proceedings 4th JCKBSE , IOS Press, Amsterdam (2000) 87–92
6. Zvárová J., Peleška, J., Hanzlíček P. and Zvára, K.: Enhanced Care of Hypertensive Patients Using Internet. Technology and Health Care 6, Vol. 9, IOS Press (2001), 487–488

7. Hanzlíček P.: Development of Universal Electronic Health Record in Cardiology. In: Surjan G., Engelbrecht R., McNair P. (eds.) Health Data in the Information Society, IOS Press, Amsterdam (2002) 356–360
8. Špidlen J., Hanzlíček P., Říha, A., Zvárová J.: Flexible Information Storage in MUDR^{II} EHR. Proceeding of the IJM EuroMISE 2004 (submitted)
9. Ohno-Machado L., Gennari, J. H., Murphy S.,N., Jain N.,L., Tu S., W., Oliver D., et al.: The GuideLine Interchange Format: A model for representing guidelines, Journal of the American Medical Informatics Association 5 (1998) 357–372
10. Boxwala A., A., Greenes R.,A., Debel S., R.: Architecture for Multipurpose Guideline Execution Engine, Proc. AMIA Symp. (1999) 701-705.
11. Peleg M., Boxwala A.,A., et al.: GLIF3: The Evolution of Guideline Representation Format, In: http:/smi-web.stanford.edu/projects/intermed-web/guidelines (2000)
12. Veselý A., Zvárová J., Peleška, J., Buchtela D., Anger Z.: Medical Guidelines Presentation and Comparing with Electronic Health Record. Proceeding of the IJM EuroMISE 2004 (submitted)
13. Říha, A., Svátek, V., Němec P., Zvárová J.: Medical Guideline as Prior Knowledge in Electronic Healthcare Record Mining. In: Zanasi A., Brebbia C. A., Ebecken N. F. F. E., Melli P. (eds.): Data Mining III, WIT Press (2002) 809-818

Managing Information Models for E-health via Planned Evolutionary Pathways[1]

Helmut Duwe

Institute of Aerospace Medicine, German Aerospace Center (DLR),
Cologne, Germany
Helmut.Duwe@dlr.de

Abstract. E-Health is one of the many e-application areas promising to increase global connectivity and computational power of applied computer networks. Information modelling is an essential ingredient of developing e-health applications. The theoretical foundations of information modelling are growing. System theory, semiotics, semantics, and ontology have become allied partners to build powerful information models. Phrases like "ontology-based conceptual modelling" indicate this process of sophistication. When developing practical e-health systems the management of complexity is still a major challenge. Evolutionary pathways seem to be an appropriate approach to handle complexity. Significant gaps between existing pieces of avantgarde academic theory and broad practical success stories are seen. The idea of planning evolutionary pathways as mechanism to handle information model complexity is illustrated in one area, namely the development of a study and subject record manager for health-relevant physiology research.

1 Introduction

May be the 21st century will be known in the far future as the rise of the e-ages. E-business, e-commerce, e-government, e-science, last not least e-health are ringing in the decades which will hopefully see a tremendous increase in global connectivity and computational power of computer networks.

Herewith, a major issue of the ethical debate as well as the technical agenda is the handling of the right balance between simplicity versus complexity of representing data, information and knowledge concepts. Per se simplicity is not bad and complexity is not good, contextual adequacy is required.

2 Finding the Right Adequacy between Simplicity and Complexity

The handling of information has become a multidisciplinary scientific effort. The subject matter disciplines themselves have developed vast infrastructures to handle data, information and knowledge (e.g. Medline).

[1] The views and conclusions contained in this paper are those of the author and should not be interpreted as representing the official policies of any organization.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1140–1147, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The following examples shall illustrate that existing approaches, useful as building blocks for e-health information models, indicate the need for paving the roads and bridges in-between.

A simple, but effective, mechanism to handle data for clinical trials has been set up by the CDISC (Clinical Data Interchange Standards Consortium) group. In the CDISC Operation Data Model (ODM)[2] the actual clinical data are handled as attribute-value pairs, as such a trivial technical approach.

The general approach behind this is called EAV (Entity-Attribute-Value) concept. It holds to a large extend the subject matter complexity within the data content and operates on the metadata level mainly with the very generic concepts of entity, attribute and value. It has been shown [see Brand et al] how this content internal complexity can be unfolded in Clinical Trial Data Management Systems, which are technically based on relational databases.

The unfolding of structural representations from internally complex data content to managed information objects with adequate metadata functionality is not only an issue of reliable "older" technologies like relation database systems, but also of object-oriented approaches. It is proven technique to avoid the explosion of classes via generic classes and attributes. Attributes work here as containers for typed content. The making of typologies is shifted in this case from a class typology (with explicit metadata entities) to the content structure (with more or less known internal metadata).

The examples to illustrate the potential and the handling difficulty with complex information and information models are taken from the LOINC (Logical Observation Identifier Name and Codes) and the openEHR (open Electronic Healthcare Record) project.[3]

The LOINC standardisation efforts of laboratory and clinical observations take care to find systematic names as universal identifiers. A LOINC long name has 6 parts: the component (analyte), the property measured, a time aspect, the sample type, the scale type, and where relevant the method used. A variant of an oral glucose tolerance test is written for example as `GLUCOSE^1H POST 100 G GLUCOSE PO`.[4] The precision of the naming needed can drive the appearance of the names towards what the creators of LOINC call themselves "awkward". Nevertheless, LOINC is not only a terminology listing, it is a very important starting point for more elaborated information models of all the identified and hence listed concepts.

The actors behind the openEHR (open Electronic Health Record) initiative have accumulated a rich documentation of information models, suitable to play a vital role in developing advanced healthcare record systems. Only one valuable central aspect should be mentioned here, namely the undertaking to clarify such an abstract concept as "data structure". The "Data Structures Reference Model" of openEHR contains sophisticated representations of a complex "structure" concept. Fundamental kinds of

---

[2] After all, the CDISC standards have their strength through the stringent way they require the explicit documentation of the metadata and by the richness in which they have captured the subject matter semantics of clinical trials.

[3] Further examples could have been taken from HL7 V3 RIM (Health Level 7, Version 3, Reference Information Model), due to size limitation of this paper this was not possible.

[4] H for hour, G for gramme, PO for *per ora*.

structure are identified, for example "single (values)"[5] (singularity), "list", "table", "tree" and temporal ones like "history", "event", and more. Further concepts like tree-tables and table-trees are thinkable and already implemented here and there. The challenge comes with the linking of complex subject matter issues and data structures.

3 Some Fundamentals of Modelling Theory Exploited for E-health

It is stated here that system theory and semiotics shall not be underestimated when working in the field of information modelling [see Duwe et al].
Here, only the triadic sign relationship as core concept of semiotics shall be given as absolute basics. In the triadic sign relationship the nature of language is seen as a triangle of the reality object (thing), the representing sign or symbol (ink on paper or signs on a computer display), and the mental process of having a concrete thought based on sensory inputs (interpretant). The mapping of concepts into this triangle is still a major scientific challenge. For example, what is about the reality aspect of "visions" which bear "reality" years later ? What is about brain processes without sensory input from a reality source like it happens with synaestesia.

Within the scope of these considerations it is only important to keep all three corners of the triadic sign relationship in mind, when a term and concept is discussed.

The Renovated[6] Alliance Partners

An improved modelling theory for e-health applications should assimilate some of the major advancement made in conceptual modelling, in data semantics and ontologies.

The first cycle of UML (Unified Modelling Language) modelling during system analysis and early system design is basically centred around conceptual modelling. It has to be distinguish from the specification modelling in a second cycle, more or less the late phase of the system design phase of software development.

Handling the vocabulary of a knowledge domain is the key to semantic analysis. The primary issue of data semantics and hence semantic interoperability is the clarification of meaning. When reading the themes of conferences dealing with data semantics the overlap with conceptual modelling and data ontology becomes obvious.

Gruber calls ontologies an "explicit formal specification of a shared conceptualisation". [Gruber, cited after Heese, see Heese]. This illustrates once more the closeness of the term "ontology" and "concept/conceptual modelling", finally showing that (A) conceptual modelling, (B) data semantics and (C) data ontology are highly interwoven. Further topical constructions as for example "foundational ontologies" [see Schneider] represent the peak with regard to catching abstract complexity.

---

[5] resulting in class names like SINGLE_S, TREE_S, TABLE_S, etc.

[6] This word is used to indicate the picture of "high technology interiors (informatics) in ancient buildings (ontology)".

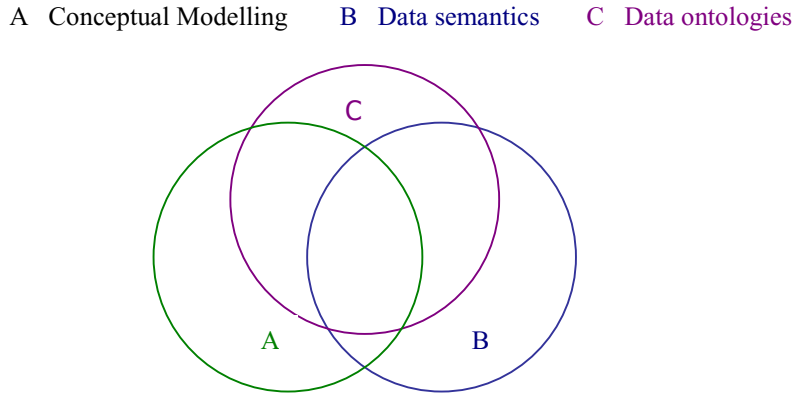


Fig. 1. Relationship between conceptual modelling, data semantics and ontologies

4 The Practical World of Information System Development in E-Health

Abstract modelling theory has to find more intensively its way into the real practical world of developing information systems for e-health.

A variety of textbooks of software engineering deliver good overviews of Software Engineering (SE) process models. The most well known kinds of SE process models are the so-called waterfall model, the V-model, the OO-models, and models with richness regarding model evolution like the incremental, the evolutionary and the spiral model. A model with increasing importance is the so-called Rational Unified Process (RUP) approach. RUP has been basically elaborated by the founding fathers of UML, which have pointed more than once to the importance of iterative and incremental development of software. There is a fruitful influence of these SE process models on each other noticeable. For example, the V-model, itself a highly structured waterfall model with integrated quality assurance, has incorporated incremental extension mechanism.

The health domain is huge. Some tentative figures from medical terminology systems shall give a first impression on the magnitude of complexity. The International Classification of Disease (ICD) contains in its four-level systematics more than 12.000 disease classes; the Medical Subject Heading (MeSH) covers around 20.000 main key words; the Logical Observation Identifier Names and Codes (LOINC) system lists more than 30.000 universal identifiers for laboratory and clinical observations, the Systematized Nomenclature of Medicine (SNOMED) covers a complex set of 200.000 terms, and the Unified Medical Language System (UMLS) project works with more than 600.000 concepts and more than a million terms. The estimation of concepts with relevance for medicine yields figures as high as 50 million.

The Institute of Aerospace Medicine (DLR-IAM) organizes and performs studies in the fields of human physiology, biology and medical technology. Scientific studies vary in the degree of complexity. An individual study is usually carried out by one institution, whereas a multi-centre study is performed by two or more institutions based on a common study protocol. Integrated studies comprise several related individual studies which are conducted by one or several institutions sharing test subjects, staff and resources at one study site. To support a wide range of study types the DLR-IAM is developing together with the Institute of Information processing and Computer supported new Media (IICM) at Technical University of Graz an open Electronic Study Record Manager (openESRM)[7]. The openESRM is based on an open Medicine Telematics Platform. This platform provides comprehensive security mechanisms like authentication, authorization, digital signature.

The Study Protocol Information Model of RCRIM[8] was studied as input for an own study protocol information model. Putting all available parts of the RCRIM Protocol Model together indicates a little bit the growing complexity of modelling a subject matter like "Clinical Trials".

Displaying complete information models like RCRIM study information model or the HL7 V3 RIM tends to cover whole walls. Evolutionary extensions of one concept adds a further dimension of complexity to this "wall paper" phenomenon.

According to the idea of evolutionary pathways each identified "element" of such an information model will get a "genealogy vector". For each concept a multi-generation design shall be prepared that marks the way from simple to complex.

The staging shall be briefly illustrated with the OpenESRM concept of "Study", being here (besides "Study Record") the "root concept" (or root class). The prospective genealogy of the concept "Study" starts in the baseline generation with a simple associate construction, more or less a kind of mind mapping exercise.

The second generation has been designed with a data skeleton class, to delegate the basic data set just like it is done with the ProtocolSkeleton class of the RCRIM model. But, the StudySkeleton class is a data type candidate to be used throughout the entire life cycle, for example as super-class for the ProtocolSkeleton too (RCRIM has done the delegation so fare only for the Study Protocol). To reduce redundancy, generic sets of data are extracted and keep available for whatever other concept (class) is interesting to use it.

The third generation introduces the Collection concept as used within the "PlannedStudy" class of the RCRIM Protocol Model. The fourth generation of the Study class is designed as mix of singularity attributes and collection attributes. Furthermore, the problem of modelling "static" and "dynamic" features of a phenomenon is here reflected very intensively.

"Study" had been defined as a specific "process". The Workflow IT community has extensively reflected the term "process". But, a typical definition here just states that processes are chains of activities. The settlement of an adequate model of

[7] The description of the ESRM concepts are based on DLR-IAM internal papers, major contributions to the concepts have been mainly made by Doris Wilke and Erwin Bartels, especially the identification and description of the process activities is taken (with minor deviations) from the work of Doris Wilke.

[8] Regulated Clinical Research Information Model, the main player are here the CDISC group, the HL7 group, and FDA (the U.S. Federal Food and Drug Administration)

"process" is currently tried as part of a "cinematographic concept" of a "dynamic world (segment)". Here, a "process" should be modelled as "series of moving static still images". "Images" are then "situations", or "snapshots of reality" as McCarthy defines situations [see: McCarthy].

The author of this paper intends to elaborate the fifth generation of the "Study" class as illustration of advanced ontological concepts (going in the direction of Luc Schneider and utilizing the work of the new Leipzig Institute of Formal Ontology and Medical Information Science). The fifth generation of the concept "Study" is up to now thus a stage with an uncertain degree of feasibility. Nevertheless, the linking of realizable and experimental stages of an concept implementation is an important issue of technology assessment.

Last not least, the continuity of one and same concept is aligned with the explication of the subject matter, keeping the explication as close as possible to the use to normal natural language and domain language. The technical explication can and must use artificial constructs to remove ambiguity (with formal languages), for example use such term constructs for different aspects of the concept "Study" as StudyStructure (static), StudyProcess, StudyProcessStructure.

5 The Methodological Quintessence of Evolutionary Information Modelling

Already in the early nineties groups like Hahn et al [see Hahn] discussed the problem of concept versioning and concept drift. They proposed tracking methods to follow evolutionary concepts in dynamic settings and favoured an incremental concept build-up [see Klenner 1994, Hahn 1994]. Therefore, features like versioning, incremental development, iterative improvement have not only to be considered as elements of general software engineering, but are also important as explicit technique for conceptual information modelling in applied domains.

It becomes quite obvious that other techniques have to complement the methodological tool set of information modelling. The most important elements shall be an advanced definition approach [see Swartz] and an applied context theory of meaning.

6 Outlook

Linking information modelling of health systems, basic life science and clinical practice to gain evidence based feedback loops should be a scientific subject topic put forward as issue on a e-health road map. Here, questions and answers concerning the need for and affordability of semantic grid computing [see De Roure] for e-health should have their place.

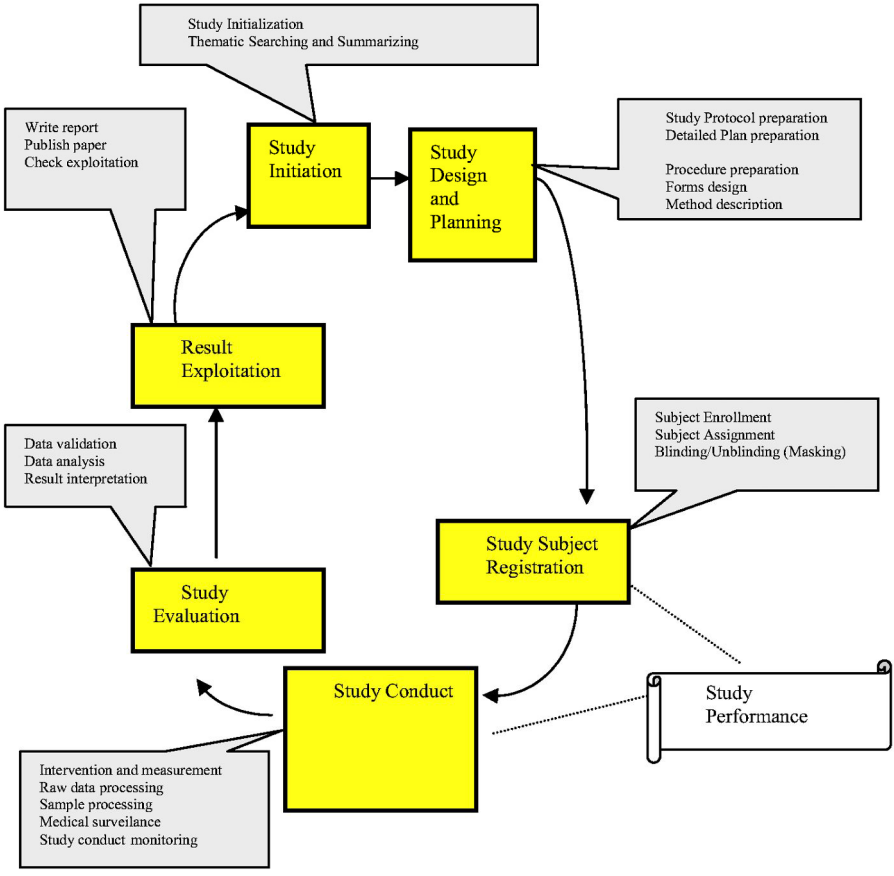


Fig. 2. The life cycle of a study and the main activities

To formulate an e-health road map for Common Europe is a science policy topic. An action item on this "road" could be the set up of a Model Library of European Generic Health System Components. Building an e-health system with a global generic backbone and integrating national as well as cultural variations is a global task to be put forward to gain benefits for all (e-health for all). After all, numerous groups around the globe may contribute to the effort that some day several millions of specific and inter-related health concepts will be available as computer-based re-usable model modules.

Acknowledgement. The author thanks Andrea Nitsche for critical-constructive proofreading.

References

1. Duwe H, Schwartmann D, Bartels E: The Relevance of System Modelling and Semantic Analysis for the Development of E-Health Applications, International Conference on E-Health in Common Europe, Kraków, 2003
2. Clinical Data Interchange Standards Consortium (CDISC), http://www.cdisc.org
3. Logical Observation Identifier Names and Codes (LOINC). User Guide, edited by McDonald et al, http://www.regenstrief.org/loinc
4. openEHR (Electronic Health Record), http://www.openEHR.org
5. Schneider L: Designing Foundational Ontologies. The Object-Centered High Level Reference Ontology OCHRE as a Case Study, in: I.Y. Song et al. (Eds.), ER 2003, LNCS 2813, 2003. pp. 91-104
6. Regulated Clinical Research Information Model, http://www.hl7.org/Special/committees/rcrim/rcrim.htm
7. Friedman LM, Furberg CD, DeMets DL: Fundamentals of Clinical Trials, St. Louis, 1996
8. Brandt CA et al: Metadata-driven creation of data marts from an EAV-modeled clinical research, International Journal of Medical Informatics, 65, 2002, pp. 225-241
9. Hahn U, Klenner M: Tracking the evolution of concepts in dynamic worlds, in: D. Karagiannis (Ed.), DEXA '94 – Database and Expert System Applications, Proceedings of the 5th International Conference, Athens, Greece, September 7-9, 1994, Berlin, pp. 410-419
10. McCarthy: Actions and Other Events in Situation Calculus, http://www-formal.stanford.edu/jmc/
11. Swartz N: Definitions, Dictionaries, and Meanings, http://www.sfu.ca/philosophy/swartz/definitions.htm
12. De Roure D, Jennings NR, Shabolt: The Semantic Grid. A Future e-Science Infrastructure, http://www.semanticgrid.org

An Attributable Role-Based Access Control for Healthcare

Dirk Schwartmann

German Aerospace Center (DLR), Cologne, Germany
Dirk.Schwartmann@dlr.de

Abstract. Role Based Access Control (RBAC) has the potential for reducing the complexity and total cost of security administration. Even though RBAC implementations aim on administrating large scale systems, they have a shortcoming in common. They do not allow to define attributable roles and permissions. But such roles are very common in our thoughts and language. When we say "attending physician of patient x", we mean a role attending physician with all associated permissions to fulfill the treatment of patient x. Because the resulting permissions only differ in the restriction to a particular patient, it is desirable that attributes like "patient x" are used in roles and permissions to restrict the rights to access only data related to that patient. This paper shows how attributes can be applied to RBAC, in order to reduce the total number of role- and permission-objects in security administration.

1 Introduction

One of the most challenging problems in modern healthcare is the security administration of large networked systems containing patient data. Today's security administration is usually based upon access control lists for each user, leading to large time exposures for updating access control lists on the data, creating new and deleting old users.
Role Based Access Control (RBAC) has the potential to reduce the complexity and total cost of security administration. Access Rights are defined by User-to-Role and Role-to-Permission assignments. This reduces the need to administrate rights at data level (e.g. files on a file-system). Even though RBAC implementations aim on administrating large scale systems, they have a shortcoming in common: they do not allow to define attributable roles and permissions. However, such roles are very common in our thoughts and language. We say "attending physician of patient a" and "quality assurance officer in project x". It is desirable that attributes like "patient a" restrict the rights to access only data related to that patient to fulfill the least-rights-principle. That means the user should possess the minimum rights to fulfill the current task. A static role "attending physician" alone would break this rule. A person in this role would have access to data of all patients. So, the problem must be solved in a different way. The simple solution would be the introduction of a static role "attending physician of patient x" with appropriate permissions to restrict access to the data of this

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1148–1155, 2004.
© Springer-Verlag Berlin Heidelberg 2004

patient within all patient records. But this is not a very good idea. Assume the amount of administration in a large hospital with hundreds of patients. For each new patient a new attending physician role would be created.
In August 2001 the National Institute of Standards and Technology (NIST) published a proposed standard for RBAC that includes the fundamental set of RBAC components (see [1]). In this paper an extension to the standard with attributable permissions and roles will be introduced. At the time of role activation the attribute values are passed on to it and are subject to permission checks. So, different attribute values lead to different access rights. As a result a single attributable role "attending physician" will be enough for a hospital. Thus, the amount of roles and permissions is reduced dramatically resulting in less error-prone administration and in an enormous reduction of the total costs of security administration.

2 Overview of RBAC

Role-based access control (RBAC) is a proven alternative to traditional access control mechanisms like discretionary access control (DAC) and mandatory access control (MAC). RBAC allows to represent the natural organizational view upon access rights. Roles are used to describe the functions of individuals with associated permissions. In RBAC, permissions are defined as operations on objects. Users are assigned to roles and are granted permissions this way. During sessions users activate roles to fulfill job functions with distinct rights.
Figure 1 gives an overview of the conceptual items and their interrelations in hierarchical RBAC of the proposed NIST standard. The hierarchical RBAC extends the so-called core RBAC with role hierarchies. General role hierarchies provide support for an arbitrary partial order to serve as the role hierarchy which models inheritance of permissions and, in the opposite direction, user membership among roles. Because it is needed further down in this paper the definition of general role hierarchies of the proposed NIST standard is presented here:

Definition 1. *General Role Hierarchies*
$RH \subseteq ROLES \times ROLES$ is a partial order on ROLES called inheritance relation, written as $\succeq$, where $r_1 \succeq r_2$ only if all permissions of r_2 are also permissions of r_1, and all users of r_1 are also users of r_2.
Formally: $r_1 \succeq r_2 \Rightarrow authorized_permissions(r_2) \subseteq authorized_permissions(r_1) \wedge authorized_users(r_1) \subseteq authorized_users(r_2)$

3 Authorization on Life-Long Health-Records

The aim of health-records for every single patient that are to be used in shared care throughout the whole life of each person, bears some problems. Envisioning the existence of a (distributed) electronic health record for each living person, the authorization mechanism must allow access to each patient-record in a treatment case. It is clear that this cannot be fulfilled by introducing a static role "treating physician of patient x" for each patient x in practice. The number of roles in

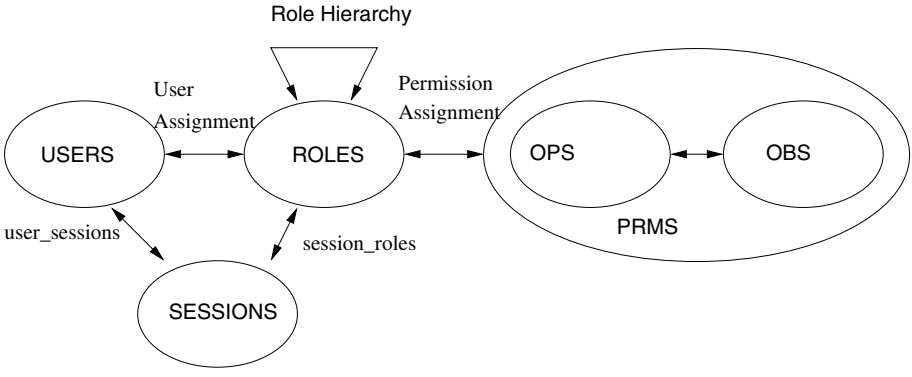


Fig. 1. Hierarchical RBAC

IT-systems would steadily increase. As a consequence security administration would become a difficult task and no matter how efficient concrete authorization sub-systems would be implemented, the large number of roles would have a very negative impact on the time required to make concrete permission decisions. Introducing a single role "treating physician" to permit access to every health-record is not a solution. Not only the possible infringement of a person's individual rights through access to a (distributed) life-long health record outside a treatment case is problematic. From the sight of security this role is in fact a great danger. An unauthorized person that possesses such a role could access healthcare data of every living individual.

4 Attributable Permissions and Roles in RBAC

In this section an attributable version of RBAC will be introduced. Attribute-definitions associated to permission objects are used to define classes of permissions. The assignment of permissions with attribute-definitions to roles constrains role activation. During activation within a session conforming Attribute-Values must be given in a role-activation-context. The attribute-values are evaluated by the permissions during run-time. Figure 2 gives an overview of the connections between the conceptual items in the RBAC-Extension introduced herein. In the following sections the items of the concept are discussed.

4.1 Attribute-Definitions and Attribute-Values

An attribute-definition holds the unique name and type of the attribute within the authorization subsystem. An attribute-definition can be shared among several permissions. Typical examples for attribute-definitions in healthcare would be patient-identification and treatment-case. Permissions with attribute-definitions allow to define classes of permissions analogous to classes in object-oriented design. This approach reduces the number of permission objects to administrate and allows to introduce conforming attribute values.

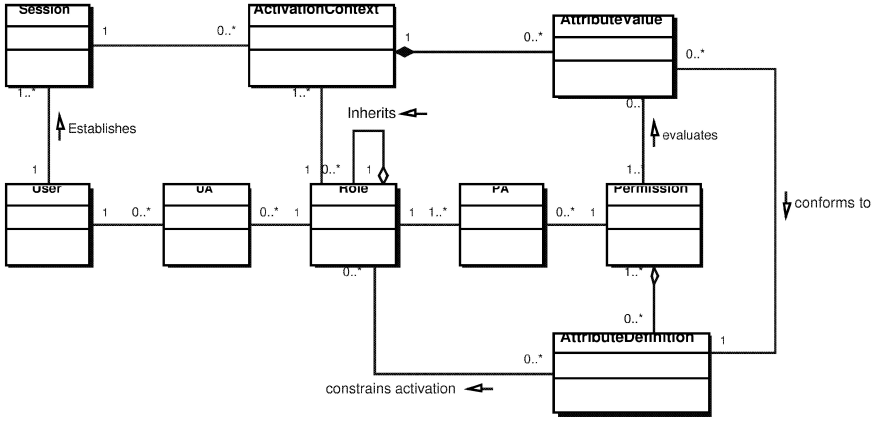


Fig. 2. Extended RBAC with Attributes

Definition 2. *Attribute-Definition*
An attribute-definition must allow at least a single attribute-value. As a consequence the set of possible attribute-values for an attribute-definition must not be empty.

An attribute-value simply consists of an association to an attribute-definition and of a value conforming to that definition. Formally, the set of all attribute-definitions and all attribute-definitions of a permission is defined as follows:

Definition 3. *Set of all attribute-definitions*
Let AD be all attribute-definitions of the authorization sub-system.

Definition 4. *Attribute-definitions of a permission*
$ADP : PRMS \rightarrow 2^{AD}$,
$\forall p \in PRMS : ADP(p) = \{ad \in AD \mid ad \text{ is part of aggregate } p\}$

4.2 Activation-Context

Activation of a role is constrained by all attribute-definitions that is implicitly attached to it through permission assignments.

Definition 5. *Direct attribute-definitions of a role*
$DADR : ROLES \rightarrow 2^{AD}$,
$\forall r \in ROLES : DADR(r) = \{ad \in AD \mid ad \in ADP(p) \wedge (p, r) \in PA\}$

Also all attribute-definitions of inherited roles are inherited and thus become attributes of a role. Now the complete set of attribute-definitions of a role can be defined:

Definition 6. *Attribute-Definitions of a Role*
$ADR : ROLES \rightarrow 2^{AD}$,
$\forall r \in ROLES : ADR(r) = \bigcup_{r' \in ROLES : r \geq r'} DADR(r')$

Upon construction this set holds all attribute-definitions from the role hierarchy following the role inheritance structure as defined by the general role hierarchy of the proposed RBAC standard.

Role activation then looks like this: for each attribute-definition of a role exactly one conforming attribute value must be given in the activation-context. Sets are possible through multiple instances of activation-contexts. For example, to allow a physician to be attending physician of patients A, B and C, an activation-context is created for each patient. Activation-contexts serve only a single role and are created and filled before that role can be activated. Activation of a role requires that the activation-context holds attribute-values for all attribute-definitions of that role. Formally the set of attribute-values in an activation context and the constraints on role activation are defined as follows:

Definition 7. *Attribute-values of an attribute-definition*
$\forall ad \in AD$ *let* $AV(ad)$ *be the (non-empty) set of all possible attribute-values of* ad.

Definition 8. *Activation contexts of a role in a session*
$\forall s \in SESSSIONS \forall r \in ROLES : AC_s(r) :=$ *Set of activation-contexts for role* r *in session* s.

Definition 9. *Attribute-values of activation-contexts for a role* r *in session* s
$\forall ac \in AC_s(r) \forall ad \in AD :$

$$value_{s,r}(ac, ad) = \begin{cases} av \in AV(ad) & \text{if } ad \in ADR(r) \text{ and } ac \text{ has value for } ad \\ \emptyset & \text{if } ad \in ADR(r), \text{ but } ac \text{ has no value for } ad \\ \emptyset & \text{otherwise} \end{cases}$$

Definition 10. *Activation constraint on roles*
Let ac *be an activation-context for role* r *The role* r *can be activated in a session* s *if and only if:*
$\exists ac \in AC_s(r) \forall ad \in ADR(r) : value_{s,r}(ac, ad) \neq \emptyset$

A role can only be activated within activation-contexts that fulfill the last definition.

4.3 Restriction of Attribute-Values

In most cases the security policy requires that the possible attribute-values shall be restricted. The user-assignment object can hold a list of allowed attribute-values for each attribute-definition of a role. The list may also be inverted to define only forbidden ones, or a combination of both lists can be used. For example, the role "emergency-doctor" would not restrict access to health-records of patients, while the user-assignment object of the role "attending physician" might hold only the IDs of the patients which the physician is currently treating.

The restriction on attribute-values using the user-assignment object is easy and still very powerful. Sophisticated implementations might use X509 attribute certificates [4] to define the allowed value-set. An approach to use such certificates for RBAC is already existing in another project that uses attribute certificates to administrate RBAC completely [3].

4.4 Role-Delegation in Attributable RBAC

The delegation of roles from one user to another requires two steps. First, a user-assignment must be created. Second, the according attribute-value-lists must be copied from the delegating user's user-assignment object to the user-assignment object of the delegee.

4.5 Checking Permissions in Attributable RBAC

To check permissions within a specific session the authorization engine uses the authorization-context(s) of activated roles. The algorithm is as easy as follows:

1. Retrieve all activation-contexts from the session that belong to activated roles. The set shall be $ActivatedAC_s$. continue with
2. For each activation-context $ac \in ActivatedAC_s$ do
   a) Get role from ac. Let r be the role.
   b) Get assigned permissions for role r. Let PA the set of assigned permissions.
   c) For each $p \in PA$ do
      i. Get all attribute-definitions from p. Let D be this set.
      ii. For all $d \in D$ retrieve the value from ac building a set V.
      iii. use set V to evaluate the permission upon the information space

The difference between standard RBAC permission checks and the algorithm above is that one has to retrieve the values from the activation context before the permission can be evaluated upon the information space. But the values can easily be cached at the permission objects after the first time. Compared to regular permission objects only one indirection is still necessary, then. This is because several activation-contexts can share a permission object with different values assigned. Because the computational cost of this is constant (pointer arithmetic) this indirection can be neglected in practice.

5 Example Reducing Role- and Permission-Objects

Let us consider a small part of a role hierarchy in an hospital that is conforming to the least-privilege-principle and uses static roles and permissions. That means roles and permission without attribute-definitions, as introduced herein. In specific the least-privilege-principle implies that there would be roles and permissions for each single patient. Figure 3 shows this situation.
With the attributable RBAC approach introduced in this paper the roles and permissions reduce to only two instances as shown in Fig. 4. The benefit for

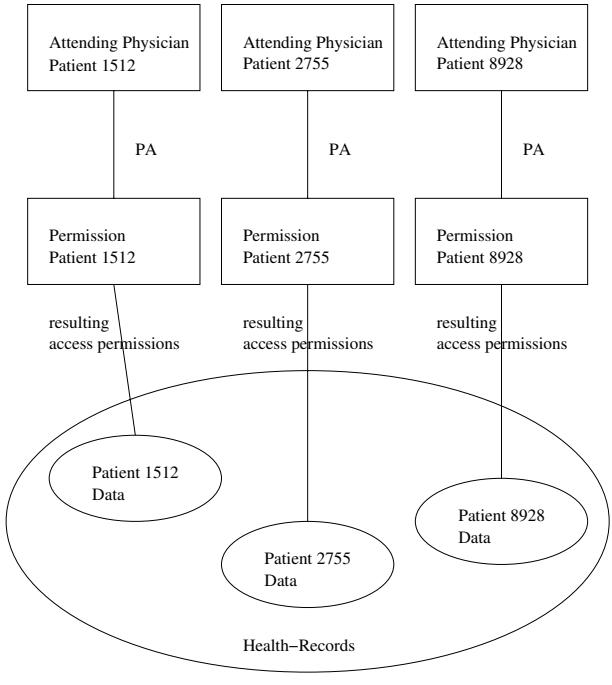


Fig. 3. Extended RBAC with Attributes

security administration becomes clear if one thinks about a large hospital with thousands of health-records for patients. If for example a revision of the hospital's security-policy would require to change the permission objects of all patients then either an administrative batch job would be necessary to adapt the permission objects of all patients or it would take a very long time to update the authorization rules. With attributable RBAC such changes would only affect a small number of objects. In the example case only one object compared to thousands.

6 Summary and Outlook

The extension of RBAC with attributes on permissions and roles is very effective. In large-scale systems the usually extremely high number of permissions and roles can be reduced to a much smaller amount which is easier to administrate and thus less error-prone.

Another advantage of the attributable RBAC is that it is ready for working together with attribute certificates. The newest health professional cards in healthcare are already smartcards with X.509 certificates and asymmetric key pair. This will allow physicians to generate valid attribute certificates holding attribute-values according to attributable RBAC as described here. Many tasks can easily be done then. The implementation of referals, second opinions, de-

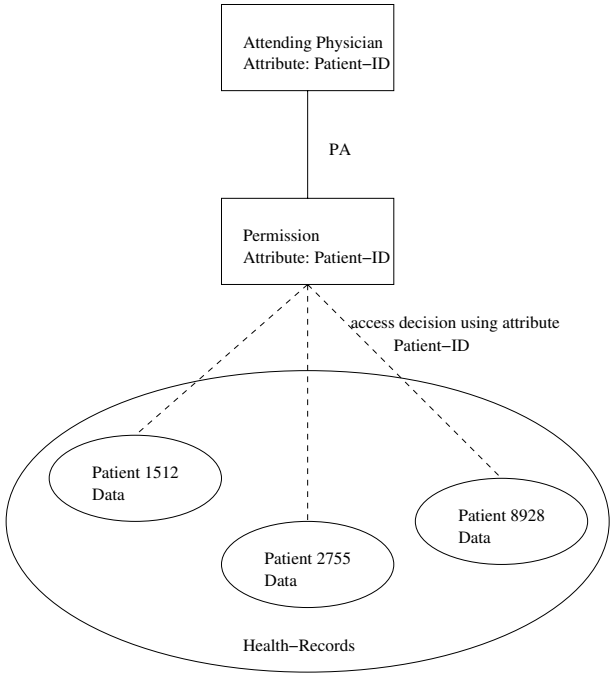


Fig. 4. Reducing role- and permission-objects with attributable RBAC

legation of cases, etc. can all be done with the same technical approach using secure attribute-certificates without the need to administrate the authorization rules by hand.

References

1. David F. Ferraiolo and Ravi Sandhu and Serban Gavrila and D. Richard Kuhn and Ramaswamy Chandramouli: Proposed NIST Standard for Role-Based Access Control. ACM Transactions on Information and System Security, Vol. 4 (2001) 224–274
2. W. Essmayr and E. Kapsammer and R.R. Wagner and A.M. Tjoa: Using Role-Templates for Handling Recurring Role Structures. 12th IFIP WG 11.3 Conference on Database Security, Chalkidiki, Greece, July 15-17, 1998
3. D.W.Chadwick, O.Otenko: The PERMIS X.509 Role Based Privilege Management Infrastructure ISI, University of Salford, Salford, M5 4WT
4. Farrell, S., and R. Housley: An Internet Attribute Certificate Profile for Authorization, RFC 3281, IETF, April 2002.

Aspects of a Massively Distributed Stable Component Space

Klaus Schmaranz[1] and Dirk Schwartmann[2]

[1] Institute for Information Systems and Computer Media (IICM), Graz, Austria
klaus@iicm.edu
[2] Institute of Aerospace Medicine, Cologne, Germany
dirk.schwartmann@dlr.de

Abstract. Modern healthcare requires massively distributed virtual electronic patient records. Recent component-based approaches are the best available solutions for the development of such systems. However, all of today's solutions have one very important problem: they do not support consistent and robust object-addressing for the dynamic case, although this is one of the most crucial features in massively distributed environments. The goal that has to be achieved is to replace physical addresses by stable and robust logical handles, managed by a distributed lookup service. Naive implementations suffer from very severe scalability problems that have to be overcome without violating robustness requirements.
DOLSA (=**D**istributed **O**bject **L**ookup **S**ervice **A**lgorithm) is designed to provide a lookup-based object addressing mechanism that is robust against any kind of object-, server- and network-dynamics. The algorithm is fully scalable in respect to the number of managed mappings, the number of lookup-requests, the number of distributed servers and the number (and frequency) of object movement operations.

Keywords: Robust Globally Unique Handles, Distributed Object System, Distributed Component System, Distributed Lookup Service, Dinopolis, MTP

1 Introduction

One of the main problems in today's heterogeneous and massively distributed information spaces is found in the way objects are addressed. As has been stated in [5], all addressing mechanisms that are in use today, share the same problem: most algorithms are not robust against object-movement. The few implementations that deal with this problem (see e.g. [7]) do not scale well, considering the number of objects, users and especially considering the number of move-operations. During development of the *Dinopolis* distributed componentware framework (see also [5] and [3]) it became clear that consistency and robustness of object-addressing are the most crucial features in a massively distributed system. The goal that had to be achieved was the development of an algorithm

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1156–1164, 2004.
© Springer-Verlag Berlin Heidelberg 2004

providing stable, *globally unique handles* (short *GUH*s). The key features of such *GUH*s can be summarized as follows:

- One *GUH* always refers to one and the same object, no matter whether and how often an object changes its physical location. This also includes the case that the original location may no longer exist and one or more different systems have taken over its responsibility.
- A *GUH* that is used for one object must never be used for any other object, even if the original object is deleted. Otherwise it could happen that an object is unintentionally replaced by a different one.
- *GUH*s can be stored anywhere, e.g. on the users' harddisks when bookmarking objects. Therefore update-operations on *GUH*s by means of push-technologies are unthinkable.
- Closely and loosely synchronized replication (see [5]) of objects have to be supported by the *GUH*-resolving mechanism.

There are two main families of algorithms that are robust against object-movement:

1. Forwarding algorithms that rely on traces which are left whenever objects are moved to different locations.
2. Algorithms based on lookup-service strategies that rely on updates of the lookup-service whenever objects are moved.

It can easily be seen that pure forwarding algorithms do not scale at all for frequent object movements. Further, servers that go offline would break the consistency of any traces that objects left on them. Finally, the requirement for transparent component replication would not work with forwarding anyhow. Therefore the algorithm of choice for implementing the concept of *GUH*s has to be a lookup-service based algorithm. Considering a huge number of *GUH*s in large and highly dynamic distributed systems, a *naive* lookup-algorithm (e.g. a central lookup-service) would not scale either. Hence an algorithm for implementing an arbitrarily distributable lookup-service has to be found. Defining *GUH*s with an internal hierarchical structure like hostnames in DNS (see [4]) is not realistic too, because the resolve order would then be defined by the *GUH*s rather than by the lookup service. What is required is that the resolve-order has to be dynamic and self-organizing in a way so that administrative aspects are not neglected. For reasons of robustness, stability, scalability and availability as well as for some administrative reasons the *DOLSA* algorithm presented in this paper is a combination of the advantages of both algorithm families mentioned above. DOLSA implements a self-organizing (but influencable!), arbitrarily distributed lookup-service with robust caching and robust short-term forwarding for scalability reasons.

In the following only the final version of the resolving algorithm together with the *GUH*-structure is described. A detailed description of all the decisions that lead to this algorithm would be beyond scope of this paper as well as an in-depth discussion of some special operations. Interested readers can find these details in [6].

2 Structure of the Distributed Object Lookup Service

The structure of the distributed lookup service is straightforward:

- There exists an arbitrarily large number of *Object Lookup Servers* (short *OLS*).
- Each *OLS* is responsible for *GUH*-to-address mappings for an administrator-defined area of responsibility.
- *OLS*s do not know about each other and do not form any kind of hierarchy.

A general *handle* is a pair of IDs:

1. *OLS-ID* (=object lookup-service ID). Every *OLS* has such a unique ID.
2. *LOID* (=local object ID). Every object gets such an ID and this ID is unique within the lookup-space of one *OLS*.

As can easily be seen, this pair results in a general globally unique handle. Please note that such a simple handle is not enough for the dynamics addressed by *DOLSA*, therefore this is not yet a *GUH* in our sense. The final structure of a *GUH* is described in section 3. Although the hostname of an *OLS* would already fulfill the requirement to be a unique ID, it is not useful for the dynamic case of *OLS*-splitting in combination with object movement. Therefore the *OLS-ID* is an arbitrary, unique ID managed by a special *Server Lookup-Service* (short *SLSvc*) which maps the ID to an actual hostname on request. The ID itself does not contain any administrative structures like e.g. the resolve-order, as is the case with DNS. Whenever a lookup-server is moved, an update of the ID-to-hostname mapping is performed in the *SLSvc*. With this indirection *GUH*s become completely independent from hostnames.

For availability and scalability reasons the *SLSvc* itself is distributed across the network in a semi self-organizing manner. In order to avoid getting sidetracked here let us just assume that a scalable and reliable, distributed, self-organizing *SLSvc* is realistic.

3 Structure of *GUH*s and the *DOLSA* Base-Algorithm

The *DOLSA* algorithm is based on the following structure of *GUH*s: A *GUH* always consists of three handles, each of them being a pair $[OLS-ID, LOID]$ as already described. The three handles are:

1. *BPH* (=birthplace handle)
2. *MBPH* (=moved-birthplace handle)
3. *CH* (=current handle)

The following definitions apply to the single handles of a *GUH*:

- The *BPH* always contains a non-empty entry refering to the birthplace of an object.

- The *BPH* is defined when an object is created and it must never change during the whole lifetime of an object.
- It is guaranteed that the *BPH* can always be resolved.
- The *BPH* and only the *BPH* is taken when comparing two handles for equality.
- The *MBPH* only contains a non-empty entry if the birthplace-*OLS* of an object was taken offline and therefore one or more different *OLS*s have taken over its responsibilities. Otherwise the *MBPH* is empty.
- It is not guaranteed that the *MBPH* can always be resolved. Resolving is not possible if the so-called moved-birthplace-*OLS* was taken offline. In this case it is guaranteed that the *SLSvc* can resolve the *BPH*.
- The *CH* only contains a non-empty entry if the object was moved across *OLS*-responsibility boundaries. Otherwise the *CH* is empty.
- It is not guaranteed that the *CH* can always be resolved. Resolving is not possible if the object has moved on. In this case the *MBPH* or the *BPH* are taken for resolving, as will be described later.

With this definition of a *GUH* the *DOLSA* base-algorithm can be sketched. Two different views exist that make up the whole algorithm:

1. The view of the distributed lookup-service as a whole that has to react on different dynamic aspects to guarantee robust *GUH*-resolving.
2. The view of the requestor that holds a *GUH* in hands and wants it to be resolved.

3.1 The View of the Distributed Lookup-Service

The core-functionality of *DOLSA* can be found in the organization of the distributed lookup-service. At the moment let us assume that the starting point for the following description is an existing but empty *SLSvc*. It is further assumed that one host of the *SLSvc* is known and that the *SLSvc* is contacted through this known host.

Adding an *OLS* to the Distributed Service. If a new *OLS* is going to be added to the distributed service, the following steps have to be performed:

- The new *OLS* has to contact the *SLSvc* to register its hostname.
- The *SLSvc* calculates a unique *OLS-ID*, registers the appropriate *OLS-ID*-to-hostname mapping and returns the *OLS-ID*.

Changing an *OLS*'s Hostname. If an *OLS* changes its hostname, it contacts the *SLSvc* and sends an update request. The *SLSvc* updates the registered *OLS-ID*-to-hostname mapping. The *OLS-ID* remains untouched.

Moving an Entry to a Different *OLS*. Moving objects around inside the area of responsibility of an *OLS* is trivial. The more interesting situation occurs when an object is moved from the area of responsibility of one *OLS* to another *OLS*. In this case the following actions are necessary:

1. Triggered by the source-*OLS* the destination-*OLS* calculates a *LOID* that the entry will obtain when moved. This *LOID* is passed on to the source *OLS*.
2. The source-*OLS* marks the entry to be moved and sets an appropriate temporary forwarder.
3. As soon as the entry in the destination-*OLS* is active, the destination-*OLS* informs the birthplace (or moved-birthplace respectively) that the entry was moved. The birthplace-*OLS* (or moved-birthplace-*OLS* respectively) updates its appropriate birthplace-entry to reflect the changes.
4. As soon as the birthplace updated its entry, the destination-*OLS* notifies the source-*OLS* which in turn can drop the temporary forwarder.

In case that the source-*OLS* is the birthplace of the object, a shortcut of the algorithm without network-traffic in step 4 can be implemented.

Dealing with an *OLS* Going Offline. If an *OLS* goes offline and its lookup tables therefore have to be moved to one or more existing *OLS*s the following actions have to take place:

1. The source-*OLS* that will go offline contacts the *SLSvc* and reports the start of the move operation as well as all destination-*OLS*s that will take over its responsibilities.
2. The tables are moved to the appropriate destination-*OLS*s.
3. The source-*OLS* reports the end of the move operation to the *SLSvc* and can then go offline.

If the *SLSvc* is contacted to resolve an *OLS* that is just in the middle of performing such a move it returns the original *OLS*'s hostname together with the destination *OLS*s' hostnames, *OLS-ID*s and a flag that alerts the requestor that a move operation is just taking place.

If the *SLSvc* is contacted to resolve an *OLS* after moving of the whole content has been finished, only the destination *OLS*s' hostnames and *OLS-ID*s are returned.

The responsibility of each of the destination *OLS*s depends on the type of the entry that it receives:

Standard-entry: If a standard-entry (i.e. not a birthplace-entry, see below) is moved to the destination, it is treated as already described in section 3.1.

Birthplace-entry: If a birthplace-entry is moved to the destination (i.e. the corresponding object had its birthplace in the area of responsibility of the source *OLS*) the *BPH* of the entry is stored in the destination *OLS*'s birthplace mapping table. Further, a standard move-operation, as described in section 3.1, takes place.

In case that an *OLS* is already the moved-birthplace for one or more other *OLS*s the same algorithm applies. So-called *moved-birthplace-entries* are treated the same as birthplace-entries. For moved-birthplace-*OLS*s going offline the *SLSvc* has the responsibility to compact the resulting move-chains to a one level uncertainty-indirection.

Splitting up an *OLS*. If for whatever reason (e.g. massive request-overload) an *OLS* has to be split up without going offline itself, this is considered the same case as moving many entries from one *OLS* to another. Therefore no *MBPH* comes into play, because the birthplace server is still online.

This behaviour is the reason why the *GUH* contains the *CH*, otherwise the load on the original-*OLS* could not be reduced by splitting it up. In this situation network traffic could not be gotten under control any more and the algorithm would not scale. By introducing the *CH* the original-*OLS* is only contacted once per *GUH* and then the traffic calms down automatically.

3.2 The View of the Requestor

From the point of view of the requestor two different instances exist in principle which have to be contacted for resolving: *OLS* and *SLSvc*. For simplicity reasons let us assume for the moment that the *SLSvc* is a server with a known hostname. In principle resolving a *GUH* is a two-step process:

1. For a given *GUH* ask the *SLSvc* for the hostname of the *OLS* which knows the *GUH*-to-address mapping. In most cases the answer will contain the hostname of exactly one *OLS*.
2. Contact the obtained *OLS* and ask for resolving of the *GUH*.

With dynamic aspects taken into account resolving a *GUH* looks as follows:

1. The Requestor contacts the *SLSvc* and passes the *GUH* on to it.
2. The *SLSvc* decides on the content of the *GUH*, which *OLS-ID* is relevant for resolving:
   - ***CH* is non-empty:** In this case the *SLSvc* looks up its internal tables, whether the *OLS-ID* in the *CH* points to an *OLS* that is registered as being online. If yes, the *SLSvc* goes on with step 3. Otherwise the decision process goes on with the next step described below.
   - ***MBPH* is non-empty:** In this case the *SLSvc* looks up its internal tables, whether the *OLS-ID* in the *MBPH* points to an *OLS* that is registered as being online. If yes, the *SLSvc* goes on with step 4. Otherwise the decision process goes on with the next step described below.
   - ***BPH* resolving:** If the *GUH* is a valid handle at all, the *BPH* contains all information necessary for resolving. In this case the *SLSvc* looks up its internal tables which *OLS-ID* the *BPH* refers to. Three cases are possible:

a) The birthplace-*OLS* is online and therefore the *OLS-ID* can directly be resolved to a hostname. In this case the *SLSvc* goes on with step 5.
b) The birthplace-*OLS* is no longer online and its content is now handled by one or more moved-birthplace-*OLS*s. In this case the *SLSvc* goes on with step 6.
c) The birthplace-*OLS* is just in progress of moving its content to one or more moved-birthplace-*OLS*s and then going offline. In this case the *SLSvc* goes on with step 7.

3. The hostname of the current-*OLS* is returned to the requestor together with a flag that *CH*-resolving has resulted in the given hostname.
The following tasks have to be performed then:
   a) The requestor contacts the given *OLS* and passes on the *GUH* with a resolve request for the *CH*.
   b) If the *OLS* can resolve the *CH*, the mapped address is returned and resolving is finished.
   c) If the *OLS* cannot resolve the *CH* (=entry was moved to a different *OLS*), it returns an appropriate error. The requestor invalidates the *CH*-part in the *GUH* and continues with step 1.
4. The hostname of the moved-birthplace-*OLS* is returned to the requestor together with a flag that *MBPH*-resolving has resulted in the given hostname.
The following tasks have to be performed then:
   a) The requestor contacts the given *OLS* and passes on the *GUH* with a resolve request for the *MBPH*.
   b) If the *OLS* is still the home of the entry it returns a *CH* (the *CH* is usually different from the *MBPH* after a birthplace-move!) together with the mapped address. The requestor updates the *CH*-part in the *GUH* and resolving is finished.
   c) If the requested entry already moved on to a different *OLS*, only a *CH* is returned. In this case the requestor updates the *CH*-part in the *GUH* and continues with step 1.
5. The hostname of the birthplace-*OLS* is returned to the requestor together with a flag that *BPH*-resolving has resulted in the given hostname.
The following tasks have to be performed then:
   a) The requestor contacts the given *OLS* and passes on the *GUH* with a resolve request for the *BPH*.
   b) If the *OLS* is still the home of the entry it returns the mapped address and resolving is finished.
   c) If the requested entry already moved on to a different *OLS*, only a *CH* is returned. In this case the requestor updates the *CH*-part in the *GUH* and continues with step 1.
6. The list of *MBPH*s together with the corresponding hostnames which could be the moved-birthplace-*OLS*s of the entry is returned to the requestor. A flag is set that there is uncertainty about the current *OLS* which manages the given *GUH*.

The requestor then follows in principle the same procedure that was described in step 4 with one difference: According to the uncertainty-approach the moved-birthplace-*OLS*s are contacted one by one, until the handle can be resolved. The *MBPH*-part of the *GUH* is updated with the *MBPH* of the *OLS* that is finally able to resolve the request. Please note that uncertainty-resolving has to be done only once because then the *MBPH* is fixed. Therefore no scalability problems occur.

7. The hostname of the birthplace-*OLS* is returned to the requestor together with a list of *MBPH*s and corresponding hostnames which could be the moved-birthplace-*OLS*s of the entry. An according flag is set to signal that moving entries are resolved.
   In this case the requestor performs in principle the same procedure that was described in step 6, with the only difference that the birthplace-*OLS* (i.e. the original host) is contacted first and asked for resolving. The uncertainty-approach only has to be taken for entries that already moved on before.

For sketching the base of *DOLSA* it was assumed that the *SLSvc* "exists and is somehow distributed". However, clever distribution of the *SLSvc* is one of the most crucial points in order to make *DOLSA* useful in practice. Nonetheless a detailed description of all distribution aspects would be beyond scope of this paper. Interested readers can find an in-depth discussion in [6].

4 Conclusion

The *DOLSA* base-algorithm, as presented in this paper, is the first algorithm meeting the requirements of massively distributed and highly dynamic component systems such as *Dinopolis* (see [3], [5]). It provides globally unique object handles that are robust against object movement. All critical aspects of scalability are taken into account and *DOLSA* scales well in respect to the number of object lookup servers distributed across the network, the number of objects managed by the object lookup servers, objects which change their locations frequently and the number of lookup-operations.

Attentive readers will have noticed that race-conditions during object movement can occur. The obvious candidates for such problems are *resolve-to-access delays*, *lookup-server-update delays* and *fast-moving objects*. To deal with these race-conditions, special algorithms for moving objects and for catching up on fast-moving objects were developed. Again, interested readers can find a detailed discussion of these aspects in [6].

Although security aspects were not especially mentioned in this discussion, because they would have been way beyond the scope of this paper, these aspects are extremely important within the area of medical applications. All security considerations that could affect *DOLSA* were taken into account:

- *GUH*s are designed in a way so that they allow electronic signatures for interrelations, even if objects move.

– *DOLSA* is designed not to interfere with authentication, authorization and encryption mechanisms.

One very important point remains to be mentioned: *DOLSA* does not dictate, which kind of mappings are stored. *GUH*s can be mapped to "anything". Therefore the algorithm is not only suited for distributed component systems, it is also possible to manage today's standard hyperlinks with *DOLSA*. *GUH*s also support a textual representation that is compatible to today's URLs. Thus it would be easy to replace today's inconsistent URLs by consistent and robust *GUH*s. The well-known browsers would only have to implement one simple additional lookup operation before loading a page and the well-known servers would only have to report restructuring operations of their areas to an *OLS*.

References

1. Andrews K., Kappe F., Maurer H., Schmaranz K.: *On Second Generation Hypermedia Systems*, Proceedings ED-MEDIA 95, Graz (1995), 75–80.
2. Aly F., Bethke K., Bartels E., Novotny J., Padeken D., Schmaranz K., Schwartmann D., Wilke D., Wirtz M.: *Medical Intranets for Telemedicine Services: Concepts and Solutions*, Proceedings G7 Meeting "The Impact of Telemedicine on Health Care Management", Regensburg (1998), available online at `http://www.uni-regensburg.de/Fakultaeten/Medizin/Uch/g7/program/mon.htm`.
3. Dallermassl C., Haub H., Maurer H., Schmaranz K., Zambelli P.: *Dinopolis - A Leading Edge Application Framework for the Internet and Intranets*, Proceedings WebNet 2000, San Antonio, TX (2000), 111–116.
4. Mockapetris P., Dunlap K. J.: *Development of the domain name system*, Proceedings ACM SIGCOMM 1988, Stanford, CA (1988), 123–133.
5. Schmaranz K.: *On Second Generation Distributed Component Systems*, J.UCS Vol.8, No.1, 97–116 (2002).
6. Schmaranz K.: *Dinopolis – A Massively Distributable Componentware System*, available online at `http://www.dinopolis.org/documentation/publications/habilthesis2002_dinopolis-a_massively_distributable_componentware_system.pdf`
7. *ObjectSpace's Home page*, available online at `http://www.objectspace.com`.

Demonstrating Wireless IPv6 Access to a Federated Health Record Server

D. Kalra [1], D. Ingram[1], A. Austin [1], V. Griffith [1], D. Lloyd [1], D. Patterson [2], P. Kirstein [3], P. Conversin [4], and W. Fritsche [5]

[1] University College London, CHIME, Holborn Union Building, Highgate Hill, London, N19 3UA
[2] The Whittington Hospital, Highgate Hill, London N19 5NF
[3] University College London, Department of Computer Science
[4] 6WIND, Paris, France
[5] IABG, Germany
d.kalra@chime.ucl.ac.uk

Abstract. This paper describes the practical implementation of a federated health record server based on a generic and comprehensive architecture, deployed in a live clinical setting and accessed from wireless and IPv6 network test-beds. The authors, working at the Centre for Health Informatics and Multiprofessional Education (University College London), have built up over twelve years of experience within Europe on the requirements and information models that are needed to underpin comprehensive multi-professional electronic health records. The UCL federated health record server is running in the Department of Cardiovascular Medicine at the Whittington Hospital in north London. Through the EU 6WINIT project this infrastructure has been modified to enable secure wireless access via IPv6.

1 Research and Development in Electronic Health Records

Realising the electronic health record has been at the heart of the EU Health Telematics Framework Programmes. Considerable research has been undertaken to explore the user requirements for adopting EHRs (e.g. published by the GEHR [1-5] and EHCR Support Action projects [6]), resulting in architecture formalisms to capture healthcare data comprehensively and in a manner which is medico-legally rigorous and preserves the original clinical meaning, (e.g. GEHR and the CEN standards ENV12265 [7] and ENV13606 [8]). These results recognise that personal health data is often very sensitive and always to be regarded as confidential.

Other research has identified the requirements to support the communication of EHRs within federated communities of healthcare enterprises to support shared patient care across sites (the Synapses project [9-11]) and middleware architectures to integrate across R&D projects (SynEx [12]). Outputs from these projects have recently been consolidated within a PhD thesis based on the UCL demonstrator [13].

More recently parallel threads of research in Europe and Australia have been united through the openEHR Foundation. This Foundation is an independent, not-for-profit organisation and community, facilitating the creation and sharing of health

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1165–1171, 2004.
© Springer-Verlag Berlin Heidelberg 2004

records by consumers and clinicians via open-source, standards-based implementations (see www.openehr.org).

A new European (CEN) standard is presently being developed to define the way in which clinical and EHR systems can interoperate to exchange patient records in a systematic and medico-legally sound manner. This work, drawing on thirteen years of European R&D, is being undertaken in close working relationship with openEHR and with HL7 [14].

UCL Information Architecture

The Synapses approach to distributed healthcare records applied the methodology of database federation to a standard and comprehensive schema, the Federated Healthcare Record (FHCR) information architecture, mediated and managed through a set of middleware services [11]. This has required a scalable model for representing any conceivable health record information, created by any profession, speciality or service, whilst recognising that the clinical data sets, templates etc. will be diverse, complex and will change frequently as clinical practice and medical knowledge advance.

A dual model approach has been adopted, which distinguishes a Reference Model, used to represent the generic properties of health record information, and Archetypes (conforming to an Archetype Model), which are meta-data used to represent the specific characteristics of the various kinds of clinical data that might need to be represented to meet the requirements of each particular profession, speciality or service. The Reference Model represents the global characteristics of health record entries, how they are aggregated, and the context information required to meet ethical, legal and provenance requirements. Archetypes each define (and effectively constrain) legal combinations of the building-block classes defined in the Reference Model for particular clinical domains and contexts by specifying particular record component names, data-types and prescribed value ranges, and values.

2 Federated Health Record Middleware Services

The federated health record (FHR) components are delivered via a set of middleware services that enable a requesting service (e.g. a client clinical application, or another middleware service such as a decision support agent) to access electronic health record information from a diversity of repository servers (feeder systems). The feeder systems may be on-site at an institution or connected remotely through telecommunications services. These actions take place in accordance with the user's role-based privilege and are registered in an access audit trail. The London demonstrator is utilising the following UCL-developed component services:

Federated Health Record services: a scalable run-time FHR environment supporting distributed access to record components from new and legacy feeder systems.

Archetype Object Dictionary Client and services: a means of facilitating feeder system sign-up, of mapping each feeder system schema and of navigating the federation environment.

Persons Look-up services: storing a core demographic database to search for and authenticate staff users of the system and to anchor patient identification and connection to the patient's federated healthcare record.

Expert Advisory (Decision Support) services: for anticoagulation management, to calculate the patient's next treatment regimen and next monitoring interval. This service is provided through specific agents called from a dedicated client and these return data to this client.

Web-based applications: to provide end-user clinical views and functions.

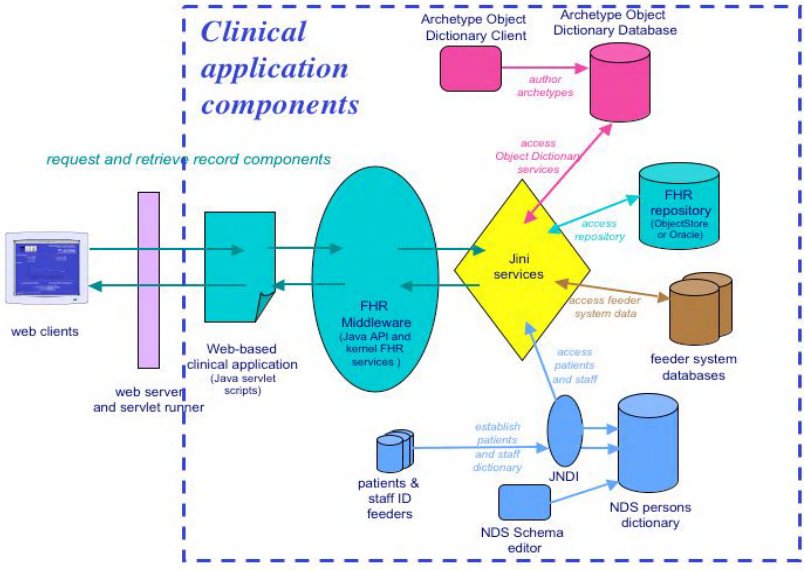


Fig. 1. Core FHR components handling the request for and retrieval of patient records

All of the main components are written in Java, and are deployed within a middleware environment managed through Novell Directory Services and JINI, an open standard service-integration technology. This overall approach will allow the ongoing development of flexible and portable applications with high-level graphical user interfaces to be made where such applications can inter-operate across diverse architectures and infrastructures. New web-based clinical applications have been written using Java servlets to provide end user access to the patient records held within the FHR server. Further details of this information architecture and implementation may be found in [13, 15].

3 The 6WINIT Project

The IPv6 Wireless internet INITiative (6WINIT) project (2000-2002) was a European IST Framework V initiative involving telecoms companies, equipment manufacturers, solutions/software providers, universities and hospitals. Its objectives were to validate the introduction of the new mobile wireless Internet in Europe, based on a

combination of the new Internet Protocol version 6 (IPv6) and the new wireless protocols (GPRS and UMTS). The UCL north London demonstrator was one of three 6WINIT clinical sites. 6WIND and IABG provided most of the wireless and IPv6 components used in this demonstrator. These were installed and configured locally by the Computer Science department of UCL.

IPv6 Transition Mechanisms (tunnelling). Native IPv6 services from the main application server were tunnelled via IPv4 from CHIME (north London) to Computer Science (central London) using UCL's main University Intranet. The 6WIND IP Edge Device implements the v4-v6 migration mechanisms over a dual stack entirely developed by 6WIND. An Edge device was configured and installed at CHIME, and another at UCL Computer Science.

Security Mechanisms. A virtual private network (VPN) was required to provide an authenticated and encrypted channel of communication between the clinical applications server and the end-user specifically supporting the use of mobile devices connected via wireless LAN, GPRS or UMTS. The IABG Road Warrior was installed and configured to provide a VPN extension to mobile users based on IPsec, using a Linux client for the mobile user and the 6WINDGate as an IPSec gateway.

Certificate-based key management. The CHIME clinical applications were adapted also to communicate with a certificate server at UCL Computer Science (developed by University of Murcia).

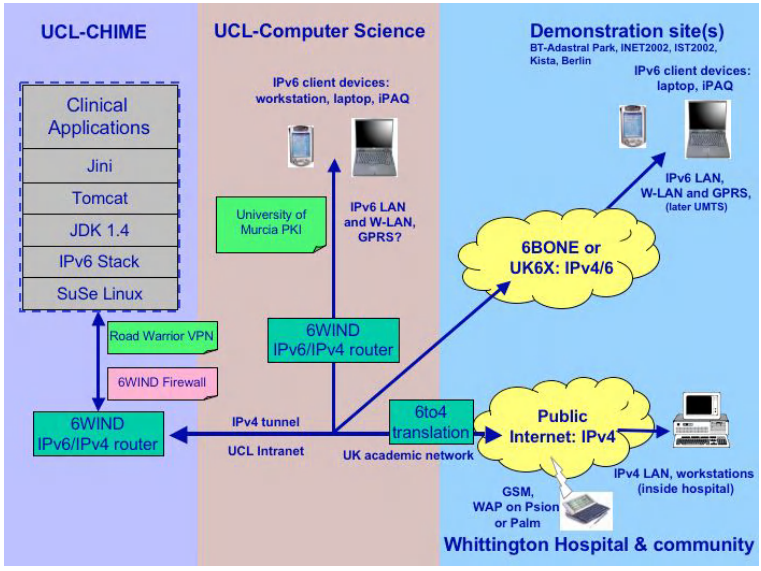


Fig. 2. Network architecture of the London demonstrator site

Figure 2 above shows the principal clinical application (health record) services, located at CHIME, being delivered via an IPv6 stack infrastructure, sent to UCL Computer Science and routed forward to the public Internet and to new IPv6 networks (6BONE or UK6X). The communications pathway involves the use of some IPv4 networks, such as the UCL Intranet connecting CHIME in north London to Computer Science in central London, and the public Internet. IPv6 access to the CHIME FHR

services and clinical applications has been validated from central London (WLAN), Ipswich (WLAN and GPRS, provided by BT), Berlin (GPRS, provided by DTAG) and Kista (UMTS, provided by Ericsson).

4 North London Demonstrator Setting

The north London demonstrator at which the UCL FHR services and clinical applications are being evaluated comprises: the Department of Cardiovascular Medicine at the Whittington hospital, 2-4 community-based cardiology clinics, Several GP practices and community pharmacies in north London and patients managing their own healthcare at home. The north London demonstrator vision is to deliver the seamless shared care of patients with cardiovascular illness, in a managed care environment. The clinical applications so far in use are summarised below.

Anticoagulant Application

This application provides a set of HTML web clients to enable the management of anticoagulation therapy by clinical staff (or patients). The system incorporates drug dosing decision support and recommends monitoring intervals between blood tests. It has been used daily by staff at the Whittington Hospital since summer 2000, running clinics with up to 120 patients per day. It is also accessed from outside the hospital by two community pharmacists, and it is hoped to include other pharmacists, GPs and patients as users in the near future.

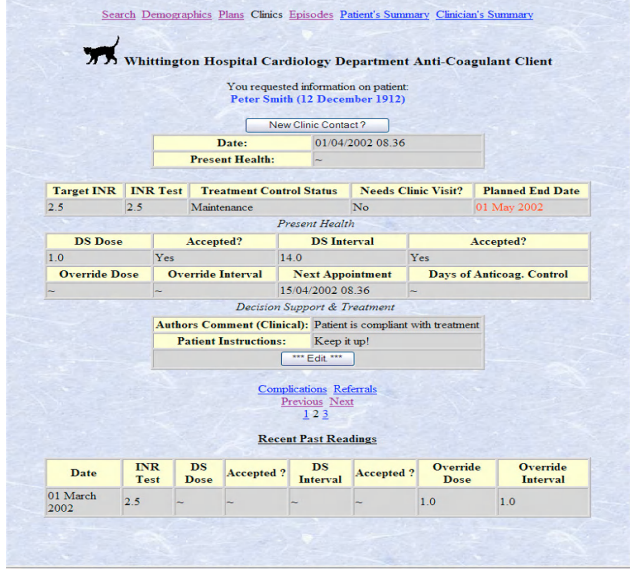


Fig. 3. Anticoagulant client - viewing a clinic contact

Rapid Access Chest Pain Clinic (RACPC)

This application provides clinicians working both inside and outside the Whittington hospital with access to the record of patients having non-acute or sub-acute chest pain (i.e. possible heart disease) symptoms. The primary clinical application has been hosted on the same FHR server as the anticoagulant system, and shares the same core middleware services. The application has been running live for nearly one year, and is being used by nurse specialists providing a Rapid Access Chest Pain Clinic service and a heart failure assessment service within the hospital and at patients' homes.

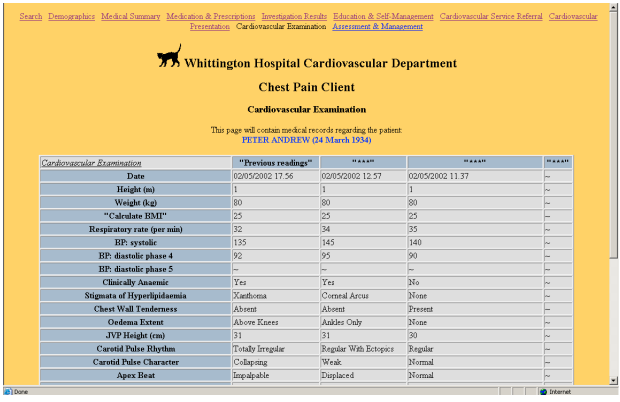


Fig. 4. RACPC client – viewing cumulative physical examination data

5 Conference and Technical Review IPv6 Demonstrations

Four proof of concept demonstrations of the 6WINIT infrastructure and clinical applications were made during the project:

- 2 technical reviews including Ericsson's IPv6 test bed in Stockholm;
- 2 international conferences: INET 2002 (Washington DC) and IST 2002 conference in Copenhagen.

These demonstrations have been documented in project deliverables and reports (available from http://www.cs.ucl.ac.uk/research/6winit). The IST conference in Copenhagen (November 2002) provided a valuable end-to-end validation of the reliability of connection, the network performance and security. Although the IPv6 wireless access point was serving several simultaneous IPv6 Cluster demonstrations, each web screen downloaded from London within about half a second. There were no timeout problems or periods of inaccessibility, and the scaling, resolution and performance of the Compaq iPAQ (adapted for IPv6) was excellent.

Ideally the UCL team had hoped to be able to deploy the validated 6WINIT infrastructure in Whittington Hospital. This was not possible, due to the requirements for the hospital to comply with national NHS and local vendor policies. Nevertheless, the Department of Cardiovascular Medicine managed to deploy several elements of the infrastructure on to internal departmental network. In practice this has meant an

IPv4 clone of the applications, web server, JINI and FHR services. It has not been possible to deploy the 6WIND edge device or the Road Warrior within the department. However, the view of the authors is that the demonstrations given, for example at IST 2002, are a realistic vision of what is both clinically valuable and technically acceptable for future distributed health care.

References

1. Ingram D., Southgate L., Kalra D., Griffith S., Heard S. and others. The GEHR Requirements for Clinical Comprehensiveness. European Commission, Brussels; 1992; The Good European Health Record Project: Deliverable 4. 144 pages.
2. Ingram D., Hap B., Lloyd D., Grubb P. and others. The GEHR Requirements for Portability. European Commission, Brussels; 1992; The Good European Health Record Project: Deliverable 5.
3. Ingram D., Lloyd D., Baille O., Grubb P. and others. The GEHR Requirements for Communication Capacity. European Commission, Brussels; 1992; The Good European Health Record Project: Deliverable 6.
4. Ingram D., Murphy J., Griffith S., Machado H. and others. GEHR Educational Requirements. European Commission, Brussels; 1993; The Good European Health Record Project: Deliverable 9.
5. Ingram D, Southgate L, Heard S, Doyle L., Kalra D. and others. The GEHR Requirements for Ethical and Legal Acceptability. European Commission, Brussels; 1993; The Good European Health Record Project: Deliverable 8. 68 pages.
6. Dixon R., Grubb P.A., Lloyd D., and Kalra D. Consolidated List of Requirements. EHCR Support Action Deliverable 1.4. European Commission DGXIII, Brussels; May 2001. 59pp. Available from http://www.chime.ucl.ac.uk/HealthI/EHCR-SupA/del1-4v1_3.PDF.
7. Hurlen P., Editor, Project Team 1-011. ENV 12265: Electronic Healthcare Record Architecture. CEN TC/251, 1995.
8. Kay S. and Marley T., Editors, Project Team 1-026. ENV 13606: EHCR Communications: Part 1 Electronic Healthcare Record Architecture. CEN TC/251; 1999.
9. Kalra D., Editor. The Synapses User Requirements and Functional Specification (Part A). EU Telematics Application Programme, Brussels; 1996; The Synapses Project: Deliverable USER 1.1.1a. 6 chapters, 176 pages.
10. Grimson W. and Groth T., Editors. The Synapses User Requirements and Functional Specification (Part B). EU Telematics Application Programme, Brussels; 1996; The Synapses Project: Deliverable USER 1.1.1b.
11. Grimson J., Grimson W., Berry D., Stephens G., Felton E., Kalra D., Toussaint P., and Weier O.W. A CORBA-based integration of distributed electronic healthcare records using the synapses approach. IEEE Trans Inf Technol Biomed. Sep 1998; 2(3):124-38.
12. Sottile P.A., Ferrara F.M., Grimson W., Kalra D., and Scherrer J.R. The holistic healthcare information system. Toward an Electronic Health Record Europe '99. Nov 1999; 259-266.
13. Kalra D. Clinical Foundations and Information Architecture for the Implementation of a Federated Health Record Service. PhD Thesis. University of London, 2003.
14. Kalra D, Freriks G, Lloyd D, Klein G, Beale T, Heard S, Schloeffel P, Maskens A, Mennerat F, Ingram D. Towards a revised CEN standard for Electronic Health Record Communication, in Brown P (ed) procs Mobile-Health Europe 2002. Medical Records Institute. April 2002
15. Kalra D. Lloyd D. Austin T. O Connor A. Patterson D. Ingram D. Information Architecture for a Federated Health Record Server, in Mennerat F (ed) Electronic Health Records and Communication for Better Health Care Proceedings of EuroRec '01. Studies in Health Technology and Informatics 2002 Issue 87: 47-71. IOS Press. ISSN: 0926-9630

Collaborative Teleradiology

Krzysztof Zieliński[1], Jacek Cała[1], Łukasz Czekierda[1], Sławomir Zieliński[1]

[1] Dept. of Computer Science, AGH – University of Science and Technology,
al. A. Mickiewicza 30, 30-059 Kraków, Poland
{kz,cala,luke,slawek}@cs.agh.edu.pl
http://www.cs.agh.edu.pl

Abstract. The paper introduces a notion of collaborative teleradiology as a novel organisation of information exchange during a remote radiological consultation. A general interaction pattern between medical experts taking part in a collaborative teleradiology session has been defined. This session has been compared to interaction typically performed in WWW centric teleradiology systems. Next technical implication of a proposed information exchange paradigm has been elaborated. Finally two prototype tools TeleNegatoscope and TeleDICOM that supporting of the collaborative teleradiology working environment have been described in more details.

1 Introduction

Most of the existing teleradiology [1] and PACS systems [2] nowadays are oriented on asynchronous communication between a radiology expert and other medical staff. The applied communication pattern assumes communication via a central database where results of radiology investigations are stored. These radiology data could be accessed either locally or remotely via computer network by radiology experts in order to make diagnoses which are stored back into the database and can be further accessible for authorized medical personnel. This co-operation pattern has been recently enhanced with WWW technology application that provides standardised uniform platform for remote access from any computer attached to a computer network to the data stored in the database. The application of WWW technology does not remove two important characteristic features of the discussed pattern i.e. lack of real-time interactivity between a radiology expert and other doctors, and sensitivity to computer network quality of service (QoS) parameters.

Progress in distributed computer system architectures and software technologies [3,4,5] offers possibility of developing new communication patterns suitable for teleradiology applications which do not suffer from the pointed out deficiencies. These patterns make it possible to achieve on-line communication between experts in a medical reference centre and doctors in a peripheral hospital and collaborative synchronous work on the same radiology data.

This paper introduces a notion of *collaborative teleradiology*, which assumes that a group of medical experts located in different clinics or hospitals work synchronously on the same radiology data sharing access to them and exchanging comments by using computer network. In that sense it brings a teleworking environment in medicine on higher, more advanced, level. Collaborative teleradiology addresses not

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1172–1179, 2004.
© Springer-Verlag Berlin Heidelberg 2004

only telediagnosis but also teleconsultancy issues by providing tools and mechanisms enabling on-line exchange of opinions. The key challenge from the technical point of view is how to implement a collaborative distributed system efficiently in context of constrained communication and computational resources along with high demands of medical applications.

The structure of the paper is as follows. First, the collaborative teleradiology system architecture requirements are presented and discussed in details. Features of this system are then compared to the web centric approach. Particular attention has been put to network QoS requirements, which result from different large data set sharing models. Next, synchronization techniques, which are essential to any collaborative distributed system, are presented. Finally, an example collaborative teleradiology application called TeleDICOM is presented. The system is under development and will be introduced to a medical practice in Krakow Center of Telemedicine. The paper ends with conclusions.

2 Web-Centric Approach vs. Collaborative Systems

Collaborative teleradiology could be understood as a new radiology medical data-sharing pattern for telemedicine. To understand better what collaborative work means is it necessary to refer to the definition of collaborative computing systems [6] and compare it to a web centric approach.

In the context of teleradiology, collaborative computing means application sharing [7] that allows:

- taking an existing single user application and making it shareable,
- broadcasting graphics, mouse movements, and edits to all participants,
- usage of telepointer and “Master” pointer,
- integration with audio, video and text chat connections.

A single-user application that could be made shareable over a network could be, for instance, a DICOM images viewer commonly used in any digital radiology system. This simple concept requires construction of highly interactive, collaborative environment, which should provide abstraction of a session (i.e., a group of objects associated with some common communication pattern), and support full-duplex multi-point communication among arbitrary number of connected application entities.

The primary functionality provided by the collaborative teleradiology application is to send digital images to all of the participants within a communication session. Each of the images is represented as a shared byte array, so that any modification made by a session member is automatically reported to all of the session members. This ability should be enhanced with synchronization facilities, e.g. possibility to ensure mutually exclusive modification of the shared images. In teleradiology application, such a modification of image could be understood as e.g.: zooming, making adjustments to Hounsfield scale window, marking regions of interest etc.

The idea of a collaborative teleradiology application is depicted in Fig.1. It is possible to identify four basic steps of its activity:

1. The same radiology image is retrieved from a database and distributed over the teleradiogy session members' computers.
2. The image is displayed on every computer screen of teleradiology session members when teleradiogy session is started.
3. All session members are notified about modification of the image made by one of them, so that the state of consulted data is synchronised.
4. Results of the teleradiology session (e.g. its recording) are stored back in a database to be accessed when necessary.

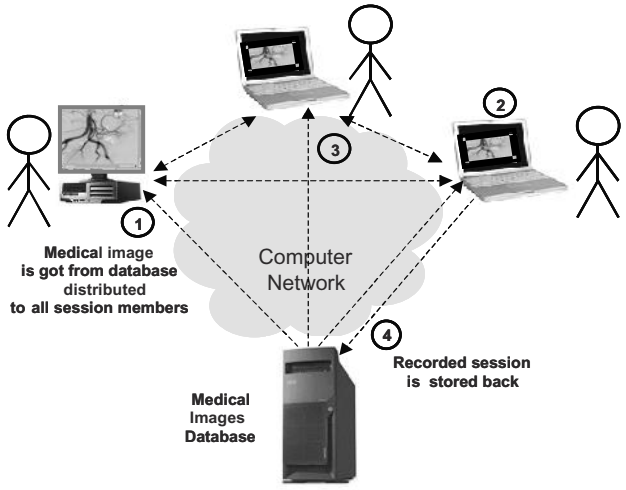


Fig. 1. Collaborative teleradiology application model

Such a scenario, with possibility of audio and video communication between the session members, would follow our definition of collaborative teleradiology.

Collaborative teleradiology could be easy compared with the web centric approach. Basic steps of the web-centric teleradiology application are depicted in Fig.2 and could be described as follows:

1. A selected radiology image is fetched from a database by dedicated software installed on a web server and sent to a user's web browser.
2. A radiology expert analyses the image.
3. A diagnosis in form of comments and added annotations is posted back to web server and stored in the database.
4. The diagnosis is read by other medical personnel when needed.

The basic difference between the collaborative and web centric approaches is that the latter does not provide interactivity, and comments could not be exchanged on-line between medical experts. This makes the analysis difficult and less efficient.

The presented general model of collaborative teleradiology applications could be further refined by defining many simple subcases. Such subcases would be taking into account QoS constrains and scalability of the system. We will continue this considerations in the following sections.

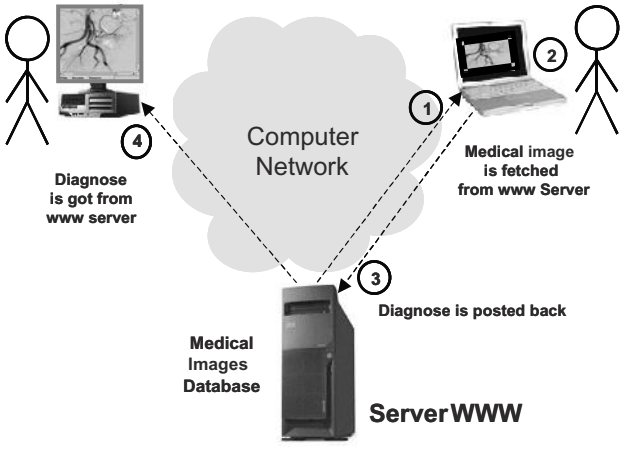


Fig. 2. Web-centric teleradiology

3 Models of Data Sharing

Radiology images are commonly represented according to DICOM standard [8]. The size of files containing radiology data stored in that format ranges from a few megabytes up to many tenths of megabytes. This is why high-speed broadband computer networks are required for fast on-line access to DICOM images. When using ISDN lines, the transfer of a DICOM file could take from a few minutes up to half an hour or even more. This is a real constraint of teleradiology application in the area where modern and expensive network infrastructure is not present. In that sense teleradiology application is network QoS (Quality of Service) dependent.

The solution to this problem is to reduce the amount of data transmitted during teleradiology sessions. For that purpose in Web centric applications the DICOM images are annotated by smaller in size JPEG pictures used for preview of radiology data. Doctors are using JPEG pictures to select interesting DICOM files, which have to be transmitted from a database to a DICOM viewer application or a web browser with DICOM viewer plug-in present on client machine. Unfortunately, the transmission of the selected DICOM files still takes a lot of time when it is performed over slow computer network. The discussed technique is used in commercially available radiology software e.g. ACOMWEB [9].

The network throughput constraint is even more serious for collaborative teleradiology applications, when on-line DICOM files transmission is assumed. The solution is to separate the collaborative teleradiology session into two phases:

1. Distribution of DICOM images to the computers of the teleradiology session members. This operation could be made off-line, e.g. at night.
2. Collaborative processing of DICOM images, which concerns the points 2,3 and 4 of collective teleradiology model disscused in Section 2.

Separation of these two phases makes sense, because the second phase is not as data transfer intensive as the first one. Transmission of synchronization and control information needed to implement collaborative processing of DICOM images generates data stream that could be transmitted even over ISDN lines.

The complexity of collaborative teleradiology applications is substantially dependent on the number of session members. From that point of view, two scenarios could be identified:

1. Dual-party teleradiology session, when only one radiology expert in a reference center and one doctor in a remote hospital are involved in collaboration.
2. Multi-party teleradiology session, with more than two members involved.

The effectiveness of the implementation of a multi-party teleradiology session depends on multicast communication support provided by the network infrastructure. The availability of reliable multicast communication has a great influence on the scalability of implementation of synchronization mechanisms and will be elaborated further in the following section.

4 Synchronisation of States of Multiple Distributed Teleconsultation Applications

One of the features that a user would expect from a teleconsultation application is synchronization of the state of remote machines. From a user point of view, the applications must ensure that any actions performed by one of cooperating parties will be noticed immediately by all the others. To make the matter more complex, the users of the applications do not frequently have the same privileges. As an example, consider a scenario when a real life teleconsultation is observed by many remotely taught students, who definitely would not have rights to modify the diagnosis. In such case, the teleconsultation system would need to implement some access restrictions[1].

Another issue, which seems to be important, is permitting the priviledged users to gain a lock on selected objects. In such case, the network of collected applications should for a while perform a „one-to-many" transmission scenario, with one source of messages and many sinks[2]. Therefore, the system needs a powerful middleware platform to send messages between its entities.

In today's highly distributed systems, two general forms of message-oriented middleware are commonly found: point-to-point and publish-subscribe. The first, often called message queuing, is used to move data between its source and destination, where the messages wait in a queue for application to pick them up. The

[1] Another important issue is the confidentiality of consulted patient's personal data.

[2] Moreover, the system should provide a fault recovery procedure in case when a client holding locks fails.

second kind is generally used when a data item needs to be sent to multiple clients. It can occur either if there is one or more sources of data.

The CORBA architecture, which is an industry standard, defines two ways of decoupled, event-driven communication, realised by Event and Notification Services, respectively. The OMG Event Service supports asynchronous exchange of event messages between objects. It introduces the following concepts:

- *event channels*, which function as brokers of event messages,
- *event suppliers*, which supply event messages, and
- *event consumers*, consuming i.e. receiving event messages.

The OMG Notification Service extends OMG Event Service by supporting a few important additional capabilities, e.g. filtering events the client wants to receive and possibility to specify various quality of service properties.

The Java Shared Data Toolkit software allows developers to add collaboration features to applets and applications written in the Java programming language. By implementing a multipoint data delivery service the toolkit can aid implementing network-centric applications, such as shared whiteboards or chat environments. The area of usage the JSDT includes also networking multimedia applications.

The Java Message Service (JMS) is a standardised set of programming interfaces to middleware. The middleware packages differ not only in their implementation details, but also in their interfaces. Since numerous middleware vendors include support for JMS in their products[3], it has become a way to escape from the vendor „lock-in" that many distributed systems are suffering from. In a way, it bridges the communication between wide variety of different software components (both servers and clients), e.g. firewalls, mail servers, application servers and so forth, which are standard entities for most of today's distributed systems. JMS supports both message queuing and publish-subscribe communication patterns.

In general, JMS could be useful for communication between different, loosely coupled system entities that make use of asynchronous communication. JMS supports one-to-one, one-to-many and many-to-many configurations.

The medical teleconsultation systems described later in this paper fall into the first and third category, respectively, so it seems that they could be served by JMS well. However, a teleconsultation application needs to be highly interactive. This feature is expected from the system even in the case of low-bandwidth connection between peripheral and referential hospital. Since that, one of the most important things to do during the design and implementation phases is to reduce any transmission overhead.

5 TeleDICOM System

The applications used in the Krakow Centre of Telemedicine formerly gave only the possibility to send digitised patients' data to the consulting centre and view it independently with no interaction. The only connection between the both sides was a

[3] This popularity of JMS is partly a result of a requirement of JMS compliance, which needs to be fulfilled by any software component in order to get a J2EE certification.

voice (or sometimes also video) channel, often realised via separate technical solutions (e.g. telephone line). The experience drawn from the TeleNegatoscope application allowed designing a more advanced environment.

TeleDICOM is going to offer many more advanced features introduced by the collaborative teleradiology paradigm. The most important of them is that it will operate on DICOM images offering fully diagnostic quality and precise textual description. DICOM format is *de facto* standard in the modern medical world; the images are very often created directly by medical equipment, e.g. a computer topography. The idea of the TeleDICOM environment originates from whiteboard-like applications where many participants share the same area making changes in its appearance or just observing modifications performed by the others. In TeleDICOM, a DICOM image delivered to all participants early enough will form the background. Operations performed on the image by one of the participants will be propagated to all the others. The available operations are supposed to be helpful during a consultation allowing users e.g. to mark fragments of the image (by placing different shapes there). The operations will also include actions typical for DICOM data, e.g. modifying the Hounsfield scale window parameters, zooming the image or running animation mode if the image is a multi-sliced one (e.g. coronarography examination).

The application has also an educational role allowing medicine students to participate in real consultations.

One of the most important design issues was how to organise flow of data in a multiparty scenario effectively. Taking into consideration the problems with multicast routing and reliable multicast, the decision was to use a centralised star topology with so called *rendez-vous* points to which all participants would direct their requests, and which would notify all participants about actions performed by the others. The whole communication is accomplished by using unicast transmission. Such a *rendez-vous* point, called session server, can be integrated into one of the applications participating in the consultation, or be run independently. The former solution is preferred in small, dual-party sessions and the latter in the multi-party ones. The logical placement of the session server in such case should be determined taking into account the location of all participants and aiming at ensuring them nearly equal distance to the central node of the star architecture in terms of available bandwidth and transmission delay. The concept of the session server in a session containing several participants is depicted in Fig. 3.

Security of medical information cannot be forgotten and the authors of TeleDICOM kept this in mind. All centres using the application have to get certificates used for authentication. The medical data (DICOM files and diagnoses) are encrypted by using SSL while being sent. The tests performed by authors proved that encryption of control data during consultation phase, severely influences its interactivity making it suffer from unacceptable delays, unless very fast computers are used. It must be however emphasised that with no knowledge about medical data itself, information intercepted by an eavesdropper is practically useless. In cases when a very strong security is required transmission can be led via Virtual Private Networks (VPNs) with fast encryption mechanisms provided on lower layers.

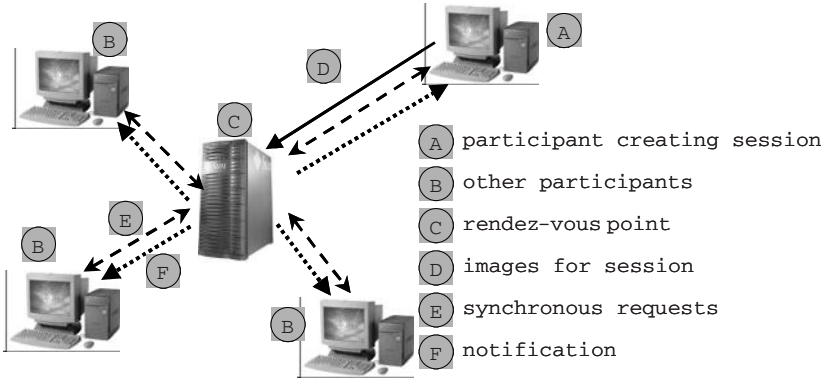


Fig. 3. The concept of rendez-vous point (session server) in multiuser session

6 Conclusions

The collaborative teleradiology opens a new, very challenging, field of multidisciplinary research in the area of telemedicine. The progress in this domain is possible due to the application open standards in distributed computing, which create a solid foundation for not only experimental, but also commercial, fully distributed, applications. Collaborative telemedicine put applied distributed software engineering in the context of very demanding applications in terms of required QoS, security and synchronised visualization techniques.

There is also open area for ergonomical and sociological studies to understand the requirements of the user interface design for group communication and on-line data sharing better. This research needs construction and deployment of a whole family of software tools supporting the proposed collaborative model and collecting experiences from the pilot systems introduced to the clinical practice.

References

1. NetRAAD system, University Health Care Company, http://www.uhc.com.pl
2. Web PACS, Universal PACS Inc., http://www.unipacs.com/webpacs.html
3. CORBAservices: Common Objects Services Specification, OMG, 1998
4. Remote Method Invocation Specification, Sun Microsystems, 1997
5. Enterprise Java Beans Specification, version 2.1, Sun Microsystems, 2002
6. Collaborative computing, Collab Worx Inc., http://www.collabworx.com/Support/resources
7. Java Shared Data Toolkit, Sun Microsystems, http://java.sun.com/products/java-media/jsdt/
8. The DICOM Standard, National Electrical Manufacturers Association.
9. Acom.WEB, Universitätsklinikum Essen, http://www.uni-essen.de/cardio/acomweb/acomweb_info.html

Some Remarks on CSE Education in Germany

Hans-Joachim Bungartz

IPVS, Universität Stuttgart, Universitätsstraße 38, D-70569 Stuttgart, Germany
bungartz@ipvs.uni-stuttgart.de

Abstract. The way Computational Science and Engineering is taught today in Germany, often, seems to depend more on local boundary conditions than on the real needs. Actually, some very fundamental questions are still open or are, at least, answered in very different and not always satisfying ways. Among these are the issues of the right focus of such a program or of the appropriate balance between methodology and real applications, as well as the question whether to implement both undergraduate and graduate programs or a graduate one, only. In this contribution, we want to discuss these and other related topics, considering selected existing CSE study programs in German-speaking countries.

1 Introduction

The last decade has seen the birth of a lot of new study programs in Germany dealing with Computational Science and Engineering (CSE), Scientific Computing, or any kind of Computational Something, where Something stands for an established discipline from science or engineering such as physics or mechanics, e.g. This welcome development reflects the grown importance of computer-aided methods in general and numerical simulation or high-performance computing in particular. As a matter of fact, at the moment, the implementation of all kind of new programs is further supported by a suddenly rather dynamic situation concerning study programs, which is mainly due to the forthcoming or ongoing, more or less radical changes the Bologna agreement entails.

However, although these CSE-related programs all have similar strategic objectives, we can observe a variety of appearances – with differences that are sometimes far from being just minor ones. First, in Germany, CSE is not an established discipline in the sense that there exist schools or faculties of this name where the respective programs are housed naturally. Here, we sometimes really get jealous of these "Schools of Computing" or "Departments of Computational Sciences" existing elsewhere! As a consequence, the rootedness of a CSE program in an existing faculty and its (and that's the point) focus as regards content, in most cases just depend on those of its initiator(s). Hence, on the one hand, we have more application-driven programs like *Computational Mechanics (COME)* in München [3] or *Computational Mechanics of Materials and Structures (COMMAS)* [7] and *Computational Physics* [8] in Stuttgart, and, on the other hand, there are more methodology-dominated programs such as *Scientific Computing* in Berlin [2] or *Computational Science and Engineering (CSE)* in München [4].

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1180–1187, 2004.
© Springer-Verlag Berlin Heidelberg 2004

A second issue is how to organize the CSE education – within an existing program or as a program of its own, and there either for graduate students only or for undergraduates, too. There are examples for all models: The faculties of mechanical engineering and civil engineering in Stuttgart, e.g., offer a study focus *CSE* [9] or *Modelling and Simulation* [10], resp., within their classical diploma programs. In contrast to that, the already mentioned *COME*, *COMMAS*, and or *CSE* as well as *Computational Engineering (CE)* [6] in Erlangen are all completely new programs, the former three being examples of pure master programs and the latter one consisting of both a bachelor and a master part.

Furthermore, we observe differences with respect to the program language. While several programs such as the ETH's *Rechnergestützte Wissenschaften* [1] hold to the German language, most pure master programs have been implemented as so-called international, i.e. English-based programs, since they want to (and actually do) attract international students, primarily, whereas the German-language programs cover most of the education of students from the German-speaking countries, at present. This raises the question of the target groups a CSE program wants to address. Apart from their nationality, this is also a question of disciplines. If a CSE master is designed by a mechanical engineering faculty for candidates holding a bachelor's degree in mechanical engineering, the profile of the students will be by far more homogeneous than it is in the case of the *CSE* program in München already mentioned, where just a bachelor's degree in some field of science or engineering is required.

This article aims at discussing these questions against the background of the needs of a modern CSE education as well as at opposing pros and cons of the respective model chosen. Although there will be statements for or against this or that way, it is, however, clear that solutions are always subject to boundary conditions such as local faculties, academic staff, and legal restrictions. Hence, there is no optimal way to be depicted, but just a collection of ideas that will, hopefully, initiate some reflections and discussions. The remainder of this paper is organized as follows. In Sect. 2, we discuss the topic of how to design CSE courses or programs and, in particular, whether to put a more concept-based or a more application-driven focus. The following Sect. 3 deals with the question whether CSE can just be seen as some kind of a specialization in the sense of either a minor topic within existing programs or some master graduate program accessible to students with some more or less conventional (first) degree, or whether a sound CSE education implies a complete restructuring of existing programs. Another issue, which will be addressed in Sect. 4, is the program language. Afterwards, in Sect. 5, we turn to the crucial question of degrees that qualify to enter the programs, and we discuss the chances and problems of a broader target group. Finally, some concluding remarks will close the discussion.

2 Simulation Technology or Grand Challenges?

Looking at the development of CSE, there have always been two driving forces of progress: general algorithmic improvements on the one hand and real applications on the other hand. The former – think of fast solvers, mesh refinement approaches, or parallelization strategies – are based on concepts such as the

multilevel paradigm and have turned out to be of great importance for real-life problems, but they, typically, neither stem from nor are restricted to a specific field of application. The primary motivation rests on notions such as speed of convergence, orders of accuracy, or speedups, e.g. In contrast to that, the other driving force is fed by the wish to solve important practical problems, to tackle the "grand challenges". The history of finite element methods provides a nice example for this duality. While engineers, especially from the field of civil or structural engineering, were eagerly looking for ways to compute more and more complicated buildings (for example, the roof of the Olympic stadium in München was one of the early constructions designed with the help of nonlinear computations – done in Stuttgart, by the way), numerical analysts were more interested in approximation spaces, error orders, or strategies of adaptive mesh refinement. Both streams were successful in their endeavours and allowed for significant steps forward, but the crucial break-throughs were only possible when applications and fundamentals or general concepts and concrete problem scenarios met.

This double-track development is reflected in many CSE-related programs in Germany. For example, consider the two master programs on computational mechanics mentioned above, *COME* [3] at the Technische Universität München and *COMMAS* [7] at the Universität Stuttgart. Both curricula do without any core courses offered by members of the math faculties; in both of them, we only find one or two courses contributed by computer science (in *COME* this formerly was, in *COMMAS* it presently is the author's job). When, during an evaluation board meeting, for instance, the question arises whether this is really enough in a program carrying the word "computational" on its banner, the standard answer is of the type "of course, we teach numerical methods, too" or "but in the exercises of my course, students have to program, too". And this argument is followed by a shrugging reference to upper credit bounds that limit each kind of elbow-room for additional topics. To avoid misunderstandings: both *COME* and *COMMAS* are excellent programs – as proven by personal experience and external evaluation – and there can be no doubt that those who are responsible for curricula and courses did and do an extraordinarily good job. Nevertheless, the question remains whether we really come up with the interdisciplinary character of CSE by practicing an "if there is any need for some subjects of discipline XYZ, I'll deal with that, too" credo, whether interdisciplinary or, better, transdisciplinary really just means that the existing group of teachers prefers to cover a broader spectrum of topics themselves instead of enlarging the group of contributing experts. Of course, if the answer is no, it is no better to do it the other way round, regarding Scientific Computing as something very similar to numerical analysis, as many mathematicians sometimes tend to think, or as being basically a synonym of supercomputing, as especially those guys sitting on the TOP 500 machines in the big computing centres sometimes argue.

The main risk of a too much application-oriented curriculum seems to be a deficit in simulation technology, i.e. in teaching general algorithmic concepts or paradigms. Numerical schemes – sometimes rather outdated, sometimes really up-to-date – are presented as the state-of-the-art, but neither derived nor justified. Note that I'm not speaking of extensive proofs where everyone (sometimes

including the lecturer) gets lost. If at all, students learn about hierarchical bases or multilevel solvers, e.g., just as being efficient tools for their specific tasks, but, at the end, they still lack of a deeper understanding of the way those work and of their potential. They learn to program and to use the few lines of the conjugate gradient algorithm, they – perhaps – learn about its main properties, but they don't really learn what happens in detail. All their doing is just caught in the narrow environment of their concrete flow problem to solve, e.g. The same holds for computer science. In many application-driven CSE-related programs, "computational" just means that some programming course is included, which by far does not reflect the needs of modern CSE. As a consequence, students are often unable to identify, carry over, or generalize the underlying concepts. Again, consider the other way round: The big risk of a pure simulation-technology-driven education with predominant mathematical and informatical components is the classical ivory tower scenario where sophisticated algorithms are derived, proven, and applied to examples as realistic as flow in a cubic motor. Obviously, such an unbalanced approach would not look better at all.

A better alternative is the really transdisciplinary approach of *CE* [6] in Erlangen and *CSE* [4] in München. In the latter case, six faculties jointly run this master program. There is a math column, provided by the math faculty, a computer science column, provided by the informatics faculty, a scientific computing column focusing on simulation *technology*, and an application column, provided by experts from the respective faculties, where a broad spectrum of lines of specialization such as computational mechanics, fluid mechanics, electrodynamics, physics, or chemistry can be chosen. Of course, at the end, there are no experts in Computational ... for one specific field, but there are students who have learned the fundamentals of a key technology and to apply the latter to some highly relevant field of application. Critics see a drawback of this model in its possible superficiality, the well-known "hearing a lot but understanding nothing in depth". Nevertheless, we need the balance of concepts and applications.

How important and fruitful a really transdisciplinary CSE education is, can be seen every year at the *Ferienakademie* [5], a joint venture of the Technische Universität München, the Universität Erlangen-Nürnberg, and the Universität Stuttgart. The *Ferienakademie* is a kind of summer school for the best and most motivated students, financed by industry from the two German states Bayern and Baden-Württemberg, by the participating universities themselves, and by their respective alumni organizations. During two weeks end of September, a dozen of seminar-type courses are held in Southern Tyrol, each with two professors from two of the organizing universities, one assistant, and about fifteen student participants who, essentially, carry the program via their presentations and discussions. The course topics cover the whole spectrum of modern technical universities. The course *Numerical simulation: from models to visualizations* co-organized twice by the author addresses students from basically all disciplines involved in CSE. Actually, in 2003, the participating students represented Mathematics, Technomathematics, Computer Science, Physics, Computational Physics, Chemistry, Computational Engineering, Electrical Engineering, Chemical Engineering, Software Engineering, and Computational Mechanics.

Of course, it is not always easy to handle such a mixture: computer scientists don't have the physical background to follow all discussions on turbulence in detail, a chemical engineer wonders what he can profit from a mathematician's presentation who solves the Laplace equation in a tricky way ("I have to solve flow problems!"), and computer scientists are astonished when a mechanical engineer asks basic questions on cache memories and on the RISC microprocessor architecture. Nevertheless, those experiences of different scientific cultures, different ways of giving a talk, different educational backgrounds, and different ways of tackling a given problem are crucial for real progress.

Finally, the given faculty structure at German universities is, to some extent, outdated, since the partitioning of today's scientific world should be definitely different from the one of a century ago. Instead of combining all groups involved in CSE activities in a powerful unit (call it school or department or however you want), there are computational minorities in most of the faculties, and there are permanent misunderstandings or even battles between pure and applied mathematicians, between floating-point and e-something computer scientists, between supercomputing engineers and experimental or production technology ones, and so on. As a result, there are CSE nuclei everywhere (which is not yet a drawback, of course) with non-coordinated curriculum activities (which definitely is one). Available expertise is not used, just because no one knows about it or, perhaps even worse, since everyone is convinced of a "we can do it on our own". Nevertheless, joint activities bringing together all experts from different faculties is definitely the way a transdisciplinary field such as CSE should be taught.

3 Basic Education or Specialization?

But let us turn to the appropriate format of CSE education, starting with a few examples. In Stuttgart, students of *Mechanical Engineering*, a classical 9-semester German diploma program, have to choose one specialization after their Vordiplom, i.e. after the second year. The alternatives are *Product Management and Design*, *Production Technology*, *Microsystem and Precision Engineering*, *General Mechanical Engineering*, and, recently, also *Computational Science and Engineering* [9]. Those choosing the latter get at least about 36 ECTS credits of a more or less directly CSE-related education, and they have to make two student research projects in this field. Students of the *Civil Engineering* diploma program in Stuttgart face a similar situation, since *Modelling and Simulation* [10] has been newly added to the three classical specializations *Structural Engineering*, *Traffic Engineering*, and *Water Engineering*. Here, an important part of the years 3–5 is dedicated to this CSE specialization. By the way, in the first case, almost all courses are provided by members of the mechanical engineering faculty; in the second case, this is true apart from a few electives imported from mathematics. Hence, these are two more examples of a more or less "in-house" solution without a real integration of expertise from other disciplines. In contrast to that, for the *Computer Science* diploma program in Stuttgart, the alternative approach of a real inter-faculty cooperation was chosen. The new minor *Scientific Computing* [11], one of fifteen possible choices including *Linguistics*, *Music Theory*, *Business Administration*, or *Mathematics*, consists of a math column and

an applications column completely provided by the mathematics department or several engineering faculties, resp. This ensures that all computer science students who want to specialize in CSE really sniff some air from mathematics and at least one field of application.

Instead of implementing specializations within existing programs, the creation of new study programs is the other possible strategy. In the majority of cases, at least at the moment, these new programs are mere graduate ones, typically master programs. As before, the underlying idea is that CSE knowledge is regarded as an extension of a conventional engineering education, e.g., giving to it some computerization and, thus, some modern flavour. Hence, the general assumption here is that all candidates already have some first relevant degree, typically a bachelor degree or a diploma degree from one of the German universities of applied science (Fachhochschulen). Next, the *Rechnergestützte Wissenschaften* program [1] of the ETH Zürich (hitherto a diploma program, now also a bachelor program, and in the future a bachelor-master combination) is some kind of a hybrid type. Although being a complete diploma program, the specific education starts in the third year only, since the first two years are identical to either mathematics, physics, chemistry, computer science, or some engineering subject. Finally, an example of a completely newly designed CSE program is the already mentioned *Computational Engineering* [6] in Erlangen, consisting of both a bachelor and a master component. This is, of course, the most consequent implementation of CSE education, which also offers most chances of a curriculum really tailored to the interdisciplinary needs of numerical simulation.

It is not clear which is the best strategy to combine a high-quality education really required by both the industrial and the academic job market with an economic usage of university resources. However, it is quite obvious that doing numerical simulations on a high level requires knowledge, skills, and experience in parts of mathematics, computer science, and in the respective field of application. And it, therefore, must be doubted whether this objective can be reached with just one of the "feel-the-flavour-type" programs within one discipline.

4 National/German or International/English?

The program language is another difference we discover when comparing the various CSE-related programs in Germany. Since one big motivation for master programs was to attract international students to our universities, most of them were designed as international programs with all or, at least, the vast majority of the courses being offered in English language. *CSE*, *COMMAS*, and *COME* are typical representatives of this development. The strategic drawback is that, at least at present, this approach keeps most of the German students out of these programs. *CE* in Erlangen is the hybrid model, offering most of the bachelor program in German (in order to integrate German students into the CSE education), but then offering the master basically in English language (in order to integrate foreign students into the CSE education). And, finally, the Swiss *Rechnergestützte Wissenschaften* program is completely given in German. This is definitely one of the not that frequent examples of a CSE study program that has been completely designed and implemented for the home students.

The language issue goes beyond this paper's scope. If, after the bachelor-master system, also the English language will gain general acceptance at German universities, it will be answered for CSE, anyway. On the other hand, if we continue distinguishing between national and international study programs, we shall see both German-language and English-language programs in the future.

5 Target Groups and Applicants

Of course, the language question is highly related to the question of whom a CSE program wants to address, primarily. But this target group issue is not restricted to countries or languages, there is also a disciplinary aspect. The two computational mechanics programs *COME* and *COMMAS* considered here address a quite narrow group of students – basically those with a bachelor degree in some mechanics-related field such as civil or mechanical engineering. This strategy allows to start immediately with special courses and to reach an impressive depth of studies within the three semesters. Nevertheless, the lecturers complain the obvious heterogeneity of the students (which is, here, probably more due to different previous levels of education than due to disciplines).

In contrast to this, the *CSE* program is open to a broad spectrum of degrees, since its regulations just require a bachelor degree in a related field including civil, mechanical, chemical, or electrical engineering, physics, chemistry, but also computer or aerospace engineering. Thus, the knowledge in some specific field of application may differ a lot among the participants of the program. However, the basic knowledge concerning advanced mathematics and basic programming skills is more depending on countries and universities than on disciplines. Hence, the umbrella character of *CSE* can handle this, providing a sound education in mathematical and informatical fundamentals of scientific computing and assigning the subject of application to each student according to his or her background. This way, mechanical engineers do computational fluid mechanics or computational structural mechanics, whereas physicists specialize in computational physics, e.g.

Especially for international target groups, one of the most urgent problems is applicant selection. For CSE, the situation is probably even harder due to its interdisciplinary character, which entails that there will always be candidates with different degrees. At present, the are a lot of discussions concerning reasonable strategies for making the process of applicant selection both faster and cheaper, resp., in order to avoid preparation semesters which are expensive both for students and universities, and to improve the hit rate. For that, the existing mechanisms of entrance exams or phone interviews etc. will have to be augmented by blended learning approaches combining online self-study components with special compact courses immediately before the first semester starts.

6 Concluding Remarks

In this paper, some topical issues concerning CSE programs in Germany were discussed. With several examples of relevant programs, the questions of the focus

and structure of a CSE program, of different ways to implement such programs, and the question of target groups concerning both nationalities and disciplines were addressed. Unfortunately, the opinion that CSE is nothing but a private subdiscipline that has recently emerged and that can be completely dealt with internally is rather widespread throughout many faculties – even if most of those responsible will always emphasize its interdisciplinary character in public. This, to some extent, often has strange and even negative consequences for the design of CSE specializations or CSE programs. There are even cases where a so-called CSE specialization is more a misleading package than something really deserving its name. However, for integrated specializations as well as for independent programs, involving all contributing disciplines – here, first of all, mathematics, computer science, and the various fields of applications have to be mentioned – is an essential precondition of a CSE education that is successful in the longer term and that really fulfils the needs of CSE as a discipline.

References

1. ETH Zürich. *Rechnergestützte Wissenschaften – diploma, bachelor, and master programs.* www.rw.ethz.ch/.
2. Freie Universität Berlin. *Scientific Computing – international master program.* www.math.fu-berlin.de/teach/webstud/studium/msc.htm.
3. Technische Universität München. *Computational Mechanics (COME) – international master program.* www.come.tum.de/.
4. Technische Universität München. *Computational Science and Engineering – international master program.* www.cse.tum.de/.
5. Technische Universität München, Universität Erlangen-Nürnberg, Universität Stuttgart. *Ferienakademie.* www5.in.tum.de/FA/.
6. Universität Erlangen-Nürnberg. *Computational Engineering – bachelor and international master programs.* www10.informatik.uni-erlangen.de/CE/.
7. Universität Stuttgart. *Computational Mechanics of Material and Structures (COMMAS) – international master program.* www.msc.commas.uni-stuttgart.de/.
8. Universität Stuttgart. *Computational Physics – bachelor program.* www.physik.uni-stuttgart.de/studium/bachelor/.
9. Universität Stuttgart. *Computational Science and Engineering – specialization within the diploma program* Mechanical Engineering. www.uni-stuttgart.de/mawesen/studienrichtungen.html/.
10. Universität Stuttgart. *Modelling and Simulation – specialization within the diploma program* Civil Engineering. www.iwb.uni-stuttgart.de/stuko/index.html.
11. Universität Stuttgart. *Scientific Computing – minor within the diploma program* Computer Science. www.informatik.uni-stuttgart.de/fakultaet/lehre/informatik/.

The Computational Science and Engineering (CS&E) Program at Purdue University

Thomas Downar and Tomasz Kozlowski

School of Nuclear Engineering,
Purdue University,
West Lafayette, Indiana 47907-1290
U.S.A
{downer, tomasz}@ecn.purdue.edu
http://www.cse.purdue.edu

Abstract. Purdue University has established an interdisciplinary graduate degree program in Computational Science and Engineering (CS&E). Both M.S. and Ph.D. degrees are offered. The program provides students with the opportunity to study a specific science or engineering discipline along with computing in a multi-disciplinary environment. The aim of the program is to produce a student who has learned how to integrate computing with another scientific or engineering discipline.

1 Introduction

In the Fall of 1995, Professors John Rice and Elias Houstis from the Computer Science Department at Purdue created the Computational Science and Engineering (CS&E) interdisciplinary graduate specialization program. Since 2000, Professor Tom Downar from the Schools of Engineering has been the director of CS&E which now has 80 students enrolled from 17 participating departments. Detailed information about the Purdue CS&E program is available at www.cse.purdue.edu.

The CS&E program is open to students admitted to a home department but interested in performing research at either the M.S. or Ph.D. level with a computational science and engineering specialization. The student's participation in the program is indicated on his/her transcript as a specialization depending on the student's school:

- Agriculture: COMPUTATIONAL SCIENCE
- Engineering: COMPUTATIONAL ENGINEERING
- Liberal Arts: COMPUTATIONAL SCIENCE
- Pharmacy: COMPUTATIONAL SCIENCE
- Science: COMPUTATIONAL SCIENCE

Seventeen departments offer CS&E specializations which are listed in Table 1 together with their Heads. Each department has a representative on the CS&E graduate committee which administers the CS&E program.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1188–1195, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Table 1. CS&E participating department list

DEPARTMMENT	NAME
AERO & ASTRO ENGINEERING	T. N. Farris
AGRICULTURAL ECONOMICS	Sally Thompson
AGRONOMY	Craig A. Beyrouty
BIOLOGICAL SCIENCES	David Asai
CHEMISTRY	Ian Rothwell
COMPUTER SCIENCES	Susanne Hambrusch
EARTH & ATMOSPHERIC SCIENCES	Harshvardhan
ELECT & COMP ENGINEERING	Leah Jemieson (Interim Head)
FOOD SCIENCE	P. Nelson
INDUS & PHYS PHARMACY	Stephen R. Byrn
MATHEMATICS	Leonard Lipshitz
MECHANICAL ENGINEERING	E. Daniel Hirleman
MED CHEMISTRY & MOLECULAR PHARMACOLOGY	Richard F. Borch
NUCLEAR ENGINEERING	Lefteri H. Tsoukalas
PHARMACY PRACTICE	Steven R. Abel
PHYSICS	Andrew Hirsch
PSYCHOLOGICAL SCIENCES	Thomas Berndt
STATISTICS	Mary Ellen Bock

Funding for the CS&E program is provided by the Provost in the form of a reoccurring grant which is used to fund the salary of the CS&E program coordinator and provide for incidental expenses such as travel funds and honorarium for CS&E seminar speakers.

CS&E students are expected to have a strong interest in computation and its application to science and engineering. Their undergraduate training is expected to have given them a strong foundation in several areas of science, engineering, and computing. It is, however, likely that many will not have both the interdisciplinary breadth and depth needed for the CS&E Program. Some students will fill in their deficiencies by taking some background courses and others will exploit their talents to catch up on missing background while taking graduate courses. The course CS 501 is specifically designed to allow students who did not major in computer science or computer engineering to learn quickly key concepts from upper level undergraduate computer science courses.

Advising CS&E students is handled by the CS&E staff (for routine matters), the Advisory Committee (for education program matters), and the major professor (for research direction). Documents are available to guide the students about requirements

and procedures. The routine registrar and graduate school forms are signed only by the CS&E staff unless agreed otherwise.

The non-educational support of a CS&E student is normally provided by the home department. In some cases (e.g., Ph.D. fellows) this support may be given by the sponsoring department. Examples of such support include:

- Business office support for financial matters,
- Advice and support for matters not directly related to education, e.g., disciplinary actions, visa problems,
- Office space, provided on the same basis as for regular students in the home department,
- Supplies, computing, copying, and similar items, provided on the same basis as for regular students in the home department.

The student enrollment in CS&E is shown in Figure 1.

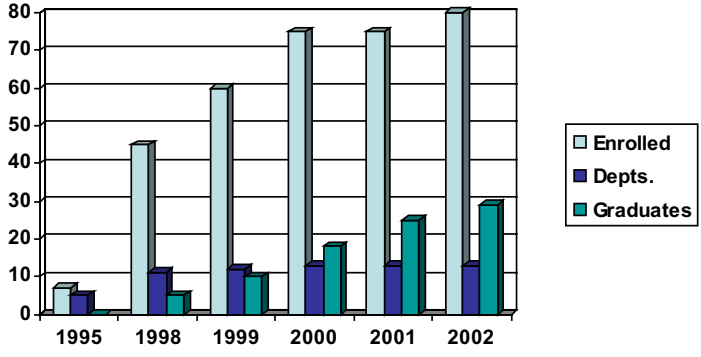


Fig. 1. Student enrolment in CS&E program

2 Admission to CS&E Program

Admission to CS&E program is administered through the student's home department. The Graduate School requires a Bachelor's degree and an overall grade point average (GPA) of at least 3.00/4 (A=4, B=3, C=2, D=1, F=0). The Graduate School also requires non-native speakers of English to obtain a score of at least 550 on the test of English as a Foreign Language (TOEFL). However, the home department might have higher TOEFL requirement.

There are two basic cases for admission to CS&E program:

1. A student admitted to a participating department at Purdue applies to the CS&E program. The CS&E Admissions Committee reviews the application and decides on the admission.
2. An applicant not at Purdue can apply directly to the home department requesting to be considered for the CS&E specialization or CS&E program. The CS&E Ad-

missions Committee reviews these applications and makes recommendations to the home department for final approval. The home department notifies the applicant of the result.

3 CS&E Curriculum

The CS&E curriculum is defined through the CS&E core courses and CS&E relevant courses specified by the participating departments. The basic graduate training in computational science and engineering is provided by the core courses listed below:

- INTRODUCTION TO CS&E
  - CS 501: Introduction to CS&E
- COMPUTATIONAL MATHEMATICS
  - CS 514: Numerical Analysis
  - CS 515: Numerical Analysis of Linear Systems
  - ME 581: Numerical Methods
- HIGH PERFORMANCE COMPUTING
  - CS 525: Parallel and Vector Computing
  - EE 563: Programming Parallel Machines
- INTELLIGENT COMPUTING
  - CS 572: Heuristic Problem Solving
  - EE 570: Programming Techniques for Artificial Intelligence
- SCIENTIFIC VISUALIZATION
  - CS 530: Scientific Visualization
  - EE 628: Computer Graphic Simulations and Visualization
- COMPUTATIONAL OPTIMIZATION
  - CS 520: Computational Methods in Analysis
  - EE 580: Computational Optimization
  - IE 535: Linear Programming
  - MA 521: Computational optimization

The CS 501 course was created specifically for the CS&E program to help students acquire the background for the other core courses. The specific objectives of CS 501 are to familiarize students from various disciplines in the science of problem solving. This includes basic computational tools (e.g. MATLAB, Mathematica, etc), computational kernels of important problems and core techniques for solving the problems. CS 501 emphasizes computer simulations as an important design tool and the role of high performance computing in large scale "Grand Challenge" type problems.

In addition to the core courses, the CS&E Program maintains a list of CS&E-relevant courses offered by the participating departments. These courses have a substantial computational component or relevance and are accessible and of potential interest to students from other departments. This list provides guidance in selecting CS&E courses for plans of study.

3.1 The CS&E Specialization for the M.S. Degree

The M.S. specialization of the CS&E Program may consist entirely of courses or of course work plus a project or M.S. thesis up to nine hours. At least thirty hours of credit are required for the degree, the equivalent of 10 normal courses. The plan of study must contain total three courses in the CS&E field of which at least two are CS&E core courses from different groups. No M.S. Degree examination is required. The plan of study is worked out by a student and advisor; it must be approved by the graduate committees of both the home department and CS&E program. Further details on the M.S. requirements are determined by the participating home departments.

3.2 The CS&E Specialization in the Ph.D. Degree

The Ph.D. plan of study is determined by the student and the major professor to prepare the student for research. The plan of study must also meet Graduate School requirements and be approved by the graduate committees of both the CS&E Program and the home department. The graduate committee of the CS&E Program expects about one third of the courses on the plan of study to be in the computing field. The following Ph.D. degree requirements are expected:

- The plan of study must contain at least four CS&E courses, including two CS&E core courses from different groups.
- The CS&E program does not require any general examination.
- GPA requirements: No course with a grade of C or lower may appear on the plan of study. The GPA in CS&E courses must be at least 3.25. The overall GPA in graduate courses on the plan of study must be at least 3.25.
- Qualifying examinations: Students are not required to pass any CS&E examination unless the home department or the Ph.D. advisory committee requests it.
- Preliminary examinations: These are given by the Ph.D advisory committee.
- The Chair of the Ph.D. advisory committee must be a member of the CS&E program faculty. The committee must contain one member from the home department and one CS&E faculty member not in the home department.

Further details on the Ph.D requirements for each participating department are provided in the CS&E graduate manual.

3.3 CS&E Specialization Requirements from Participating Departments

Each participating department also specifies requirements for their individual schools that the student must satisfy. As an example, details of the program requirements from the Aeronautics and Astronautics Engineering School are provided below.

The specialization in Computational Engineering is available to students in the School of Aeronautics & Astronautics wishing to develop a strong background in computational methods of engineering analysis and design. The CS&E specialization

is available to both M.S. and Ph.D. students. M.S. students may take either the thesis or non-thesis option.

The CS&E specialization for MS-A&AE. The student must choose a major and a minor area from among the approved School list. The CS&E specialization is available as a minor area under the interdisciplinary area option.

Minimum requirements for the thesis option are:

- Completion of twenty-one credit hours of graduate course work on an approved plan of study, which must include nine credit hours in the major area, six hours in the minor area and six hours in mathematics. Six of the CS&E credit hours must be taken in courses outside the major area.
- Attainment of a minimum cumulative index of 3.0 for graduation.
- Completion of nine credit hours of thesis-related research. Completion of a thesis to the satisfaction of the student's examination committee in compliance with the School and University format requirements.
- Public oral presentation of the research results.
- Passage of a closed oral examination by the examination committee.

Minimum requirements for the non-thesis option are:

- Completion of thirty credit hours of graduate course work on an approved plan of study, which must include twelve credit hours in the major area, six hours in the minor area and six hours in mathematics. Six of the CS&E credit hours must be taken in courses outside the major area.
- Attainment of a minimum cumulative index of 3.0 for graduation.
- Certification of the degree by the student's examination committee.

The CS&E specialization for Ph.D.-A&AE. The student must choose a major and a minor area from among the approved School list. The CS&E specialization is not considered a major or minor area for the Ph.D. degree program.

Minimum requirements for the Ph.D. degree are:

- Completion of eighteen credit hours of graduate course work beyond the M.S. degree on an approved plan of study.
- Attainment of an index of 3.0 for all courses on an approved plan of study.
- Successful completion of qualifying examinations in the major and minor areas and in mathematics. In lieu of the minor area qualifying exam, six hours of course work in the minor area may be taken, subject to obtaining a grade of "B" or better. Courses used to fulfill the minor area requirement may not be counted toward satisfaction of the CS&E course requirements.
- Successful completion of an oral preliminary examination given by the student's examination committee.
- Completion of a thesis to the satisfaction of the student's examination committee in compliance with the School and University format requirements.
- Public oral presentation of the research results.
- Passage of a closed oral examination by the examination committee.

4 CS&E Seminar Program and Fellowships

Another feature of the CS&E program is the seminar program. Students are required to attend 3 CS&E seminars each semester. An example of the seminars from years 2002-2003 is shown in Table 2 below.

Table 2. Representative CS&E seminary speakers

Speaker	Title of Talk
Gene Golub, Computer Science, Stanford University	Solution of Non-Symmetric, Real Positive Linear Systems
Frank Tip, IBM TJ Watson Research Center	Practical Extraction Techniques for Java
Hironori Kasahara, Waseda University, Tokyo, Japan	Multigrain Parallel Processing in Japanese Millennium Project IT21 "Advanced Parallelizing Compiler"
Ricardo Bianchini, Rudger University	Application Transformations for Energy-Efficient Device Control
Yousef Saad, University of Minnesita	Parallel Algebraic Recursive Multilevel Solvers (pARMS)
Steven F. Ashby, Lawrence Livermore National Laboratory	Terascale Scientific Simulation
Ed Angel, University of New Mexico	Large Scale Rendering Projects
Ahmed K. Noor, NASA Langley Research Center	Perspectives on Multiscale Modeling, Simulation and Visualization
Vince Mousseau, Los Alamos	Physics-Based Preconditioned Jacobian-free Newton-Krylov Solutions
Li-Shi Luo, ICASE	Lattice Gas and Lattice Boltzmann Methods: Past, Present, and Future

An additional service provided by CS&E is to award fellowships and scholarships to the top CS&E students. Since 1995, CS&E has awarded 30 fellowships to CS&E students, to include 5 U.S. Department of Education GAANN (Graduate Education in Areas of National Need) Ph.D. Fellowships.

5 Summary

Future plans for CS&E include the development of a second introductory core course, CS 502 and closer collaboration with other computational educational programs at Purdue. Next year, Purdue will begin a Computational Life Sciences program and plans are underway to begin an Information Systems Science and Engineering graduate educational program. Plans are to administer all Purdue interdisciplinary computational educational programs as part of the Computer Research Institute at Purdue.

References

1. Graduate Policies and Procedures Manual, Purdue University, 2003. http://www.purdue.edu/GradSchool/faculty/resources/policies.html
2. Computer Science and Engineering Manual, Purdue University, 2003. http://www.cse.purdue.edu/
3. Computing Research Institute, Purdue University, 2003. http://www.cri.purdue.edu/

Adapting the CSE Program at ETH Zurich to the Bologna Process

Rolf Jeltsch and Kaspar Nipp

Seminar for Applied Mathematics, ETH Zurich, CH-8092 Zurich, Switzerland
jeltsch@math.ethz.ch, nipp@math.ethz.ch

Abstract. In 1997 a new interdisciplinary Diploma program in Computational Science and Engineering (CSE) was started at ETH Zurich. We report on the changes of the curriculum due to the Bologna Declaration of June 19, 1999 by the European Ministers of Education. The new Bachelor program in CSE at ETH Zurich started in October 2003 and the Master program in CSE will follow in October 2005. We will describe both programs in some detail as well as the design principles. We will also discuss the boundary conditions imposed by ETH and we will give some preliminary experiences.

1 Introduction

In the 2002 International Conference on Computational Science, ICCS, held in Amsterdam, April 21 - 24, we have reported on the CSE Program at ETH, see [1]. This program was started in 1997 and was leading to a Diploma degree in CSE. In brief, students could join the CSE curriculum after 4 semesters of studies in a subject related to CSE (e.g., engineering, computer science, mathematics, physics, science). The CSE curriculum consisted of additional 4 semesters of studies followed by a diploma thesis to be completed in 4 months. The details of the curriculum are given in [1]. We report on the changes of the curriculum due to the Bologna Declaration of June 19, 1999 by the European Ministers of Education. ETH Zurich wanted to be among the first to change the diploma curricula into Bachelor and Master programs. The new Bachelor program started in the fall of 2003 with about 15 students and the Master program will be started in the fall of 2005. We will describe here the basic design principles of the new programs. For the design of the new Bachelor/Master curricula in CSE we still had the same understanding of the new discipline Computational Science and Engineering (CSE) and what we feel a student in this subject should learn in the program.

2 The Bachelor and Master Curriculum in CSE at ETH

2.1 Design Principles and Boundary Conditions

The transformation of the previous diploma studies into the new two stages program had to follow the objectives we had intended for the diploma studies and some boundary conditions imposed on us by the ETH system.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1196–1201, 2004.
© Springer-Verlag Berlin Heidelberg 2004

a) Aim of Curriculum

When designing a new curriculum one should think of what students should know and which abilities they should have acquired when successfully finishing the program. Our principles for the Bachelor and Master program are basically the same as for the Diploma curriculum in CSE, see [1].

The major objectives and design principles are:

- The emphasis is on Science and Engineering and not just on computing for the sake of computing, i.e., we want that the students know what they are computing.
- Computing has to be an essential tool in any field taught.
- Students should make an in-depth study in at least one or maybe two fields.
- Students should obtain a broad view and should have knowledge of many applications.
- Students should acquire the ability to work in a team.
- Students should have the ability to work together with people from different backgrounds (e.g., they must be able to talk to scientists and/or engineers who know the application area and they must be able to talk to mathematicians and computer scientists since they cannot solve all problems themselves).
- Students should have the ability to enter quickly into an ongoing research project of a team, make a contribution and transmit it successfully to the team.
- Students should acquire communication abilities.
- The overall study time for the Bachelor and Master programs should not be longer than for other students at ETH Zurich, i.e., 4 years and an additional 4 months for the diploma thesis.

b) ETH Boundary Conditions

Let us first discuss the two major new boundary conditions.

- With the change from the diploma curriculum to the two stages Bachelor/Master program, the major exam after 4 semesters has been abandoned. In our diploma program this used to be the natural point to motivate students from different studies to join the CSE program. This had the advantage to still have two years to educate them in the direction of CSE. It did have, however, the disadvantage that students entered the program with rather diverse backgrounds leading to an individually designed program for each student to *fill the gaps*.
- ETH kept the first major exam to be done by the students two semesters after entering ETH. This exam is now called basic exam and is much the same in most Bachelor programs.

Therefore, it was natural to start the CSE training directly after the first year of education. Considering the whole period including the master program this gave us an additional two semesters where the selection of courses could

be tailored to the education in CSE. In addition, since students have no longer strongly diverse backgrounds the *filling of the gaps* could be abolished.

There is a third dramatic boundary condition, namely, that any two stages Bachelor/Master program should fit the old Diploma program in length. In classical diploma studies one considered the first two years as the basic education terminated by the above mentioned major exam after 4 semesters. Then the second 4 semesters were used to train the students in advanced topics, i.e., in the CSE core courses and in their field of specialization. With the two stages program the Bachelor exam just cuts the time interval for the advanced education exactly in the middle. This meant that we had to make sure that the advanced courses already started in the Bachelor curriculum. For this reason we have designed the Bachelor/Master program as a whole. Concerning the core courses, students have to take all of them during their Bachelor/Master studies. Concerning the Fields of Specialization, we have introduced a minor field in the Bachelor program and a major field in the Master program. Clearly this combined design of the Bachelor/Master program could prolong the Master studies for students from outside. Another problem turned out to be that for mobility reasons students should finish their Bachelor degree by the end of the summer semester and not only in the fall. Hence, the Bachelor thesis should preferably be done during the final semester of the Bachelor studies. This thesis is in the range of a term paper which means that it is about 160 hours of work. Two positive changes at ETH have improved the curriculum in CSE. One is the increase of teaching staff so that we can now offer more courses which are geared directly to this group of students. The other is that we could introduce new fields of specialization, namely, *Computational Astrophysics* and *Computational Chemistry and Biology* . The latter replaces Computational Chemistry. In addition, we have introduced a new core course *Computational Statistics*.

2.2 Curriculum

a) The Bachelor Curriculum in CSE

The CSE Bachelor curriculum consists of two years of studies (4 semesters, second and third year) and is based on knowledge acquired in a first year of basic studies at ETH Zurich or elsewhere. The basic exam after the first year is counted for 60 credit points (ECTS). In the following two years the students have to get 120 ECTS, half of them in mandatory Basic Courses, the other half mainly in eligible Core Courses, Fields of Specialization, Elective Courses and with a Bachelor Thesis, respectively.

The **Basic Courses** provide basic knowledge in Mathematics, Computer Science and Science and Engineering. They are mandatory. The corresponding examinations are combined in 5 blocks supposed to be taken after the third, fourth and fifth semester, respectively. The list of basic courses is following:

- Analysis III
- Complex Analysis

- Data Bases
- Programming Techniques for CSE
- Numerical Mathematics for CSE
- Computer Networks
- Parallel Numerical Computing
- Optimization Techniques
- Quantum Mechanics
- Stochastics
- Statistical Physics
- Chemistry for CSE
- Physics I and Physics II
- Fluid Dynamics for CSE

The **Core Courses** are concerned with computational mathematical methods and with more advanced topics in computer science. Students have to take 2 of them in the third year and pass a combined exam. The list of core courses is following:

- Theory and Numerical Techniques of Differential Equations
- Computational Statistics
- Software Engineering
- Visualization / Graphics

The **Fields of Specialization** provide a profound knowledge in application areas with an emphasis on computational aspects. In the third year, students have to take 2 courses in a chosen field of specialization and pass the exams. The list of specialization fields is following:

- Astrophysics
- Physics of the Atmosphere
- Chemistry and Biology
- Fluid Dynamics
- Control Theory
- Robotics
- Theoretical Physics

The **Elective Courses** are complementary to the Fields of Specialization and to the Core Courses and should have an emphasis on computational aspects. Students in the third year have to take 2 courses out of a large list and pass the exams. In their third year, students have to attend twice the **CSE Case Studies Seminar** where each semester invited speakers from ETH, from other universities as well as from industry give a 2x45 minutes talk on an applied topic. Beside of attending the scientific talks students are asked to give short presentations (10 minutes) on a published paper out of a list (containing articles from, e.g., Nature, Science, Scientific American, etc.). In the case studies ETCS are obtained on a passed/failed basis. The **Bachelor Thesis** concludes the curriculum. It should deepen the knowledge in a specialization area, provide a

first contact with applications and should enable students to learn to approach applied problems computationally. It should also enable them to work within a scientific group. The bachelor thesis is typically written in a topic of a core course or a field of specialization. It requires about 160 hours of work and ECTS are obtained for an accepted paper.

Table 1. The CSE Bachelor Curriculum

	Year	ETCS (at least)
Basic Studies	**Semester 1/2**	**60**
Bachelor CSE	**Semester 3/4**	**65**
	Basic Courses	65
	Semester 5/6	**55**
	Core Courses	17
	Field of Specialization	6
	Elective Courses	6
	Case Studies	6
	Bachelor Thesis	8

After having successfully finished their studies students will be given the degree Bachelor of Science ETH in Computational Science and Engineering **(BSc ETH CSE)**. The main objective of the Bachelor curriculum is to prepare the students for the continuation of their studies in high quality Master curricula.

b) The Master Curriculum in CSE

The CSE Master curriculum at ETH Zurich consists of one year of studies (2 semesters, fourth year) followed by a Master thesis. The Master curriculum is based on the CSE Bachelor curriculum and its objective is to prepare students for a successful professional career in research in industry and/or on a university level. The Master students have to get 90 ECTS mainly in Core Courses, in eligible Fields of Specialization and Elective Courses as described in Subsection a) and with a Term Paper and a Master Thesis, respectively. In the **Core Courses** students have to take those 2 courses not covered for the Bachelor degree and pass a combined exam. Students have to choose a second **Field of Specialization** out of the list given in Subsection a) (in addition to the one chosen for the Bachelor degree) and have to take 4 courses within their two fields of specialization and pass the exams. Students have to take 2 **Elective Courses** and pass the exams. Students have to attend twice the **CSE Case Studies Seminar** as described in Subsection a) for the Bachelor curriculum. The **Term Paper** should be computational and application oriented work in a team in order to deepen the knowledge in a specialization area. It should require about 160 hours and ECTS are obtained for an accepted paper. The **Master Thesis** takes 4

months and is advised by a professor teaching in a core course or in a field of specialization. It concludes the CSE Master studies and should teach students to work independently on a scientific topic. ECTS are obtained for an accepted thesis. After having successfully finished their studies students will be given the degree Master of Science ETH in Computational Science and Engineering **(MSc ETH CSE)**.

Table 2. The CSE Master Curriculum

	Year	ETCS (at least)
Master CSE	**Semester 7/8**	**60**
	Core Courses	17
	Field of Specialization	16
	Elective Courses	6
	Case Studies	6
	Term Paper	8
	Semester 9	**30**
	Master Thesis (4 months)	30

3 Conclusions

As the program has started just a few weeks before the writing of this manuscript we can report only on preliminary experiences. One is that due to the fact that the Bachelor program starts after one year of studies at ETH we have now students which entered ETH after high school with the aim to study CSE. In the fall of 2003 14 students have entered the program. Again all students are highly motivated. Currently, there are no foreign students. Hence, the teaching is not done in English. It will be no problem, however, to do so if necessary. For more information on the CSE Bachelor/Master curricula see the webpage `www.cse.ethz.ch`. Reference [2] may be ordered from the Seminar for Applied Mathematics, ETH Zurich, CH-8092 Zurich.

References

1. Jeltsch, R., Nipp, K.: CSE Program at ETH Zurich: Are we doing the right thing? Eds. P.M.A. Slott, C.J. K. Tan, J.J. Dongarra, A.G. Hoekstra, Proceedings of the Conference: Computational Science - ICCS 2002, Vol III, pp. 863 - 871.
2. CSE Computational Science and Engineering, Annual Report 2002/2003, R. Jeltsch, K. Nipp, W. van Gunsteren, edits., ETH Zurich, 2003.

Computational Engineering and Science Program at the University of Utah

Carleton DeTar[3], Aaron L. Fogelson[2], Chris R. Johnson[1], Christopher A. Sikorski[1], and Thanh Truong[4]

[1] School of Computing, University of Utah, Salt Lake City, UT 84112, USA,
{crj, sikorski}@cs.utah.edu,
http://www.cs.utah.edu/{~crj,~sikorski}
[2] Department of Mathematics, University of Utah, Salt Lake City, UT 84112, USA,
fogelson@math.utah.edu,
http://www.math.utah.edu/~fogelson
[3] Department of Physics, University of Utah, Salt Lake City, UT 84112, USA,
detar@physics.utah.edu,
http://www.physics.utah.edu/~detar
[4] Department of Chemistry, University of Utah, Salt Lake City, UT 84112, USA,
truong@chemistry.utah.edu,
http://www.chem.utah.edu/faculty/truong/index.html

Abstract. We summarize the Computational Engineering and Science program at the University of Utah. Program requirements as well as related research areas are outlined. To obtain the MS degree in CES, a student must complete courses and present original research in scientific computing, scientific visualization, mathematical modeling, and the case studies in CES. The outlined research areas include scientific visualization, computational combustion, computational physics, computational chemistry, mathematical and computational biology, and computational medicine.

1 Computational Engineering and Science Program

The grand computational challenges in engineering and science require for their resolution a new scientific approach. As one report points out, "The use of modern computers in scientific and engineering research and development over the last three decades has led to the inescapable conclusion that a third branch of scientific methodology has been created. It is now widely acknowledged that, along with traditional experimental and theoretical methodologies, advanced work in all areas of science and technology has come to rely critically on the computational approach." This methodology represents a new intellectual paradigm for scientific exploration and visualization of scientific phenomena. It permits a new approach to the solution of problems that were previously inaccessible.

At present, too few researchers have the training and expertise necessary to utilize fully the opportunities presented by this new methodology; more importantly, traditional educational programs do not adequately prepare students to take advantage of these opportunities.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1202–1209, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Too often we have highly trained computer scientists whose knowledge about engineering and sciences is at the college sophomore, or lower, level. Traditional educational programs in each of these areas stop at the sophomore level — or earlier — in the other area. Also, education tends to be ad hoc, on the job and self-taught.

This situation has arisen because the proper utilization of the new methodology requires expertise and skills in several areas that are considered disparate in traditional educational programs. The obvious remedy is to create new programs that do provide integrated training in the relevant areas of science, mathematics, technology, and algorithms. The obvious obstacles are territorial nature of established academic units, entrenched academic curricula, and a lack of resources.

At the University of Utah the School of Computing (located in the College of Engineering), with the Departments of Mathematics and Physics (located in the College of Science) have established a graduate program that we consider a first step towards the modernization of the University's curriculum in what we call "Computational Engineering and Science" (CES). Our goal is to provide a mechanism by which a graduate student can obtain integrated expertise and skills in all areas that are required for the solution of a particular problem via the computational methodology.

We have recently created an M.S. CES degree program. If the M.S. CES program is successfel, we will consider expanding the program to a Ph.D. in CES. Our program is designed mostly for students in the Colleges of Engineering, Mines, and Science. However, in principle any graduate student at the University can participate.

To obtain the MS degree in CES, a student must complete courses and present original thesis research (for the thesis option) in each of the following areas:

I. Introduction to Scientific Computing
II. Advanced Scientific Computation
III. Scientific Visualization
IV. Mathematical Modeling
V. Case Studies in CES
VI. Elective course
VII. Seminar in Computational Engineering and Science

Of the above items, only V and VII are truly new requirements. Numerical Analysis has been taught in our departments for several years. Mathematical modeling has been spread over a large number of courses in the current Mathematics curriculum; the new course has been designed particularly for the CES program and has replaced one or more other courses in the students' load. The situation in regard to courses I and III in the Computer Science curriculum is very similar.

All courses are designed for first- or second-year science and engineering graduate students who have a knowledge of basic mathematics and computing skills. The most innovative aspect of our CES program is course V., Case Studies in CES. This course consists of presentations by science and engineering faculty from various departments around the campus. These faculty, all active in computationally intensive research, introduce students to their own work over intervals

of, typically, three weeks. The course provides students with a reasonably deep understanding of both the underlying science and engineering principles involved in the various projects and the practical issues confronting the researchers. It also provide a meeting place for faculty and graduate students engaged in CES activities in various departments throughout campus. In the CES seminar (VII) students are required to report on their own CES activities to their peers.

This approach, then, will serve the students by at once focusing their activities and bringing together in one place several essential components of CES that were previously spread over a larger and less clearly defined part of the existing curriculum. In addition, the program will help students learn to work with researchers in other disciplines and to understand how expertise in another field can help propel their own research forward.

The program is administered by a director and a "steering committee" consisting of two members each from the Departments of Mathematics, Physics and the School of Computing. The committee is advised by a board of faculty active in CES research.

We have already graduated 6 MS students and currently have 21 students participating in the program. We believe that the program will grow to about 30 students by Fall 2004. Our graduates continue their careers in national laboratories, university research laboratories, and technology companies.

2 Computational Engineering and Science Research

The University of Utah has a rich pool of faculty who are active in the various areas involved in CES. These include fields such as computational fluid dynamics, physics and chemistry, earthquake simulation, computational medicine, pharmacy, biology, computational combustion, materials science, climate modeling, genetics, scientific visualization and numerical techniques. A few examples of some of the current research in these areas follow.

2.1 Scientific Visualization

Common to many of the computational science application areas is the need to visualize model geometry and simulation results. The School of Computing at the University of Utah has been a pioneer and leader in computer graphics and scientific visualization research and education. Some of the first scientific visualizations were invented and displayed here, such as the use of color for finite element analysis and the use of geometrical visualizations of molecules. Utah is home to the Scientific Computing and Imaging (SCI) Institute [1]. SCI Institute researchers have innovated several new techniques to effectively visualize large-scale computational fields [2,3,4,5,6,7,8,9].

2.2 Computational Chemistry

Prof. Thanh N. Truong from the Department of Chemistry is leading major efforts in collaborations with Dr. Julio Facelli from the University of Utah Center

for High Performance Computing, Prof. Thom Cheatham from the Medicinal Chemistry Department, Prof. Chuck Wight also from the Chemistry Department, and Prof. James Lewis from Brigham Young University to develop an integrated extendable web-based simulation environment called Computational Science and Engineering On-line (CSEO). A number of graduate students from our CES program are involved in this project. CSEO allows computational scientists to perform research using state-of-the-art tools, querying data from personal or public databases, discuss results with colleagues or attend a workshop, and access resources beyond those available locally from a web browser. CSEO will provide a problem-solving environment for integrating different fields of computational science and engineering, thus facilitating inter-disciplinary research and particularly multi-scale simulations of complex scientific problems. Currently, CSEO supports an integrated environment for multi-scale modeling of complex reacting systems. Specifically, it allows results from fundamental quantum chemistry simulations to be used to calculate thermodynamic and kinetic properties of a chemical reaction, which subsequently are used in the simulation of a combustion reactor. A unique feature of CSEO is in its framework that allows data to flow from one application to another in a transparent manner. Efforts to extend the CSEO environment to computational biology and materials science are also being made. CSEO can be accessed at http://cseo.net.

2.3 Computational Combustion

The Center for the Simulation of Accidental Fires and Explosions (C-SAFE) (www.csafe.utah.edu), now in its 7th year, was created at the University of Utah through the Department of Energy's ASCI Program (www.llnl.gov/asci/alliances). C-SAFE focuses on creating a software based simulation system that can accurately simulate the complex interactions between large fires and structures within the fire, especially structures containing explosive materials. This system is designed to effectively use the computational resources available at the National Laboratories which include supercomputers consisting of 1000's of processors [10]. To produce accurate simulations of the chemistry and engineering physics involved, the software requires a combination of fully coupled non-linear solvers, optimization, computational steering, and visualization, along with validation with experimental data [11].

To create such a system, the combined talents of a highly skilled multidisciplinary team are required. Faculty, post-docs, staff members, and students from Chemistry, Chemical Engineering, Mechanical Engineering, and the School of Computing have worked together to produce this software. When it is completed, the software will provide a valuable means of improving and validating the safety of structures and explosive materials when exposed to large fires [12].

2.4 Mathematical and Computational Biology

The Mathematics Department at the University of Utah is home to one of the world's largest and most active research groups in Mathematical Biology. The

work in this group centers on understanding the biological mechanisms that regulate the dynamics of important physiological, biochemical, biophysical, and ecological interactions. A major focus of research, lead by Aaron Fogelson, is in using mathematics and computation to understand how the complex biochemical and biophysical components, especially the fluid dynamics, of platelet aggregation and coagulation interact in hemostasis (normal blood clotting) and thrombosis (pathological blood clotting within blood vessels). This fascinating area has tremendous practical importance because thrombosis is the immediate cause of most heart attacks and strokes. Because the models of clotting are very complex (they involve fluid dynamics, fluid-structure interactions, chemical kinetics, chemical and mass transport), they pose substantial computational challenges, and have required the development of novel numerical methods and software to meet these challenges. This software has been and is being applied to a wide range of biological problems in which fluid flow plays an important role. A second major research area, led by James Keener, involves modeling and three-dimensional computation of electrical waves in cardiac muscle. The goal is to understand normal signal propagation and the coupling between the electrical stimuli and cardiac muscle contraction, to understand the mechanisms underlying the onset of pathological arrythmias, and to understand at the cellular and tissue level how defibrillation works so as to help optimize defibrillation strategies. Some of the other work in the group includes studies of information processing in the primary visual cortex, models of territoriality in interacting animal populations, and studies of invasion of ecosystems by new species.

2.5 Computational Medicine

An interdisciplinary team of nationally recognized research centers at Utah involving the Cardiovascular Research and Training Institute (CVRTI), the Scientific Computing and Imaging (SCI) Institute, the Center for Advanced Medical Technology (CAMT), and the Neurosurgery Department are working together to tackle large-scale computational problems in medicine.

Every year, approximately 500,000 people die suddenly because of abnormalities in their hearts' electrical system (cardiac arrhythmias) and/or from coronary artery disease. While external defibrillation units have been in use for some time, their use is limited because it takes such a short time for a heart attack victim to die from insufficient oxygen to the brain. Lately, research has been initiated to find a practical way of implanting electrodes within the body to defibrillate a person automatically upon onset of cardiac fibrillation. Because of the complex geometry and inhomogeneous nature of the human thorax and the lack of sophisticated thorax models, most past design work on defibrillation devices has relied on animal studies. We have constructed a large-scale model of the human thorax, the Utah Torso Model [13,14,15], for simulating both the endogenous fields of the heart and applied current sources (defibrillation devices). Using these computer models, We are also able to simulate the multitude of electrode configurations, electrode sizes, and magnitudes of defibrillation shocks. Given the large number of possible external and internal electrode sites, magnitudes, and configurations, it is a daunting problem to computationally test

and verify various configurations. For each new configuration tested, geometries, mesh discretization levels, and a number of other parameters must be changed.

Excitation currents in the brain produce an electrical field that can be detected as small voltages on the scalp. By measuring changes in the patterns of the scalp's electrical activity, physicians can detect some forms of neurological disorders. Electroencephalograms, EEGs, measure these voltages; however, they provide physicians with only a snapshot of brain activity. These glimpses help doctors spot disorders but are sometimes insufficient for diagnosing them. For the latter, doctors turn to other techniques; in rare cases, to investigative surgery.

Such is the case with some forms of epilepsy. To determine whether a patient who is not responding to medication has an operable form of the disorder, known as temporal lobe epilepsy, neurosurgeons use an inverse procedure to identify whether the abnormal electrical activity is highly localized (thus operable) or diffused over the entire brain [16,17].

To solve these two bioelectric field problems in medicine, we have created two problem solving environments, SCIRun and BioPSE [18,19,20].

Using SCIRun and BioPSE, scientists and engineers are able to design internal defibrillation devices and source models for the epileptic foci, place them directly into the computer model, and automatically change parameters (size, shape and number of electrodes) and source terms (position and magnitude of voltage and current sources) as well as the mesh discretization level needed for an accurate numerical solution. Furthermore, engineers can use the interactive visualization capabilities to visually gauge the effectiveness of their designs and simulations in terms of distribution of electrical current flow and density maps of current distribution.

2.6 Computational Physics

Several computational science opportunities are offered by the Physics Department. Students can gain experience in computational physics on platforms ranging from small Beowulf clusters to supercomputers at the national laboratories to the most powerful, special-purpose computers available in the world.

Experimental High Energy Astrophysics. The High Resolution Fly's Eye and the High Energy Gamma Ray research groups collect terabytes of observational data that require processing on high performance parallel computers. Monte Carlo simulations of detector performance and of theories also require intensive computation. The goal of this research is to understand the origins and production mechanisms of the mysterious, energetic particles. Professors Pierre Sokolsky, David Kieda, Paolo Gondolo, Charles Jui, Kai Martens, and Wayne Springer participate in this work.

Computational Astrophysics. Computer simulations allow us to experiment with the Universe at large. For example, by varying the amount of dark matter in a simulated galaxy, and then comparing it to real, observed galaxies, we can

place constraints on the nature of dark matter. Ben Bromley has run simulations to study the effects of dark matter in colliding galaxies. With Scott Kenyon, Bromley is also working on models for the formation of rocky planets like Earth. Key issues include how rocky planets form, and whether they are common around other stars like our Sun. Recently, Bromley has joined Richard Price in simulating the collision between two black holes. Colliding black holes produce gravity waves which be will detectable by observatories (e.g., LIGO and LISA) in the near future. Only with computer models will we be able to accurately interpret the gravity wave data.

Solving the Strong Interactions. Large scale *ab initio* simulations of quantum chromodynamics, the theory of interacting quarks and gluons, have proven to be an indispensible guide to our understanding of the masses and structure of the light elementary particles, the decays of heavy mesons, and the cooling of the quark-gluon plasma, which existed in the early universe. Professor Carleton DeTar is carrying out this work.

2.7 Future Directions

The field of Computational Engineering and Science holds rich possibilities for future development. The *computational paradigm* has taken hold in nearly every area in science and engineering. Its use is also becoming more common in many fields outside of science and engineering, such as the social sciences, architecture, business, and history. Its success hinges on researchers' ability and willingness to transcend traditional disciplinary barriers and share expertise and experience with a large group of colleagues who may have been perceived previously as working in unrelated fields. Such boundary-crossing enriches researchers' work by providing new computational opportunities and insights. Building the interdisciplinary spirit will enable the solution of problems that were previously inaccessible. At the University of Utah, we are excited by such prospects and have taken the first step in initiating a mechanism to educate the next generation of scientists and engineers. One concrete future step we are taking is the creation of a Ph.D. degree in Computing that will have a Scientific Computing track. We hope to start admitting students into the new Ph.D. Computing degree program in Fall 2004.

For more information on our CES program visit www.ces.utah.edu or contact Christopher Sikorski (sikorski@cs.utah.edu).

References

1. Scientific Computing and Imaging Institute; http://www.sci.utah.edu/.
2. Y. Livnat and C.D. Hansen. View dependent isosurface extraction. In *IEEE Visualization '98*, pages 175–180. IEEE Computer Society, October 1998.
3. Y. Livnat, S.G. Parker, and C.R. Johnson. Fast isosurface extraction methods for large image data sets. In A.N. Bankman, editor, *Handbook of Medical Imaging*, pages 731–745. Academic Press, San Diego, CA, 2000.

4. G.L. Kindlmann and D.M. Weinstein. Hue-balls and lit-tensors for direct volume rendering of diffusion tensor fields. In *Proceedings of the IEEE Visualization 99*, pages 183–189, 1999.
5. G.L. Kindlmann and D.M. Weinstein. Strategies for direct volume rendering of diffusion tensor fields. *IEEE Trans. Visualization and Computer Graphics*, 6(2):124–138, 2000.
6. C.R. Johnson and A.R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, September/October 2003.
7. J. Kniss, G. Kindlmann, and C. Hansen. Multidimensional transfer functions for interactive volume rendering. *IEEE Transactions on Visualization and Computer-Graphics*, 8(3):270–285, 2002.
8. D. Breen and R. Whitaker. A level-set approach to 3d shape metamorphosis. *IEEE Transactions on Visualization and ComputerGraphics*, 7(2):173–192, 2001.
9. C.R. Johnson, Y. Livnat, L. Zhukov, D. Hart, and G. Kindlmann. Computational field visualization. In B. Engquist and W. Schmid, editors, *Mathematics Unlimited – 2001 and Beyond*, volume 2, pages 605–630. Springer-Verlag, 2001.
10. J.D. de St. Germain, J. McCorquodale, S.G. Parker, and C.R. Johnson. Uintah: A massively parallel problem solving environment. In *Ninth IEEE International Symposium on High Performance andDistributed Computing*, pages 33–41. IEEE, Piscataway, NJ, Nov 2000.
11. R. Rawat, S.G. Parker, P.J. Smith, and C.R. Johnson. Parallelization and integration of fire simulations in the uintah pse. In *Proceedings of the Tenth SIAM Conference on Parallel Processing for Scientific Computing*, March 12-14 2001.
12. T.C. Henderson, P.A. McMurtry, P.J. Smith, G.A. Voth, C.A. Wight, and D.W. Pershing. Simulating accidental fires and explosions. *IEEE Computational Science and Engineering*, 2:64–76, 2000.
13. C.R. Johnson, R.S. MacLeod, and P.R. Ershler. A computer model for the study of electrical current flow in the human thorax. *Computers in Biology and Medicine*, 22(3):305–323, 1992.
14. C.R. Johnson, R.S. MacLeod, and M.A. Matheson. Computational medicine: Bioelectric field problems. *IEEE COMPUTER*, pages 59–67, October 1993.
15. R.S. MacLeod, C.R. Johnson, and M.A. Matheson. Visualization of cardiac bioelectricity — a case study. In *IEEE Visualization '92*, pages 411–418, 1992.
16. D.M. Weinstein, L. Zhukov, and C.R. Johnson. Lead-field bases for EEG source imaging. *Annals of Biomedical Engineering*, 28:1–7, 2000.
17. D. Weinstein, P. Krysl, and C. Johnson. The bioPSE inverse EEG modeling pipeline. In *ISGG 7th International Conference on Numerical Grid Generation in Computation Field Simulations*, pages 1091–1100, Mississippi State University, 2000. The International Society of Grid Generation.
18. SCIRun: A Scientific Computing Problem Solving Environment. Scientific Computing and Imaging Institute (SCI), http://software.sci.utah.edu/scirun.html, 2002.
19. BioPSE: Problem Solving Environment for modeling, simulation, and visualization of bioelectric fields. Scientific Computing and Imaging Institute (SCI), http://software.sci.utah.edu/biopse.html, 2002.
20. C.R. Johnson, S. Parker, D. Weinstein, and S. Heffernan. Component-based problem solving environments for large-scale scientific computing. *Journal on Concurrency and Computation: Practice and Experience*, (14):1337–1349, 2002.

A Comparison of C, MATLAB, and Python as Teaching Languages in Engineering

Hans Fangohr

University of Southampton, Southampton SO17 1BJ, UK
fangohr@soton.ac.uk

Abstract. We describe and compare the programming languages C, MATLAB and Python as teaching languages for engineering students. We distinguish between two distinct phases in the process of converting a given problem into a computer program that can provide a solution: (i) finding an algorithmic solution and (ii) implementing this in a particular programming language. It is argued that it is most important for the understanding of the students to perform the first step whereas the actual implementation in a programming language is of secondary importance for the learning of problem-solving techniques. We therefore suggest to chose a well-structured teaching language that provides a clear and intuitive syntax and allows students to quickly express their algorithms. In our experience in engineering computing we find that MATLAB is much better suited than C for this task but the best choice in terms of clarity and functionality of the language is provided by Python.

1 Introduction

Computers are increasingly used for a variety of purposes in engineering and science including control, data analysis, simulations and design optimisation. It is therefore becoming more important for engineering students to have a robust understanding of computing and to learn how to program. In this paper, we outline the difficulties in learning and teaching programming in an academic context including the choice of the programming language. In section 2, we suggest a distinction between the algorithmic problem-solving part of computer programming and the efforts to implement the algorithm using a particular programming language. In section 3, we describe and compare MATLAB, C and Python as potential teaching languages and report our experience of them in an Engineering Department in section 4 before we conclude.

2 Teaching Objectives

We understand the subject of "computing" to broadly represent the usage of computers and numerical methods to solve scientific and engineering problems. In the curriculum, we aim to go beyond the usage of dedicated software packages and to enable students to write their own computer programs to provide insight

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1210–1217, 2004.
© Springer-Verlag Berlin Heidelberg 2004

into the functionality of any software (at least in principle) that they may encounter. In this section, we derive our requirements for programming languages to be used in education.

Universal programming building blocks: To analyse the programming process, we list the main ingredients for any computer program (this can be done more formally but is sufficient for our purposes here): (i) statements that do something (for example adding two numbers, perform a Fourier transform, read a sensor), (ii) blocks of statements, (iii) loops that repeat blocks (for-loops, for-each-loops, while-loops, repeat until loops), (iv) conditional execution of blocks (if-then statements, case, switch) and (v) grouping of blocks into modules (functions, procedures, methods).

This lists a set of commands that is sufficient to describe sequential algorithms, and by grouping statements into modules, moderately large and well-structured programs can be written within this framework.[1] Virtually all programming languages provide constructs that correspond to the listed items.

Problem-solving process. Computer programs are generally used to solve a given problem. We divide the process of writing a computer program into two parts:

1. finding the algorithmic solution (we will call this the "algorithmic problem-solving part") and
2. implementing the algorithm in a particular language (the "implementation part").

For example, to compute an approximation A of the integral $I = \int_a^b f(x)\mathrm{d}x$ using the composite trapezoidal rule with n subdivisions of the interval $[a, b]$: $A = \frac{h}{2}\left(f(a) + f(b) + 2\sum_{i=1}^{n-1} f(x_i)\right)$ with $h = \frac{b-a}{n}$ and $x_i = a + ih$, the two parts of the solution are:

1. the algorithmic solution which can be written in some form of pseudo-code:

```
user provides f, a, b, n
compute interval width h=(b-a)/n
set initial area=0.5*h*(f(a)+f(b))
for each point x=x_i with i increasing from 1 to n-1
    compute area under f(x)
        x = a + i*h          (this is the current x)
        dA = f(x)*h          (this is the new contribution)
    area = area + dA         (update the total area)
return area to user
```

2. the implementation which expresses the algorithmic solution in some programming language. For example figure 1 shows MATLAB code[2] that performs the calculation (for $f(x) = \exp(-x^2)$ and $a = 0$ and $b = 1$).

[1] We recognise the importance of object orientation but have excluded such features from the list above for clarity of argument.

[2] There are more efficient, general and elegant ways to code this but these are unlikely to be used by beginners and irrelevant to the situation described here.

```
% input from user
a = 0.0;    b = 1.0;    n = 100;

h = (b-a)/n;
area = 0.5*h*( exp(-a^2) + exp(-b^2) );

for i=1:n-1
   x = a+i*h;
   area = area + h*exp(-x^2);
end

fprintf('The value of the approximation is \%f\n', area)
```

Fig. 1. A MATLAB program to approximate $\int_0^1 \exp(-x^2)\mathrm{d}x$

While in this example the algorithmic problem-solving is relatively straight forward (because the problem was posed in form of an equation), in general this is the main challenge the students face: the conversion of a problem described vaguely and informally in natural language into a sequence of instructions that break the problem in many small parts that a computer can solve subsequently. The implementation part can be complicated and time consuming but does not (or at least should not) contain major intellectual challenges.

The boundary between the algorithmic problem-solving and the implementation is, of course, not clearly defined. However, it is clear that different implementations of a problem-solving algorithm in different languages share the same underlying algorithm (which we use here as the definition for the algorithmic problem solving part). In the teaching practice, the algorithmic problem-solving and implementation tasks are often entangled simply because the students need to test an algorithm they invented by implementing it.

Teaching objectives in computing: We argue that the primary target in teaching computing is to enable the students to convert engineering problems into pseudo-code. This is a challenging task that requires analytical thinking and creativity. The conversion of this pseudo-code into a program written in one programming language is of secondary importance because it is, in principle, an algorithmic procedure and requires less intellectual effort.

Consequently, the choice of the teaching language should be governed by which language provides the best support to the student in performing the implementation part of the problem-solving task. (The remainder of this paper addresses this question.) Once students are confident in the algorithmic problem-solving part, they can learn new programming languages as required to port the algorithmic solutions to their current working environment.

3 Overview of Programming Languages Used

The C programming language: The C programming language [1] is a low-level compiled language (sometimes classified as a 3$^{\mathrm{rd}}$ generation language) that is widely used in academia, industry and commerce. Fortran falls into the same category but while Fortran is still commonly used in academia it appears to be

overtaken by C (and C++) in many industrial applications. C++ provides a different programming paradigm than C but for the purpose of this work, C++ is more similar to C than it is to MATLAB or Python. The main advantage of compiled low-level languages is their execution speed and efficiency (for example in embedded systems).

MATLAB: The MATLAB programming language is part of the commercial MATLAB software [2] that is often employed in research and industry and is an example of a high-level "scripting" or "4^{th} generation" language.

The most striking difference to C and other compiled languages is that the code is interpreted when the program is executed (an interpreter program reads the source code line by line and translates it into machine instructions on the fly), *i.e.* no compilation is required. While this decreases the execution speed, it frees the programmer from memory management, allows dynamic typing and interactive sessions. It is worth mentioning that programs written in scripting languages are usually significantly shorter [3] than equivalent programs written in compiled languages and also take significantly less time to code and debug. In short, there is a trade-off between the execution time (small for compiled languages) and the development time (small for interpreted languages).

An important feature for teaching purposes is the ability of MATLAB (and other interpreted languages) to have interactive sessions. The user can type one or several commands at the command prompt and after pressing return, these commands are executed immediately. This allows interactive testing of small parts of the code (without any delay stemming from compilation) and encourages experimentation. Using the interactive prompt, interpreted languages also tend to be easier to debug than compiled executables.

The MATLAB package comes with sophisticated libraries for matrix operations, general numeric methods and plotting of data. Universities may have to acquire licences and this may cost tens of thousands of pounds.

Python: Python [4] is another high-level language and at first sight very similar to MATLAB: it is interpreted, has an interactive prompt, allows dynamic typing and provides automatic memory management (and comes with in-built complex numbers).

We have included Python in this work because it provides several advantages over MATLAB in the context of teaching: (i) Python has a very clear, unambiguous and intuitive syntax and uses indentation to group blocks of statements. (ii) Python has a small core of commands which provide nearly all the functionality beginners will require. (iii) Python can be used as a fully object-orientated language and supports different styles of coding. (iv) The Python interpreter is free software (*i.e.* readily available), and Python interpreters for virtually all platforms exist (including Windows, Linux/Unix, Mac OS).

It is worth noting that although Python has been around for only approximately 10 years, it is a relatively stable language and used increasingly in industry and academia (currently including organisations such as Philips, Google, NASA, US Navy and Disney). It also provides the framework for creating and managing large modularised codes. Commonly used extension modules provide access to

```
#include<stdio.h>
#include<math.h>

int main( void ) {
  int n = 100;   double a = 0.0;   double b = 1.0;

  int i;
  double h = (b - a) / (double) n;
  double area = h*0.5*( exp( -a*a ) + exp( -b*b ) );
  double x;

  for (i=1; i<n; i++) {
    x = a + i*h;
    area = area + h*exp( -x*x );
  }

  printf("The value of the approximation is %f\n", area);

  return 0;
}
```

Fig. 2. C program to approximate $\int_0^1 \exp(-x^2)\mathrm{d}x$

compiled libraries including high performance computation [5] and visualisation tools.

4 Teaching Experience

In this section, we present for each of the languages in question a short program that performs the numerical integration as given in the equation in section 1. The source codes shown are neither commented, optimised for speed and elegance nor do they explore advanced features of the respective language. This is to save space and to represent the coding style that students would initially use. We then report and discuss experiences in teaching the three languages to first and second year undergraduate engineering students and to postgraduate students from different backgrounds.

Using C: Figure 2 shows one C program that performs the same computation as the MATLAB program shown in figure 1. In comparison to the MATLAB program, this code is longer and carries a substantial overhead (such as the include statements, the wrapping of the main code into the `main` function and the return of the exit status). It is necessary to declare variables and their types before any statements are executed. Eventually, the student needs to compile the code (and link to the mathematics library) before it can be executed. Typical problems students experience while programming in C are: (i) Indentation of for-loop (and other blocks) and scope (as defined by curly braces) do not agree, thus the for-loop executes incorrect commands. (ii) Missing semicolons, curly braces, parentheses around if-statement tests in combination with moderately useful error messages from the compiler stop the compilation process. (iii) Passing of values of wrong type in function calls or printing numbers using the wrong format

```
import math
n = 100;    a = 0.0;     b = 1.0

h = (b-a)/n
area = 0.5*h*( math.exp(-a**2) + math.exp(-b**2) )

for i in range(1,n):
    x = a+i*h
    area = area + h*math.exp(-x**2)

print "The value of the approximation is", area
```

Fig. 3. Python program to approximate $\int_0^1 \exp(-x^2)\mathrm{d}x$

identifier token; both problems result in wrong numerical results that students find difficult to interpret and rectify. While these are not particularly challenging to the experienced programmer it can be observed that beginners struggle when they come across this framework, whereas there are significantly less difficulties when they start using MATLAB or Python.

Using MATLAB: Figure 1 shows the example program. We generally find that it is much easier for students to start programming using MATLAB than it is using C because MATLAB addresses many of the issues raised above.

It is good practice to split code into small functional units where-ever possible to modularise programs and to be able to re-use the functions individually for different projects. One of the recurring problems that students experience in learning programming using MATLAB is the convention of storing only one (globally visible) function in a file. This can result in a large number of files and initially the usage of functions is experienced by students as being counterproductive: several files have to be displayed to see all the source code at the same time (thus making it hard to follow the programme flow on the screen). A related issue is that in MATLAB the name of the function as specified within the source file should be identical to the file name of the source file containing that function. (The filename determines the globally visible name of the function.) This is often overlooked (by the students) and a source for many errors.

Using Python: Figure 3 shows a Python implementation of the integration problem. By importing the `math`ematics module in the first line, the use of name spaces is nicely demonstrated.[3]

The `range` command in line 7 returns a list of integers (ranging from 1 up to but not including `n`) which `i` refers to in subsequent iterations. The for-loop used here is actually a for-*each* loop because for each element in the list of integers, the body of the loop is executed. For-each-loops are in general more powerful than the for-loop, although this is not exploited here. (Note that the for-loop in MATLAB is also a for-each loop although the syntax makes this less obvious.)

It is of significant benefit in teaching programming that the block of statements in the body of the for-each loop is limited solely by indentation because

[3] This also addresses issues with unintented use of global variables in MATLAB which can confuse beginners but which are outside the scope of this report.

```
import math

def integrate( f, a, b, n):
    h = (b-a)/n
    sum = (f(a) + f(b))*0.5

    for i in range(1,n-1):
        x = a+i*h
        sum = sum + f( x )
    return sum*h

def f1(x):
    return math.exp(-x**2)

def f2(x):
    return math.sin(-x**4)

# main program starts here:
n = 100; a = 0.0; b = 1.0
print "The approximation of f1 is ", integrate( f1, a, b, n)
print "The approximation of f2 is ", integrate( f2, a, b, n)
```

Fig. 4. A Python program demonstrating how to pass functions as arguments

much time is spent encouraging students to ensure that the actual block delimiters (curly braces {} in C, the `for` and `end` keywords in MATLAB) are in agreement with the chosen indentation (because our perception of the program structure is led by indentation). Any disagreement between the two is likely to represent an error.

Python offers an environment which addresses most of the problems we observe in teaching C and MATLAB. Students experience comparatively few problems in starting to program in Python and develop to enjoy and experiment within the intuitive environment.

Advanced example: It is outside the scope of this work to explain many of the reasons why Python is generally appealing as a first programming language (see for example [6]). Instead, we present the implementation of one logical extension of the integration programmes in Python and compare this with C and MATLAB.

Students should experience that it is advantageous to build a library of small generic functions that can be used and re-used to solve bigger problems rather than to write specific code to solve one, and only one, problem. As part of this, it is sensible to code the integration routine in a function which takes the following parameters: the function to integrate (`f`), the lower and upper integration limits (`a` and `b`) and the number of subdivisions (`n`). A Python program that first defines this function with the name `integrate` and then integrates two functions `f1` and `f2` is shown in figure 4.

Note that the function to be integrated is passed to the `integrate` function just like any other object, and that `f` is evaluated inside the `integrate` function as if a function `f` would be defined (whereas in fact `f` refers to either `f1` or `f2`). This is in stark contrast to C where the passing of functions as arguments needs

the introduction of pointers (and where the type of theses pointers take the argument list of the function they point to into account), and where the pointer has to be de-referenced to call the function. In MATLAB, the situation is not as complicated as in C but the function `integrate` would have to be rewritten to use the `feval` command to evaluate the function rather than to evaluate it intuitively.

The technical reason for the ease with which functions can intuitively be passed and used in Python, is that everything in Python (including numbers, functions and classes) are objects. Note that it would be easy to replace the function calls to the `integrate` function in the last lines of figure 4 by this for-each loop over the list of functions `f1` and `f2`:

```
for f in [f1,f2]:
    print "The approximation is", integrate( f, a, b, n)
```

Each object stores its name in an attribute `__name__`, so the above code can be modified to print not only the value of the integral of `f1` and `f2` respectively, but also the name `f1` and `f2`:

```
for f in [f1,f2]:
    print "The approximation of",f.__name__,"is",integrate( f, a, b, n)
```

A corresponding solution in MATLAB is far less elegant, and it requires significant effort (which we believe is outside the scope of first year students) to achieve this in C.

Acknowledgments. The author thanks A. J. Chipperfield, R. P. Boardman, M. Molinari and J. M. Generowicz for useful discussions.

References

1. Kernighan, B.W., Ritchie, D.M.: The C Programming Language. Prentice Hall Software Series (1988)
2. The Mathworks: Matlab (2003) `www.mathworks.com`.
3. Prechelt, L.: An empirical comparison of seven programming languages. IEEE Computer **33** (2000) 23–29
4. van Rossum, G.: Python tutorial. Centrum voor Wiskunde en Informatica (CWI), Amsterdam. (1995) `www.python.org`.
5. NumPy Team: Numerical Python (2003) `www.pfdubois.com/numpy`.
6. Donaldson, T.: Python as a first programming language for everyone. In: Western Canadian Conference on Computing Education. (2003) `www.cs.ubc.ca/wccce/Program03/papers/Toby.html`.

Teaching Computational Science Using VPython and Virtual Reality

Stephen Roberts[1], Henry Gardner[2], Shaun Press[3], and Linda Stals[1]

[1] Mathematical Sciences Institute,
Australian National University, Canberra ACT 0200, Australia,
[Stephen.Roberts,Linda.Stals]@anu.edu.au
[2] Department of Computer Science,
Australian National University, Canberra ACT 0200, Australia,
Henry.Gardner@anu.edu.au
[3] Research School of Information Sciences and Engineering,
Australian National University, Canberra ACT 0200, Australia,
Shaun.Press@anu.edu.au

Abstract. The Australian National University has two new complementary computational science programs, the Bachelor of Computational Science and the eScience graduate program. Students from the eScience program have developed sophisticated visualisation projects which have then been used to educate current and prospective undergraduate students in the Bachelor program. In this paper we will discuss the use of VPython combined with a 3D visualisation theatre, the Wedge, to produce hands-on computational science tutorials which we use to motivate computational science. We will briefly describe the use of VPython in our outreach tutorials, in particular a bouncing ball and gas simulation tutorial. The wedge virtual reality environment is also described as is the porting of VPython to the Wedge. Overall we provide a glimpse into our coordinated approach to using high level visualisation and virtual reality in the promotion of computational science.

1 Introduction

In this paper we describe how we use a combination of VPython and the "Wedge" (a virtual reality theatre) to introduce computational science concepts to K-12 students and teachers through to undergraduate and postgraduate students.

The Australian National University (ANU) is home to two novel and complementary education programs in computational science and related technologies, the undergraduate Bachelor in Computational Science (BComptlSci) degree program (and an associated major as part of the Bachelor of Science) [2] and the eScience graduate program [6]. Based on these educational programs we have built an outreach program for year 9-12 students and teachers. This program is exemplified by a hands-on VPython tutorial in which we introduce computational modelling by showing how to produce a simple gas simulation code. This tutorial culminates with students viewing the results of their coding in a

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1218–1225, 2004.
© Springer-Verlag Berlin Heidelberg 2004

3D virtual reality theatre, the Wedge, which shows how advanced visualisation can be used to complement scientific discovery (and be fun). The development of these outreach projects and other undergraduate projects have been a driver for some of the educational outcomes in the graduate program. For instance the porting of VPython to the Wedge was completed as a master level project with the eScience program.

In the subsequent sections we will give a more detailed description of these efforts to use sophisticated visualisation tools and environments to educate and enthuse people in the area of computational science.

2 The Undergraduate BComptlSci Program

The ANU Bachelor of Computational Science took in its first students in 2001. The degree is a joint initiative of the Departments of Mathematics and Computer Science, with support from the Department of Physics. The structure of the undergraduate program is representative of the various "churches" which make up computational science: Half of its program is delivered from a combination of courses in mathematics and computer science. The other half of the degree program is associated with a chosen application area such as physics, chemistry, environmental sciences, etc. In this way, students can accumulate enough credits to take an Honours year in a chosen discipline but also to have a healthy mix of programming, numerical mathematics, high-performance scientific computing and related technologies. Check the web site [1] for details of the core courses and structure of the degree program.

In 2003 we had our first three students graduate from the three-year undergraduate program. All three have decided to continue and undertake an extra Honours year in the program. We are averaging about ten students enrolling in the degree each year. The students in our program are developing very good problem-solving skills and they have been employed over the summer break to develop teaching modules and case studies. The number of students enrolled in our specific program is small, but the number of students in our core courses are healthy, supported by students undertaking a wide range of other majors, ranging from the classic areas of physics and chemistry, to earth and environmental sciences, to biology and bioinformatics.

We have found that the concept of "computational science" is quite hard to describe to even our own students. To help attack this problem we run regular seminars in which speakers from a wide range of areas speak on the use of computational science, both in research and in the wider community.

3 The Graduate eScience Program

Starting in 2001, the Computer Science department at ANU also introduced a new graduate program in "eScience" with the objective of delivering "conversion" course-work programs for students with a background in science and engineering [6]. On the one hand the program attempts to provide them with

a range of computing skills which will help them "convert" to being computational scientists in their chosen disciplines. On the other hand, the eScience program recognises that many graduates will aspire to having careers in the IT industry. Because of this, it attempts to provide them with a rigourous introduction to modern programming languages, software engineering, computer graphics, networking, human-computer interaction and high-performance computing optimisation. In this latter case, the computational science orientation of the program serves to act as a "pedagogical hook" to empathise with the backgrounds of the students and to, hopefully, afford better learning. The eScience program has a present enrolment of about 40 and altogether some 20 students will have graduated by the end of 2003. Graduates are finding employment in universities and local industry or are undertaking further studies.

4 Python and VPython in Computational Science Education

The understanding of computational science in the wider community is extremely poorly understood. To help remedy this, we have been involved in numerous outreach activities to raise the profile of computational science to year 9-12 students and teachers.

As part of our computational science program we have been building up a library of teaching modules and tutorials, [3]. These modules use a range of computing environments, but we have found that VPython provides a particularly good environment for people with little previous programming experience. In this section we will describe our experiences using VPython to develop a simulation of a gas. As part of this process we have also found that these teaching modules provide very nice project topics for our graduate eScience students. We describe such a project in the subsequent sections in which VPython has been ported to the Wedge, a virtual reality theatre. But for now let us provide a short introduction to VPython.

Visual Python (VPython) is a 3-D graphics system that is an extension of the Python language. Its main usage has been in the area of demonstration of physical systems in Physics, Chemistry, and Engineering. VPython [8] was written by Dave Scherer under the supervision of Bruce Sherwood and Ruth Chaby. The software is released under the GNU Public License and is available from the VPython website `http://www.vpython.org`.

The VPython module is written in C++ and built to fit seamlessly into the Python Environment. It is imported as a Python module and provides the user with the ability to model 3 dimensional scenes. The strength of VPython is that the programmer need not worry about the mechanics of how to build the display environment. Unlike other graphics environment (Java3D^{TM} for example), the programmer only has to specify the layout of the shapes used in any particular model, and VPython handles the rest. Fig. 1 show the results of our 50 line gas simulation VPython code.

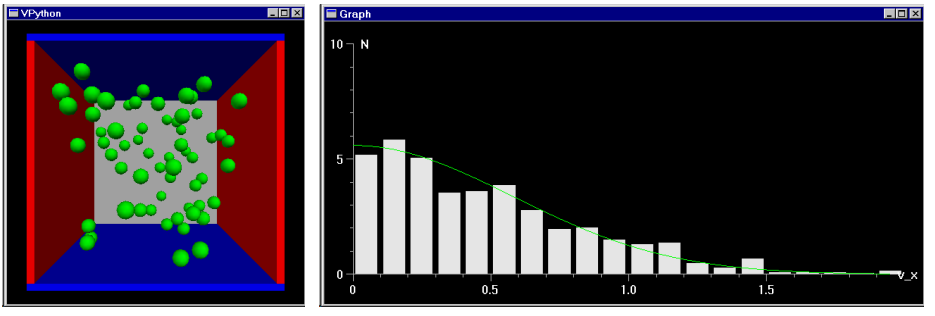


Fig. 1. Screen dump of the gas simulation tutorial showing particles inside a box, together with a graph showing the distribution of speeds of the simulated gas compared to the theoretical distribution

4.1 Bouncing Ball and Gas Simulation Teaching Modules

We have had great success in using the VPython bouncing ball example program as a basis for a tutorial for year 9-12 students and teachers. In the tutorial (see [4]), we lead the participants, who often have little or no programming experience, from creating a ball and wall in VPython, to creating a simulation of a ball bouncing in a box. This is a nice example as it introduces non-trivial programming concepts, such as objects, loops, conditionals, and numerical techniques such as Euler's method, all of which are motivated by the underlying physical problem.

As an extension we also have a gas-simulation tutorial [5], which starts from the bouncing ball code and ends with a simulation of a gas in which the distribution of the velocities of the gas can be observed and compared against the expected Maxwellian distribution. Fig. 1 shows the VPython visualisation of the gas. We find that the second tutorial allows participants to see how sophisticated models can be developed to a stage where useful intuition of the real system can be gained.

We have presented the bouncing ball tutorial to a number of different groups including year 9-10 students, year 11-12 students, Science/Mathematics teachers and Career Advisors. The overall response has been very positive; after attending a tutorial herself one of the teachers arranged for her students to be given the tutorial and we have been asked again, for the second year, to give the tutorial in the National Science Teachers Summer School. From our point of view it is very rewarding to listen to the participants discuss the tasks amongst themselves, ask us numerous questions and occasionally give shouts of joy as some problem is solved. For the most part the participants pick up on the ideas very quickly and keep their enthusiasm throughout the tutorial.

To date, all of the sessions have been held in computer labs at the university campus. In theory the tutorial can be given at the schools and other locations as all of the necessary software is available online. In fact, at the end of some of

the tutorials a CD was handed out containing a copy of VPython and the online tutorial.

We usually have about three staff members assisting with a group of 20-30 participants so any technical problems can be dealt with quickly and the participants can be guided through trouble spots. More importantly, as the tutorial progresses the participants often ask questions about how to extend the process to other contexts and it is at this point that we have the best opportunity to promote the role of computational science in the wider community.

The participants tend to pick up the notation used in VPython very quickly. They will often play around with the size, colour and speed of the balls, which is not specifically discussed in the tutorial. The main aspect of VPython that is not intuitive is the use of indents; if a participant is having problems compiling the code it is often because it was not indented correctly.

Observing the various groups we found that it is best if the participants have some elementary mathematics background. In particular, understanding vectors and the physics behind velocity is important. Many of the Year 9-10 students were not familiar with the 3D coordinate system and had some problems understanding the notation used in VPython. With some guidance they were able to complete the task, but more through a trial and error type of approach.

We have also found that it is important to stop the participants from jumping straight to the end of the instructions and searching for the answer. Many tend to take this approach rather than reading the directions given, so in the beginning it is necessary to "encourage" the participants to go through the tutorial step-by-step. Once they realise that they will get to build something in the tutorial they seem more willing to read the details to get a better understanding of the building blocks.

5 The Wedge

Virtual reality can be used as an important tool in scientific discovery, but can also be a great help in motivating the study of computational science. We use a 3D virtual reality theatre co-designed by one of the authors, Henry Gardner, as a core of the eScience program and as an environment for our outreach program. The Wedge [9] is a two-walled, walk-in, virtual reality theatre. It was originally built as a cheap alternative to the CAVE [7] for visualising scientific and engineering data. Participants look towards the vertex of the two screens and wear light, shutter-glasses which synchronise with the time-multiplexed images being calculated by the computer. The false-stereo effect makes images appear to float in space in front of the screens. Like the CAVE, the Wedge enables participants to see each other, the theatre and themselves as well as having a compelling experience of being "immersed" in the 3D graphics. Because of this, it is a good environment for collaborative virtual reality even though only one viewer has the ideal viewpoint. Built originally at ANU, Wedges have also been installed at two other Australian universities and in two science museums. A portable version of the Wedge has been used for K-12 outreach in country areas of New South Wales

and has been taken to conferences (one of which was on an island in the Great Barrier Reef!). A view of the largest Wedge constructed (the "WedgeOrama") is shown in Fig. 2.

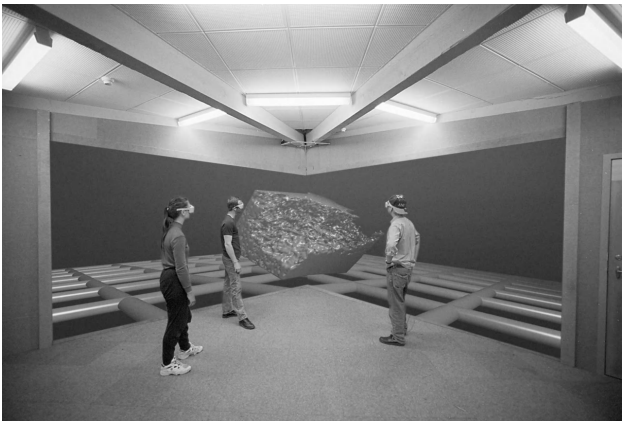


Fig. 2. View of the largest Wedge installation: the "WedgeOrama"

It is not possible to take any graphics application and display it in a straightforward fashion on the Wedge. Instead, care needs to be taken to ensure that, the software interface to the graphics hardware (commonly using the OpenGL™ application programming interface) can be controlled to ensure that the scene is rendered in stereo and that each stereo buffer for each eye is rendered with the correct perspective transformation. The software also needs to communicate the eye coordinates to the graphics renderer – either by assuming a default viewer position or by interpreting the signals from an external, tracked, device (such as a hat) worn by the lead viewer. Combining all of this with a remote, tracked, mouse and a keyboard extender and an 8-speaker surround sound system makes programming the ANU Wedge a sophisticated exercise in interaction design.

Over the years a number of software interfaces have been written to the Wedge. Students undertaking projects in the eScience program usually make use of one which is based on Java™ and Java3D™. This interface, known as the Tracked Interactive Wedge Interface or TIWI [11], has been developed to a sophisticated level but programmers need to have taken courses in Java™ and in computer graphics before they are able to begin programming with it. It is a further, exciting, pedagogical hook for our eScience teaching program but it does not translate readily to mainstream computational scientists.

As an example of a sophisticated Wedge application, Fig. 3 shows an eScience student using the Wedge to design paths in 3D space which are later used to position sounds and so construct a fully 3D sound scene. This application was written by the student for their eScience project [10].

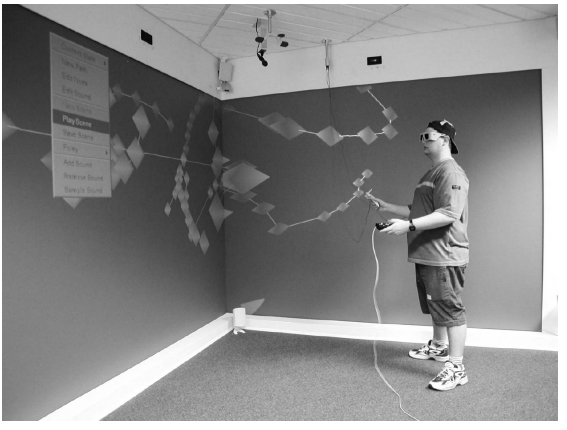


Fig. 3. An eScience student operating an interactive sound scene spatialisation program in the Wedge theatre

6 Porting VPython to the Wedge

As part of an eScience masters project, one of the authors, Shaun Press, ported VPython to the wedge. This provides an example of how the use of the Wedge and VPython provides a vehicle for our education program, in this case at the graduate level.

The porting of VPython turned out to be an interesting challenge. The basic rendering in VPython uses OpenGL, but as it turned out, VPython has only a limited dependence upon OpenGL. It utilises OpenGL functions for drawing and colouring shapes, but very little else. In a standard OpenGL implementation the majority of these tasks would be carried out via OpenGL calls. In VPython most of these functions are calculated via matrix multiplication operations to create the final product, which is the physical vertices for the shape.

To port VPython to the Wedge required a modification to the VPython rendering routines. Rather than rendering a single scene to the screen, the scene had to be transformed and rendered 4 times. Each scene was broken up into an eye/screen pair (ie Left Eye/Left Screen, Right Eye/Right Screen etc). The steps for rendering are now

- Switch stereo on
- Calculate eye position and eye offset for each eye/screen pair.
- Transform the screen for each eye/screen pair
- Render scene in OpenGL stereo buffers
- Swap OpenGL buffers to front

Code for handling Mouse and Keyboard input was also added. Support for the Domino interface was added, with it's major purpose to mimic the zoom and rotate functions of a desktop mouse.

In addition to porting VPython to the Wedge, we have also made a number of other extensions to the VPython. Additional geometries (Pyramid, Ellipsoid) have been added with other additions designed to extend the capabilities of VPython. At the time of writing a Particle Cloud object is being tested, with the intention that it be included in the next release of VPython. This object will help with the modelling of particle systems with VPython.

7 Conclusion

Our aim is to continue to leverage our investments in visualisation and VR to provide sophisticated projects for our eScience graduates which can be used in our outreach and undergraduate programs.

Acknowledgements. We would like to thank Sally Lloyd and Hugh Fisher for their help in developing the VPython tutorial and in porting VPython to the Wedge. The BComptlSci program was supported as part of the Australian Partnership for Advanced Computing Education program and the eScience program has been supported by the Australian Government's Science Lectureships Initiative.

References

1. ANU Computational Science Program. Bachelor of Computational Science Core Courses. `http://www.maths.anu.edu.au/bcomptlsci/core`, 2003.
2. ANU Computational Science Program. Bachelor of Computational Science Home Page. `http://www.maths.anu.edu.au/bcomptlsci`, 2003.
3. ANU Computational Science Program. Teaching Modules and Tutorials. `http://www.maths.anu.edu.au/comptlsci/tutorials.html`, 2003.
4. ANU Computational Science Program. VPython Bouncing Ball Tutorial. `http://www.maths.anu.edu.au/comptlsci/Tutorial-Gas/tute-bounce.html`, 2003.
5. ANU Computational Science Program. VPython Gas Simulation Tutorial. `http://www.maths.anu.edu.au/comptlsci/Tutorial-Gas/tute-gas.html`, 2003.
6. ANU eScience Program. Home Page. `http://eScience.anu.edu.au`, 2003.
7. Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. De-Fanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *SIGGRAPH 93*, pages 135–142, New York, August 1993. Association of Computing Machinery.
8. David Scherer and Bruce Sherwood and Ruth Chaby. VPython Home Page. `http://www.vpython.org`, 2003.
9. Henry Gardner and Rod Boswell. The wedge virtual reality theatre. In *Proceedings of the Apple University Consortium Conference*, pages 23–26, New York, August 2001. Association of Computing Machinery. `http://auc.uow.edu.au/conf/conf01/downloads/AUC2001_Proceedings.pdf`, pp 2.1-2.6.
10. Rod Harris. escience project report, 2001.
11. David Walsh. The tracked interactive wedge interface. `http://ephebe/tiwi/index.html`, 2003.

Student Exercises on Fossil Fuels, Global Warming, and Gaia

Bert W. Rust

National Institute of Standards and Technology, 100 Bureau Drive, Stop 8910, Gaithersburg, MD 20899-8910, USA
bert.rust@nist.gov

Abstract. In a recent series of tutorial papers, Rust [7,8,9,10] modelled measured time series for global temperatures and fossil fuel CO_2 emissions with related combinations of polynomials, exponentials, and sinusoids, using least squares fits that can be done by students well grounded in practical statistics. The analysis suggested that temperatures cycle around a monotonically increasing, accelerating baseline with a period of ≈ 65 years and that the exponential growth rate of CO_2 emissions varies inversely with this cycle. The *Gaia hypothesis* [5] suggests that the biosphere adjusts atmospheric greenhouse gases to maintain an optimal temperature for life. The previous analysis is here extended with a series of fitting exercises designed to demonstrate that the above described inverse relation represents a *Gaian* feedback.

1 Introduction

Figure 1 gives a plot of annual global total fossil fuel CO_2 emissions, measured in millions of metric tons of carbon [MtC], for the years 1856-1999. These data, which were compiled by Marland, et. al. [6], can be found at *http://cdiac.ornl.gov /trends/emis/em_cont.htm*. The dashed curve is a nonlinear fit [9] of the model

$$P(t - t_0) = P_0 e^{\alpha(t-t_0)} \ , \tag{1}$$

where $P(t)$ is the total emission in year t, $t_0 = 1856.0$, and P_0 and α are free parameters. The observations undulate systematically around the fit, and a Fourier analysis of the residuals indicated a cycle with period ≈ 65 years. The amplitude of the cycle appeared to increase in time with the same exponential rate as the fitted curve. This suggested a model of the form

$$P(t - t_0) = \left\{ P_0 + P_1 \sin\left[\frac{2\pi}{\tau}(t - t_0 + \theta)\right] \right\} e^{\alpha(t-t_0)} \ , \tag{2}$$

with free parameters $P_0,\ P_1,\ \alpha,\ \tau,\ \theta$. Fitting this model [10] gave estimates

$$\begin{gathered} \hat{P}_0 = 133.8 \pm 5.0\ [MtC]\ , \quad \hat{P}_1 = 25.4 \pm 1.6\ [MtC]\ , \\ \hat{\tau} = 64.9 \pm 1.5\ [\mathrm{yr}]\ , \quad \hat{\theta} = 26.6 \pm 2.7\ [\mathrm{yr}]\ , \quad \hat{\alpha} = 0.02814 \pm .00034\ [\mathrm{yr}^{-1}]\ . \end{gathered} \tag{3}$$

The fit, plotted as a solid curve, explains 99.73% of the variance in the data.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1226–1233, 2004.
© Springer-Verlag Berlin Heidelberg 2004

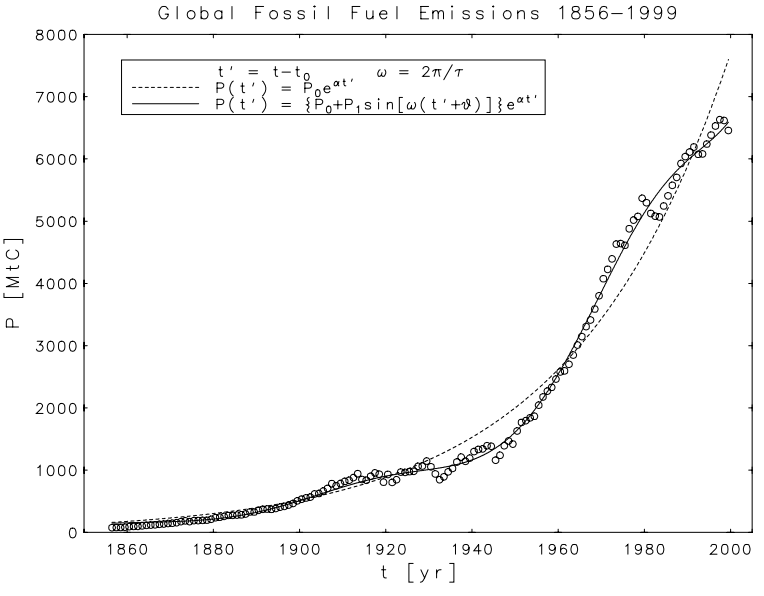


Fig. 1. Annual global total fossil fuel CO_2 emissions are plotted as discrete circles. The dashed curve is the fit of the model (1), and the solid curve is the fit of (2).

Figure 2 is a plot of annual global average tropospheric temperature anomalies (in °C) for the years 1856-2001. The anomalies were obtained from the measured temperatures by subtracting the average temperature for 1961-1990. These data, which were calculated by P. D. Jones, et. al. [2], can be obtained online at *http://www.cru.uea.ac.uk/cru/cru.htm*. Low order polynomials fits [7, 8] suggested an accelerating warming with a monotonically increasing baseline. If $T(t)$ is the anomaly in year t, then warming with constant acceleration can be modelled by

$$T(t - t_0) = T_0 + C_2(t - t_0)^2 \ , \tag{4}$$

with free parameters T_0 and C_2. Including a linear term did not significantly reduce the sum of squared residuals [8]. A slightly better fit was obtained [9] for

$$T(t - t_0) = T_0 + C_2 \exp[0.01407(t - t_0)] \ , \tag{5}$$

where the rate 0.01407 [yr^{-1}] is one half of the $\hat{\alpha}$ in (3). The residuals for both fits oscillated with period $\tau \approx 63$ years [8,9,11] which is close to the $\hat{\tau}$ in (3). Also, the oscillation was a half cycle out of phase with the emissions cycle, with maximum emissions corresponding to minimum temperatures, and vice versa. This temperature cycle was previously noted by Schlesinger and Ramankutty [14], and the inverse correlation between variations in temperature and emissions was previously noted by Rust and Kirk [13]. Their work was extended by Rust and Crosby [12] who argued that the correlation arises from a *Gaian* feedback by which increasing temperatures reduce the growth in fossil fuel production.

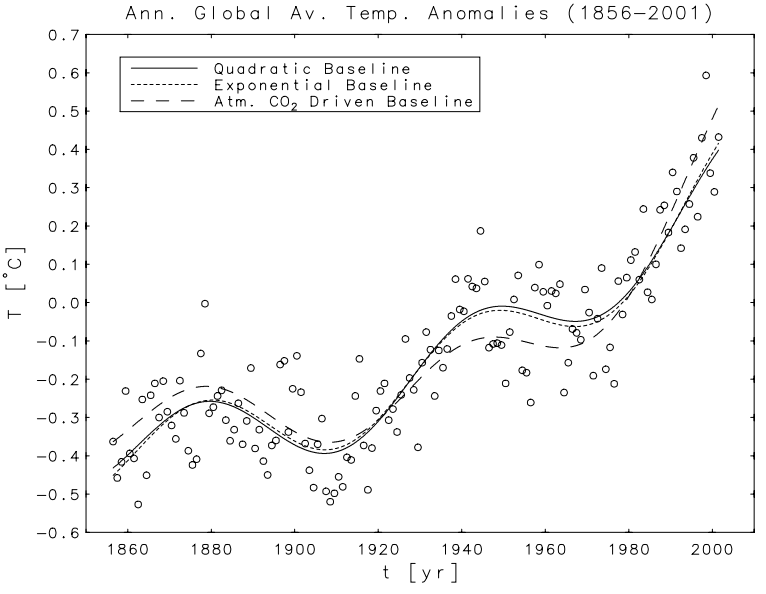


Fig. 2. The circles are observed annual global average temperature anomalies. The solid curve is the fit of the quadratic/sinusoidal model (6), the finely dashed curve is the fit of the exponential/sinusoidal model (7), and the coarsely dashed curve is the fit of the atmospheric-CO_2-driven/sinusoidal model (10).

Adding the temperature cycle to the models (4) and (5) gives

$$T(t-t_0) = T_0 \; + \; C_2(t-t_0)^2 + A_1 \sin\left[\frac{2\pi}{64.9}(t-t_0-5.85)\right] \quad , \tag{6}$$

and

$$T(t-t_0) = T_0 + C_2 \exp\left[0.01407(t-t_0)\right] + A_1 \sin\left[\frac{2\pi}{64.9}(t-t_0-5.85)\right] \quad , \tag{7}$$

with free parameters T_0, C_2, and A_1. The period of the sinusoid is exactly the $\hat{\tau}$ in (3), and the phase constant has been set to $\hat{\theta} - (\hat{\tau}/2) = -5.85$ to make the oscillation exactly one half cycle ahead (or behind) the one for the emissions. Fitting these two models [10] gave the parameter estimates in columns 2 and 3 of Table 1 and the solid and finely dashed curves in Fig. 2.

2 Exercises 1 and 2: Precedence and Causality

A 64.9 year cycle in both the temperatures and the emissions cannot be a coincidence. CO_2 is a greenhouse gas, so its accumulation in the atmosphere is expected to cause some warming. But is the cycle an emissions effect vectored into the temperatures, or is it a *Gaian* feedback by which increasing temperatures

Table 1. Estimates and coefficients of determination for models (6), (7), and (10).

Param.	Quad. Baseline(6)	Expon. Baseline (7)	CO_2-driven Baseline (10)
$\hat{T}_0$	$-0.380 \pm .013$	$-0.510 \pm .017$	$-3.25 \pm .14$
$\hat{C}_2$	$(3.27 \pm .14) \times 10^{-5}$	$0.1094 \pm .0045$	$(4.66 \pm .21) \times 10^{-6}$
$\hat{A}_1$	$0.105 \pm .013$	$0.104 \pm .013$	$0.090 \pm .013$
R^2	0.8071	0.8081	0.7802

supress fossil fuel consumption? One indication of causality would be a definite time delay between the cycles in the two records. If the phase constants -5.85 in (6) and (7) are replaced by free parameters ϕ_q and ϕ_e, then two 4-parameter linear least squares fits give estimates $\hat{\phi}_q = -4.75 \pm 1.17$ and $\hat{\phi}_e = -4.80 \pm 1.18$. These estimates indicate a nominal lag of approximately one year between a temperature extremum and the opposite extremum in the emissions. But -5.85 falls inside the $\pm 1\sigma$ uncertainty intervals for both estimates, so that lag is not statistically significant. Also, for both fits, an F-test indicates that the additional free parameter does not produce a statistically significant reduction in the sum of squared residuals. Thus the data cannot reliably determine precedence for the two cycles.

3 Exercise 3: Atmospheric Accumulation of CO_2

High quality measurements of global atmospheric CO_2 concentrations, made by C. D. Keeling, et. al. [4], can be found at *http://cdiac.ornl.gov/ftp/maunaloa-co2/maunaloa.co2*. Yearly averages for 1959-2001 are plotted in Fig. 3. It is generally accepted that $\approx 50\%$ of the fossil fuel CO_2 emissions remains in the atmosphere, so a model for atmospheric accumulation $c(t)$ can be written

$$c(t - t_0) = c_0 + \gamma \int_0^{t-t_0} P(t')dt' \ , \tag{8}$$

where c_0 is the concentration at epoch 1856.0, $P(t')$ is the emissions total in year t', and γ is the fraction remaining in the atmosphere. Substituting (2) gives

$$c(t - t_0) = c_0 + \gamma \int_0^{t-t_0} \left\{ \hat{P}_0 + \hat{P}_1 \sin\left[\frac{2\pi}{\hat{\tau}}(t' + \hat{\theta})\right] \right\} e^{\hat{\alpha}t'} dt' \ , \tag{9}$$

where $\hat{P}_0$, $\hat{P}_1$, $\hat{\tau}$, $\hat{\theta}$, and $\hat{\alpha}$ are the estimates given in (3). It is an easy linear least squares exercise to fit this model, with free parameters c_0 and γ, to Keeling's

data. The integral could be evaluated in closed form, but it is easier to compute it with a numerical integrator. The only complication is that the units of the atmospheric measurements are [ppmv], but the units of $\hat{P}_0$ and $\hat{P}_1$ are [MtC]. The conversion factor [15] is 1 [ppmv] = 2.130×10^3 [MtC]. The strategy used here was to convert the measurements to [MtC] to obtain the estimates $\hat{c}_0 = (6.2899 \pm .0058) \times 10^5$ [MtC] and $\hat{\gamma} = 0.572 \pm .032$. The fit, which explained 99.87% of the variance, was converted back to [ppmv] units to be plotted as the smooth curve in Fig. 3. Clearly the 64.9 year cycle was totally smoothed out by the cumulative quadrature, so its appearance in the temperature record cannot be due to its presence in the emissions record.

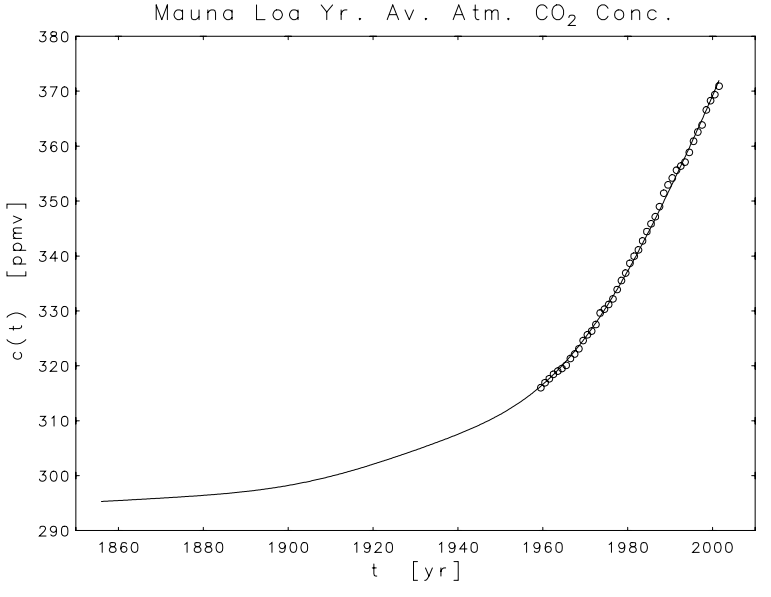


Fig. 3. The circles are the yearly average atmospheric CO_2 concentrations in units of part per million by volume [ppmv] for the years 1959-2001. The curve was obtained by fitting (9) to the data and extending the fit back to 1856.0.

4 Exercise 4: The Atmospheric Temperature Connection

Conventional wisdom holds that the response of global temperatures to increasing atmospheric CO_2 can only be estimated within the context of sophisticated climate models. But reconstructions [1,3] from Antarctican ice cores of atmospheric temperature changes and atmospheric CO_2 concentrations for the past 160,000 years have shown that the two quantities have varied proportionately, and in lock step, through a temperature range of length 12.5°C and a concentration range of length 100 ppmv. This suggests a model of the form

Table 2. Parameter estimates and coefficients of determination for three fits of (12) to the CO_2 emissions, using temperature feedback models (6), (7), and (10).

Param.	Quad. Baseline(6)	Expon. Baseline (7)	CO_2-driven Baseline (10)
$\hat{P}_0$	78.1 ± 3.2	93.5 ± 3.7	69.7 ± 4.4
$\hat{\alpha}$	$0.03862 \pm .00050$	$0.03804 \pm .00051$	$0.03758 \pm .00068$
$\hat{\beta}$	$1.481 \pm .053$	$1.345 \pm .052$	$0.994 \pm .053$
R^2	0.9963	0.9958	0.9930

$$T(t - t_0) = T_0 + C_2\, c(t - t_0) + A_1 \sin\left[\frac{2\pi}{64.9}(t - t_0 - 5.85)\right] \quad . \tag{10}$$

where $c(t - t_0)$ is the function plotted as a smooth curve in Fig. 3, and T_0, C_2, and A_1 are free parameters, with T_0 here being not $T(t_0)$ but rather the temperature anomaly for an atmosphere with no CO_2. This model assumes that the 64.9 year cycle truly is a weather/climate effect that is superposed on the CO_2-driven baseline. Fitting it to the temperature anomalies is another easy linear least squares problem, given the software written for the previous exercise. The parameter estimates are given in the last column of Table 1, and the fit is plotted as the coarsely dashed curve in Fig. 2.

5 Exercises 5-7: The Gaian Feedback

Since the 64.9 year temperature cycle is not caused by the CO_2 emissions, the inverse cycle in the emissions must be caused by a negative, temperature dependent feedback. Rust and Kirk [13] modelled this feedback with the differential equation

$$\frac{dP}{dt} = \left(\alpha - \beta\frac{dT}{dt}\right) P \,, \quad P(t_0) = P_0 \ , \tag{11}$$

where α, β, and P_0 were free parameters. With the shift of the zero point to t_0, the solution becomes

$$P(t - t_0) = P_0 \exp\left\{\alpha(t - t_0) - \beta\left[T(t - t_0) - T(0)\right]\right\} \quad . \tag{12}$$

Three exercises, which use the cylcle to confirm the feedback, are to fit this equation to the emissions data using the estimates from Table 1 in their corresponding models (6), (7), and (10) for $T(t - t_0)$. These are easy nonlinear least

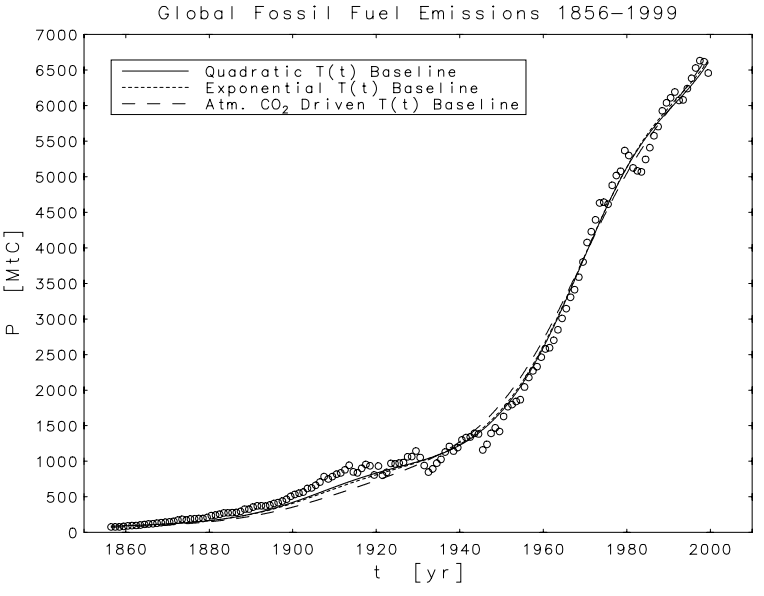


Fig. 4. Three temperature feedback model fits to the fossil fuel emissions record.

squares problems. The results are given in Table 2 and Fig. 4. None of the fits is quite as good as the solid curve in Fig. 1, which had $R^2 = 0.9973$, but all are good enough to lend strong support to the feedback hypothesis.

6 Exercise 8: For Advanced Students

With the fit of the CO_2-driven version (10) of the model (12), in the last exercise above, the analysis has completed a full circle from the purely mathematical representation (2) for the emissions to the dynamical representation (11). The agreement between the two fits, i.e. between the solid curve in Fig. 1 and the coarsely dashed one in Fig. 4, is not perfect, but it is good enough to argue not only that the feedback hypothesis is correct, but also that the intermediate hypotheses (8) and (10) are least approximately correct. These results, taken together, suggest the following dynamical models for $P(t)$, $c(t)$ and $T(t)$:

$$\begin{aligned} \tfrac{dP}{dt} &= \alpha P - \beta \left\{ \eta P + A \cos\left[\tfrac{2\pi}{\tau}(t+\phi)\right] \right\} P \;, \quad P(t_0) = P_0 \;, \\ \tfrac{dc}{dt} &= \gamma P \qquad\qquad\qquad\qquad\quad , \quad c(t_0) = c_0 \;, \\ \tfrac{dT}{dt} &= \eta P + A \cos\left[\tfrac{2\pi}{\tau}(t+\phi)\right] \qquad , \quad T(t_0) = T_0 \;, \end{aligned} \tag{13}$$

with unknown parameters α, β, γ, η, A, τ, ϕ, P_0, c_0, and T_0. The exercise, left to the student, is to fit these three equations simultaneously to the three data sets.

References

1. Barnola, J. M., Raynaud, D., Korotkevich, Y. S., Lorius, C.: Vostok ice core provides 160,000-year record of atmospheric CO_2. Nature, **329** (1987) 408-414
2. Jones, P. D., Moberg, A.: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. Jour. Climate, **16** (2003) 206-223
3. Jouzel, J., Lorius, C., Petit, J. R., Genthon, C., Barkov, N. I., Kotlyakov, V. M., Petrov, V. M.: Vostok ice core: a continuous isotope temperature record over the last climatic cycle (160,000 years). Nature, **329** (1987) 403-408
4. Keeling, C. D., Whorf, T. P.: Atmospheric carbon dioxide record from Mauna Loa. in Online Trends: A Compendium of Data on Global Change, http://cdiac.ornl.gov/trends/co2/contents.htm, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN (2003)
5. Lovelock, J. E., Margulis, L.: Atmospheric homeostasis by and for the biosphere: the gaia hypothesis. Tellus, **26** (1974) 2-10
6. Marland, G., Boden, T. A., Andres, R. J.: Global, Regional, and National CO_2 Emissions. in Trends: A Compendium of Data on Global Change, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN (2000)
7. Rust, B. W.: Fitting nature's basic functions Part I: polynomials and linear least squares. Computing in Science & Engineering, **3 nr 5** (2001) 84-89
8. Rust, B. W.: Fitting nature's basic functions Part II: estimating uncertainties and testing hypotheses. Computing in Science & Engineering, **3 nr 6** (2001) 60-64
9. Rust, B. W.: Fitting nature's basic functions Part III: exponentials, sinusoids, and nonlinear least squares. Computing in Science & Engineering, **4 nr 4** (2002) 72-77
10. Rust, B. W.: Fitting nature's basic functions Part IV:the variable projection algorithm. Computing in Science & Engineering, **5 nr 2** (2003) 74-79
11. Rust, B. W.: Separating signal from noise in global warming. Computing Science and Statistics, **35** (2003) in press
12. Rust B. W., Crosby, F. J.: Further studies on the modulation of fossil fuel production by global temperature variations. Environment International, **20 nr 4** (1994) 429-456
13. Rust B. W., Kirk, B. L.: Modulation of fossil fuel production by global temperature variations. Environment International, **7** (1982) 419-422
14. Schlesinger, M. E., Ramankutty, N.: An oscillation in the global climate system of period 65-70 years. Nature, **367** (1994) 723-726
15. Watts, J. A.: The Carbon Dioxide Question: Data Sampler. in Clark, W. C., editor, Carbon Dioxide Review: 1982, Oxford University Press, New York (1982) 431-469

Teaching Scientific Computing

B.A. Shadwick[1,2]

[1] Institute for Advanced Physics
10875 US Hwy. 285, Suite 199, Conifer, C0 80433 USA
[2] Center for Beam Physics
LBNL, Berkeley, CA 94720, USA
BAShadwick@IAPhysics.org

Abstract. We consider the wide scale use of computers in scientific research and the implications for the educational curriculum. We examine the practice of scientific computing from the perspective of obtaining robust numerical results. We explore the philosophy underlying this process and isolate the essential elements. We discuss in detail practices that are designed to minimize the likelihood of errors and also survey a set of supporting skills. Throughout, examples are used to illustrate the key ideas.

1 Introduction

Less than a generation ago, the routine use of computers in scientific research was largely restricted to specialists. CPU time was metered and in reasonably tight supply for all but the largest institutions. Modern tools and the tremendous computing power[1] that is routinely available to researchers has resulted in widespread, almost casual, use of computing in the physical sciences. A greater proportion of scientists are part-time practitioners of numerical computing, and numerous software packages make it seemingly easy to the casual user to do elementary numerical computations: to evaluate functions for graphing or for other purposes; to numerically (or symbolically) evaluate integrals; to solve ODEs and simple PDEs and so forth. Regrettably many of these scientists have little understanding of the mechanics underlying these numerical results. The scale of problem that the non-specialist is willing to take on scales with the growth in desktop computing power. This trend began with the general availability of desktop systems some two decades ago. It exists as a consequence of the natural human tendency to categorize; we make the (artificial) distinction between what we do on our desktop computers and what "computational physicists" do with supercomputers. We tend to view users of truly large-scale computing as specialists who have the detailed knowledge to perform these grand calculations, whereas the researcher who does the occasional computation using the a computer which is also used for email and writing papers does not need (or often even

[1] Of course, a college freshman might not consider today's computers necessarily that remarkable. Context is important.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1234–1241, 2004.
© Springer-Verlag Berlin Heidelberg 2004

want) such specialized knowledge. The code that runs today in a few minutes on a workstation would have taken hours (if not days) on the central systems commonly in use twenty-five years ago. If we cannot distinguish specialists from non-specialists through the kinds of problems the address, then how can we differentiate them? Of course, there is no distinction and herein lies the potential problem.

As in society at large, scientists have come to view computers as mere appliances. Fortunately, it is also becoming more common for graduate and undergraduate programs to include courses in numerical methods. At the same time there is still a perception (probably less so in engineering than in, say, physics) that only students who plan to specialize in numerical work need to take such courses.

It is increasingly common for students to take on "large" computational projects with less-than-adequate supervision. Most faculty members would be reluctant to suggest a student tackle an analytical problem in an area where the advisor has little or no experience. Unfortunately, this same reticence does not extend to suggesting a student undertake numerical work, even when neither the advisor nor the student has significant computational experience. While there is something to be said for the "sink or swim" philosophy, failing to recognize the inherent complexity and subtlety of obtaining correct computational results does a grave disservice to students.

These observations argue that all students need more than a cursory introduction to numerical methods. Though impractical, computational methods should ideally be viewed as a mathematical tool as important as calculus, and receive similar weight in the curriculum. There is, for lack of a better term, a core process (and an accompanying mindset) by which we obtain computational results that we can believe. By concentrating on this crucial core, it is still possible to construct a course of study which the curriculum will bear.

We begin with a detailed examination of these processes, while providing examples which illustrate the key ideas. We turn next to broad survey of some essential skills and specialized knowledge required. By the end, we will have identified the essential elements, and hopefully have outlined a means to impart our process and our mindset to our students.

2 Getting the Right Answer

In this section we will examine the process by which we can obtain computational results in which we have confidence. To provide a reference, the discussion will be geared towards the task of obtaining numerical solutions of differential equations. Partial differential equations are ubiquitous in the physical sciences and students are often faced with formidable PDEs from which they must wrest a numerical solution. While the philosophy presented here is generic to scientific computing; some of the details will be different when considering numerical tasks other than solving differential equations.

Perhaps this section could be more accurately titled "Avoiding Wrong Answers" as that is actually what one does. It is quite rare for one to be able to claim that a numerical calculation is "right," since such a claim would imply an independent, exact (*i.e.*, an analytical) solution to the problem in question. If such a solution exists, then why the recourse to numerical solutions? So what do we meant by "the right answer"? The best we can normally accomplish is a belief that our computer code is not wrong. That is, at best, we find no evidence that the result of our calculation is wrong; we have successfully performed a series of tests — typically running the code with sets of parameters for which something is known about the solution (more on this below) — which show that the code is functioning as we expect. We then extrapolate this ostensively correct behaviour to cases where we know less about (but have more interest in) the solution. This is what we mean by our use of the term "right."

We do not wish to be overly pessimistic about computational results, but it is essential that we are mindful of the nature of the endeavour. In some respects, numerical computation has more in common with experimental science (in terms of assessing the outcome) that it does with analytical work[2]. While numerical work is "theoretical," the required mindset more closely matches that of the experimentalist than the theoretician. An analytical calculation always "works"; one either finishes the calculation, or reaches a point where further progress appears impossible. The point where difficulties arise is clear and the calculator is typically aware of the step which serves as the stumbling block. An experiment, however, has a different character. In some cases, even though the apparatus is functioning as the experimentalist expects, there is no signal, *i.e.,* the experiment doesn't work. Even when there is a signal, the experimentalist must determine whether this signal has the meaning intended when the experiment was designed.

The only demonstrable evidence that something is badly wrong with a numerical calculation occurs when the code crashes (*i.e.,* the computer operating system terminates the program due to bad behaviour) or the output contains some quantity of `NaN`s. Absent these difficulties, some output is always obtained. Thus we are always faced with the question "Does the output of this code mean what we think it means?" The importance of recognizing this question is inherent in *all* numerical computation and cannot be over emphasised. We are always at risk of being misled by the output of our codes. There is a natural predisposition to believe all of those digits before us have some intrinsic meaning. Compiler errors and errors in the internal implementation of the machine instructions are thankfully quite rare.[3] The output of a computer program is almost certainly in

[2] The dreadful term "numerical experiment" *should not* come to mind in this context. Numerical computations are no more or no less than analytical theory; in both cases, the goal is to determine the predictions of a particular model. The term "numerical experiment" seems intended to elevate the stature of computing beyond that of other theoretical tools. Numerical "experiments" are not experiments. Unlike a *real* experiment, which always entertains the possibility of the discovery of new physical phenomena beyond what is already known, the *output of a numerical computation can contain no more physics than the input.*

[3] The infamous Pentium division bug of the mid 1990's notwithstanding.

precise agreement with what the programmer instructed (or *thought* he instructed) the computer to calculate. It is easy, especially for beginning practitioners, to confuse this correctness with the question of whether or not the algorithm as implemented (or as envisioned) does indeed solve the problem under consideration, given the current set of computational parameters. As Acton points out [1], most errors in computational results are ultimately due to human errors. Only by recognizing this fact can we instill practices and attitudes which will work to mitigate such errors.

We minimize the likelihood of errors by performing tests which indicate whether the numerical solution is *consistent* with the solution of the problem at hand. We will now discuss some techniques for building confidence in the results of a computation, as well as some general guidelines for avoiding problems.

Validation. The construction of effective tests is critical to verify proper operation of numerical codes. There are two classes of tests: those designed to test that the algorithm implemented in the code solves the problem it was designed to solve (referred to as *benchmarks*), and those designed to test the implementation logic (this class of tests is often referred to as *regression tests* by software engineers). While both classes of tests overlap in terms of the functionality they probe, the philosophy behind the two classes is markedly different.

Regression tests are primarily used to verify that alterations to a program do not result in unintended consequences. Regression tests are used to identify changes in behaviour over time; their purpose is to verify that a given set of inputs produces the same outputs today as yesterday. The cases used for regression testing do not have to produce meaningful output, only consistent output. The parameters do not need to be physically reasonable, nor does the resolution need to be fine enough to provide converged results. Consequently, regression tests can be very short runs that take small amounts of CPU time. Due to their nature, regression tests cannot be used to validate a code in an absolute sense. After a particular version of a code has been validated with benchmarks, the regression tests are run, the output is deemed to be correct, and these test results are then used for future comparisons. It is not uncommon for even moderate-sized codes to be run through a nightly set of regression tests. In any event, after code modifications, regression testing should be performed before any "production runs" are made.

Benchmarks, on the other hand, are meant to verify that a program solves its intended problem. This is normally accomplished by considering parameter regimes (*e.g.*, small amplitude excitation) where either exact or approximate solutions exist, or where specific properties of the solution are known. For example, a linearized version of the model equations may be exactly soluble or parameters (or initial conditions) may exist which effectively reduce the dimensionality of the problem. The most valuable tests are those where an exact solution exists in a certain limit; these tests tend to be quite sensitive to algorithmic errors. Additionally, there are often analytical constraints on the solution (for example conservation of energy) which should be respected by the numerical solution to

a level consistent with the algorithm. Generally, numerical solutions should be tested for agreement with all known analytical properties.

Recently, we encountered a rather serious example of poor benchmarking. There is a simple analytical relationship between the values of put and call options (so called put-call parity), yet we discovered that the software for computing options prices used by a trading desk in an investment bank produced results that significantly violated this constraint. This was a production setting: the software was used for negotiating contracts and to determine prices both for internal auditing purposes. An error of this nature should never have gone undetected by any reasonable testing procedure.

Convergence. Verifying convergence of the numerical solution is critically important; except for certain classes of ODEs and linear PDEs, there is no general theorem regarding the convergence of the solution of finite difference equations as the discretization parameters tend to zero.[4] While there is no guarantee that the numerical solution will converge to the analytical solution (assuming such exists), a necessary condition for the numerical solution to be consistent with the analytical solution is that the numerical solution tends to a limit as the finite difference cell size tends to zero. Moreover, verifying that the convergence rate is consistent with the order of the discretization is a valuable test of a code and can uncover subtle errors. The most important reason for studying convergence is as a means to determine the resolution necessary to produce results of acceptable accuracy. It is quite difficult to determine, *a priori*, the necessary resolution. An example of this presented itself recently in our work related to the propagation of an intense laser pulse in a long plasma. One would normally suppose that a resolution of, say, ten to twenty grid-points per laser cycle would be sufficient to obtain reasonable results. Studying the convergence of the solution led to the surprising discovery that a resolution on the order of one hundred grid-points per cycle was necessary to obtain the correct results. At lower resolutions, the solution appeared quite reasonable, but in fact, the laser was coupling too strongly to the plasma and transferring energy considerably more rapidly than was physically correct. While the answer appeared plausible, it contained a *qualitative error*: the graph of laser energy *versus* time had the wrong concavity. This error was only uncovered by performing a series of runs that essentially produced a fully converged result. Often one hears comments to the effect "I doubled the number of grid-points and the solution only changed by ten percent, so that's good enough." While the asymptotic order of the error terms will be known from formal analysis, convergence rates will only match asymptotic estimates for sufficiently "small" grid sizes. Only by examining the changes in the solution over a series of step sizes is it possible to determine if the step sizes involved are such that the asymptotic error scaling applies. Furthermore, knowing that asymptotic scaling applies, in conjunction with observed convergence rates, allows for global error estimates.

[4] Clearly such a theorem would imply an existence theorem for the solution which explains the dearth of results for nonlinear PDEs.

Algorithms. It is an exceedingly bad idea to modify the model equations solely for the convenience of numerical solution (an exception to this injunction is the conversion of a PDE to integral form from which a weak solution is obtained). For the most part, simplifications which appeal to the analytical eye do not necessarily make the numerical solution "easier" to obtain. Arithmetical complexity is not generally an important concern. It is certainly of questionable worth to reduce the validity of the model (say by linearizing certain terms) to "make it simpler for the computer." The complexity of the calculation is only at issue as it relates to amount of CPU time used to obtain a solution. Rarely is "making the equation easier to program" a valid consideration.[5] In any case, it makes the most sense to implement the full model. After that code is properly working one can explore any "simplifications" which may prove advantageous. Experience suggests that it is universally easier to impose new assumptions than to remove existing ones.

As a general principle, there must always exist a control parameter (which is part of the input data) that in some limit takes the discretized equations into formal agreement with the model equations. Grid sizes and time steps introduced in the finite difference solutions of differential equations are examples of such control parameters. As an illustration of the danger of *uncontrolled approximations*, consider the following: the wave equation for a laser pulse propagating in a plasma was modified to facilitate solution, and the essence of the modification was that the formal expression for an inverse of the differential operator was approximated in Fourier space by expansion in a power series. Only if *all* terms were kept was there formal agreement with the unmodified equations. The code in question kept only the first term in this expansion, with the result that the group velocity of the laser pulse, after a long propagation distance, was grossly incorrect (as was the final spatial position of the pulse). In the absence of a control parameter, there was no easy way to discover that the program was, in fact, solving a very different equation than the author of the program intended. Provided the propagation distance was small enough, this error could be considered tolerable. Of course, as is virtually always the case, the original restriction on propagation distance imposed by this modification was forgotten, and the code was used in a regime where it was invalid. This discrepancy was only found when an independent group compared the (published) erroneous solutions with the results from a code which made no such approximations. It is worth noting that reliable algorithms for solving the unmodified equations were readily available, but apparently unknown to the researchers. (In the final analysis, in addition to producing the wrong answer, the modifications in question made the equations *harder* to solve numerically.) Modification of model equations should not be thought of as a numerical technique; there are likely assumptions about the solution and/or input parameters which are necessary for solutions of the modified model to approximate solutions of the original model. Above all, modifying an equation should be considered a last, desperate, resort.

[5] With the use of various code generation techniques, arguably, simplicity of implementation is *never* a valid consideration.

3 General Knowledge

In addition to the techniques discussed in the previous sections, there also exists a general body of knowledge and skills with which researchers involved in computation must have at least a passing acquaintance.

Numerical Analysis. General numerical analysis is well covered in any number of references. The texts by Acton [1,2] are both eminently readable and genuinely insightful. They serve as an excellent starting point for acquiring an understanding of floating point arithmetic, round off error, numerical stability of algorithms, numerical integration, solving linear and non-linear equations and the like. This said, students must be aware that the "cookbook" approach to problem solving will inevitably fail. For example, any one of a number of "standard" methods can be used to solve a set of ODEs (assuming, of course, that there are not critical properties which must be exactly mirrored in the numerical solution), but with the exception of certain standard PDEs, off-the-shelf PDE solvers simply don't exist. Algorithms have to be constructed specifically for the equation in question.

To be successful, it is necessary to understand the mathematical basis for the various algorithms one uses. That all algorithms solve some problem exactly is an important realization. In the case of differential equations, one obtains numerically the exact solution of a slightly modified differential operator. This modified operator reduces to the original equations as the time-step and/or grid spacing tends to zero. In the case of systems of linear equations [3,4] the numerical result is a solution to a system with a slightly modified coefficient matrix. In all cases studying the modified equations can lead to significant insight into characteristics of the numerical solution. These techniques, collectively known as *backwards error analysis*, are the bases for constructing the structure preserving methods.

Software Engineering. There has been much improvement in software engineering practices in recent years. Ideally, practitioners of scientific computing should, as much as possible, avail themselves of the current "best practices" in software engineering. While not wanting to wade into the partisan arguments over which programming language is best suited to scientific computing, it is nonetheless worth noting that object-oriented design and coding practices now dominate commercial software development and for very good reason. Adhering to good design principles (encapsulation, abstraction, structured programming, *etc.*) can greatly reduce the possibility of bugs as well as improve productivity. Regardless of which language is used, it is critical to have a thorough understanding of how to properly use the chosen language. In particular, it is important to understand that most languages will have *undefined behaviours* — language constructs which will compile but whose consequence at runtime is not defined by the language standard — any of which can result in hard-to-find bugs, and therefore, must be actively avoided.

In addition to understanding the proper use of the programming language, the use of some form of source code revision control is also essential. A fun-

damental aspect of science is the reproducibility of results. All to often, this hallmark is missing from computational work. Research codes are constantly changing, and only through tracking source code changes is it possible to revisit previous work. A critical element of this reproducibility is the ability to know which version of the code produced a particular result. For example, each time one of our own codes is run, the parameters, along with the current version of the code, are stored and the run is assigned a unique identifier. In addition, all hard (and soft) copies of results are labeled with this same identifier. This identifier, in conjunction with version control, allows any run (especially those used in publications) to be reproduced at any time.

4 Conclusions

Few students will master all the material we have discussed here during their graduate careers, nor should we expect this to occur. This material should be regarded as general framework that we encourage our students to fill in as appropriate. The question of what should comprise the undergraduate curriculum is particularly challenging, given all the forces competing for lecture time. At the undergraduate level, beyond the usual elementary numerical analysis, it is most important to emphasize the difficulty of validating any computation. This mindset is critical; if we can condition our students to ask "is this right?" then they will experience much greater success, while simultaneously encountering less frustration.

Acknowledgements. The advice contained is this paper is the consequence of over two decades of practical computing. Over that time, the author has benefited from countless discussions about software and numerical methods with colleagues and collaborators alike and with pleasure acknowledges John C. Bowman, W. F. Buell, Andrew Charman, E. H. Esarey, Richard Fitzpatrick, C. B. Schroeder, W. F. Shadwick, J. D. Talman, and G. M. Tarkenton. This work was supported by the US DoE under contract No. DE-AC03-76SF0098 and by The Institute for Advanced Physics.

References

1. Acton, F.S.: REAL Computing Made Real: Preventing Errors in Scientific and Engineering Calculations. Princeton University Press, Princeton, NJ (1996)
2. Acton, F.S.: Numerical Methods That Work. Mathematical Association of America, Washington DC (1990)
3. von Neumann, J., Goldstine, H.H.: Numerical inverting of matrices of high order. Bull. Amer. Math. Soc. **53** (1947) 1021–1099
4. Turing, A.M.: Rounding-off errors in matrix processes. Quart. J. Mech. Appl. Math. **1** (1948) 287–308

Creating a Sustainable High-Performance Scientific Computing Course

E.R. Jessup and H.M. Tufo

[1] University of Colorado, Boulder, CO 80309, USA,
jessup@cs.colorado.edu,
http://www.cs.colorado.edu/~jessup
[2] University of Colorado, Boulder, CO 80309, USA,
tufo@cs.colorado.edu,
http://www.cs.colorado.edu/~tufo

Abstract. We describe our experiences with computational science and engineering (CS&E) education at the University of Colorado at Boulder in order to illustrate the difficulties associated with such a course of study. Our CS&E offerings began with a course on high-performance scientific computing (HPSC) developed under a CISE Educational Infrastructure grant from the National Science Foundation. The course was taught from 1991 through 1998 when various practical concerns ended its run. The course was revived in 2003 to meet the demands of several emerging programs in computational science. In this paper, we outline the rise, fall, and restoration of the HPSC course. We identify the technological developments that presently make such a course less labor intensive and more sustainable.

1 Introduction

The Department of Computer Science at the University of Colorado at Boulder (UCB) was an early player in the world of computational science and engineering (CS&E) education for undergraduates. Its course in high-performance scientific computing (HPSC) was introduced in 1991. The course provided an introduction to the use of high-performance computing systems in scientific and engineering applications. Its development was supported by the National Science Foundation under a CISE Educational Infrastructure grant awarded in 1990. The course enjoyed a successful seven year run before succumbing to the pressures of technological change and decline of student interest. In this paper, we describe our experiences with the HPSC course at UCB. In section 2, we discuss the design and development of the original HPSC course. Analysis of its decline is provided in section 3. In section 4, we discuss the reasons for the rebirth of HPSC. Finally, in section 5, we discuss the changes in technology and local CS&E curricula that will now allow us to offer the course over the long term.

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1242–1248, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 The Rise of HPSC

The original HPSC course was offered as a two semester sequence each academic year from 1991-2 through 1997-8. The course was based on the philosophy that students were best turned into effective users of supercomputers by introducing them to practical use of the machines. That is, students learned through real application problems and not just toy problems or simple numerical examples. The students learned which architectures were appropriate for which problems and how best to map the problems onto those machines. The students also learned how to assess the performance of their programs, including how to interpret what the compiler did and how to use the clock properly. Finally, the students learned to use the color, perspective, and animation capabilities of suitable visualization tools to display and interpret the large amounts of data typically generated by supercomputer programs.

Over the years, students in the course used DEC and SGI workstations and an Intel iPSC/2 hypercube multiprocessor located on the UCB campus as well as a variety of supercomputers at remote sites. The latter, which were accessed via the Internet, included a Cray Y-MP located at the National Center for Atmospheric Research (NCAR), an Intel Paragon at the National Oceanic and Atmospheric Administration (NOAA), and a Thinking Machines CM-2 at the National Center for Supercomputing Applications (NCSA) among others. Time on those machines was donated by the centers.

The first semester began with introductions to three different machines and to performance measurement. Students also studied the computational and scientific visualization tools Matlab [1] and an early AVS product [2]. The concepts and tools learned in the first few weeks were then applied to solution of numerical problems in molecular dynamics and computerized tomography. A midterm project required evaluation of the various architectures for solving several specified numerical problems. A final project required the student to port a molecular dynamics code to a new architecture and to evaluate the results. The second semester covered one or more additional architectures and an application in advection. In that semester, students completed a final project on an advanced topic of their choosing. Both semesters were offered as three hour lecture courses supplemented by a three hour supervised lab.

Because no appropriate textbook was available, we wrote all of the course materials. Our series of single-topic tutorials and short reference guides eventually became a textbook [3]. A laboratory manual and manuals for the machines and computational tools are still available on the Web [4].

The only prerequisite for the course was one semester of numerical computation at the undergraduate level. During its tenure, the course, which was listed at both the graduate and undergraduate levels, attracted juniors, seniors, and first year graduate students majoring in computer science, physics, applied mathematics, or chemical, mechanical, or aerospace engineering.

The HPSC course was a success on several levels. The student evaluations were routinely positive. Several of our graduates went on to jobs at national laboratories (mainly NCAR and NOAA in Boulder) or into graduate programs

relevant to computational science or engineering. They reported that the material they'd learned in the course was helping them in their jobs or studies. Other graduates became research assistants at UCB in the field of HPSC and ultimately took on employment in that area.

Our course was exported to other colleges and universities. As part of our proposal to NSF, we named several collaborating institutions. Professors from those institutions aided us in course development and received equipment purchased with the NSF funds to help support their own HPSC courses. The collaborators were chosen from institutions in our region with some emphasis on schools with significant minority enrollment and on schools that would not otherwise have the resources necessary for teaching HPSC. The ultimate goal of the collaboration was to put HPSC courses modeled on ours into those institutions, and that goal was reached at all of the schools.

For two years, we offered (separate) two-week summer workshops for students and for faculty from other institutions during which we covered the first semester's materials. The faculty members attending the workshops were subsequently able implement scientific computing courses like ours at their own institutions.

We were also aware of courses based fully on ours offered at other universities not associated with our program. Instructors at other institutions were using a sampling of our materials in related courses. Interest was demonstrated by the large number of downloads from our course materials Web site and continuing textbook sales.

The course received formal recognition in the form of a 1995 Undergraduate Computational Science Education Award from the Department of Energy.

3 The Decline of HPSC

Although the HPSC course was very well-received, it was not without its problems. While we were able to use a network of workstations via MPI [5] in the last two offerings of the course, MPI was not fully developed, and it was not yet considered standard. As a result, codes had to be ported explicitly to each new architecture, using constructs particular to that machine. Furthermore, supercomputers came and went at a rapid rate. Sometimes a computer we used in one offering of the course was not available for the next year's offering. As a result, the process of keeping the course materials up to date was difficult and time-consuming.

While the topics covered by the original materials served us well at UCB, they were not always perfect for courses at other institutions. Expanding the offerings represented another substantial investment of effort, and it was not obvious how to obtain funding for further course development. While many funding agencies have excellent programs for course initiation, they do not have programs for course continuation and maintenance.

While the UCB model of HPSC education was appropriate at many other schools, it did not work everywhere. In those days before ubiquitous Internet,

smaller schools complained of the lack of fast Net access and of the dearth of funding to purchase and support workstations and visualization tools (especially the AVS program we used for 3D animation). We considered the problem of downsizing the materials for use at those schools but again ran into lack of support for such efforts.

Perhaps the most serious problem confronting the HPSC course was a local one. The extremely tight curriculum in the Engineering College (which includes the Department of Computer Science) made it difficult for students to find time for the HPSC elective. Further, the number of Computer Science majors interested in HPSC was relatively small, and students in other departments were not always aware that the course existed. As a result, first semester class enrollments ranged from only 11 to 17 over the years (small numbers at our university). The second semester class was typically even smaller. In some semesters, finding students at all required a significant amount of recruiting. Before the 1998 offering, the instructor opted not to recruit, no students enrolled, and the course was canceled. That semester marked the end of the original version of the HPSC course. Its run had lasted seven years.

The decline of HPSC at UCB mirrored a general lessening of interest in the topic in the educational community as a whole. The years 1992-1997 saw a large number of special conferences and conference sessions devoted to HPSC education. That time period also saw large-scale development of materials for computational science by other authors (e.g., DOE's Undergraduate Computational Engineering and Sciences Project [6]). Such activity decreased markedly in subsequent years, reviving only recently.

4 The Rebirth of HPSC

The beginning of the twenty first century has brought with it an increased demand for computational skills in a variety of disciplines. The source of that demand extends from academic researchers to employers in industry and the national laboratories. As a result, the seeds of CS&E education have spread to numerous engineering and applied science departments at UCB and have gradually developed into small pockets of CS&E training within those departments. In an effort to formalize this training, several departments have begun to construct undergraduate and graduate educational programs to teach their students the specific CS&E skills required by their disciplines. Because of the difficulty of creating new degree programs at UCB, all of these training programs have been offered as supplements to or tracks within current degree offerings.

The material and the practical skills once taught in our HPSC course are at the core of many of these new CS&E programs. One of the first approved programs came from the Department of Applied Mathematics in 2003. Instead of developing their own HPSC course or spreading the material over several core courses, the Applied Mathematics faculty asked us to revive our HPSC course. To address their discipline-specific content concerns we moved the open course

project from the second semester to the first and allowed significant flexibility in project selection.

To address technological developments since its first offering, we embarked on a redesign of the course. First, several structural changes were effected. To further increase the potential pool of students the numerical analysis prerequisite was removed and now appears only on the recommended course list. Since the project had been moved to the first semester we chose to make HPSC a one semester course instead of two. However, as our intention was that it remain a hands-on project-based course, we kept the original four credit hour design (three hours of lecture and three hours of supervised lab per week).

In addition to structural changes, rapid changes in technology and demands from employers in industry and the national laboratories needed to be reflected in the course content and tools. Parallel programming skills are currently in high demand. As MPI is now the de facto standard for writing message-passing parallel programs and can be either be easily installed on or purchased with a computing system, we concentrate on learning MPI programming in the first eight weeks of the sixteen-week course. Given that mixed model programming paradigms are being driven by the increasing use of shared memory multiprocessing technology, we then spend time learning the basics of OpenMP [7] and writing hybrid MPI/OpenMP programs. Because the recent success of the Japanese Earth Simulator program has sparked renewed interest in vector programming, we also provide a brief introduction to vector computing. The majority of the remaining time is spent examining parallel architectures and algorithm development in more detail, concentrating on mesh, tree, and hypercube architectures and developing architecture specific algorithms for commonly encountered problems (e.g., sorting, matrix and graph algorithms, FFT).

5 The Sustainability of HPSC

The original HPSC course fell victim to a variety of problems: the rapid evolution of computer architectures, the difficulty of maintaining course materials, and the decline of local demand were the most serious.

That course relied on a combination of local and remote computing resources. The fact that many of the parallel systems were housed off-site and changed rapidly made keeping the course materials current and preparing for next year's offering difficult. The intervening five years (1998-2003) saw the advent of Beowulf cluster computing systems [8]. Because these systems are built using commodity off-the-shelf components and employ open source software, they are extremely cost effective, providing computing capacity at less then $300 per GigaFlop (e.g., [9]) and requiring no (expensive) service contracts. (To put this in perspective, in 1998 the world's fastest computer, ASCI Option Red, cost about $30,000 per GigaFlop [10].) Given the reduced cost and the fact that the open source software, in particular the Linux operating system, has become relatively mature and sufficiently stable, we purchased a 128-processor Beowulf cluster and the HPSC course relies solely on this platform for code development *and*

production runs. Making long term predictions about the the future of high-performance computing is extremely difficult. However, we firmly believe that these cluster solutions will be around for next decade and remain the most cost effective solution for designing systems of less than a thousand processors.

The maturity and stability of the software and hardware environment have not only made life much easier for the instructor but also for the students. The fact that most students are already familiar with Linux and, hence, our computing environment, means that we were able to accelerate the rate at which we covered material. This accelerated rate allows us essentially to present all of the computational material previously taught in two semesters in one, thus removing any objections to the decision to make this a one semester course.

In the interest of providing a course appropriate to a variety of disciplines, the physics-based application problems are no longer covered. Scientific visualization also receives less coverage primarily due to time constraints, although freeware visualization (e.g., VTK [11]) is available on the cluster. The original intensive slate of programming assignments and laboratory exercises now provides an effective method for rapidly introducing and learning basic and intermediate MPI programming techniques.

The maintenance of course materials is no longer the problem it once was. While many of the original course materials are now dated and their emphasis on applications does not fit the current model of instruction, there is no longer a need to produce homemade materials to replace them. In the current course, we use two books on MPI [12,13] and one on parallel algorithms and architectures [14]. A variety of other sources exist that could be used to supplement or replace these materials should the need arise.

Like its predecessor, the new version of the HPSC course is appropriate for export to other institutions. In particular, the availability of the original on-line materials (updated and possibly supplemented with on-line lectures) and of well-developed open source software make it possible to "package" the entire course in an affordable box. Furthermore, the availability of less expensive systems and free software make an HPSC course more accessible to smaller institutions. Even two networked workstations running MPI are sufficient to illustrate many important concepts in parallel computing.

Finally, we expect that the now broader interest in CS&E at UCB will help us to maintain healthier enrollments. We have requested that Applied Mathematics feed junior and senior level undergraduates and graduate students from the undergraduate and graduate numerical analysis courses and other appropriate courses into the new HPSC course. Although the partnership is still in its infancy, we anticipate that Applied Mathematics will supply a steady stream of 8-16 students per year. Additional demand is ultimately expected from emerging CS&E programs in Astrophysical and Planetary Sciences, Aerospace Engineering, Atmospheric and Oceanic Sciences, Physics, Civil and Environmental Engineering, and Computer Science. To increase the appeal to Computer Science graduate students, we have ensured that the course serves to satisfy the breadth requirement for the master's degree.

References

1. The MathWorks: MATLAB and Simulink for Technical Computing (2003) http://www.mathworks.com/.
2. AVS/Advanced Visual Systems: Data visualization software and solutions (2003) http://www.avs.com/.
3. Fosdick, L.D., Jessup, E.R., Schauble, C.J.C., Domik, G.: An Introduction to High-Performance Scientific Computing. MIT Press, Cambridge, MA (1996)
4. Fosdick, L.D., Jessup, E.R., Schauble, C.J.C., Domik, G.: Computer Science – Course Materials, CSCI 4576 (2003) http://www.cs.colorado.edu/ftp/pub/HPSC/README.html.
5. MPI The message Passing Interface Standard (2003) http://www.mcs.anl.gov/mpi/.
6. The Undergraduate Computational Engineering and Sciences Project Homepage (2003) http://www.krellinst.org/UCES/.
7. OpenMP: Simple, Portable, Scalable SMP Programming (2003) http://www.opemmp.org/.
8. Beowulf.Org The Beowulf Clustering Site (2003) http://www.beowulf.org/.
9. Terascale Cluster - Research Computing - computing.vt.edu (2003) http://computing.vt.edu/research_computing/terascale/.
10. ASCI Red Web Site (2003) http://www.sandia.gov/ASCI/Red/.
11. VTK Home Page (2003) http://www.vtk.org/.
12. Pacheco, P.: Parallel Programming with MPI. Morgan Kaufmann, San Francisco, CA (1997)
13. Gropp, W., Lusk, E., Skjellum, A.: Using MPI: Portable Parallel Programming with Message-Passing Interface. MIT Press, Cambridge, MA (1999)
14. Leighton, F.T.: Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes. Morgan Kaufmann, San Francisco, CA (1991)

CSE without Math? A First Course in Modeling and Simulation

Wolfgang Wiechert

Department of Simulation, Faculty of Mechanical Engineering,
University of Siegen, Germany
wiechert@simtec.mb.uni-siegen.de,
http://www.simtec.mb.uni-siegen.de

Abstract. There are several good reasons to accompany a course in computational science and engineering from the very beginning with a general lecture on modeling and simulation. On the other hand first year undergraduate students only have elementary knowledge on mathematics and natural sciences. Consequently, a first lecture must be conceived that requires only basic mathematical and physical skills. The course on 'Foundations of Modeling and Simulation' given at the university of Siegen fulfills this requirements. It is the first in a series of 7 lectures on 'Modeling and Simulation' which later on will go much deeper into the technical details. Basic didactic considerations, examples and experiences from the first course are presented. In particular an extensive example of a fairground looping swing is discussed.

1 Introduction

No course in computational science and engineering (CSE) or in modeling and simulation (M&S) can be given without a substantial amount of mathematical contents. For that reason teaching CSE without mathematics seems to be a contradiction. However, there are some good reasons to start education in CSE with an introductory course in which mathematical subjects have been strongly reduced:

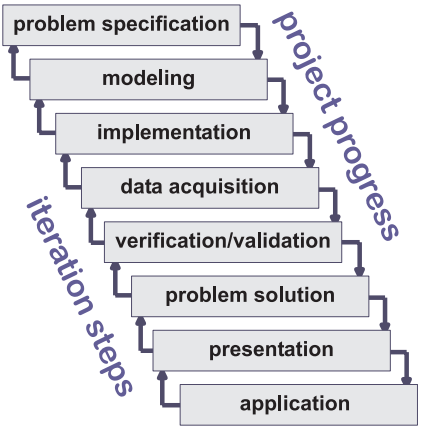


Fig. 1. General steps in a simulation project

Virtual and real world: Simulation projects require a close cooperation of researchers from different disciplines. The steps that frequently are the most critical for the success of a simulation project are project definition, model validation and communication of the results (cf. Fig. 1). These are exactly those steps where the virtual world meets reality. To sensitize the students for the

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3039, pp. 1249–1256, 2004.
© Springer-Verlag Berlin Heidelberg 2004

problems arising at this interface they should be led to their first practical simulation project as quickly as possible. In particular the aspect of communication in multi disciplinary teams, the specification of the right project goals and a reasonable balance between algorithmic performance and overall time to deliver the results must be taught.

Multi disciplinarity: Due to its multi disciplinary nature CSE covers a wide range of methods from mathematics and computer science to the applications in the natural sciences and engineering. For a student it is hard to survey all this different subjects, their connections and interdependencies. Consequently, the students need to learn very early how to integrate subjects from different disciplines in the focus of a certain application problem. Thus example projects which require little in depth knowledge of mathematics, computer science and natural sciences should accompany a CSE course from the very beginning.

Heterogeneous audience: There are only a few universities were a complete CSE program with undergraduate and graduate courses is established. A good alternative is to establish CSE as a specialization in classical study programs like engineering, mathematics or physics or to establish a graduate course which can be attended with different prerequisites. In each case it is a big problem to integrate all the students with their different origins and to synchronize their knowledge in the basic disciplines. In this situation a first course on M&S should introduce into the subject without relying on advanced skills in mathematics, computer science and physics. Accompanying simulation projects will give the students the opportunity to profit from their different skills.

CSE in practice: Not only at the university but also in industry CSE methods have become an important part of research and development processes. Project managers in industry know that simulation is a potential tool to understand complex systems, to evaluate experiments, to build virtual prototypes or to judge the potential economic benefit of a development. On the other hand this creates a need for a basic M&S education even for non CSE specialists. Group leaders and managers should at least be able to judge under which conditions simulation may be a useful tool, what the basic problems are and what the costs of a simulation project will be.

2 Simulation Courses at the University of Siegen

2.1 Different Prerequisites in a Heterogeneous Audience

The Department of Simulation at the University of Siegen is a part of the engineering faculties (mechanical and electrical engineering, computer science). The lectures on M&S are attended by students with quite different preconditions taking part in different study programs:

1. M&S can be chosen as an optional course for all students of electrical and mechanic engineering. Both have basic mathematical skills from undergraduate courses and also a good knowledge of the physics of mechanical and electrical systems. On the other hand engineers are typically less trained in computer science and programming.

2. These prerequisites are quite different for the computer science students. Although they also have several mathematical lectures in the undergraduate courses they are usually not familiar with the basics of analysis, numerical algorithms and differential equations. Moreover, it requires a high effort for them to understand the laws of physics.
3. Lectures on M&S are a mandatory part of the Siegen international graduate course in mechatronics. M&S is scheduled in the third semester of the mechatronics course and thus these students already have good knowledge of programming and control engineering.
4. A combination of engineering and economics can be studied at Siegen. It is quite probable that these students in their future industrial career will have to decide on the initiation of simulation projects from an economical viewpoint. However they have only very restricted mathematical skills from the undergraduate courses whereas their knowledge of physics is sufficient.
5. Finally, it is also possible for the students from other departments to participate in the courses. However only physicists and mathematicians with basic physics courses have good preconditions for a course on M&S. Only very engaged students with less preconditions are able to follow the lecture.

2.2 Lectures

The lectures on M&S are currently attended by about 50 students per semester. The course has been broken down to a series of 7 short half semester lectures with four hours per week. Part of the lectures is directly done in the computer laboratory. The courses have already been described in more detail in [2]. From these 7 lectures the different students can choose those units that best fit their requirements and skills. In each case the first course on M&S is mandatory. This first part is also the subject of the present contribution. Whereas the lectures 2-6 are more advanced with respect to the mathematical and computer science demands the first one is really a CSE course with (almost) no mathematics.

The Siegen courses on M&S are embedded in the activities of the FOMAAS research center. Founded in 1992 it is concerned with CSE with a special emphasis on the optimization of complex systems. More details about FOMAAS can be taken from [4,3]. Under the auspices of FOMAAS several other lectures can be chosen by the students which complement the basic CSE education. They are dealing for example with finite elements, computational fluid dynamics, electro magnetic field computation, optimization or distributed software systems.

3 A First Course in Modeling and Simulation

3.1 Goals

The course on 'Foundations of M&S' has the following goals:

1. To present an overview of the applications of M&S in academia and industry and to explain the different aims and preconditions of simulation projects.

2. To establish a first understanding on the general course of a simulation project (see Fig. 1). In particular a rough impression of the financial and manpower efforts of a simulation project is given.
3. An overview on the major methods and tools to carry out continuous time simulations based on ordinary differential equations is given. In particular the students get a survey on some commercially available simulation tools.
4. The basic principles of stochastic simulation are presented. The difference between stochastic and deterministic simulation with respect to accuracy and computing time is emphasized.
5. A general introduction into discrete event simulation which is usually combined with stochastics is given and a survey of major tools is done.

3.2 How Much Math Is Required?

With respect to the mathematical prerequisites some compromises must be made in a first course because only two hours in this semester are available:

1. The numerical solution of ODEs is introduced with the Euler algorithm which is easy to understand for the whole audience. Although no specialist will use this algorithm in practice it can well be used to explain many basic problems of numerical ODE solution like accuracy, step size control or instability. More details on the numerical ODE solution are postponed to the second part of the lecture which gives an in depth treatment of this topic. Instead, the preconfigured ODE solvers from MATLAB and SIMULINK are used to demonstrate that they are superior to the Euler method and that it is worthwhile to develop specialized algorithms for certain problems.
2. Although this is a major part of any CSE program no introduction to spatially distributed systems and PDE solution is given. The reason is that even a basic introduction to this domain will take a whole courses' time. Nevertheless, if more time is available then some basics on PDEs can be integrated into the course. The emphasis here should be on a prototypical solution algorithm like finite differences and a discussion of the computational complexity that is associated with PDE problems. Then a commercial finite element tool like FEMLAB could be used to demonstrate the basic steps in the solution of a PDE problem. In fact this is done in the third part of the lectures at the University of Siegen. More details on these subjects are given in the acompanying lectures (see Section 2.2).
3. Typically, students from the engineering departments have little knowledge of probability theory and statistics. Because this is required for stochastic simulation a 'survival kit' in probability and statistics is integrated in the course. This kit contains a basic knowledge of distributions, statistical estimates and tests. The emphasis is not on a broad coverage of topics but on some fundamental results concerning the precision and convergence behavior of stochastic algorithms.
4. What integrators are for ODE solution that are random number generators for stochastic simulation. However, only the basic congruence generator is

presented as a prototype for more advanced modern generators. The basic problems of pseudo random numbers are shortly discussed. Finally, the most basic methods for the generation of non uniform distributions (transformation, rejection) are presented.

4 The Looping Swing: An Instructive Example

As a basic example of a complete simulation project the looping swing is now discussed in more detail. This example requires only little preliminary knowledge from mathematics and physics. It is an important part of the first course on M&S and takes about ten hours including two hours of computer exercises. It must be pointed out that this example does *not* reduce a simulation project just to the process of solving a given differential equation system with numerical methods (this is just one of many steps in Fig. 1) but tries to show the course of a simulation project in a realistic application context. The typical steps of a simulation project as shown in Fig. 1 are discussed in the lecture. The following remarks concentrate on the educational and didactic aspects of the looping swing example:

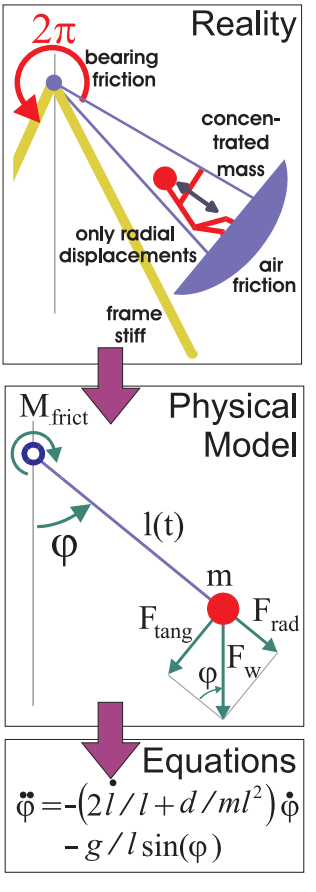


Fig. 2. Modeling of the looping swing: From the real system to the mathematical model

Project specification: The project goal is to find out if it possible to construct a fairground swing that can be driven around the top in a 360^o loop (Fig. 2). In order to make such an apparatus profitable for the operator it should not take too much time until the first looping is managed by an average person. Moreover, for economic reasons the operator must know a suitable initial swing velocity that is required to keep the total swinging time within a reasonable limit. From an educational viewpoint it is very important to emphasize that an operator of a fairground swing is typically not an academically skilled person and thus the communication with this client will be rather difficult for a computational scientist. This should give an impression to the students of how important the initial project specification is for the overall progress.

Mathematical modeling: By common definition simulation must always been based on a mathematical model of the system under investigation. Modeling is usually subdivided into a first step of building an idealized physical representation of the real system and a second step of generating the mathematical equations from this system. As pointed out in [2] these two steps should be

sharply separated in a lecture. In the running example one major educational goal is to demonstrate the dependency of model complexity on the aim of the simulation project. In this case it is not necessary to predict the dynamics of the looping swing with a high precision. For this reason a very simple representation of the complex system can be used. After a thorough discussion the whole system is reduced to a mathematical pendulum with a time dependent length. Here the important concept of the scope of a mathematical model can be clearly demonstrated. Due to the very simple model (Fig. 2) only the radial movements of the person in the swing are taken into account while tangential movements are neglected. This means that the model is not able to describe the acceleration of the swing starting with an initial angle of zero and zero velocity. Having drawn a picture of the idealized physical model the generation of the mathematical equations is straight forward from the law of angular momentum (Fig. 2).

Implementation: To implement a simulator on a very basic level the Euler method is quickly introduced by deriving it from a differential quotient approximation of the time derivative. The well-known deficiencies of this primitive ODE solver are rather an advantage in a first course on simulation than a disadvantage because the shortcomings of the Euler method will immediately turn out in the next steps of the project. In the next unit of the lecture which is concerned with tools for continuous time simulation the Euler method is compared with other integrators contained in the MATLAB and SIMULINK packages. This in turn motivates to learn more about numerical algorithms in the subsequent lectures.

Data acquisition: It is frequently disregarded that the acquisition of data to parameterize a model can be one of the most time consuming parts of a simulation project in practice. This can be illustrated even with the simple swing model. It is discussed how empirical data, material constants, separate calculations, data sheets of component manufactures, parameters from comparable systems, interviews with the operator, or experiments can be used as data sources in the present example.

Verification: The term verification means to prove the correctness of the simulator implementation relative to the idealized physical model. In the example the plausibility of the model is tested by doing extensive parameter and initial value variation studies. In each case it is checked if the model behaves in the expected way. Because the mathematical pendulum can be approximated for small angles by a linear model an analytical solution is available. This solution can also be taken to check the correctness of the implementation.

Validation: The term model validation means to prove the correct representation of the real system by the simulation with respect to the accuracy requirements and the experimental frame given in the project specification. Model validation is another very critical step in a simulation project which is closely related to the very fundamentals of natural science [7]. Two aspects of model validation are considered in the example. The first aspect is the friction model

contained in the system equations. By comparing different friction models (constant, linear, quadratic, mixed) it turns out that there is a difference but this difference is not important for the problem solution. Thus a linear friction model can be safely assumed although it might not exactly represent the reality. The friction coefficient must be chosen with a reasonable order of magnitude. The second aspect of validation is the estimation of model parameters which could not be obtained from available data. Two methods to estimate the linear friction coefficient from experiments are shortly discussed which are the direct usage of the analytical solution or general parameter fitting procedure.

Problem solution: In reality swings are accelerated by the same mechanism that is used in iceskating to perform a pirouette. To increase the rotational velocity the center of mass has to be moved closer to the center of rotation. Unlike in iceskating this cannot be continuously done in a looping swing. However, the swing length can be increased without an effect on the rotational energy at the turning points of the swing. This is also the recipe to build up a control strategy for the looping swing. Interestingly, this strategy is not that easy to implement because further restrictions due to human limitations like smoothness and bounded accelerations must be recognized. Moreover, the control strategy should account for the 'sportiveness' of the swing driver. A parametrized analytical control function is constructed based on piecewise scaled trigonometric functions. Simulation experiments can now be carried out to solve the original problem. It turns out that a well trained driver is able to achieve a looping after 15 swings.

Result presentation: The presentation of the results of a simulation study is another very important step in practice because they must be communicated to a customer who may not be familiar with the basics of simulation. Thus it is very important which graphical visualisation is used to demonstrate the results. To this end several different visualizations for the dynamics of the looping swing are introduced and compared with respect to their possible audience (Fig. 3). Clearly, the virtual rea-

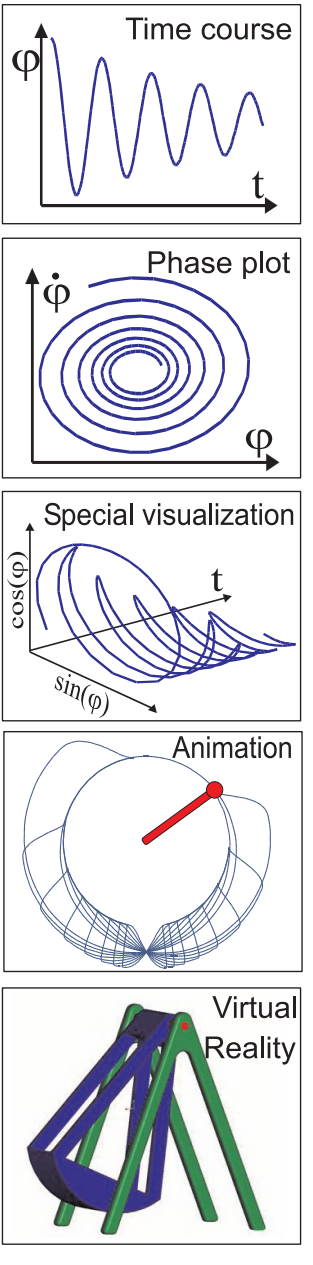


Fig. 3. Different visual representations of the simulation results

lity representation (bottom) is totally 'oversized' because this is essentially a 2-dimensional system. Nevertheless it will be the most convincing visualisation for an operator of a fairground swing.

Application: In both academic and industrial applications simulation projects have the tendency to be never finished because once the results have been presented and applied to the real system new problems will arise. In the case of the boat swing this might be the problem of how to safely fix the swing driver to the swing body. In many situations it may be the case that some important details of the real system have been disregarded. In the present example it may be the case that the swing representation by the mathematical pendulum is too much simplified and a more detailed physical pendulum model has to be taken.

5 Experiences

In the first years of the introductory course on the 'Foundations of M&S' about 100 students with very different scientific backgrounds have attended it. The chosen mathematical contents are well accepted and – as the first exams and project works show – can be applied by the students in practice. By conception the students have the opportunity to choose any other course of the series M&S 2-7 afterwards. In fact many students with a sound mathematical background (like the mechatronics students) participate in 3-4 courses while the others rather choose one of the more practical courses like 'M&S 7: Simulation tools'. The Siegen concept of a modular education in computational science for students from different disciplines will be extended and more institutionalized in the next years by installing CSE as a specialization that can be chosen in different courses like engineering, computer science or mathematics.

References

1. W. Wiechert, Lecture Notes on Modeling and Simulation, `http://www.simtec.mb.uni-siegen.de`.
2. W. Wiechert, W., The Role of Modeling in Computational Science Education. Future Generation Computer Systems, 2003, pp.1363-1374.
3. FOMAAS (Forschungszentrum für Multidisziplinäre Analysen und Angewandte Systemoptimierung): University of Siegen, `http://www.fomaas.uni-siegen.de`.
4. W. Wiechert, W. Becker, H.A. Eschenauer, B. Freisleben, M. Grauer, D. Hartmann, H.-J. Reinhardt, and G. Thierauf, eds. FOMAAS Status-Report 2000-2002. 2002. ISBN 3-9807356-3-x
5. W. Wiechert, Eine Checkliste für den Aufbau einer Simulationstechnik Vorlesung, in: K. Panreck, F. Dörrscheid, 15. Symposium Simulationstechnik, Paderborn, SCS Publishing House, 2001, pp. 151-156.
6. B.F. Gore, The Child's Swing. Journal of Physics, 1970. 38: p. 378-379.
7. W. Wiechert, R. Takors, Validation of metabolic models: Concepts, tools, and problems, in: H.V. Westerhoff, B.N. Kholodenko (Eds.), Metabolic Engineering in a Post Genomic Era, Horizon Scientific Press, 2003.

Author Index

Abad, A. IV-327
Abawajy, J.H. I-27, III-66, III-339
Abdalhaq, B. II-220
Abe, S. III-1046
Abrahamyan, L. IV-548
Abramson, D. I-148, I-164
Acacio, M.E. I-541
Acosta-Elías, J. I-561, I-615
Adamiak, R.W. III-1013
Adelmann, A. II-459
Adriaansen, T. III-905
Ahn, J. I-192
Ahn, S. I-349
Ahn, T.-C. II-188
Akay, B. II-722
Akçelik, V. III-717
Akdim, B. II-260
Akker, H.E.A. Van den IV-637
Al-Badarneh, A. II-606
Albuquerque, P. IV-540
Alemani, D. IV-540
Alexandrov, V.N. IV-475, IV-491
Alique, Á. I-523
Alonso, J.M. I-482
Altas, I. IV-364
Amaral, J.F.M. III-1273
Amaral, J.L.M. III-1273
Anai, H. IV-286
Anderluh, J. IV-851
Angelova, D. IV-531
Antonio, A. de I-433, I-545, III-575, III-1004, IV-252
Antoulas, A. III-740
Araújo, S.G. III-1265
Ardaiz, O. I-67
Arenas, A. III-1078
Arodź, T. I-462
Artoli, A.M. IV-548
Asensio, M.I. II-586
Asensio-Pérez, J.I. I-495
Asmussen, S. IV-795
Astalos, J. I-124
Austin, A. IV-1165
Austin, K. III-1054
Babuška, I. III-756
Bacak, G. IV-376
Bachmeir, C. III-187
Bae, S.H. III-234
Bai, Y. IV-900
Baik, S. II-678
Bajaj, Chandrajit II-419, III-756
Bakar, N. III-379
Bala, J. II-678
Bała, P. I-99
Balas, L. II-618
Balázs, F. I-318
Baldridge, K.K. I-75, I-148
Baliś, B. I-107, III-26
Balk, I. IV-1012, IV-1045
Balogh, Z. III-583, III-599
Balos, K. III-114
Banaś, K. II-155
Banaszak, J. IV-687, IV-694
Bandera, G. I-409
Bang, Y.-C. I-582, I-602, I-611
Bang Nielsen, M. III-921
Bansevicius, R. IV-278
Barbosa, J.L.V. III-2
Barreiro, E. IV-139
Bass, J. III-756
Bassi, A. III-202
Batty, M. III-1109
Batura, T. III-545
Beck, M. III-202
Bein, D. III-1233
Beletskyy, V. I-635, III-398
Benedyczak, K. I-99
Benoit, A. III-289
Benouahmane, B. II-557
Benyon, D. III-936
Bergamaschi, L. IV-434
Bernard, M. III-986
Bernatík, R. III-631
Bernholdt, D.E. III-1
Bernsen, N.O. III-946
Berthold, J. III-297
Bertoldo, Alberto II-614
Bertolotto, M. IV-948
Beseri, T. IV-376

Bešter, J. III-141, III-179
Bevzushenko, M. IV-1049
Beyls, K. III-448
Bianco, M. II-614
Bidarra, J. III-1025
Bielak, J. III-717
Bierbrauer, M. IV-859
Bijl, H. IV-604
Biros, G. III-717
Blažič, B.J. III-141, III-179
Blesa, M. I-429
Boada, I. IV-98
Bogdanov, A.V. I-239
Bojanic, S. I-644
Boku, T. IV-629
Bollapragada, R. IV-1017
Bonelli, A. II-459
Bonnefoi, P.-F. IV-80
Borgosz, Jan II-507
Boss, M. III-1070
Bote-Lorenzo, M.L. I-495
Boufflet, J.P. II-163
Boukhanovsky, A.V. I-239
Bourchtein, A. IV-667
Bourchtein, L. IV-667
Brasser, P. IV-637
Breitkopf, P. II-163
Bretto, A. IV-343
Broeckhove, J. I-51, II-443
Brogan, D. III-789
Brotzge, J. III-670
Browne, J.C. III-756
Bruno, G. I-75
Brunton, R. III-835
Bruti Liberati, N. IV-771
Brutzman, D. III-835
Brzezinski, J. II-475, III-82
Bubak, M. I-107, I-462, III-26, III-34, III-472, III-848
Buchtela, D. IV-1133
Budinska, I. III-599
Budzyńska, L. III-1241
Bungartz, H.-J. IV-394, IV-1180
Buono, N. Del IV-459
Burak, D. I-635
Burgos, P. II-204
Burnecki, K. IV-868
Burns, J. III-1094
Bushehrian, O. II-699
Byeon, O.H. III-250
Bylina, B. II-578
Bylina, J. II-578
Byun, Y. I-334

Caballer, M. III-54
Čada, V. IV-123
Cai, G. II-404, IV-956, IV-981, IV-989
Cai, J.-M. II-435
Cai, W. III-856
Cai, X. III-889
Cała, J. IV-1172
Caliari, M. IV-434
Calleja, M. IV-483
Calteau, A. IV-700
Camahort, E. IV-197
Campos, C. IV-197
Cánovas, Ó. III-158
Cant, S. I-478
Cantillo, K. I-523
Capitelli, M. IV-580, IV-588
Čapkovič, F. III-529
Carlucci Santana, R.H. I-537, I-661
Carmichael, G.R. III-709
Carvalho, R. III-1109
Cascón, J.M. II-130, II-586
Castro, F. IV-189
Catalyurek, U. III-805
Cebrat, S. IV-687, IV-694, IV-709, IV-765
Cepulkauskas, A. IV-278
Cermak, M. II-412
Cetnarowicz, K. III-607
Cha, W. I-648
Chai, T. III-709
Chandresakar, V. III-670
Chang, J.H. IV-1088
Changhuang, C. I-59
Chanussot, F. III-202
Chapman, B. III-408
Chaturvedi, A. III-779
Chen, H. I-59, I-470, I-623
Chen, L. I-310, I-507
Chen, M. III-725
Chen, R. III-889
Chen, W. III-1131
Chen, Y. III-388, III-414
Chen, Z. I-657, IV-931
Cheng, C. IV-972
Cheng, L. III-141
Chi, J. III-779

Chin Jr., G. II-540
Cho, G. I-648
Cho, H.-S. I-381
Cho, J.-W. I-43
Cho, J.H. II-687
Cho, K. III-1117
Cho, S. I-255
Cho, S.-P. I-590
Cho, T.-S. IV-1095
Choe, T.-Y. II-598
Choe, Y.H. IV-1095
Choi, J.K. I-607
Choi, I.-B. I-18
Choi, J. I-445
Choi, K.H. III-234
Choi, M.-H. II-300
Choi, S. I-499
Choi, S.J. II-121
Choi, W.-H. I-569
Choo, H. I-565, I-573, I-582, I-602, I-611, III-266, III-1138
Chopard, B. IV-540
Chover, M. IV-107, IV-173
Chrastek, R. IV-41, IV-49
Chun, K.J. III-250
Chung, C.H. I-687
Chung, J. I-349
Chung, M.G. III-1193
Chung, M.Y. I-573, III-1201
Chung, T.M. III-250
Chung, W.-G. III-1170
Ciereszko, A. II-451
Ciesielski, M. II-348
Çinar, A. II-523
Ciskowski, P. II-196
Coen, J.L. III-725
Cole, M. III-289
Cole, M.J. III-701
Coll, N. IV-98
Combarro, E.F. I-666, II-658
Conversin, P. IV-1165
Coppex, F. IV-742
Corchado, E. II-204, II-650
Corchado, J. II-650
Corradini, A. III-913, III-946
Correia, M.V. II-380
Cortés, A. II-220
Cortes, T. I-10
Costa, C.A. da III-2
Créput, J.-C. III-638, III-1102
Cruz Guzmán, J. de II-552, II-602
Cuyt, A. I-695, II-557
Cyganek, B. II-507, II-548
Czekierda, Ł. IV-1172

D'Apice, C. IV-351
D'Hollander, E.H. III-448
Daescu, D. III-709
Danion, F. IV-1056
Danon, L. III-1078
Daoud, D.S. II-582
Darema, F. III-662
Datta, A.K. III-1233
Datta, D. III-1209, III-1224
Daubin, V. IV-700
Davinelli, M. IV-1073
Demeester, P. III-149
Demkowicz, L.F. III-756
Deng, H. I-507
Deng, Q. I-507
Deng, S. I-627
Denis, C. II-163
Deschrijver, D. II-443
DeTar, C. IV-1202
Dew, R. I-397
Dhaene, T. II-443
Dhar, P. II-284
Dhoedt, B. III-149
Dhooge, A. II-42
Dias, A. III-1025
Díaz, I. I-666, II-658
Díaz-Guilera, A. III-1078
Diele, F. IV-426
Dimitriadis, Y.A. I-495
Diomede, P. IV-580
Dirnstorfer, S. IV-394
Doallo, R. I-132
Dobrucky, M. I-124
Dolk, D. III-779
Dong, H. I-174
Dong, S. I-287
Dongarra, J. III-432, IV-270
Dooren, P. Van III-740
Dopico, A.G. II-380
Douglas, C.C. III-701, III-725
Dove, M.T. IV-483
Downar, T. IV-1188
Drake, D. III-835
Dreżewski, R. III-654
Droegemeier, K. III-670

Droz, M. IV-742
Duan, X. II-260
Dubey, P. I-657
Dubu, D. I-490
Dudek, M.R. IV-687, IV-694
Dudkiewicz, M. IV-687, IV-694
Dumitrescu, D. II-670
Dumitriu, L. III-497, III-560
Duplaga, M. IV-1118
Dupuis, A. IV-556
Dureisseix, D. IV-612
Duwe, H. IV-1140
Dydejczyk, A. II-638
Dzemyda, G. I-652
Dziewierz, M. III-583
Dziurzanski, P. III-398

Eberspächer, J. III-187
Efendiev, Y. III-701
El-Aker, F. II-606
El-Amrani, M. II-89
Ellahi, T.N. III-210
Enticott, C. I-148
Epanomeritakis, I. III-717
Erciyes, K. I-357
Erdogan, N. I-184, I-437
Ergenc, T. IV-364
Ermolaev, V. IV-1049
Escobar, J.M. II-642
Esparza, O. III-174
Evangelinos, C. III-685
Ewing, R. III-701

Fahringer, T. II-459, III-18, III-42, III-424, III-456
Falcou, J. III-314
Fan, J. II-435
Fangohr, H. IV-1210
Fantozzi, S. IV-1073
Farhat, C. III-693
Farreras, M. I-10
Federl, P. II-65, II-138
Feng, Y. II-26, III-756
Fernández, J. II-658
Ferragut, L. II-130, II-586
Fic, G. II-654
Fidanova, S. IV-523
Fogelson, A.L. IV-1202
Forné, J. III-174
Forster, F. III-90
Frączak, M. II-451
Franca, L.P. III-725
Fredriksson, M. III-946
Freitag, F. I-10
Frijns, A.J.H. IV-661
Fritsche, W. IV-1165
Fudos, I. IV-258
Funika, W. I-462, III-472
Fyfe, C. II-650

Gabrijelčič, D. III-179
Gaffiot, F. IV-1037
Galán, R. I-523
Galis, A. III-141, III-166
Gall, M. II-467, IV-673
Gallivan, K.A. III-740
Gao, F. IV-1056
García, F.J. III-158
García, J.M. I-541
García, P. III-98
García-Loureiro, A.J. II-590
Gardner, H. IV-1218
Gargiulo, G. IV-351
Garic, S. I-148
Garstecki, Ł. III-74
Gashkov, I. IV-370
Gashkov, I.B. II-561
Gashkov, S.B. II-561
Gava, F. III-50, III-331
Gavidia, D.P. III-970
Gdaniec, Z. III-1013
Gecow, A. III-552
Gelas, J.-P. III-202
Geyer, C.F.R. III-2
Ghattas, O. III-717
Ghim, H. I-43
Giczi, D. I-83
Gillibert, L. IV-343
Gilmore, S. III-289, III-306
Gimenez, J. I-466
Ginting, V. III-701
Glasa, J. IV-18
Głut, B. II-163, II-646
Goh, K.-I. III-1038
Goh, Y.K. II-332
Goldstein, R.A. IV-718
Gomes, A.J.P. IV-221
Gómez-Sánchez, E. I-495
González, P. I-132
González-Castaño, F.J. I-326

González-Yuste, J.M. II-642
Gordon, B.W. II-292
Górriz, J.M. II-50
Goscinski, A. I-397
Gou, J. I-503
Govaerts, W. II-42
Gracio, D.K. II-540
Grama, A. III-740
Gray, L. III-756
Greenberg, J.P. I-75
Gregory, R.W. IV-1056
Griffith, V. IV-1165
Grochowski, M. III-623
Grønbæk, K. III-921
Gronek, P. II-638
Groşan, C. II-674, III-1257, III-1281
Großmann, S. III-865
Groumpos, P. II-515
Gschaider, B.F.W. IV-564
Guan, Y. IV-964
Gubała, T. III-34
Guimerà, R. III-1078
Gunn, C. III-905
Guo, W. II-388
Gurov, T.V. IV-507

Ha, J.S. IV-90, IV-131
Habala, O. I-124
Haber, R.E. I-523
Haddow, P.C. III-1249
Hadjarian, A. II-678
Haley, P.J. III-685
Hammami, M. I-310
Hampton, S.S. II-268
Han, H.S. III-258
Han, K. I-271, I-279, I-334, II-309, II-317
Han, K.-H. I-598
Han, N. II-309
Han, S. I-499, I-586, III-1185
Han, S.K. I-255
Han, S. II-427
Han, Z. I-623
Hanzliček, P. IV-1133
Harakaly, R. III-202
Havran, V. IV-164
Hayes, M. I-478
Henze, D. III-709
Heritage, J.P. III-1224
Hernández, V. I-482, III-54
Herrero, P. I-433, I-545, III-575, III-1004, IV-252
Herruzo, E. I-409
Hertzberger, B. I-515, III-771
Heyfitch, V. IV-1004
Hieb, M. III-835
Higaki, H. I-691
Hilaire, V. III-638
Hilbers, P.A.J. IV-661
Hillston, J. III-289
Hippold, J. II-146
Hluchy, L. I-124, I-425, III-599
Hochreiter, R. I-83
Hoekstra, A.G. IV-548
Hofacker, I.L. IV-728
Hoffmann, C. III-740
Hoffmann, G.A. II-682
Holmgren, S. II-9
Honeger, C.C. IV-564
Hong, I. I-611
Hong, J. III-1185
Hong, M. II-491, II-499
Hong, M. II-300
Hong, V. IV-34
Horák, B. III-631
Horan, P. I-397
Horie, K. IV-915
Houstis, E. III-693, III-732
Hsieh, S.-Y. I-223
Hu, J. I-174
Hu, Q. IV-923
Hu, Y. II-404, IV-956, IV-981, IV-989
Huang, G. IV-907
Huang, L. I-627, III-408
Huang, W. I-206
Huang, Y. III-1209, III-1224
Huerta, J. IV-197
Huh, E.N. III-1162, III-1193
Huttunen, P. III-10
Hwang, D.-U. I-255
Hwang, I.-C. I-43
Hwang, I.-S. I-594
Hwang, S. I-116, I-445
Hyun, Y.K. IV-181

Iglesias, A. IV-229, IV-237, IV-303
Ihm, I. II-419
Ikonen, J. III-10
Im, E.-J. I-116
Imamiya, A. III-897

Imre, S. I-318
Ingram, D. IV-1165
Ipanaqué, R. IV-303
Ivanova, T. IV-388
Ivanovska, S. IV-499
Iwadate, Y. I-519
Ixaru, L.G. IV-443
Izaguirre, J.A. II-268

Jacobs, R.L. II-332
Jaeschke, M. III-218
Jalili-Kharaajoo, M. II-662, II-708, II-713, II-717
Jang, J. III-124
Janicki, A. IV-803
Jankowski, G. I-449
Januszewski, R. I-449
Jaworek, K. IV-1103
Jean, K. III-166
Jeanblanc, M. IV-787
Jee, J.-G. I-342
Jeltsch, R. IV-1196
Jenkins, K. I-478
Jeong, C.-S. III-873
Jeong, H.Y. III-266
Jeong, K. I-116
Jeong, S.-J. III-1146
Jeong, Y. III-281
Jessup, E.R. IV-1242
Jezierski, J. I-532
Jiang, J. II-113
Jiang, J.Q. II-666
Jiang, L. III-58
Jiang, M. IV-923
Jiang, Y. III-591
Jigang, W. I-200
Jiménez, J.E. I-523
Jin, H. I-623
Jing, Z. I-373
Jinshu, S. I-373
Jo, G.-S. III-513
Joan-Arinyo, R. IV-139
Johannesson, L. III-946
Johns, C. III-725
Johnson, C.R. III-701, IV-1202
Jolivet, V. IV-115
Jones, G. III-701
Joo, B.-g. III-1185
Joshi, A. III-732
Jost, G. I-466
Jovanovic, V. III-748
Jugravu, A. III-18
Jung, H. I-43, II-317
Jung, J.J. I-247, III-513
Jung, K.-Y. I-295
Jurczyk, T. II-646
Jyung, K.J. III-274

Kacsuk, P. IV-475
Kahng, B. III-1038
Kalkuhl, M. II-228
Kallivokas, L.F. III-717
Kalra, D. IV-1165
Kamiński, M. II-171
Kaminski, M. IV-1029
Kan, Y. I-373
Kang, C. I-590
Kang, D. I-421, II-113
Kang, H.-K. IV-1110
Kang, K. III-1117
Kang, S.-S. I-670
Kang, Y. III-1117
Kapałka, M. III-34
Kar, T. II-260
Karaivanova, A. IV-499
Karniadakis, G.E. III-827
Kasperska, E. II-634
Katarzyniak, R. III-567
Katsionis, G. III-962
Katz, M. I-75
Katzfey, J. I-164
Kawulok, L. III-218
Kechadi, M.T. III-210
Kendall, R. III-408
Khan, J.I. III-978, III-1008
Khanal, M. II-364
Kharitonovich, A. IV-1049
Khrustaleva, E.Y. II-73
Kim, B.J. IV-1065
Kim, C. II-491
Kim, D. III-1038
Kim, D.-H. I-43
Kim, D.-W. III-363, III-371
Kim, D.H. II-687, II-695
Kim, E.J. III-717
Kim, H. I-43, I-577, I-594
Kim, H.-K. II-179
Kim, H.S. IV-1095
Kim, J.-H. I-598
Kim, J.-S. I-569, III-1170

Kim, J.A. I-255
Kim, J.H. IV-1088
Kim, J.O. I-687
Kim, J.R. I-519
Kim, K. III-1185
Kim, K.J. I-631
Kim, M. I-582, I-602, III-1217
Kim, S.-M. IV-1095, IV-1110
Kim, S.S. III-274
Kim, T.K. III-250
Kim, W. II-212, II-491, II-499
Kim, W.-H. IV-1110
Kim, Y. I-499, I-639
Kim, Y.-C. III-1209, III-1224
Kim, Y.H. IV-1081, IV-1095
Kirstein, P. IV-1165
Kisiel-Dorohinicki, M. III-646
Kitowski, J. I-458, III-583
Kleijn, C.R. IV-637
Klicker, K.R. II-540
Klie, H. III-805
Knight, D. III-748
Ko, Y.-B. I-598
Königsmann, J. III-946
Koestler, H. IV-410
Koetsier, J. II-650
Kokosiński, Z. I-215
Kokoszka, P. IV-827
Kolingerová, I. IV-123, IV-147
Kołodziej, M. I-215
Kominek, P. III-1241
Kommineni, J. I-164
Kondratieva, P. IV-164
Kong, J. II-388
Koo, J. I-349
Korczak, J.J. IV-835
Kormentzas, G. III-195
Kosacki, K. III-623
Koshigoe, H. II-105
Kou, G. IV-931
Koukam, A. III-638, III-1102
Kowalczuk, M. IV-687, IV-694
Kowarschik, M. III-440
Kozlak, J. III-638, III-1102
Kozlowski, J. III-521
Kozlowski, T. IV-1188
Krammer, B. III-464
Kramp, G. III-921
Krawczyk, H. I-557
Krawczyk, K. III-583, III-599
Kremens, R. III-725
Krishnan, A. II-284
Krumm-Heller, A. III-905
Kudela, H. II-356
Kuksin, A.Y. IV-596
Kulikov, G.Y. II-73, II-565, II-570
Kulikowski, J.L. III-488
Kulvietiene, R. IV-278
Kulvietis, G. IV-278
Kuo, T.W. III-226
Kupczyk, M. I-91
Kurc, T. III-805
Kurdziel, M. I-462
Kurgan, E. II-244, II-252
Kurose, J. III-670
Kurzyniec, D. I-140
Kutner, R. II-467, IV-673
Kuznetsov, Y.A. II-42
Kwak, B.M. IV-1088
Kwarciany, K. I-215
Kwiatkowska, T. IV-787
Kwon, C.H. III-234
Kwon, O. I-499
Kwon, S.Y. IV-1065
Kwon, Y. I-607
Kyriazis, I. IV-258

Labarta, J. I-10, I-466
Laclavik, M. III-599
Ladevèze, P. IV-612
Lai, K.K. IV-876
Lamantia, F. IV-779
Lambert, S. III-583
Landman, I. IV-1045
Langer, M. IV-1021, IV-1037
Lannocca, M. IV-1073
Latash, M.L. IV-1056
Lawenda, M. III-1013
Lawrence, S. III-764
Lawson, S.W. III-936
Lazarov, R. III-701
Lee, B.-H. III-1201
Lee, B.G. III-258
Lee, C. II-300
Lee, D.-Y. II-691
Lee, D.C. I-594, I-631
Lee, G.-S. I-389
Lee, H. I-639, II-499
Lee, H.-H. I-365
Lee, H.-J. III-1209

Lee, H.K. III-281, III-1124, III-1217
Lee, H.P. II-666
Lee, H.S. II-121
Lee, J. I-590
Lee, J.-B. II-179
Lee, J.-D. I-18
Lee, J.-H. I-295
Lee, K.-H. I-381
Lee, K.-J. I-569
Lee, K.-W. I-381
Lee, S. I-577, I-639
Lee, S.-G. IV-1081
Lee, S.-H. I-365, III-371
Lee, S.J. IV-1065
Lee, T.-D. III-873
Lee, T.-J. I-573, III-1201
Lee, W.-G. III-363
Lee, Y.-H. I-683
Leemput, P. Van IV-572
Lees, M. III-881
Lefèvre, L. III-202
Leone, P. IV-540
Lermusiaux, P.F.J. III-685
Leszczynski, J.S. II-348
Lewis, R.W. II-622
Li, J. III-388, III-414
Li, J. III-889
Li, J. IV-892
Li, M. I-174
Li, S. I-441, IV-1056
Li, X. I-441
Li, Y. II-26, II-574
Li, Z.-M. IV-1056
Liang, Y.C. II-666
Liao, W. III-709
Lim, D. I-279
Lim, H.W. I-474
Lim, J. III-266
Lim, S. II-396
Lin, D. IV-295
Lin, W. II-58, II-81, II-574
Lin, W.D. I-549
Lipinski, P. IV-835
Lipowski, A. IV-742
Lipowsky, R. III-1062
Lisik, Z. IV-1021, IV-1037
Lisowska, A. IV-1
Liu, C.S. III-226
Liu, H. III-58
Liu, J. IV-892
Liu, L. II-372
Liu, R. II-372
Liu, Z. II-26, III-408
Lloyd, D. IV-1165
Lluch, J. IV-197
Lobry, J.R. IV-679
Löf, H. II-9
Loffeld, O. II-228
Logan, B. III-881
Loitière, Y. III-789
Lombardo, S. III-106
Longo, S. IV-580, IV-588
Lopez, L. IV-459
Loulergue, F. I-401
Low, M.Y.H. III-856
Lu, F. IV-964, IV-972
Lu, J. I-421, II-113
Lu, T. IV-653
Lu, X. I-287
Lucor, D. III-827
Luengo, F. IV-229, IV-237
Lüthi, J. III-865
Lukac, R. IV-72
Lukasik, J. III-1102
Luna, J.M. I-561
Luo, S.P. II-58
Luo, Y. II-404, IV-940, IV-956, IV-981, IV-989, IV-997
Luo, Z. III-897
Luque, E. II-220
Lust, K. IV-572
Luszczek, P. IV-270
Luzón, M.V. IV-139

Ma, Y. III-978
MacDonald, D. II-650
Macedo Mourelle, L. de III-1289
Machì, A. III-106
Machiraju, R. III-764
Machmoum, A. II-89
Mackiewicz, D. IV-687, IV-694
Mackiewicz, P. IV-687, IV-694
Maeng, S.-R. I-43
Magaña, E. III-133
Magdoń-Maksymowicz, M. IV-750, IV-758
Majerowski, A. IV-673
Majewska, M. III-583
Maksymowicz, A.Z. IV-758
Malarz, K. II-638

Malawski, M. III-34, III-848
Malowiecki, M. III-537
Mandel, J. III-725
Manohar, P. III-1178
Manos, K. III-962
Mantiuk, R. IV-264
Mao, Y. I-470
Marangi, C. IV-426
Marco, J. IV-245
Margalef, T. II-220
Mari, J.-L. IV-156
Marín, M. III-480
Marinov, D. III-187
Markvoort, A.J. IV-661
Martín, M.J. I-132
Martínez, G. III-158
Martucci, S. IV-418
Mat Deris, M. III-379
Matossian, V. III-805
Matsuhisa, T. IV-884, IV-915
Mazurkiewicz, J. II-236
McCarthy, J.J. III-685
McGregor, J.L. I-164
McLaughlin, D. III-670
Mead, J. IV-451
Meer, H. De III-90
Mehta, M. III-946
Mehta, S. III-779
Melnik, R.V.N. IV-645
Meng, Z. IV-923
Merkulov, A.I. II-73
Mesquita, A. III-1265
Metaxas, D. III-813
Meyer, A. II-146
Meyer, N. I-91, I-449, III-1013
Meyer, R.R. I-326
Miatliuk, K. IV-1103
Michelson, G. IV-41, IV-49
Michopoulos, J. III-693, III-732, IV-621
Midura, J. III-114
Mielikäinen, T. I-231
Mihajlović, M.D. II-97
Mihaylova, L. IV-531
Mijalković, S. II-97
Mikolajczak, R. I-449
Min, B.-M. I-342
Mock, S. I-75
Moltó, G. I-482
Momot, J. III-623
Mondéjar, R. III-98
Montañés, E. I-666, II-658
Montenegro, R. II-642
Montero, G. II-642
Moon, K. III-124
Moon, S. I-334
Moore, S. III-432
Moore, T. III-202
Moreno, F. IV-213
Morey, J. III-996
Morgado, J.F.M. IV-221
Morozov, I.V. IV-596
Morse, K.L. III-835
Morzy, T. I-532
Mouriño, J.C. I-132
Mucci, P. III-432
Müller, M.S. III-464
Mukherjee, B. III-1209, III-1224
Mun, J.H. IV-1095
Mun, Y. III-281, III-1124, III-1217
Muñoz, J.L. III-174
Murawko-Wisniewska, K. IV-264
Murzin, F. III-545
Murzyn, K. II-325

Nachev, G. III-187
Napieralski, A. IV-1029
Navarro-Moldes, L. I-615
Ndoundam, R. II-610
Nedea, S.V. IV-661
Nedjah, N. III-1289
Nenortaite, J. IV-843
Néron, D. IV-612
Neuman, L. III-521
Nguyen, N.T. III-537
Ni, J. I-619
Nikolow, D. I-458
Nipp, K. IV-1196
Nobile, F. III-756
Nöh, K. II-228, II-594
Noël, A.G. IV-335
Noh, B. I-639
Noh, B.-N. I-365
Noh, M. III-281
Noh, S.-C. I-631
Nordén, M. II-9
Norman, G.E. IV-596
Novikava, S. IV-1103
Nowak, G. II-654
Nowiński, A. I-99
Nowiński, K.S. I-99

Nowicka, A. IV-687, IV-694
Nowicki, E. II-483
Nowostawski, M. III-552
Nunes, L.M. II-380

O'Conor, I. IV-1037
O'Hare, G. III-954
Ochmańska, E. I-527
Oden, J.T. III-756
Odintsov, E. IV-1049
Oguara, T. III-881
Oh, H.S. I-342
Oh, S.-C. II-212
Oh, S.-K. II-179, II-188, II-691
Okoń, M. III-1013
Oltean, M. III-1257, III-1281
Oltean, M. II-670, II-674, III-1257, III-1281
Ono, M. I-691
Onuma, K. IV-629
Orlando, S. I-263
Ortobelli, S. IV-779
Oziewicz, Z. II-602

Pacheco, M.A.C. III-1273
Pachowicz, P. II-678
Pachter, R. II-260
Pairot, C. III-98
Palak, B. I-91
Palus, H. IV-34
Pan, G. II-427
Pan, Y. I-441
Papadopoulos, P. I-75
Papageorgiou, E. II-515
Paprzycki, M. I-490
Parashar, M. III-58, III-805
Pareschi, L. II-276
Park, B. I-271
Park, B.-J. II-188
Park, C.-I. III-363, III-371
Park, C.-I. II-598
Park, D.-H. I-295
Park, G.-L. I-453
Park, H. I-116, I-445, I-499
Park, H.-K. III-1224
Park, H.J. III-274
Park, J.D. III-266
Park, K. I-499, III-1185
Park, N. III-124
Park, S. I-499, II-419
Parsa, S. II-699
Parthasarathy, S. III-764
Pasenkiewicz-Gierula, M. II-325
Paternoster, B. IV-418, IV-443
Patrikalakis, N.M. III-685
Patterson, D. IV-1165
Paulus, D. IV-34
Pavlovsky, I. IV-1045
Pawlaczyk, L. IV-49
Payne, D.A. II-540
Paz-Ramos, M.A. II-532
Pecheanu, E. III-497
Pedroza, A.C.P. III-1265
Pedrycz, W. II-179, II-188
Pegueroles, J. III-174
Pęka lski, A. IV-765
Peña, J.M. III-347
Peng, Y. IV-931
Perego, R. I-263
Pérez, F. III-347
Pérez, F.A. II-130
Pérez, G. I-561
Pérez, M.S. III-347
Perng, N.C. III-226
Perrieère, G. IV-700
Petcu, D. I-490
Philips, B. III-670
Pieczyńska-Kuchtiak, A. III-567
Pineda, U. I-561
Pingali, K. III-797
Piotrowski, M. I-557
Plank, J. III-202
Plata, O. I-409
Plataniotis, K.N. IV-72
Platen, E. IV-771
Plemenos, D. IV-80, IV-115
Pllana, S. III-42, III-456
Płóciennik, M.P. I-91
Polak, N. IV-687, IV-694
Polgár, B. III-1153
Politi, T. IV-467
Polymenakos, L. III-195
Pommereau, F. III-322
Popenda, Ł. III-1013
Porras, J. III-10
Porwik, P. IV-1
Pothoven, S. III-986
Poulingeas, P. IV-115
Pramanik, S. III-889
Preißner, O. III-929

Press, S. IV-1218
Presti, G.L. I-156
Preston, M. III-670
Primet, P. III-202
Prodan, R. II-459
Prudhomme, S. III-756
Prusinkiewicz, P. II-65, II-138
Prykarpatskyy, Y. IV-653
Pucci, G. II-614
Puhalskii, A. III-725
Pullen, J.M. III-835
Puntonet, C.G. II-50
Purvis, M. III-552

Qi, H. I-503
Qi, L. III-388
Qi, X. I-405
Qiu, X. III-1209, III-1224
Quevedo, J.R. I-666
Quintero-Marmol-Marquez, E. II-532
Quirós, R. IV-197

Ra, I.H. III-234
Rabenseifner, R. I-1
Rabiei, M. III-379
Rachev, S. IV-779
Radomski, J.P. IV-709
Ragni, S. IV-426
Rajtar, T. III-1013
Rakthanmanon, T. I-303
Ramos, J.F. IV-107
Ranilla, J. I-666, II-658
Rasmus, S. IV-795
Ravantti, J. I-231
Ray, A. I-200
Re, G.L. I-156
Rebollo, C. IV-173
Redl, C.E.P. IV-564
Regucki, P. II-356
Reilly, R. III-954
Remolar, I. IV-173
Rendell, A.P. II-17
Resch, M.M. III-464
Reynolds, P. III-789
Rhee, S.-B. I-683
Ribelles, J. IV-173
Richie, J.T. II-726
Richman, M.B. III-678
Ridwan, A.M. II-284
Rinaldi, N. IV-1021
Ro, Y.M. IV-1110
Roberts, S. IV-1218
Robinson, A.R. III-685
Robles, V. III-347
Robshaw, M.J.B. I-474
Rodgers, G.J. III-1054
Rodionov, A.S. I-565
Rodionova, O.K. I-565
Rodríguez, E. II-642
Rodríguez, M. II-204
Róg, T. II-325
Roh, S.-B. II-691
Rojek, G. III-607
Román, J.E. III-54
Romero, S. IV-213
Rosanna, M. IV-351
Ruchaud, W. IV-80
Ruede, U. IV-410
Rünger, G. II-1, II-146
Rum, F. II-292
Ruskin, H.J. III-1094
Rust, B.W. IV-1226
Rycerz, K. III-34, III-848
Rząsa, W. I-107

Sacerdoti, F. I-75
Şahan, A. I-357
Sahingoz, O.K. I-184, I-437
Salamanca, E. III-133
Salmerón, M. II-50
Saltz, J. III-805
Sameh, A. III-740
Samulyak, R. IV-653
Sánchez, A. III-347
Sands, J. III-936
Sandu, A. III-709
Sanguesa, R. I-67
Sanjeevan, K. I-67
Santana, M.J. I-537, I-661
Santini, C.C. III-1273
Santos, J.A. II-380
Santosa, B. III-678
San Juan, J.F. IV-327
Sarafian, H. IV-319
Sas, C. III-954, III-1017
Sato, M. IV-629
Savanović, A. III-141, III-179
Sbert, M. IV-189, IV-245
Schaefer, R. III-623
Schikuta, E. I-486

Schmaranz, K. IV-1156
Schmidt, H. III-685
Schmidt, P. II-244
Schneider, R. IV-588
Schrefler, B.A. IV-612
Schubert, W. II-364
Schwartmann, D. IV-1148, IV-1156
Seaïd, M. II-89, II-276
Searcy, Mary E. II-726
Sedig, K. III-996, III-1030
Segal, C. III-497, III-560
Segovia, J. III-1004
Seidel, H.-P. IV-164
Seinfeld, J.H. III-709
Sekelsky, S. III-670
Sellarès, J.A. IV-98
Semerdjiev, T. IV-531
Semovski, S.V. IV-736
Senger, L.J. I-661
Seo, S.J. I-687
Seoane, N. II-590
Sequeira, J. IV-156
Sérot, J. III-314
Serrano, S. IV-327
Serrat, J. III-133
Seymour, K. III-432
Shadwick, B.A. IV-1234
Shakhov, V.V. III-1138
Shannon, C.E. III-701
Shen, M. III-388
Shen, M.M. III-242
Shen, P. IV-972
Shi, X.H. II-666
Shi, Y. I-657, IV-892, IV-931
Shi, Z. III-355
Shin, B.-S. II-396
Shin, D.-R. III-1170
Shin, M.H. III-234
Shin, Y.-s. IV-205
Shindin, S.K. II-570
Sidi, A. IV-402
Sikorski, C.A. IV-1202
Silver, D. III-58
Silvestri, F. I-263
Simo, B. I-124
Simpson, J. III-701
Simutis, R. IV-843
Singhal, M. II-540
Sitarz, M. II-638
Siver, A.S. IV-358
Skala, K. III-119
Skala, V. II-412
Skarmeta, A.F.G. III-98, III-158
Skitał, Ł. I-458
Sławińska, M. I-35
Słomiński, L. IV-811
Sloot, P.M.A. III-848, III-970
Słota, D. II-634
Słota, R. I-458, III-583
Slusarczyk, K. IV-1029
Smętek, M. III-472
Smolarczyk, K. IV-687, IV-694
Smutnicki, C. II-483
Snášel, V. III-631
Snider, R. III-822
Sobachkin, A. IV-1049
Sobaniec, C. III-82
Sobecki, J. III-505
Sobera, M.P. IV-637
Sohn, S. III-124
Sojat, Z. III-119
Soldatos, J. III-195
Song, J. I-577
Song, S.K. I-453
Songsiri, C. I-303
Soon, Y.H. IV-181
Sorensen, D. III-740
Soriano, M. III-174
Špidlen, J. IV-1133
Spisak, B.J. II-630
Sridhar, V. III-1178
Srovnal, V. III-631
Stadler, P.F. IV-728
Stagni, R. IV-1073
Stals, L. IV-1218
Stąpor, K. IV-41, IV-49
Stauffer, D. IV-709
Steenhoven, A.A. van IV-661
Stegailov, V.V. IV-596
Stephan, E.G. II-540
Stevens, E. I-561
Stodghill, P. III-797
Stokłosa, D. III-1013
Storniolo, P. I-156
Strazdins, P.E. II-17
Stroiński, M. I-91, I-449, III-1013
Strych, V. IV-123
Stuer, G. I-51
Su, C.-H. III-827
Subasi, D. II-582

Sudholt, W. I-148
Summer, M. III-1070
Sun, J. I-623
Sun, X. IV-956, IV-981, IV-989
Sun, X. II-404
Sunderam, V. I-51, I-140
Sung, H. I-586
Susa, H. IV-629
Suttles, D. III-889
Suzuki, N. III-1046
Suzuri, H.M. III-379
Święcicki, M. II-703
Świtonski, A. IV-41
Szatzschneider, W. IV-787
Szegő, D. I-553
Szepieniec, T. I-107
Szwarcman, M.H. III-1273
Szychowiak, M. II-475
Szymczyk, M. I-417
Szymczyk, P. I-417

Tabery, P. III-187
Taccogna, F. IV-588
Tack, G.R. IV-1065
Tadić, B. III-1086
Tadonki, C. I-679, II-610
Tae, K.S. I-453
Takahashi, D. IV-629
Tan, A. III-166
Tang, J. II-404, IV-956, IV-981, IV-989
Tanscheit, R. III-1273
Taskinoglu, E. III-748
Tchuente, M. II-610
Teixeira, M.M. I-537
Tejchman, J. II-340, II-626
Tempone, R. III-756
Testori, J. III-456
Teyssière, G. IV-827
Thambipillai, S. I-200
Thandavan, A. IV-475
Thanh, N.N. IV-1110
Theodoropoulos, G. III-881
Thompson, D. III-764
Thurner, S. III-1070
Thysebaert, P. III-149
Tian, J. IV-907
Tian, R. III-685
Tolk, A. III-835
Tomas, J. II-364
Tomaševic, V. I-644
Tong, C. II-372
Tong, W. I-511
Torres-Jimenez, J. II-532
Tosik, G. IV-1037
Trafalis, T.B. III-678
Tran, V.D. I-124, I-425
Trease, L.L. II-540
Tricio, V. II-204
Trinitis, C. III-440
Trück, S. IV-859
Truong, H.-L. III-424
Truong, T. IV-1202
Tsifakis, D. II-17
Tsishkou, D. I-310
Tsompanopoulou, P. III-732
Tufo, H.M. IV-1242
Tufte, G. III-1249
Turck, F. De III-149
Turner, S.J. III-856

Überhuber, C. II-459
Ufuktepe, Ü. IV-376
Uhruski, P. III-623
Ukkonen, E. I-231
Um, T.-W. I-607
Umemura, M. IV-629
Urso, A. I-156
Ushakov, A. IV-1045

Vaquero-González, L.M. I-495
Vargas, C. I-561
Varnuška, M. IV-147
Vayssade, M. II-163
Vázquez, P.-P. IV-245
Vega-Gorgojo, G. I-495
Venetsanopoulos, A.N. IV-26, IV-72
Venkataraman, S. III-813
Verdonk, B. I-695, II-557
Vernov, S.Y. IV-382
Veselý, A. IV-1133
Vial, J.-P. I-679
Vianello, M. IV-434
Villa, F.J. I-541
Villazon, A. III-42
Virvou, M. III-962
Visser, A. I-515, III-771
Vivó, R. IV-197
Vlček, M. IV-56, IV-64
Vodacek, A. III-725
Vogler, C. III-813

Volckaert, B. III-149
Voss, H. II-34
Vuong, S. III-889

Wais, P. II-703
Waiyamai, K. I-303
Wajs, W. II-703
Walentyński, R.A. IV-311
Walkowiak, T. II-236
Wang, D. I-174, III-414, III-1131
Wang, G. I-648
Wang, J. II-372, II-404, IV-956, IV-981, IV-989
Wang, L. IV-645
Wang, R. I-413
Wang, S. III-414
Wang, S. IV-876
Wang, S. III-1209
Wang, X. IV-940, IV-997
Wang, Y. II-404, II-427, IV-956, IV-981, IV-989
Wang, Y. IV-900
Węgiel, M. III-26
Weide, H. van der IV-851
Weidendorfer, J. III-440
Weishäupl, T. I-486
Weron, R. IV-859, IV-868
Wesela, W. II-467
Wheeler, M.F. III-805
Whitlock, P.A. IV-507
Wiechert, W. II-228, II-594, IV-1249
Wiesinger, C. I-83
Wiktorsson, M. IV-795
Winnem, O.M. IV-1126
Wismüller, R. I-462
Witek, H. I-140
Wojciechowski, T. IV-811
Wołoszyn, M. II-630
Won, Y. I-519
Workman, D. III-986
Woźniak, A. III-615
Wozniak, M. I-675
Wozny, J. IV-1021
Wu, Y. IV-295, IV-876
Wu, Z. I-59, I-441, I-470, I-627, II-427
Wu, Z. I-507
Wypychowski, J. I-99

Xhafa, F. I-429
Xia, Z. III-591
Xian, J. II-58, II-574
Xu, B. I-405, I-421, II-113
Xu, C. I-441
Xu, J. I-59
Xu, W. IV-892
Xu, X. III-978
Xu, Z. IV-940, IV-997
Xue, J. II-26
Xue, Y. II-404, IV-956, IV-981, IV-989

Yakali, H. I-515, III-771
Yakutovich, M. IV-515
Yamin, A.C. III-2
Yan, L. I-619
Yanami, H. IV-286
Yang, G. IV-295
Yang, J. I-503
Yang, L. II-26
Yang, S.S. III-1008
Yang, S.-Y. II-81
Yang, X. I-478
Yang, X.S. II-622
Yang, Z. IV-900
Yaroslavtsev, A.F. I-573
Yazici, A. IV-364
Ye, Yang II-435
Yeguas, E. IV-139
Yeomans, J.M. IV-556
Yeon, L.S. IV-181
Yijie, W. I-373
Yoo, J.H. III-258
Yoo, K.-H. IV-90, IV-131
Yoo, S.-H. III-873
Yoon, H.-W. III-1201
Yoon, K. II-212
Yoon, W.-S. I-598
You, H. III-432
Youn, H.Y. I-453, II-121
Yu, H.L. III-242
Yu, L. IV-876
Yu, Y.-H. III-513
Yu, Z. I-627
Yuan, Z. III-856

Zahradnik, P. IV-56, IV-64
Zatsiorsky, V.M. IV-1056
Zdrojewska, D. IV-264
Żebrowski, D. IV-673
Zgrzywa, A. III-521
Zgrzywa, M. III-537

Zhang, A. IV-827
Zhang, A. II-404, IV-956, IV-981
Zhang, B. II-388
Zhang, D. I-206
Zhang, H. II-622
Zhang, J. III-355
Zhang, J. III-1209, III-1224
Zhang, L. I-287
Zhang, M. IV-972
Zhang, S. III-591, IV-907
Zhang, X. IV-900
Zhang, Y. I-413
Zhao, H. III-748
Zhao, W. III-725
Zhen, W.M. III-242
Zheng, G. I-59
Zheng, W. I-174, III-1131, IV-295
Zhi, X. I-511
Zhong, S. II-404, IV-956, IV-981, IV-989
Zhong, Y. III-591
Zhou, H. III-1062
Zhou, M. IV-948
Zhou, X. III-355
Zhu, K. III-1209
Zhu, L. IV-907
Zhu, Y. III-822
Zieliński, K. III-114, III-218, IV-1172
Zieliński, S. IV-1172
Ziemkiewicz, B. IV-819
Zlatev, Z. IV-491
Zoetebier, J. I-515, III-771
Zorin, S. IV-1012, IV-1049
Zubik-Kowal, B. IV-451
Zudilova, E.V. III-970
Zuijlen, A. van IV-604
Zvárová, J. IV-1133
Zwierz, J. IV-803