Lecture Notes in Computer Science 3211

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

Editorial Board

David Hutchison
Lancaster University, UK

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

Aurélio Campilho Mohamed Kamel (Eds.)

Image Analysis and Recognition

International Conference, ICIAR 2004
Porto, Portugal, September 29 - October 1, 2004
Proceedings, Part I

Springer

Volume Editors

Aurélio Campilho
University of Porto
Institute of Biomedical Engineering, Faculty of Engineering
Rua Dr. Roberto Frias, s/n, Edif. I Poente, I 319
4200-465 Porto, Portugal
E-mail: campilho@fe.up.pt

Mohamed Kamel
University of Waterloo
Department of Electrical and Computer Engineering
Waterloo, Ontario N2L 3G1, Canada
E-mail: mkamel@pami.uwaterloo.ca

Library of Congress Control Number: 2004112583

CR Subject Classification (1998): I.4, I.5, I.3, I.7.5

ISSN 0302-9743
ISBN 3-540-23223-0 Springer 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. Violations are liable to prosecution under the German Copyright Law.

Springer 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: 11319733 06/3142 5 4 3 2 1 0

Preface

ICIAR 2004, the International Conference on Image Analysis and Recognition, was the first ICIAR conference, and was held in Porto, Portugal. ICIAR will be organized annually, and will alternate between Europe and North America. ICIAR 2005 will take place in Toronto, Ontario, Canada. The idea of offering these conferences came as a result of discussion between researchers in Portugal and Canada to encourage collaboration and exchange, mainly between these two countries, but also with the open participation of other countries, addressing recent advances in theory, methodology and applications.

The response to the call for papers for ICIAR 2004 was very positive. From 316 full papers submitted, 210 were accepted (97 oral presentations, and 113 posters). The review process was carried out by the Program Committee members and other reviewers; all are experts in various image analysis and recognition areas. Each paper was reviewed by at least two reviewing parties. The high quality of the papers in these proceedings is attributed first to the authors, and second to the quality of the reviews provided by the experts. We would like to thank the authors for responding to our call, and we wholeheartedly thank the reviewers for their excellent work in such a short amount of time. We are especially indebted to the Program Committee for their efforts that allowed us to set up this publication.

We were very pleased to be able to include in the conference, Prof. Murat Kunt from the Swiss Federal Institute of Technology, and Prof. Mário Figueiredo, of the Instituto Superior Técnico, in Portugal. These two world-renowned experts were a great addition to the conference and we would like to express our sincere gratitude to each of them for accepting our invitations.

We would also like to thank Prof. Ana Maria Mendonça and Prof. Luís Corte-Real for all their help in organizing this meeting; Khaled Hammouda, the webmaster of the conference, for maintaining the Web pages, interacting with authors and preparing the proceedings; and Gabriela Afonso, for her administrative assistance. We also appreciate the help of the editorial staff from Springer for supporting this publication in the LNCS series.

Finally, we were very pleased to welcome all the participants to this conference. For those who did not attend, we hope this publication provides a brief view into the research presented at the conference, and we look forward to meeting you at the next ICIAR conference, to be held in Toronto, 2005.

September 2004 Aurélio Campilho, Mohamed Kamel

ICIAR 2004 – International Conference on Image Analysis and Recognition

General Chair

Aurélio Campilho
University of Porto, Portugal
campilho@fe.up.pt

General Co-chair

Mohamed Kamel
University of Waterloo, Canada
mkamel@uwaterloo.ca

Local Chairs

Ana Maria Mendonça
University of Porto, Portugal
amendon@fe.up.pt

Luís Corte-Real
University of Porto, Portugal
lreal@inescporto.pt

Webmaster

Khaled Hammouda
University of Waterloo, Canada
hammouda@pami.uwaterloo.ca

Supported by

Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Portugal

INEB – Instituto de Engenharia Biomédica

Pattern Analysis and Machine Intelligence Group, University of Waterloo, Canada

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DO ENSINO SUPERIOR

Advisory and Program Committee

M. Ahmadi	University of Windsor, Canada
M. Ahmed	Wilfrid Laurier University, Canada
A. Amin	University of New South Wales, Australia
O. Basir	University of Waterloo, Canada
J. Bioucas	Technical University of Lisbon, Portugal
M. Cheriet	University of Quebec, Canada
D. Clausi	University of Waterloo, Canada
L. Corte-Real	University of Porto, Portugal
M. El-Sakka	University of Western Ontario, Canada
P. Fieguth	University of Waterloo, Canada
M. Ferretti	University of Pavia, Italy
M. Figueiredo	Technical University of Lisbon, Portugal
A. Fred	Technical University of Lisbon, Portugal
L. Guan	Ryerson University, Canada
E. Hancock	University of York, UK
M. Kunt	Swiss Federal Institute of Technology, Switzerland
E. Jerningan	University of Waterloo, Canada
J. Marques	Technical University of Lisbon, Portugal
A. Mendonça	University of Porto, Portugal
A. Padilha	University of Porto, Portugal
F. Perales	University of the Balearic Islands, Spain
F. Pereira	Technical University of Lisbon, Portugal
A. Pinho	University of Aveiro, Portugal
N. Peres de la Blanca	University of Granada, Spain
P. Pina	Technical University of Lisbon, Portugal
F. Pla	University of Jaume I, Spain
K. Plataniotis	University of Toronto, Canada
T. Rabie	University of Toronto, Canada
P. Scheunders	University of Antwerp, Belgium
M. Sid-Ahmed	University of Windsor, Canada
W. Skarbek	Warsaw University of Technology, Poland
H. Tizhoosh	University of Waterloo, Canada
D. Vandermeulen	Catholic University of Leuven, Belgium
M. Vento	University of Salerno, Italy
R. Ward	University of British Columbia, Canada
D. Zhang	Hong Kong Polytechnic, Hong Kong

Reviewers

M. Abasolo	University of the Balearic Islands, Spain
A. Adegorite	University of Waterloo, Canada
N. Alajlan	University of Waterloo, Canada
H. Araújo	University of Coimbra, Portugal
B. Ávila	Universidade Federal de Pernambuco, Brazil
Z. Azimifar	University of Waterloo, Canada
O. Badawy	University of Waterloo, Canada
J. Batista	University of Coimbra, Portugal
A. Buchowicz	Warsaw University of Technology, Poland
J. Caeiro	Beja Polytechnical Institute, Portugal
L. Chen	University of Waterloo, Canada
G. Corkidi	National University of Mexico, Mexico
M. Correia	University of Porto, Portugal
J. Costeira	Technical University of Lisbon, Portugal
R. Dara	University of Waterloo, Canada
A. Dawoud	University of South Alabama, USA
H. du Buf	University of the Algarve, Portugal
I. El Rube	University of Waterloo, Canada
L. Guan	Ryerson University, Canada
M. Hidalgo	University of the Balearic Islands, Spain
J. Jiang	University of Waterloo, Canada
J. Jorge	Technical University of Lisbon, Portugal
A. Kong	University of Waterloo, Canada
M. Koprnicky	University of Waterloo, Canada
R. Lins	Universidade Federal de Pernambuco, Brazil
W. Mageed	University of Maryland, USA
B. Miners	University of Waterloo, Canada
A. Monteiro	University of Porto, Portugal
J. Orchard	University of Waterloo, Canada
M. Piedade	Technical University of Lisbon, Portugal
J. Pinto	Technical University of Lisbon, Portugal
M. Portells	University of the Balearic Islands, Spain
A. Puga	University of Porto, Portugal
W. Rakowski	Bialystok Technical University, Poland
B. Santos	University of Aveiro, Portugal
J. Santos-Victor	Technical University of Lisbon, Portugal
G. Schaefer	Nottingham Trent University, UK
J. Sequeira	Laboratoire LSIS (UMR CNRS 6168), France
J. Silva	University of Porto, Portugal
J. Sousa	Technical University of Lisbon, Portugal
L. Sousa	Technical University of Lisbon, Portugal
X. Varona	University of the Balearic Islands, Spain
E. Vrscay	University of Waterloo, Canada
S. Wesolkowski	University of Waterloo, Canada
L. Winger	LSI Logic Canada Corporation, Canada

Table of Contents – Part I

Image Segmentation

Automatic Image Segmentation Using a Deformable Model Based on Charged Particles 1
Andrei C. Jalba, Michael H.F. Wilkinson, Jos B.T.M. Roerdink

Hierarchical Regions for Image Segmentation 9
Slawo Wesolkowski, Paul Fieguth

Efficiently Segmenting Images with Dominant Sets 17
Massimiliano Pavan, Marcello Pelillo

Color Image Segmentation Using Energy Minimization on a Quadtree Representation 25
Adolfo Martínez-Usó, Filiberto Pla, Pedro García-Sevilla

Segmentation Using Saturation Thresholding and Its Application in Content-Based Retrieval of Images 33
A. Vadivel, M. Mohan, Shamik Sural, A.K. Majumdar

A New Approach to Unsupervised Image Segmentation Based on Wavelet-Domain Hidden Markov Tree Models 41
Qiang Sun, Shuiping Gou, Licheng Jiao

Spatial Discriminant Function with Minimum Error Rate for Image Segmentation 49
EunSang Bak

Detecting Foreground Components in Grey Level Images for Shift Invariant and Topology Preserving Pyramids 57
Giuliana Ramella, Gabriella Sanniti di Baja

Pulling, Pushing, and Grouping for Image Segmentation 65
Guoping Qiu, Kin-Man Lam

Image Segmentation by a Robust Clustering Algorithm Using Gaussian Estimator 74
Lei Wang, Hongbing Ji, Xinbo Gao

A Multistage Image Segmentation and Denoising Method – Based on the Mumford and Shah Variational Approach 82
Song Gao, Tien D. Bui

A Multiresolution Threshold Selection Method Based on Training 90
J.R. Martinez-de Dios, A. Ollero

Segmentation Based Environment Modeling
Using a Single Image 98
Seung Taek Ryoo

Unsupervised Color-Texture Segmentation 106
Yuzhong Wang, Jie Yang, Yue Zhou

Image Processing and Analysis

Hierarchical MCMC Sampling 114
Paul Fieguth

Registration and Fusion of Blurred Images 122
Filip Sroubek, Jan Flusser

A New Numerical Scheme for Anisotropic Diffusion 130
Hongwen Yi, Peter H. Gregson

An Effective Detail Preserving Filter for Impulse Noise Removal 139
Naif Alajlan, Ed Jernigan

A Quantum-Inspired Genetic Algorithm
for Multi-source Affine Image Registration 147
Hichem Talbi, Mohamed Batouche, Amer Draa

Nonparametric Impulsive Noise Removal 155
Bogdan Smolka, Rastislav Lukac

BayesShrink Ridgelets for Image Denoising 163
Nezamoddin Nezamoddini-Kachouie, Paul Fieguth, Edward Jernigan

Image Salt-Pepper Noise Elimination by Detecting Edges
and Isolated Noise Points 171
Gang Li, Binheng Song

Image De-noising via Overlapping Wavelet Atoms 179
V. Bruni, D. Vitulano

Gradient Pile Up Algorithm for Edge Enhancement and Detection 187
Leticia Guimarães, André Soares, Viviane Cordeiro, Altamiro Susin

Co-histogram and Image Degradation Evaluation 195
Pengwei Hao, Chao Zhang, Anrong Dang

MAP Signal Reconstruction with Non Regular Grids 204
João M. Sanches, Jorge S. Marques

Comparative Frameworks for Directional Primitive Extraction 212
M. Penas, M.J. Carreira, M.G. Penedo, M. Mirmehdi, B.T. Thomas

Dynamic Content Adaptive Super-Resolution 220
Mei Chen

Efficient Classification Method for Autonomous Driving Application 228
Pangyu Jeong, Sergiu Nedevschi

Image Analysis and Synthesis

Parameterized Hierarchical Annealing for Scientific Models 236
Simon K. Alexander, Paul Fieguth, Edward R. Vrscay

Significance Test for Feature Subset Selection on Image Recognition 244
Qianren Xu, M. Kamel, M.M.A. Salama

Image Recognition Applied to Robot Control Using Fuzzy Modeling 253
Paulo J. Sequeira Gonçalves, L.F. Mendonça, J.M.C. Sousa, J.R. Caldas Pinto

Large Display Interaction Using Video Avatar and Hand Gesture Recognition 261
Sang Chul Ahn, Tae-Seong Lee, Ig-Jae Kim, Yong-Moo Kwon, Hyoung-Gon Kim

Image and Video Coding

Optimal Transform in Perceptually Uniform Color Space and Its Application in Image Coding 269
Ying Chen, Pengwei Hao, Anrong Dang

Lossless Compression of Color-Quantized Images Using Block-Based Palette Reordering 277
António J.R. Neves, Armando J. Pinho

Fovea Based Coding for Video Streaming 285
Çağatay Dikici, H. Işıl Bozma, Reha Civanlar

Influence of Task and Scene Content on Subjective Video Quality 295
Ying Zhong, Iain Richardson, Arash Sahraie, Peter McGeorge

Evaluation of Some Reordering Techniques for Image VQ Index Compression 302
António R.C. Paiva, Armando J. Pinho

Adaptive Methods for Motion Characterization and Segmentation of MPEG Compressed Frame Sequences 310
C. Doulaverakis, S. Vagionitis, M. Zervakis, E. Petrakis

On the Automatic Creation of Customized Video Content 318
José San Pedro, Nicolas Denis, Sergio Domínguez

Shape and Matching

Graph Pattern Spaces from Laplacian Spectral Polynomials 327
Bin Luo, Richard C. Wilson, Edwin R. Hancock

A Hierarchical Framework for Shape Recognition Using Articulated Shape Mixtures 335
Abdullah Al Shaher, Edwin R. Hancock

A New Affine Invariant Fitting Algorithm for Algebraic Curves 344
Sait Sener, Mustafa Unel

Graph Matching Using Manifold Embedding 352
Bai Xiao, Hang Yu, Edwin Hancock

A Matching Algorithm Based on Local Topologic Structure 360
Xinjian Chen, Jie Tian, Xin Yang

2-D Shape Matching Using Asymmetric Wavelet-Based Dissimilarity Measure 368
Ibrahim El Rube', Mohamed Kamel, Maher Ahmed

A Real-Time Image Stabilization System Based on Fourier-Mellin Transform 376
J.R. Martinez-de Dios, A. Ollero

A Novel Shape Descriptor Based on Interrelation Quadruplet 384
Dongil Han, Bum-Jae You, Sang-Rok Oh

An Efficient Representation of Hand Sketch Graphic Messages Using Recursive Bezier Curve Approximation 392
Jaehwa Park, Young-Bin Kwon

Contour Description Through Set Operations on Dynamic Reference Shapes 400
Miroslav Koprnicky, Maher Ahmed, Mohamed Kamel

An Algorithm for Efficient and Exhaustive Template Matching 408
Luigi Di Stefano, Stefano Mattoccia, Federico Tombari

Modelling of Overlapping Circular Objects Based on Level Set Approach 416
Eva Dejnozkova, Petr Dokladal

A Method for Dominant Points Detection and Matching 2D Object Identification 424
A. Carmona-Poyato, N.L. Fernández-García, R. Medina-Carnicer, F.J. Madrid-Cuevas

Image Description and Recognition

Character Recognition Using Canonical Invariants 432
Sema Doguscu, Mustafa Unel

Finding Significant Points for a Handwritten Classification Task 440
Juan Ramón Rico-Juan, Luisa Micó

The System for Handwritten Symbol and Signature Recognition Using FPGA Computing 447
Rauf K. Sadykhov, Leonid P. Podenok, Vladimir A. Samokhval, Andrey A. Uvarov

Reconstruction of Order Parameters Based on Immunity Clonal Strategy for Image Classification 455
Xiuli Ma, Licheng Jiao

Visual Object Recognition Through One-Class Learning 463
QingHua Wang, Luís Seabra Lopes, David M.J. Tax

Semantic Image Analysis Based on the Representation of the Spatial Relations Between Objects in Images 471
Hyunjang Kong, Miyoung Cho, Kwanho Jung, Sunkyoung Baek, Pankoo Kim

Ridgelets Frame 479
Tan Shan, Licheng Jiao, Xiangchu Feng

Adaptive Curved Feature Detection Based on Ridgelet 487
Kang Liu, Licheng Jiao

Globally Stabilized 3L Curve Fitting 495
Turker Sahin, Mustafa Unel

Learning an Information Theoretic Transform for Object Detection 503
Jianzhong Fang, Guoping Qiu

Image Object Localization by AdaBoost Classifier 511
Władysław Skarbek, Krzysztof Kucharski

Cost and Information-Driven Algorithm Selection for Vision Systems ... 519
Mauricio Marengoni, Allen Hanson, Shlomo Zilberstein, Edward Riseman

Gesture Recognition for Human-Robot Interaction
Through a Knowledge Based Software Platform 530
M. Hasanuzzaman, Tao Zhang, V. Ampornaramveth, M.A. Bhuiyan, Yoshiaki Shirai, H. Ueno

Appearance-Based Object Detection in Space-Variant Images:
A Multi-model Approach 538
V. Javier Traver, Alexandre Bernardino, Plinio Moreno, José Santos-Victor

3D Object Recognition from Appearance:
PCA Versus ICA Approaches 547
M. Asunción Vicente, Cesar Fernández, Oscar Reinoso, Luis Payá

A Stochastic Search Algorithm to Optimize an N-tuple Classifier
by Selecting Its Inputs 556
Hannan Bin Azhar, Keith Dimond

Video Processing and Analysis

A Multi-expert Approach for Shot Classification in News Videos 564
M. De Santo, G. Percannella, C. Sansone, M. Vento

Motion-Compensated Wavelet Video Denoising 572
Fu Jin, Paul Fieguth, Lowell Winger

Alpha-Stable Noise Reduction in Video Sequences 580
Mohammed El Hassouni, Hocine Cherifi

Automatic Text Extraction in Digital Video
Based on Motion Analysis 588
Duarte Palma, João Ascenso, Fernando Pereira

Fast Video Registration Method for Video Quality Assessment 597
Jihwan Choe, Chulhee Lee

Hidden Markov Model Based Events Detection in Soccer Video 605
Guoying Jin, Linmi Tao, Guangyou Xu

3D Imaging

Improving Height Recovery from a Single Image of a Face
Using Local Shape Indicators 613
Mario Castelán, Edwin R. Hancock

Recovery of Surface Height from Diffuse Polarisation 621
Gary Atkinson, Edwin Hancock

Vectorization-Free Reconstruction of 3D CAD Models from Paper Drawings 629
Frank Ditrich, Herbert Suesse, Klaus Voss

Plane Segmentation from Two Views in Reciprocal-Polar Image Space 638
Zezhi Chen, Nick E. Pears, Bojian Liang, John McDermid

Tracking of Points in a Calibrated and Noisy Image Sequence 647
Domingo Mery, Felipe Ochoa, René Vidal

Multiresolution Approach to "Visual Pattern" Partitioning of 3D Images 655
Raquel Dosil, Xosé R. Fdez-Vidal, Xosé M. Pardo

Visual Cortex Frontend: Integrating Lines, Edges, Keypoints, and Disparity 664
João Rodrigues, J.M. Hans du Buf

Estimation of Directional and Ambient Illumination Parameters by Means of a Calibration Object 672
Alberto Ortiz, Gabriel Oliver

Environment Authentication Through 3D Structural Analysis 680
Toby P. Breckon, Robert B. Fisher

Camera Calibration Using Two Concentric Circles 688
Francisco Abad, Emilio Camahort, Roberto Vivó

Three-Dimensional Object Recognition Using a Modified Exoskeleton and Extended Hausdorff Distance Matching Algorithm 697
Rajalida Lipikorn, Akinobu Shimizu, Hidefumi Kobatake

Recognition of 3D Object from One Image Based on Projective and Permutative Invariants 705
J.M. González, J.M. Sebastián, D. García, F. Sánchez, L. Angel

Wide Baseline Stereo Matching by Corner-Edge-Regions 713
Jun Xie, Hung Tat Tsui

Gradient Based Dense Stereo Matching 721
Tomasz Twardowski, Boguslaw Cyganek, Jan Borgosz

Image Retrieval and Indexing

Accelerating Multimedia Search by Visual Features 729
Grzegorz Galinski, Karol Wnukowicz, Wladyslaw Skarbek

Semantic Browsing and Retrieval in Image Libraries 737
Andrea Kutics, Akihiko Nakagawa

Robust Shape Retrieval Using Maximum Likelihood Theory 745
Naif Alajlan, Paul Fieguth, Mohamed Kamel

A Novel Shape Feature for Image Classification and Retrieval 753
Rami Rautkorpi, Jukka Iivarinen

A Local Structure Matching Approach
for Large Image Database Retrieval 761
Yanling Chi, Maylor K.H. Leung

People Action Recognition in Image Sequences
Using a 3D Articulated Object. 769
Jean-Charles Atine

CVPIC Compressed Domain Image Retrieval by Colour and Shape 778
Gerald Schaefer, Simon Lieutaud

Automating GIS Image Retrieval Based on MCM 787
Adel Hafiane, Bertrand Zavidovique

Significant Perceptual Regions by Active-Nets 795
David García-Pérez, Antonio Mosquera, Marcos Ortega,
Manuel G. Penedo

Improving the Boosted Correlogram 803
Nicholas R. Howe, Amanda Ricketson

Distance Map Retrieval 811
László Czúni, Dezső Csordás, Gergely Császár

Grass Field Segmentation, the First Step Toward Player Tracking,
Deep Compression, and Content Based Football Image Retrieval 818
Kaveh Kangarloo, Ehsanollah Kabir

Spatio-temporal Primitive Extraction Using Hermite
and Laguerre Filters for Early Vision Video Indexing 825
Carlos Joel Rivero-Moreno, Stéphane Bres

Non-parametric Performance Comparison in Pictorial Query
by Content Systems 833
Sergio Domínguez

Morphology

Hierarchical Watersheds with Inter-pixel Boundaries 840
Luc Brun, Philippe Vautrot, Fernand Meyer

From Min Tree to Watershed Lake Tree: Theory and Implementation 848
Xiaoqiang Huang, Mark Fisher, Yanong Zhu

From Min Tree to Watershed Lake Tree: Evaluation 858
Xiaoqiang Huang, Mark Fisher

Optimizing Texture Primitives Description
Based on Variography and Mathematical Morphology 866
Assia Kourgli, Aichouche Belhadj-aissa, Lynda Bouchemakh

Author Index 875

Table of Contents – Part II

Biomedical Applications

An Automated Multichannel Procedure
for cDNA Microarray Image Processing 1
Rastislav Lukac, Konstantinos N. Plataniotis, Bogdan Smolka, Anastasios N. Venetsanopoulos

A Modified Nearest Neighbor Method for Image Reconstruction
in Fluorescence Microscopy 9
Koji Yano, Itsuo Kumazawa

An Improved Clustering-Based Approach
for DNA Microarray Image Segmentation 17
Luis Rueda, Li Qin

A Spatially Adaptive Filter Reducing Arc Stripe Noise
for Sector Scan Medical Ultrasound Imaging 25
Qianren Xu, M. Kamel, M.M.A. Salama

Fuzzy-Snake Segmentation of Anatomical Structures Applied
to CT Images 33
Gloria Bueno, Antonio Martínez-Albalá, Antonio Adán

Topological Active Volumes for Segmentation
and Shape Reconstruction of Medical Images 43
N. Barreira, M.G. Penedo

Region of Interest Based Prostate Tissue Characterization
Using Least Square Support Vector Machine LS-SVM 51
S.S. Mohamed, M.M.A. Salama, M. Kamel, K. Rizkalla

Ribcage Boundary Delineation in Chest X-ray Images 59
Carlos Vinhais, Aurélio Campilho

A Level-Set Based Volumetric CT Segmentation Technique:
A Case Study with Pulmonary Air Bubbles 68
José Silvestre Silva, Beatriz Sousa Santos, Augusto Silva, Joaquim Madeira

Robust Fitting of a Point Distribution Model of the Prostate
Using Genetic Algorithms 76
Fernando Arámbula Cosío

A Quantification Tool to Analyse Stained Cell Cultures 84
E. Glory, A. Faure, V. Meas-Yedid, F. Cloppet, Ch. Pinset, G. Stamon, J-Ch. Olivo-Marin

Dynamic Pedobarography Transitional Objects by Lagrange's Equation with FEM, Modal Matching, and Optimization Techniques 92
Raquel Ramos Pinho, João Manuel, R.S. Tavares

3D Meshes Registration: Application to Statistical Skull Model 100
M. Berar, M. Desvignes, G. Bailly, Y. Payan

Detection of Rib Borders on X-ray Chest Radiographs 108
Rui Moreira, Ana Maria Mendonça, Aurélio Campilho

Isosurface-Based Level Set Framework for MRA Segmentation 116
Yongqiang Zhao, Minglu Li

Segmentation of the Comet Assay Images 124
Bogdan Smolka, Rastislav Lukac

Automatic Extraction of the Retina AV Index 132
I.G. Caderno, M.G. Penedo, C. Mariño, M.J. Carreira, F. Gomez-Ulla, F. González

Image Registration in Electron Microscopy. A Stochastic Optimization Approach 141
J.L. Redondo, P.M. Ortigosa, I. García, J.J. Fernández

Evolutionary Active Contours for Muscle Recognition 150
A. Caro, P.G. Rodríguez, M.L. Durán, J.A. Ávila, T. Antequera, R. Palacios

Automatic Lane and Band Detection in Images of Thin Layer Chromatography 158
António V. Sousa, Rui Aguiar, Ana Maria Mendonça, Aurélio Campilho

Automatic Tracking of *Arabidopsis thaliana* Root Meristem in Confocal Microscopy 166
Bernardo Garcia, Ana Campilho, Ben Scheres, Aurélio Campilho

Document Processing

A New File Format for Decorative Tiles 175
Rafael Dueire Lins

Projection Profile Based Algorithm for Slant Removal 183
Moisés Pastor, Alejandro Toselli, Enrique Vidal

Novel Adaptive Filtering for Salt-and-Pepper Noise Removal from Binary Document Images 191
Amr R. Abdel-Dayem, Ali K. Hamou, Mahmoud R. El-Sakka

Automated Seeded Region Growing Method for Document Image Binarization Based on Topographic Features 200
Yufei Sun, Yan Chen, Yuzhi Zhang, Yanxia Li

Image Segmentation of Historical Documents: Using a Quality Index 209
Carlos A.B. de Mello

A Complete System for Detection and Identification of Tabular Structures from Document Images 217
S. Mandal, S.P. Chowdhury, A.K. Das, Bhabatosh Chanda

Underline Removal on Old Documents 226
João R. Caldas Pinto, Pedro Pina, Lourenço Bandeira, Luís Pimentel, Mário Ramalho

A New Algorithm for Skew Detection in Images of Documents 234
Rafael Dueire Lins, Bruno Tenório Ávila

Blind Source Separation Techniques for Detecting Hidden Texts and Textures in Document Images 241
Anna Tonazzini, Emanuele Salerno, Matteo Mochi, Luigi Bedini

Efficient Removal of Noisy Borders from Monochromatic Documents 249
Bruno Tenório Ávila, Rafael Dueire Lins

Colour Analysis

Robust Dichromatic Colour Constancy 257
Gerald Schaefer

Soccer Field Detection in Video Images Using Color and Spatial Coherence 265
Arnaud Le Troter, Sebastien Mavromatis, Jean Sequeira

New Methods to Produce High Quality Color Anaglyphs for 3-D Visualization 273
Ianir Ideses, Leonid Yaroslavsky

A New Color Filter Array Interpolation Approach for Single-Sensor Imaging 281
Rastislav Lukac, Konstantinos N. Plataniotis, Bogdan Smolka

A Combinatorial Color Edge Detector 289
Soufiane Rital, Hocine Cherifi

Texture Analysis

A Fast Probabilistic Bidirectional Texture Function Model 298
Michal Haindl, Jiří Filip

Model-Based Texture Segmentation 306
Michal Haindl, Stanislav Mikeš

A New Gabor Filter Based Kernel
for Texture Classification with SVM 314
Mahdi Sabri, Paul Fieguth

Grading Textured Surfaces with Automated Soft Clustering
in a Supervised SOM 323
J. Martín-Herrero, M. Ferreiro-Armán, J.L. Alba-Castro

Textures and Wavelet-Domain Joint Statistics 331
Zohreh Azimifar, Paul Fieguth, Ed Jernigan

Video Segmentation Through Multiscale Texture Analysis 339
Miguel Alemán-Flores, Luis Álvarez-León

Motion Analysis

Estimation of Common Groundplane Based on Co-motion Statistics 347
Zoltan Szlavik, Laszlo Havasi, Tamas Sziranyi

An Adaptive Estimation Method for Rigid Motion Parameters
of 2D Curves 355
Turker Sahin, Mustafa Unel

Classifiers Combination for Improved Motion Segmentation 363
Ahmad Al-Mazeed, Mark Nixon, Steve Gunn

A Pipelined Real-Time Optical Flow Algorithm 372
Miguel V. Correia, Aurélio Campilho

De-interlacing Algorithm Based on Motion Objects 381
Junxia Gu, Xinbo Gao, Jie Li

Automatic Selection of Training Samples
for Multitemporal Image Classification 389
T.B. Cazes, R.Q. Feitosa, G.L.A. Mota

Parallel Computation of Optical Flow 397
Antonio G. Dopico, Miguel V. Correia, Jorge A. Santos,
Luis M. Nunes

Lipreading Using Recurrent Neural Prediction Model 405
Takuya Tsunekawa, Kazuhiro Hotta, Haruhisa Takahashi

Multi-model Adaptive Estimation for Nonuniformity Correction of Infrared Image Sequences 413
Jorge E. Pezoa, Sergio N. Torres

Surveillance and Remote Sensing

A MRF Based Segmentatiom Approach to Classification Using Dempster Shafer Fusion for Multisensor Imagery 421
A. Sarkar, N. Banerjee, P. Nair, A. Banerjee, S. Brahma, B. Kartikeyan, K.L. Majumder

Regularized RBF Networks for Hyperspectral Data Classification 429
G. Camps-Valls, A.J. Serrano-López, L. Gómez-Chova, J.D. Martín-Guerrero, J. Calpe-Maravilla, J. Moreno

A Change-Detection Algorithm Enabling Intelligent Background Maintenance 437
Luigi Di Stefano, Stefano Mattoccia, Martino Mola

Dimension Reduction and Pre-emphasis for Compression of Hyperspectral Images 446
C. Lee, E. Choi, J. Choe, T. Jeong

Viewpoint Independent Detection of Vehicle Trajectories and Lane Geometry from Uncalibrated Traffic Surveillance Cameras 454
José Melo, Andrew Naftel, Alexandre Bernardino, José Santos-Victor

Robust Tracking and Object Classification Towards Automated Video Surveillance 463
Jose-Luis Landabaso, Li-Qun Xu, Montse Pardas

Detection of Vehicles in a Motorway Environment by Means of Telemetric and Visual Data 471
Sonia Izri, Eric Brassart, Laurent Delahoche, Bruno Marhic, Arnaud Clérentin

High Quality-Speed Dilemma: A Comparison Between Segmentation Methods for Traffic Monitoring Applications 481
Alessandro Bevilacqua, Luigi Di Stefano, Alessandro Lanza

Automatic Recognition of Impact Craters on the Surface of Mars 489
Teresa Barata, E. Ivo Alves, José Saraiva, Pedro Pina

Classification of Dune Vegetation from Remotely Sensed Hyperspectral Images 497
Steve De Backer, Pieter Kempeneers, Walter Debruyn, Paul Scheunders

SAR Image Classification Based on Immune Clonal Feature Selection 504
Xiangrong Zhang, Tan Shan, Licheng Jiao

Depth Extraction System Using Stereo Pairs 512
Rizwan Ghaffar, Noman Jafri, Shoab Ahmed Khan

Fast Moving Region Detection Scheme in Ad Hoc Sensor Network..... 520
Yazhou Liu, Wen Gao, Hongxun Yao, Shaohui Liu, Lijun Wang

Tracking

LOD Canny Edge Based Boundary Edge Selection
for Human Body Tracking 528
Jihun Park, Tae-Yong Kim, Sunghun Park

Object Boundary Edge Selection for Accurate Contour Tracking
Using Multi-level Canny Edges 536
Tae-Yong Kim, Jihun Park, Seong-Whan Lee

Reliable Dual-Band Based Contour Detection:
A Double Dynamic Programming Approach 544
Mohammad Dawood, Xiaoyi Jiang, Klaus P. Schäfers

Tracking Pedestrians Under Occlusion
Using Multiple Cameras 552
Jorge P. Batista

Application of Radon Transform to Lane Boundaries Tracking 563
R. Nourine, M. Elarbi Boudihir, S.F. Khelifi

A Speaker Tracking Algorithm Based on Audio
and Visual Information Fusion Using Particle Filter 572
Xin Li, Luo Sun, Linmi Tao, Guangyou Xu, Ying Jia

Kernel-Bandwidth Adaptation for Tracking Object Changing in Size 581
Ning-Song Peng, Jie Yang, Jia-Xin Chen

Tracking Algorithms Evaluation in Feature Points Image Sequences 589
Vanessa Robles, Enrique Alegre, Jose M. Sebastian

Short-Term Memory-Based Object Tracking 597
Hang-Bong Kang, Sang-Hyun Cho

Real Time Multiple Object Tracking Based on Active Contours 606
Sébastien Lefèvre, Nicole Vincent

An Object Tracking Algorithm Combining Different Cost Functions 614
D. Conte, P. Foggia, C. Guidobaldi, A. Limongiello, M. Vento

Vehicle Tracking at Traffic Scene with Modified RLS 623
Hadi Sadoghi Yazdi, Mahmood Fathy, A. Mojtaba Lotfizad

Face Detection and Recognition

Understanding In-Plane Face Rotations Using Integral Projections 633
Henry Nicponski

Feature Fusion Based Face Recognition Using EFM 643
Dake Zhou, Xin Yang

Real-Time Facial Feature Extraction
by Cascaded Parameter Prediction and Image Optimization 651
Fei Zuo, Peter H.N. de With

Frontal Face Authentication Through Creaseness-Driven Gabor Jets 660
Daniel González-Jiménez, José Luis Alba-Castro

A Coarse-to-Fine Classification Scheme
for Facial Expression Recognition 668
Xiaoyi Feng, Abdenour Hadid, Matti Pietikäinen

Fast Face Detection Using QuadTree Based Color Analysis
and Support Vector Verification 676
Shu-Fai Wong, Kwan-Yee Kenneth Wong

Three-Dimensional Face Recognition: A Fishersurface Approach 684
Thomas Heseltine, Nick Pears, Jim Austin

Face Recognition Using Improved-LDA 692
Dake Zhou, Xin Yang

Analysis and Recognition of Facial Expression Based
on Point-Wise Motion Energy 700
Hanhoon Park, Jong-Il Park

Face Class Modeling Using Mixture of SVMs 709
Julien Meynet, Vlad Popovici, Jean-Philippe Thiran

Comparing Robustness of Two-Dimensional PCA and Eigenfaces
for Face Recognition 717
Muriel Visani, Christophe Garcia, Christophe Laurent

Useful Computer Vision Techniques for Human-Robot Interaction 725
O. Deniz, A. Falcon, J. Mendez, M. Castrillon

Face Recognition with Generalized Entropy Measurements 733
Yang Li, Edwin R. Hancock

Facial Feature Extraction and Principal Component Analysis for Face Detection in Color Images 741
Saman Cooray, Noel O'Connor

Security Systems

Fingerprint Enhancement Using Circular Gabor Filter 750
En Zhu, Jianping Yin, Guomin Zhang

A Secure and Localizing Watermarking Technique for Image Authentication 759
Abdelkader H. Ouda, Mahmoud R. El-Sakka

A Hardware Implementation of Fingerprint Verification for Secure Biometric Authentication Systems 770
Yongwha Chung, Daesung Moon, Sung Bum Pan, Min Kim, Kichul Kim

Inter-frame Differential Energy Video Watermarking Algorithm Based on Compressed Domain 778
Lijun Wang, Hongxun Yao, Shaohui Liu, Wen Gao, Yazhou Liu

Improving DTW for Online Handwritten Signature Verification 786
M. Wirotius, J.Y. Ramel, N. Vincent

Distribution of Watermark According to Image Complexity for Higher Stability 794
Mansour Jamzad, Farzin Yaghmaee

Visual Inspection

Comparison of Intelligent Classification Techniques Applied to Marble Classification 802
João M.C. Sousa, João R. Caldas Pinto

Inspecting Colour Tonality on Textured Surfaces 810
Xianghua Xie, Majid Mirmehdi, Barry Thomas

Automated Visual Inspection of Glass Bottles Using Adapted Median Filtering 818
Domingo Mery, Olaya Medina

Neuro-Fuzzy Method for Automated Defect Detection in Aluminium Castings 826
Sergio Hernández, Doris Sáez, Domingo Mery

Online Sauter Diameter Measurement of Air Bubbles and Oil Drops in Stirred Bioreactors by Using Hough Transform 834
L. Vega-Alvarado, M.S. Cordova, B. Taboada, E. Galindo, G. Corkidi

Defect Detection in Textile Images Using Gabor Filters 841
Céu L. Beirão, Mário A.T. Figueiredo

Geometric Surface Inspection of Raw Milled Steel Blocks 849
Ingo Reindl, Paul O'Leary

Author Index 857

Automatic Image Segmentation Using a Deformable Model Based on Charged Particles

Andrei C. Jalba, Michael H.F. Wilkinson, and Jos B.T.M. Roerdink

Institute of Mathematics and Computing Science
University of Groningen, P.O. Box 800
9700 AV, Groningen, The Netherlands
{andrei,michael,roe}@cs.rug.nl
http://www.cs.rug.nl

Abstract. We present a method for automatic segmentation of grey-scale images, based on a recently introduced deformable model, the charged-particle model (CPM). The model is inspired by classical electrodynamics and is based on a simulation of charged particles moving in an electrostatic field. The charges are attracted towards the contours of the objects of interest by an electrostatic field, whose sources are computed based on the gradient-magnitude image. Unlike the case of active contours, extensive user interaction in the initialization phase is not mandatory, and segmentation can be performed automatically. To demonstrate the reliability of the model, we conducted experiments on a large database of microscopic images of diatom shells. Since the shells are highly textured, a post-processing step is necessary in order to extract only their outlines.

1 Introduction

An important aspect in many image analysis and computer vision tasks is image segmentation, the process in which an image is divided in its constituent parts. Here, we shall focus on boundary-based segmentation using the recently introduced charged-particle model (CPM) [1].

The CPM is inspired by classical electrodynamics and consists of a system of charged particles moving in an electrostatic field. The charges are attracted towards the contours of the objects of interest by an electric field, whose sources are computed based on the gradient-magnitude image. The electric field plays the same role as the potential force (defined to be the negative gradient of some potential function) in the snake model, while internal interactions are modeled by repulsive electrostatic forces (referred to as Coulomb forces). The method needs an initialization step, which is much less critical than in the snake model. Unlike the active contour model, in our model charges can be placed entirely inside an object, outside on one side of the object, or they can cross over parts of boundaries. In contrast to attractive forces based on the squared gradient-magnitude image [2], which act only in small vicinities along boundaries of objects, the electric field exhibits increased capture range because of its long range attraction, and enhanced robustness of the model against boundary leakage. Due to the combined effect of external interactions of particles with the electrostatic field, and internal repelling forces between them, particles follow paths along object boundaries, and hence

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 1–8, 2004.
© Springer-Verlag Berlin Heidelberg 2004

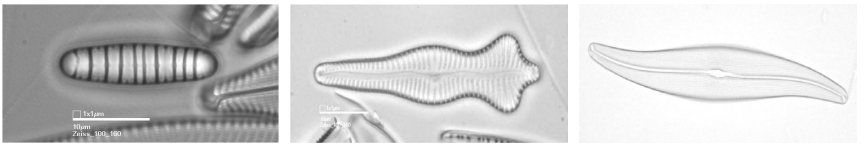


Fig. 1. Some examples of diatom shells.

converge without difficulty into deep boundary concavities or internal boundaries separating embedded objects. Moreover, the method is insensitive to initialization, and can adapt to topological changes of the underlying shape, see [1].

In this paper we present methods for automatic segmentation based on the CPM, using different strategies for automatic initialization of particles: (i) particles are spread uniformly over the image plane, (ii) particles are placed at locations of high gradient-magnitude, and (iii) particles are initialized on boundaries of the regions found by a marker-selection procedure [3]. To demonstrate the reliability of the model, we conducted experiments on a large database of microscopic images of diatom shells (see Fig. 1 for some examples).

2 The Charged-Particle Model (CPM)

The CPM consists of a system of N positively charged particles p_i with electric charges q_i, $i = 1 \dots N$, which freely move in an electrostatic field $\boldsymbol{E}$, generated by fixed, negative charges, placed at each pixel position of the input image, with charge magnitude proportional to the edge-map of the input image. Therefore, each free particle q_i moves under the influence of two forces: (i) internal Coulomb force, $\boldsymbol{F}_c$, due to the interaction of the particle with other free particles, and (ii) external Lorentz force, $\boldsymbol{F}_l$, due to the electric field generated by the fixed negative charges e_i, see Fig. 2. The resulting force $\boldsymbol{F}$ acting on a particle p_i located at position vector $\boldsymbol{r}_i = [x_i, y_i]$ is

$$\boldsymbol{F}(\boldsymbol{r}_i) = \boldsymbol{F}_c(\boldsymbol{r}_i) + \boldsymbol{F}_l(\boldsymbol{r}_i), \tag{1}$$

where $\boldsymbol{F}_c$ is the Coulomb force and $\boldsymbol{F}_l$ is the Lorentz force. Assuming that all free particles have the same positive charge $q_i = q$, it can be shown that the equilibrium equation (Eq. (1)) can be rewritten as

$$\boldsymbol{F}(\boldsymbol{r}_i) = w_1 \sum_{j \neq i}^{N} \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j|^3} - w_2 \sum_{k:\boldsymbol{R}_k \neq \boldsymbol{r}_i}^{M} e_k \frac{\boldsymbol{r}_i - \boldsymbol{R}_k}{|\boldsymbol{r}_i - \boldsymbol{R}_k|^3}, \tag{2}$$

where $w_1 = kq^2$ and $w_2 = kq$ are weights, k is a constant, and $\boldsymbol{R}_k$ is a grid-position vector. The major difference between the two terms in Eq. (2) is that the Lorentz force reflects particle-mesh or external interactions and is computed in the image domain, while the Coulomb force represents particle-particle or internal interactions. Therefore, each particle is the subject of two antagonistic forces: (i) the Coulomb force, which makes the particles to repel each other, and (ii) the external Lorentz force which attracts

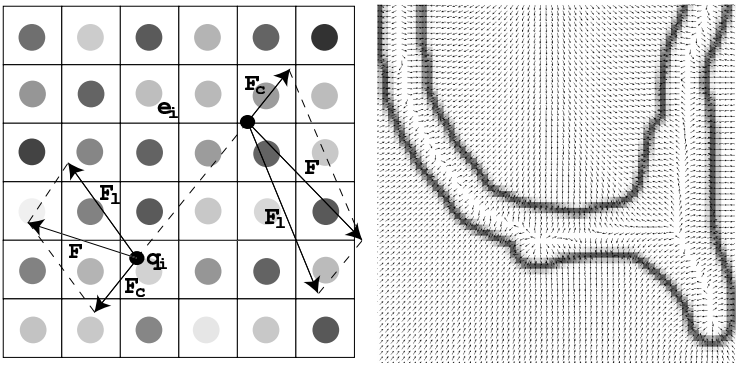


Fig. 2. The charged-particle model. *Left*: Forces acting on free particles q_i (indicated by small black dots) which move in the electric field generated by fixed charges e_i (indicated by grey dots); different grey values represent different charge magnitudes. *Right*: Example of electrostatic field $\boldsymbol{E}$ generated by fixed charges.

the particles. Since the distribution of fixed charges e_i reflects the strength of the edge map, and the electric force is "inverse-square", i.e., it decays with the squared distance, the electrostatic field has large values near edges and small values in homogeneous regions of the objects present in the input image.

2.1 Particle Dynamics

The total energy of the system is the summation of all particle energies, i.e.,

$$E_p(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \frac{1}{2} \sum_{i=1}^{N} \left(w_1 \sum_{j \neq i}^{N} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} - w_2 \sum_{k:\boldsymbol{R}_k \neq \boldsymbol{r}_i}^{M} \frac{e_k}{|\boldsymbol{r}_i - \boldsymbol{R}_k|} \right). \quad (3)$$

Having defined the energy associated with our system, we can derive its equations of motion. The standard approach is to consider the Newtonian equations of motion, and to integrate the corresponding system of differential equations in time, i.e.,

$$\boldsymbol{F}(\boldsymbol{r}_i) = w_1 \boldsymbol{F}_c(\boldsymbol{r}_i) + w_2 \boldsymbol{F}_l(\boldsymbol{r}_i) - \beta \boldsymbol{v}_i \quad (4)$$

$$\boldsymbol{a}_i = \frac{\boldsymbol{F}(\boldsymbol{r}_i)}{m_i} = \frac{d^2 \boldsymbol{r}_i(t)}{dt^2}, \quad (5)$$

where m_i is the mass of the particle p_i (we set $m_i = 1$), and $\boldsymbol{r}_i$, $\boldsymbol{v}_i$ and $\boldsymbol{a}_i$ are its position, speed and acceleration, respectively. Notice that compared to Eq. (1), Eq. (4) has an additional term, $\boldsymbol{F}_{damp}(\boldsymbol{r}_i) = -\beta \boldsymbol{v}_i$, the damping (or viscous) force which is required by the particles to attain a stable equilibrium state, which minimizes their potential energies, see Eq. (3). Eq. (5) is written as a system of coupled, first order differential equations, and solved using some method for numerical integration [4, 1]. For detailed information on the CPM, efficient methods for implement it and pseudo-code, we refer to [1].

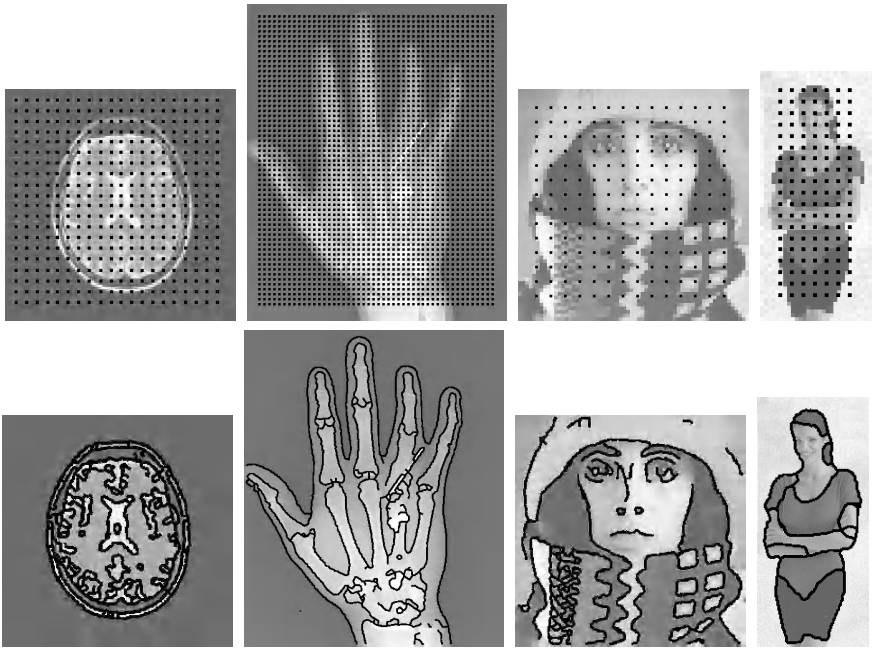


Fig. 3. Automatic segmentation. *First row*: initializations; *second row*: segmentation results.

2.2 Curve Reconstruction

So far, our particle system does not provide us with explicit representations of object boundaries. This problem can be thought of as that of curve reconstruction from unorganized points: we are given a set of points and asked to connect them into the most likely polygonal curve. If the aim is to recover only one, closed contour, the reconstruction problem can be formulated as enumerating the particles and then ordering them into a sequence which describes a closed contour along the boundary of the object. The problem is now isomorphic to the classical symmetric traveling salesman problem (STSP), and established techniques for approximating TSP can be used. However, under the more general assumption that no a priori knowledge about the underlying topology is available, curve reconstruction algorithms must be involved. Therefore, in all experiments reported below, we use the algorithms by Amenta et al. [5] to reconstruct the recovered curves.

3 Segmentation Results

3.1 Natural Images

Our first experiment is automatic segmentation of natural images using (trivial) automatic strategies for initialization. In all experiments which we report in this paper, we used the same values for the two weights w_1 and w_2 (see Eq. (2)), i.e. $w_1 = 0.6$ and $w_2 = 0.7$,

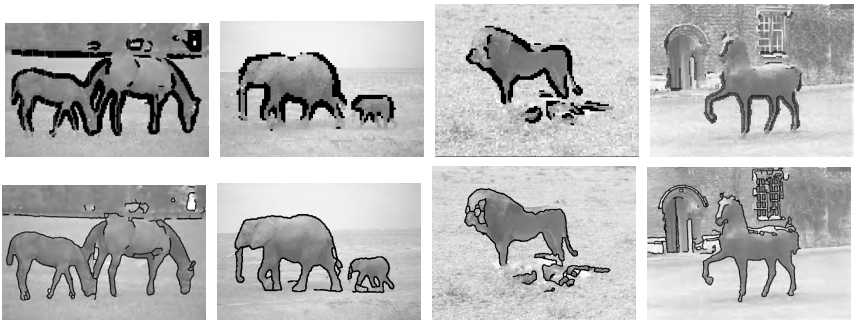


Fig. 4. Segmentation of natural images. *First row*: initializations; *second row*: results.

and all other parameters of the model were set as in [1]. The pre-processing step consists in image filtering by means of a Gaussian pyramid with three levels.

With this experimental setup, the first set of segmentation results is shown in Fig. 3. The initializations shown in this figure were performed by uniformly spreading particles over the image plane. As it can be seen, the most important structures present in these images were correctly recovered.

The second set of results is shown in Fig. 4. In this case, free particles were placed at those locations of the gradient-magnitude image with values above 10% of the maximum magnitude. Natural images are known to be particularly difficult to segment, mostly because of the background texture surrounding the main objects. Without being perfect, the segmentation results shown in both figures are quite good, even though a very simple initialization method was used.

3.2 Results for a Large Database of Diatom-Shell Images

The second experiment we conduct is automatic segmentation, on a large database consisting of 808 diatom images (see Fig. 1 for some examples).

The goal is to extract the outline of each diatom shell present in the input image. The extracted outlines, encoded as chain-codes, provide the input for identification methods such as those in [6].

The input consists of grey-scale, high-magnification images of diatom shells obtained by automatic slide scanning [7]. Ideally, each image contains a single diatom shell, but as it can be seen in the figure, diatoms may lay on top of each other, may not be in proper focus, or they can be very close to each other. Moreover, dust specks and background texture may be visible in some images. Most diatoms in images such as those in Fig. 1 present prominent outlines which can be detected either by thresholding or by edge detectors. Unfortunately, if the illumination around the diatom is not uniform, most global thresholding methods fail to find a proper threshold value. In addition, in microscopic images, diatoms exhibit the same grey levels as the background, and the histogram is unimodal [8]. This fact upsets most threshold selection methods which make the assumption that the histogram of the image is multimodal. Moreover, if the diatom is not in proper focus, the edges are blurred, and can only be partly detected by

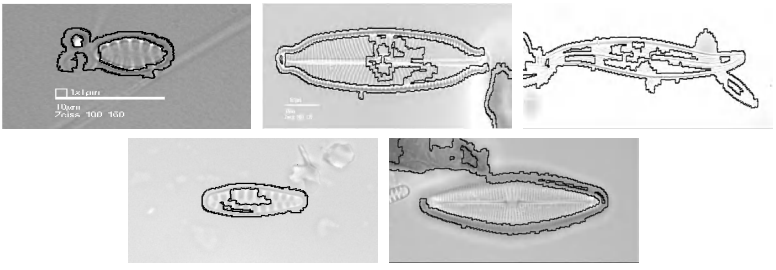


Fig. 5. Problematic diatom images for the CPM, with superimposed initializations.

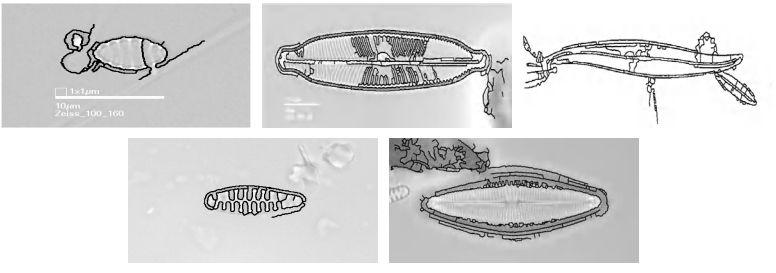


Fig. 6. Problematic diatom images for the CPM; final (erroneous) results.

most edge detection techniques. Therefore, we use a method based on morphological filtering [3] to provide marker-regions (the same method was used in [3] in the context of watershed-based segmentation), and we initialize the particles on the boundaries of these regions. To guarantee that only one closed contour per diatom is extracted, each contour obtained using a standard contour-following algorithm is flood-filled, and then, traced once again.

With this experimental setup, the method succeeded in extracting 99.4% of visually-estimated correct contours. Visual estimation was guided by the following criteria: (i) the contours should be smooth, (ii) they should correspond well with the perceived diatom outlines, and (iii) they should not enclose debris or diatom fragments. All contours that did not fulfill the above requirements were considered errors.

The initializations and final results (without the contour-tracing step) for the five cases in which the method failed are shown in Figs. 5 and 6, respectively. Four of the images shown in Fig. 5 have debris or fragments of other diatoms very close to the central diatom. The fourth image shows a very low contrast of the diatom outline, which is reflected in the weak gradient-magnitude response that is used by the CPM to compute the electric field. Nevertheless, in our opinion this is a very good result, considering that the CPM is a boundary-based method.

Fig. 7 shows some example results obtained using the CPM on difficult images on which a hybrid technique based on the morphological watershed from markers failed. This method obtained 98% (i.e. 16 errors), of correctly extracted contours, see [3].

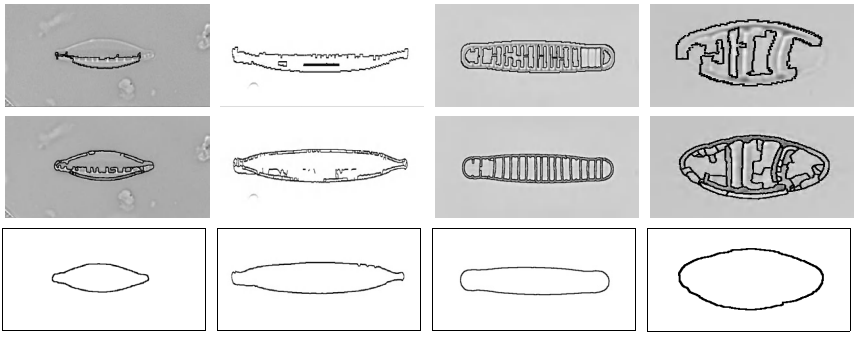


Fig. 7. Difficult diatom images (and initializations), correctly segmented by the CPM. *First row*: initializations; *second row*: reconstructed curve(s); *third row*: extracted diatom contours.

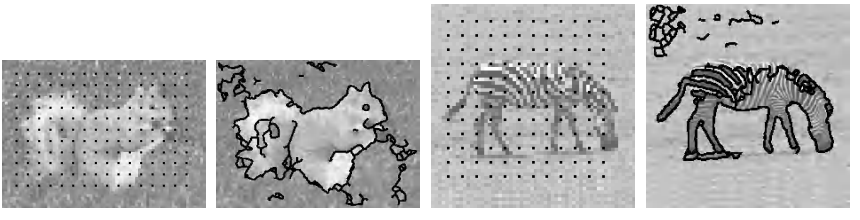


Fig. 8. The CPM may fail if highly textured regions surround the main object or belong to the main object; results obtained with the first initialization method.

3.3 Discussion

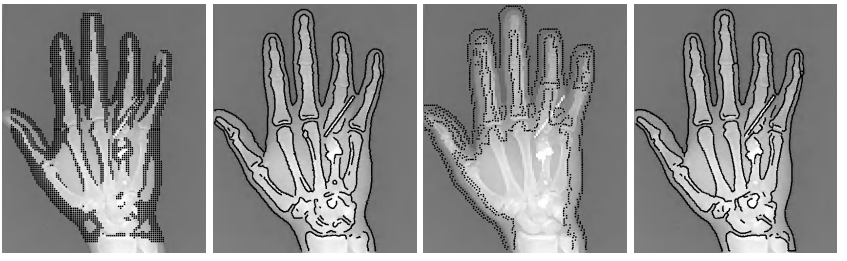
The advantages of using the second and third initialization strategies over the first one are twofold. First, the particles are already close to the final equilibrium positions, and therefore the total convergence time is smaller. Second, using the first initialization method, it may happen that some particles will be attracted towards highly textured regions, which are also regions with high response of the gradient magnitude, and therefore they will be trapped at these regions, see Fig. 8.

Fig. 9 shows segmentation results obtained using the second and third initialization strategies; see also the result in Fig. 3 obtained with the first method. The CPU timings (on a Pentium III machine at 670 MHz) for segmenting this x-ray image of 417×510 pixels were 45, 20, 25 seconds, using the first, second and third initialization methods, respectively.

4 Conclusions

The experimental results presented in this paper showed that the CPM can be used successfully to perform automatic segmentation, provided that a suitable setup has been identified.

Further investigations of the CPM are the subject of ongoing research. We shall focus on supplementing the energy formulation of the model with some information useful in



(a) (b) (c) (d)

Fig. 9. Comparative segmentation results; (a) initialization by the second method, (b) result, (c) initialization using the third method, (d) result.

the reconstruction phase. A shortcoming of the current method is that it cannot guarantee that the recovered contours (surfaces) are without gaps. Finally, many improvements of the CPM are possible. For example, instead of using Gaussian pyramids, one can use wavelet or other pyramids based on non-linear diffusion operators.

References

1. Jalba, A.C., Wilkinson, M.H.F., Roerdink, J.B.T.M.: CPM: A deformable model for shape recovery and segmentation based on charged particles. IEEE Trans. Pattern Anal. Machine Intell. (2004) in press.
2. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. Int. J. Comput. Vis. **1** (1987) 321–331
3. Jalba, A.C., Roerdink, J.B.T.M.: Automatic segmentation of diatom images. In: Proc. Comput. Anal. Images Patterns 2003. Volume 2756 of Lecture Notes in Computer Science. (2003) 369–376
4. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univ. Press, Cambridge (1988)
5. Amenta, N., Bern, M., Eppstein, D.: The crust and the β-skeleton: Combinatorial curve reconstruction. Graphical Models and Image Processing **60** (1998) 125–135
6. Wilkinson, M.H.F., Jalba, A.C., Urbach, E.R., Roerdink, J.B.T.M.: Identification by mathematical morphology. In Du Buf, J.M.H., Bayer, M.M., eds.: Automatic Diatom Identification. Volume 51 of Series in Machine Perception and Artificial Intelligence. World Scientific Publishing Co. , Singapore (2002) 221–244
7. Pech-Pacheco, J.L., Cristobal, G.: Automatic slide scanning. In du Buf, H., Bayer, M.M., eds.: Automatic Diatom Identification. World Scientific Publishing, Singapore (2002) 259–288
8. Fischer, S., Bunke, H., Shahbazkia, H.R.: Contour extraction. In du Buf, H., Bayer, M., eds.: Automatic Diatom Identification. World Scientific Publishing, Singapore (2002) 93–107

Hierarchical Regions for Image Segmentation

Slawo Wesolkowski and Paul Fieguth

Systems Design Engineering
University of Waterloo
Waterloo, Ontario, Canada, N2L-3G1
{swesolko,pfieguth}@uwaterloo.ca

Abstract. Image segmentation is one of the key problems in computer vision. Gibbs Random Fields (GRFs), which produce elegant models, but which have very poor computational speed have been widely applied to image segmentation. In this paper, we propose a hierarchical region-based approach to the GRF. In contrast to block-based hierarchies usually constructed for GRFs, the irregular region-based approach is a far more natural model in segmenting real images. By deliberately oversegmenting at the finer scales, the method proceeds conservatively by avoiding the construction of regions which straddle a region boundary. In addition to the expected benefit of computational speed and preserved modelling elegance, our approach does not require a stopping criterion, common in iterated segmentation methods, since the hierarchy seeks the unique minimum of the original GRF model.

1 Introduction

A key problem in computer vision is to distinguish between separate objects in an image scene. A critical step is that of image segmentation, which seeks to separate objects on the basis of distinct appearance. The image segmentation process is dependent on two interactive components: 1) a pixel dissimilarity criterion and 2) a framework for grouping similar pixels and separating dissimilar ones.

The focus of this paper is the pixel grouping algorithm. That is, given a specified dissimilarity criterion, what is an efficient and effective means of constructing groups of pixels or image segments? We consider hierarchical methods based on Markov/Gibbs Random Fields [5] given their ease of constructing models for segmentation [7]. Indeed, many Gibbs Random Fields methods have been introduced in recent years [3,5,7,11], however, most of these methods are computationally slow and, therefore, not practical. To increase the convergence speed of the algorithm, it is necessary at some point to move away from processing individual pixels to processing image patches or regions, which can be achieved using multiscale or hierarchical methods. In multiscale methods, information is processed from coarse-to-fine resolutions of the same image while in most hierarchical methods [3,7], an ever finer hierarchy of labels is established at the same image resolution. In coarse-to-fine problem formulations, regions are progressively defined at finer square subdivisions of higher levels. This is problematic

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 9–16, 2004.
© Springer-Verlag Berlin Heidelberg 2004

since the label relaxation which occurs at every lower level might not correctly characterize the underlying regions given the higher level constraint for each square child block.

Another way to approach this problem would be to pose it as a graph where pixels/regions are nodes and edges represent relationships between the nodes (e.g., edge gradients between pixels). Barbu and Zhu [2] propose a method which searches the space of all graph partitions (i.e., segmentations) to find the global optimum of a Baysian posterior probability. They reformulate the Swendsen-Wang (SW) algorithm [9] for graphs by allowing the algorithm to split, merge or re-group a sizeable subgraph (sets of pixels) and thus by achieving fast mixing at low temperatures, it eliminates the slow Gibbs sampling procedure. Although not hierarchical in nature, this algorithm is similar to ours in that it allows groups of pixel labels to be flipped at any one time. The major difference being the splitting of regions/subgraphs in addition to merging them.

We propose an approach in which a hierarchy is constructed from fine to coarse level. Because the regions at a level are produced as arbitrary concatenations of regions at finer levels, the resulting regions can naturally fit those of the image being analyzed, rather than the poor fit of predefined square regions in a coarse-to-fine hierarchy.

The paper is organized as follows. The second section describes the Gibbs Random Fields framework. The third section details the new region-based hierarchical approach. Section four presents results while the fifth section concludes the paper.

2 Local GRF Model

The modelling problems in this paper are addressed from the computational viewpoint by using Gibbs Random Fields to model the image segmentation process. There are two primary concerns: how to *define* an objective function for the optimal solution of the image segmentation, and how to *find* this optimal solution. For the purpose of this paper, the "exact" solution to our segmentation problem will be interpreted as the optimum solution to the optimization objective. In principle, the solution is straightforward: simulated annealing [5] is widely used in solving Gibbs problems; however, it is very slow. The principal concern of this paper is the definition of an appropriate hierarchical approach for faster annealing.

Suppose we are given an image $\boldsymbol{X}$ with labels l on a pixel lattice $\mathcal{L} = \{i, j\}$ with dissimilarity criterion $\Phi(\cdot)$. We will assume $\mathcal{L}$ has a first order neighborhood structure on a regular grid shown in Figure 1a (a second order neighborhood structure would also be feasible). The energy model is then written as follows:

$$U = \sum_{i,j} \{\Phi(\boldsymbol{X}_{i,j}, \boldsymbol{X}_{i,j+1})\delta_{l_{i,j},l_{i,j+1}} + \Phi(\boldsymbol{X}_{i,j}, \boldsymbol{X}_{i+1,j})\delta_{l_{i,j},l_{i+1,j}} + \beta\left[(1 - \delta_{l_{i,j},l_{i,j+1}}) + (1 - \delta_{l_{i,j},l_{i+1,j}})\right]\} \quad (1)$$

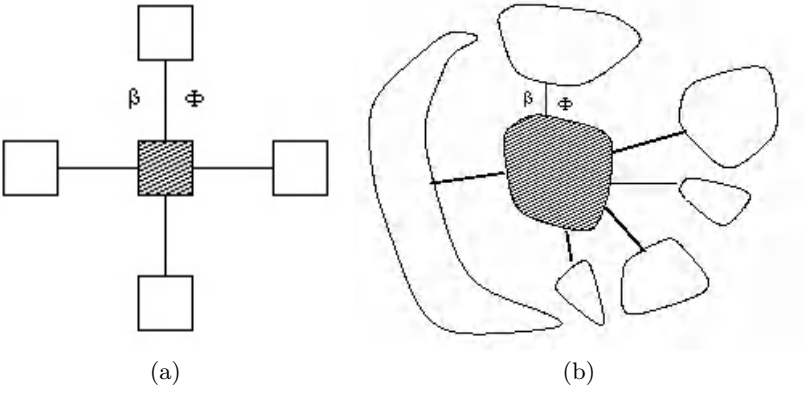


Fig. 1. Illustration of β and Φ interactions between adjacent pixels/regions: (a) first order neighborhood on a regular grid for the finest or pixel-level model, (b) region neighborhood on an irregular grid for higher level region-based model.

where β controls the relative constraints on the degree of region cohesion and fragmentation, while $\delta_{l_{i,j},l_{i,j+1}}$ is the Kronecker δ. This model operates directly on pixels and is therefore a fine or pixel level model. The δ functions ensure that the labelling is consistent. β is usually determined experimentally. This is essentially a region growing-type model [10,6] where decisions to integrate a pixel into the region are done with respect to the criterion Φ. The major difference between this local GRF model and region growing methods is that it is non-causal due to its stochastic nature (whereas in region growing algorithms the inclusion of a pixel is very much dependent on previously included pixels).

Model (1) suffers from a slow random walk of information. For example, assume we have identical pixels in an image. If we have one homogenously labelled region and some pixels out side of it labelled differently then to see *any* change in energy, all remaining pixels will have to be flipped as shown in Figure 2. This is because the term with the region coupling criterion, β, measures boundary length and the dissimilarity criterion Φ is zero. This implies that only the slowest of annealing schedules will successfully converge. One way to overcome this limitation would be to merge adjacent regions in successive higher levels stages after the annealer has converged on the finest level. This would occur only if the merging would lower the overall energy. However, having an explicit merging step in the algorithm points to a deficiency in the original model formulation. Therefore, the merging step needs to be part of the model formulation.

3 Hierarchical GRF Region Grouping

One way to overcome the slow random walk limitation would be to design the model to lower the global energy by merging similar adjacent regions by flipping all the pixel labels in a region (in an analogous fashion to flipping the pixel label

Fig. 2. The slow random walk of annealing: Given a homogeneous set of pixels, the energies of the three illustrated cases of two segmented regions (one shaded and one black) are identically equal. Therefore, there is no energy gradient to drive the solution to the optimum, simple region. Within the domain of flat energies, the annealer performs a random walk eventually finding one of the optimal endpoints. The time to converge to an endpoint grows quadratically with region size.

at the finest level). The formulation of our Gibbs/Markov model will be similar to model (1). To devise a hierarchical fine-to-coarse region-based approach, we first reformulate this model in order to define interactions between regions:

$$U^{(s)} = \sum_{r,r' \in \mathcal{R}^{(s)}, r \neq r'} \{\Phi^{(s)}_{r,r'} \delta_{l_r,l_{r'}} + \beta^{(s)}_{r,r'}(1 - \delta_{l_r,l_{r'}})\} \tag{2}$$

where r is a region indicator, s is the level on the hierarchy, $\mathcal{R}^{(s)}$ is the set of all regions r, $\Phi^{(s)}_{r,r'}$ is the dissimilarity criterion between regions r and r' and $\beta^{(s)}_{r,r'}$ is the region coupling parameter between regions r and r'. When $s = 0$, the formulation corresponds to the special case of the finest (or pixel) level model which was presented in (1) which means that $\Phi^{(s)}_{r,r'}$ and $\beta^{(s)}_{r,r'}$ define relationships between all pixels (and are non-zero only for adjacent pixels). Furthermore, the neighborhood structure is now defined on an irregular grid as shown in Figure 1b.

This model is in practice non-local in that it operates on regions rather than pixels. Indeed, the region-to-region interactions are cumulative local interactions between the pixels. This model still performs a random walk; however, the operation is now sped-up since the label comparisons now happen on a regional, multi-pixel level rather than the single pixel interactions of Model (1) thus speeding the convergence process considerably.

We assert that model (1) is by construction a special case of model (2). First, let us consider a neighborhood structure on an irregular grid in model (2). Interactions between regions on the irregular grid are governed by $\Phi^{(s)}_{r,r'}$ and $\beta^{(s)}_{r,r'}$. However, at the finest level, the pixel level, these values correspond respectively to the pixel-wise interactions Φ and β (since the edge penalty is the same for all pixel pairs) between two pixels. In other words, the irregular grid at the finest level just happens to be a regular grid because of how pixels are arranged in an image. Therefore, models (1) and (2) are equivalent for $s = 0$.

We also assert that both models are equivalent at $s > 1$ as long as the pixels/regions that are *supposed* to be merged are merged. We can say this if we are able to keep the information from level s to level $s + 1$ equivalent. We do this by constructing transition equations between levels which transfer the interactions between regions at level s to the merged regions at level $s+1$, as well

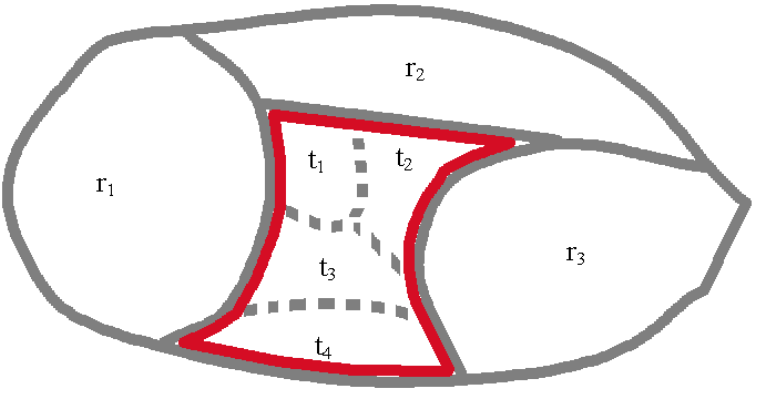


Fig. 3. Region merging: regions $t_{1...4}$ are being merged into a single region r_4 delineated by the thicker boundary.

as by choosing a conservative value for β. The transition for the dissimilarity criterion between two regions r and r' at level $s+1$ is dependent on all the individual distances between all the regions in $G_r^{(s)}$ and $G_{r'}^{(s)}$:

$$\Phi_{r,r'}^{(s+1)} = \sum_{t \in G_r^{(s)}} \sum_{t' \in G_{r'}^{(s)}} \Phi_{t,t'}^{(s)} \tag{3}$$

where t and t' are region indicators. The coupling parameter between two neighboring regions r and r' at level $s+1$ is written in an analogous fashion:

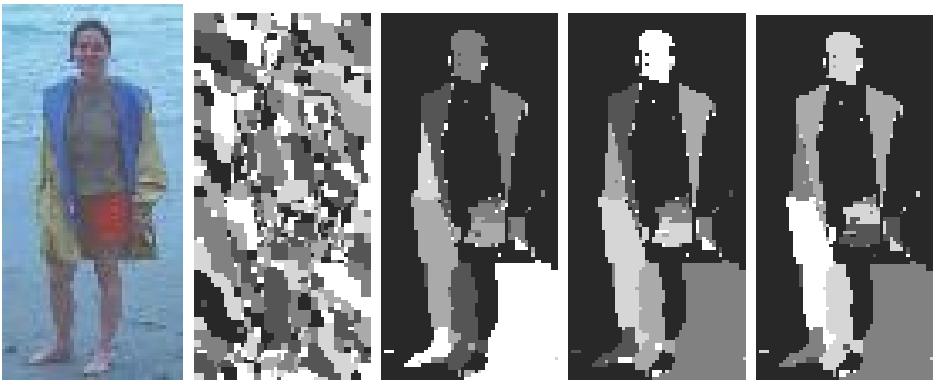
$$\beta_{r,r'}^{(s+1)} = \sum_{t \in G_r^{(s)}} \sum_{t' \in G_{r'}^{(s)}} \beta_{t,t'}^{(s)} \tag{4}$$

Therefore, we now have model (2) which governs how the labelling is done at each level s together with between level transition equations (3) and (4).

To illustrate how the transition equations work consider two sets of regions $r_{1...3}$ and $t_{1...4}$ as shown in Figure 3. Let us assume that regions $t_{1...4}$ will be merged into one region r_4. When regions $t_{1...4}$ are merged into one region, all the relationships between them governed by $\Phi_{r,r'}^{(s)}$ and $\beta_{r,r'}^{(s)}$ must be eliminated since the energy only matters near an edge. Since we are eliminating those relationships, the relationships between $t_{1...4}$ and $r_{1...3}$ will now become the relationships between r_4 and $r_{1...4}$. To accomplish this, since the model is ultimately pixel-based, all we need to do is respectively sum the appropriate $\Phi_{r,r'}^{(s)}$ and $\beta_{r,r'}^{(s)}$. For example, $\Phi_{r_2,r_4}^{(s+1)} = \Phi_{r_2,t_1}^{(s)} + \Phi_{r_2,t_1}^{(s)}$.

The segmentation algorithm is divided into two parts: a trivial image splitting part in the first step, and a region merging part in subsequent steps:

– Assign labels $\{l_{i,j}\}$ randomly to corresponding pixels $\{X_{i,j}\}$



(a) Original (b) Level 1 (c) Level 4 (d) Level 8 (e) Level 11 (final)

Fig. 4. Color image segmentation results with model (2) using $\beta = 0.015$ (image pixels were first normalized to unit length). It is clear that results beyond level 4 are just refinements of previous results. This becomes clear when the number of regions is examined: 416 regions at level 1, 82 regions at level 4, 74 regions at level 8, and 72 regions at level 11.

- Make each $\{X_{i,j}\}$ its own region
- Loop over levels: from finest (pixel) to coarsest:
  - Anneal until convergence:
    - Minimize the energy in model (2) for every region
    - Update the region's label based on Gibbs sampling
  - Apply transition equations (3) and (4)

If the temperature reduction occurs slowly enough, the annealing process converges in probability to the global minimum [5].

We assert that given an *appropriate* distance metric Φ and region coupling parameter β, the algorithm performs an accurate oversegmentation of the image at the first level by creating a multitude of small, compact regions. In practice, any oversegmentation result can be used as a precursor to the subsequent merging iterations as long as only the desired pixels were grouped (i.e., no regions that straddle borders are present in the initial and subsequent segmentations).

Model (2) shares similarities with a few other models in the literature. Zhu's region competition method [12] is similar in that it minimizes an energy function. However, it differs considerably by fostering "competition" between regions (expanding regions from seeds and allowing region splitting) instead of a careful merging strategy adopted here. Angulo's and Serra's ordered mergings algorithm [1] is similar in that it creates a hierarchy of region mergings however it does this in a morphological and not stochastic framework. Their algorithm requires heuristics for merging regions and a stopping criterion for the algorithm.

Fig. 5. Color image segmentation results using model (2) with $\beta = 130$. Depending on the initial segmentation, different results were obtained. This is most likely because the annealing schedule was too fast and the algorithm became stuck in a local minimum.

4 Results and Discussion

Results are presented on color images. The pixel dissimilarity criterion Φ was chosen to be the vector angle measure following [4] as the image has some intensity differences (e.g. shading). Results are shown in Figure 4.

Model (2) encodes only distances between individual pixels and not for example distances between region prototypes [4,12]. Therefore, regions connected by a slowly varying gradient will be merged. This is illustrated in Figure 5 where the Euclidean distance was used for Φ.

We have presented hierarchical regions, a new method for image segmentation based on Gibbs Random Fields (GRFs). In contrast to block-based hierarchies usually constructed for GRFs, the irregular region-based approach is a far more natural model in segmenting real images. By deliberately oversegmenting at the finer scales, the method proceeds conservatively by avoiding the construction of regions which straddle a region boundary. In addition to the expected benefit of computational speed and preserved modelling elegance, our approach does not require a stopping criterion, common in iterated segmentation methods, since the hierarchy seeks the unique minimum of the original GRF model.

We are currently experimenting with a variety of alternate models which might be more appropriate for image segmentation at the finest level to be able to deal with high levels of noise. Furthermore, a structured approach to estimating β for a particular application is also being investigated.

References

1. J. Angulo, and J. Serra, "Color segmentation by ordered mergings," *IEEE ICIP*, Vol. 2, pp. 125-128, Barcelona: September 2003.
2. A. Barbu and S.C. Zhu, "Graph Partition by Swendsen-Wang Cut," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004. (under review)
3. Z. Kato, M. Berthod, and J. Zeroubia, "A Hierarchical Markov Random Field Model and Multitemperature Annealing for Parallel Image Classification," *Graphical Models and Image Processing*, vol. 58, no. 1, 1996, pp. 18-37.
4. P. Fieguth and S. Wesolkowski, "Highlight and Shading Invariant Color Image Segmentation Using Simulated Annealing," *Energy Minimization Methods in Computer Vision and Pattern Recognition III*, Sophia-Antipolis, France, September 2001, pp. 314-327.

5. S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans-PAMI*, Vol. 6, No. 6, 1984.
6. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, Vol. 1, Addison-Welsey, 1992.
7. S. Z. Li, *Markov Random Field Modelling in Image Analysis*, Springer: Tokyo, Japan, 2001.
8. L. Lucchese and S. K. Mitra, "Color Image Segmentation: A State-of-the-Art Survey," *Proc. of the Indian National Science Academy (INSA-A)*, New Delhi, India, Vol. 67, A, No. 2, p. 207-221, March 2001.
9. R. H. Swendsen and J. S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Physical Review Letters*, vol. 58, no. 2, pp. 86-88, 1987.
10. A. Tremeau, and N. Borel, "A Region Growing and Merging Algorithm to Color Segmentation," *Pattern Recognition*, vol. 30, no. 7, pp. 1191-1203, 1997.
11. G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Springer-Verlag, Berlin, Germany, 1995.
12. S. C. Zhu and A. Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 9, pp. 884 -900, Sept. 1996.

Efficiently Segmenting Images with Dominant Sets

Massimiliano Pavan and Marcello Pelillo

Dipartimento di Informatica
Università Ca' Foscari di Venezia
Via Torino 155, 30172 Venezia Mestre, Italy
{pavan,pelillo}@dsi.unive.it

Abstract. Dominant sets are a new graph-theoretic concept that has proven to be relevant in clustering as well as image segmentation problems. However, due to the computational loads of this approach, applications to large problems such as high resolution imagery have been unfeasible. In this paper we provide a method that substantially reduces the computational burden of the dominant set framework, making it possible to apply it to very large grouping problems. Our approach is based on a heuristic technique that allows one to obtain the complete grouping solution using only a small number of samples.

1 Introduction

The segmentation of images is a classic problem in computer vision and pattern recognition, and recently there has been an increasing interest in graph-theoretic segmentation algorithms based on clustering [1]. In a recent paper [5], we have developed a new framework for partitional (i.e., flat) pairwise clustering based on a new graph-theoretic concept, that of a *dominant set*. An intriguing connection between dominant sets and the solutions of a (continuous) quadratic optimization problem allows the use of straightforward dynamics from evolutionary game theory to determine them [7]. The approach has proven to be a powerful one when applied to problems such as intensity, color, and texture segmentation [5, 6].

The drawback of pairwise methods, including the dominant set framework, is the requirement of comparing all possible pairs of pixels in an image. As a consequence, in practical applications it is customary to reduce the number of considered pairs by placing a threshold on the number of connections per pixel, e.g., by specifying a cutoff radius in the image plane or in the color space. However, while discarding long-range connections allows the use of efficient sparse representations, it may results in the oversegmentation of homogeneous regions.

In this paper, we present a heuristic technique that alleviates the computational burden of the dominant set framework and also avoids the side effects of sparse representations. In short, the heuristics works by first solving the grouping problem for a small random subset of pixels and then extending this solution to the full set of pixels in the image. We shall see that the notion of a dominant

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 17–24, 2004.
© Springer-Verlag Berlin Heidelberg 2004

set naturally suggests an elegant way to infer a cluster of data items from a given cluster in the sample set. To do so, we compare all pixels to those in the sample cluster and determine the membership of each pixel in linear time and space with respect to the cardinality of the sample class.

2 Dominant Sets and Their Characterization

We represent the data to be clustered as an undirected edge-weighted (*similarity*) graph with no self-loops $G = (V, E, w)$, where $V = \{1, \ldots, n\}$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \to \mathbb{R}_+^*$ is the (positive) weight function. Vertices in G correspond to data points, edges represent neighborhood relationships, and edge-weights reflect similarity between pairs of linked vertices. As customary, we represent the graph G with the corresponding weighted adjacency (or similarity) matrix, which is the $n \times n$ nonnegative, symmetric matrix $A = (a_{ij})$ defined as:

$$a_{ij} = \begin{cases} w(i,j) \ , & \text{if } (i,j) \in E \\ 0 \ , & \text{otherwise} \ . \end{cases}$$

Clearly, since there are no self-loops, all the elements on the main diagonal of A are zero.

Let $S \subseteq V$ be a non-empty subset of vertices and $i \in V$. The *(average) weighted degree* of i w.r.t. S is defined as:

$$\text{awdeg}_S(i) = \frac{1}{|S|} \sum_{j \in S} a_{ij} \ . \tag{1}$$

Moreover, if $j \notin S$ we define:

$$\phi_S(i,j) = a_{ij} - \text{awdeg}_S(i) \ . \tag{2}$$

Intuitively, $\phi_S(i,j)$ measures the similarity between nodes j and i, with respect to the average similarity between node i and its neighbors in S.

Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$. The *weight* of i w.r.t. S is

$$\text{w}_S(i) = \begin{cases} 1, & \text{if } |S| = 1 \\ \displaystyle\sum_{j \in S \setminus \{i\}} \phi_{S \setminus \{i\}}(j,i)\, \text{w}_{S \setminus \{i\}}(j) \ , & \text{otherwise.} \end{cases} \tag{3}$$

Moreover, the *total weight* of S is defined to be:

$$\text{W}(S) = \sum_{i \in S} \text{w}_S(i) \ . \tag{4}$$

Intuitively, $\text{w}_S(i)$ gives us a measure of the overall similarity between vertex i and the vertices of $S \setminus \{i\}$ with respect to the overall similarity among the vertices in $S \setminus \{i\}$. The following definition represents our formalization of the concept of a cluster in an edge-weighted graph.

Definition 1. *A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be* dominant *if:*

1. $\mathrm{w}_S(i) > 0$, *for all* $i \in S$
2. $\mathrm{w}_{S\cup\{i\}}(i) < 0$, *for all* $i \notin S$.

The two conditions of the above definition correspond to the two main properties of a cluster: the first regards internal homogeneity, whereas the second regards external inhomogeneity. The condition $W(T) > 0$ for any non-empty $T \subseteq S$ is a technicality explained in some detail in [5] and references therein.

We now describe a continuous formulation of the problem of finding dominant sets in an edge-weighted graph. Consider the following quadratic program (which is a generalization of the so-called Motzkin-Straus program [4]):

$$\begin{array}{ll} \text{maximize} & f(\boldsymbol{x}) = \boldsymbol{x}' A \boldsymbol{x} \\ \text{subject to} & \boldsymbol{x} \in \Delta_n \end{array} \tag{5}$$

where

$$\Delta_n = \{\boldsymbol{x} \in \mathbb{R}^n \ : \ x_i \geq 0 \text{ for all } i \in V \text{ and } \mathbf{e}'\boldsymbol{x} = 1\}$$

is the standard simplex of $\mathbb{R}^n$, $\mathbf{e}$ is a vector of appropriate length consisting of unit entries (hence $\mathbf{e}'\boldsymbol{x} = \sum_i x_i$), and a prime denotes transposition. The *support* of a vector $\boldsymbol{x} \in \Delta_n$ is defined as the set of indices corresponding to its positive components, that is $\sigma(\mathbf{x}) = \{i \in V \ : \ x_i > 0\}$.

The following theorem, proved in [5], establishes an intriguing connection between dominant sets and local solutions of program (5).

Theorem 1. *If S is a dominant subset of vertices, then its weighted characteristics vector $\mathbf{x}^S$, which is the vector of Δ_n defined as*

$$x_i^S = \begin{cases} \frac{\mathrm{w}_S(i)}{\mathrm{W}(S)}, & \text{if } i \in S \\ 0, & \text{otherwise} \end{cases}$$

is a strict local solution of program (5).

Conversely, if $\boldsymbol{x}$ is a strict local solution of program (5) then its support $S = \sigma(\boldsymbol{x})$ is a dominant set, provided that $\mathrm{w}_{S\cup\{i\}}(i) \neq 0$ for all $i \notin S$.

The condition that $\mathrm{w}_{S\cup\{i\}}(i) \neq 0$ for all $i \notin S = \sigma(\boldsymbol{x})$ is a technicality which deals with non-generic situations.

3 From Partial to Complete Groupings

Given an edge-weighted similarity graph, by virtue of Theorem 1 we can find a cluster of data items by first localizing a solution of program (5) with an appropriate continuous optimization technique, and then picking up the support set of the solution found. Unfortunately, continuous optimization of (5) is a computationally demanding task. In [5,6], we have used a straightforward continuous optimization technique known as *replicator equations*, a class of dynamical systems arising in evolutionary game theory [7]. Although such systems are capable

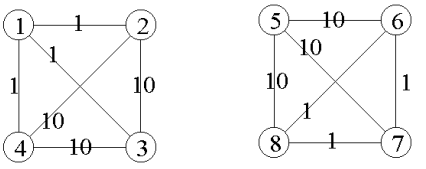


Fig. 1. An example edge-weighted graph.

of providing satisfactory results after only few iterations [5,6], applications to large images have proven to be problematic. To speed up the grouping process, our idea is to first cluster only a small number of image pixels and then extrapolate the complete grouping solution by exploiting the properties of a dominant set in a principled way.

Let $\hat{G} = (\hat{V}, \hat{E}, \hat{w})$ be the similarity graph built upon the whole dataset. Let also $G = (V, E, w)$ be the similarity graph built upon an arbitrarily chosen partial data set. We assume that G is the subgraph induced on $\hat{G}$ by $V \subseteq \hat{V}$, i.e. two nodes are adjacent in G if and only if they are adjacent in $\hat{G}$, and $w(i,j) = \hat{w}(i,j)$ for all $(i,j) \in E$.

Recall from Sect. 2 that, given a subset of nodes S and a vertex $i \notin S$, $\mathrm{w}_{S\cup\{i\}}(i)$ gives us a measure of the overall similarity between vertex i and the vertices of S with respect to the overall similarity among the vertices in S. For example, in the graph of Fig. 1 it turns out that $\mathrm{w}_{\{1,2,3,4\}}(1) < 0$ and $\mathrm{w}_{\{5,6,7,8\}}(5) > 0$ and this can be intuitively grasped by looking at the amount of edge-weight associated to vertices 1 and 5: that associated to vertex 1 is significantly smaller than that of subset $\{2,3,4\}$; conversely, that associated to vertex 5 is significantly greater than that of subset $\{6,7,8\}$.

Let $S \subseteq V$ be a subset of vertices which is dominant in the partial graph G and let $i \in \hat{V} \setminus V$. Whenever $\mathrm{w}_{S\cup\{i\}}(i) > 0$, node i is tightly coupled with the nodes in S, while if $\mathrm{w}_{S\cup\{i\}}(i) < 0$, node i is loosely coupled. The case $\mathrm{w}_{S\cup\{i\}}(i) = 0$ corresponds to a non-generic boundary situation that does not arise in practical applications, and thus can be safely ignored. According to these observations, given a subset of vertices $S \subseteq V$ which is dominant in the partial graph G, we argue for taking

$$S \cup \{i \in \hat{V} \setminus V : \mathrm{w}_{S\cup\{i\}}(i) > 0\}$$

as a cluster of data items in $\hat{G}$.

To illustrate, we proceed with an illustrative example on synthetic data. We use a point set example in two dimensions which can be easily visualized, see Figure 2 (left). The data consists of $n = 39$ points in $\mathbf{R}^2$ arranged in four globular regions. The partial data set is drawn at random, with sampling probability $p = 0.3$. The corresponding partial similarity graph is a complete graph where the similarity between points i and j is $w(i,j) = \exp(-d_{ij}^2/\sigma^2)$, d_{ij} is the pairwise Euclidean distance, and σ is a positive real number which reflects some reasonable local scale. Figure 2 (middle and right) shows the clustering result on the sample and the whole data set (respectively). Clustering the complete data set with no sampling at all, yields precisely the same result.

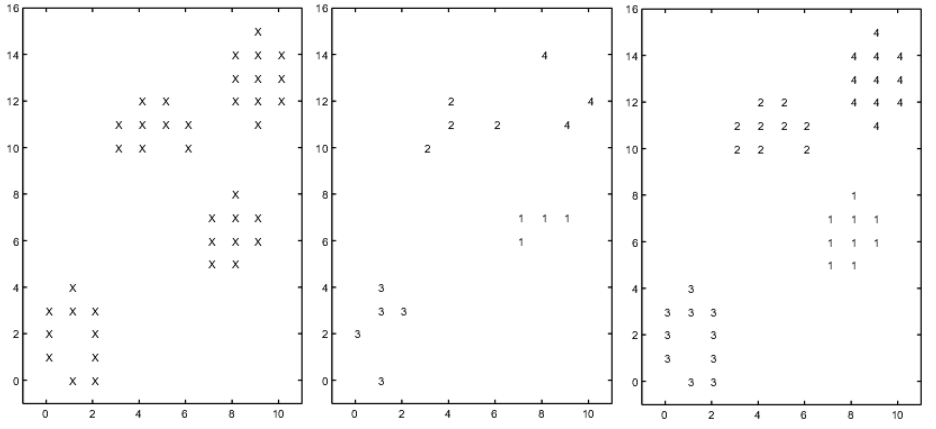


Fig. 2. Illustrative example using 39 data points. Left: The data points. Middle: Clustering result on the partial data set. Right: Clustering result on the complete data set. Parameter setting: $\sigma = 5.8$.

The following proposition allows us to efficiently implement the proposed heuristic technique.

Proposition 1. *Let $G = (V, E, w)$, $\hat{G} = (\hat{V}, \hat{E}, \hat{w})$ be two similarity graphs such that G is the subgraph induced on $\hat{G}$ by $V \subseteq \hat{V}$. Let $A = (a_{ij})$, $\hat{A} = (\hat{a}_{ij})$ be the weighted adjacency matrix of G and $\hat{G}$, respectively. Let also $S \subseteq V$ be a dominant subset of vertices in G and $\boldsymbol{x}^S$ its weighted characteristic vector. Then, we have:*

$$\mathrm{w}_{S\cup\{i\}}(i) > 0 \qquad \textit{iff} \qquad \sum_{h\in S}(\hat{a}_{hi} - a_{hj})x_h^S > 0 \tag{6}$$

for all $i \in \hat{V} \setminus V$ and $j \in S$ (the sum in (6) does not depend upon the choice of j).

Proof (sketch). From Theorem 1, it follows that $\boldsymbol{x}^S$ is a strict local solution of program (5). As a consequence, $\boldsymbol{x}^S$ must satisfies the Karush-Kuhn-Tucker (KKT) equality conditions for problem (5), i.e., the first-order necessary equality conditions for local optimality [2]. Now, let $\hat{n} = |\hat{V}|$ be the cardinality of $\hat{V}$ and let $\hat{\boldsymbol{x}}^S$ be the ($\hat{n}$-dimensional) weighted characteristic vector of S in $\hat{G}$, which can be obtained by properly extending $\boldsymbol{x}^S$ with zero-valued components for all the nodes in $\hat{V} \setminus V$. It is immediate to see that $\hat{\boldsymbol{x}}^S$ satisfies the KKT equality conditions for the problem of maximizing $\hat{f}(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}'\hat{A}\hat{\boldsymbol{x}}$, subject to $\hat{\boldsymbol{x}} \in \Delta_{\hat{n}}$. The proposition follows easily from [5, Lemma 2] and the fact that, by Definition 1, $\mathrm{W}(S) > 0$. □

Note that the sign of $\mathrm{w}_{S\cup\{i\}}(i)$ can be determined in linear time and space with respect to the cardinality of the cluster S.

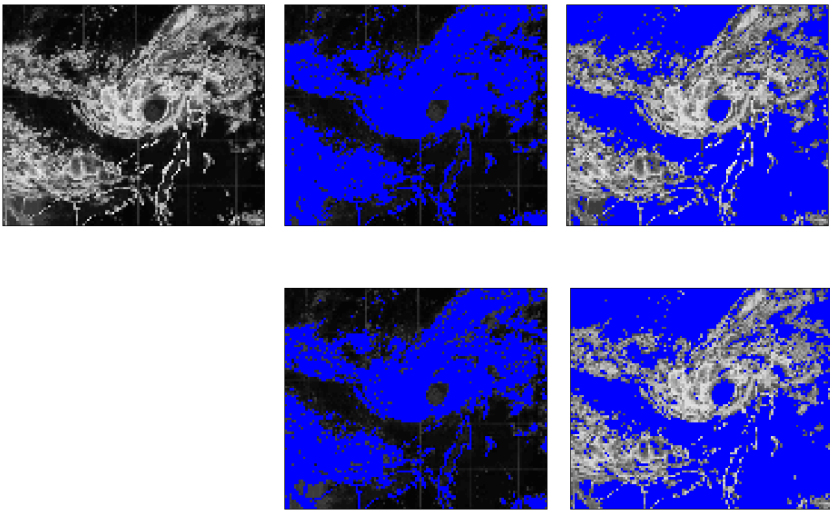


Fig. 3. Top: A 115×97 weather radar image and the components of the segmentation obtained with the complete algorithm. Bottom: the components of the segmentation obtained with the partial algorithm. Parameter setting: $\sigma = 1$.

4 Application to Image Segmentation

We applied our clustering methodology to the segmentation of brightness images. The image to be segmented is represented as an edge-weighted undirected graph, where vertices correspond to individual pixels and the edge-weights reflect the "similarity" between pairs of vertices. As customary, we defined a similarity measure between pixels based on brightness proximity. Specifically, following [5], in our experiments the similarity between pixels i and j was measured by:

$$w(i,j) = \exp\left(\frac{-\|\mathbf{I}(i) - \mathbf{I}(j)\|_2^2}{\sigma^2}\right)$$

where σ is a positive real number which affects the decreasing rate of w, and $\mathbf{I}(i)$ is defined as the intensity value at node i, normalized to a real number in the interval $[0, 1]$.

After drawing a set of pixels at random with sampling probability $p = 0.005$, we iteratively found a dominant set in the partial graph (i.e., a solution to program (5)) and then removed it from that graph. At each iteration, we also extrapolated and removed a cluster from the whole data set. The continuous optimization method we use to solve problem (5) is called *replicator equations*, a class of dynamical systems arising in evolutionary game theory [7]. We refer the reader to [5,7] for details.

Figures 3 to 5 show the results obtained with our segmentation algorithm on various natural brightness images. The major components of the segmentations

Fig. 4. Left: An image of a plane. Middle and right: the components of the segmentation. Parameter setting: $\sigma = 3.5$.

are drawn on a blue background. The leftmost cluster is the one obtained after the first iteration of the algorithm, and successive clusters are shown left to right.

Figure 3, which shows a weather radar image, has been used in [5] with the complete (i.e., with no sampling) grouping algorithm. The segmentations obtained with the complete and the partial algorithm (respectively, first and second row in Figure 3) look quite similar. In both cases, the algorithms correctly discovered a background and a foreground region. The approximation algorithm took a couple of seconds to return the segmentation. Compared with the complete algorithm, this corresponds to a time speedup greater than 15.

Figures 4 and 5 show results on a couple of 481×321 images taken from the database presented in [3]. On these images the sampling process produced a partial data set with no more than 1000 pixels, and our current MATLAB implementation took only a few seconds to return a segmentation. Running the complete grouping algorithm on the same images (which contain more than $150,000$ pixels) would be unfeasible.

Figure 4 shows the image of a plane. Essentially, the algorithm found two main regions: a large component for the background of clouds, and another one for the plane. Note also that some small parts of the cloud region are incorrectly put together with the plane cluster.

Finally, Figure 5 shows the image of a church. Here, apart from some small spurious regions, the algorithm was able to segment the image into meaningful components. It found a large component for the sky and, within the church, it distinguished between the white walls and the dark areas (the door, the balcony, and the stairs).

Due to space restrictions, we omit results on the stability of the algorithm with respect to the sampling stage. In short, the algorithm exhibits a nice tolerance to random variations in the sample data set.

5 Conclusion

In this paper, we have presented a technique for efficiently partitioning data with dominant sets for image segmentation. The heuristics is simple to implement as well as computationally efficient, and leverages the fact that the number of regions in an image is usually much smaller than the number of pixels. Experimentally, we have demonstrated the potential of our approach for intensity

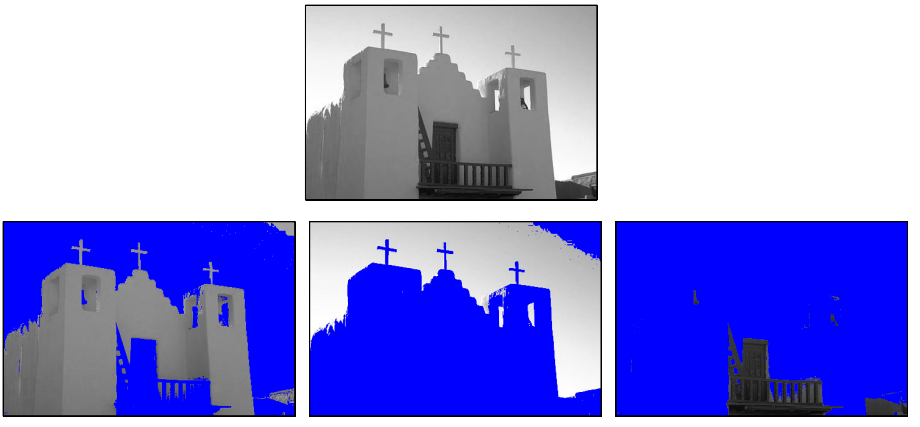


Fig. 5. Top: An image of a church. Bottom: the components of the segmentation. Parameter setting: $\sigma = 2.25$.

image segmentation. The framework, however, is general and can be applied in a variety of image analysis and recognition domains such as, for example, color, texture and motion segmentation, and the unsupervised organization of an image database. All this will be the subject of future work.

References

1. D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach.* Prentice-Hall, Englewood Cliffs, NJ, 2002.
2. D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Reading, MA, 1984.
3. D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. of the IEEE Int. Conf. on Computer Vision*, volume 2, pages 416–424, 2001.
4. T. S. Motzkin and E. G. Straus. Maxima for graphs and a new proof of a theorem of Turán. *Canad. J. Math.*, 17:533–540, 1965.
5. M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 145–152, 2003.
6. M. Pavan and M. Pelillo. Unsupervised texture segmentation by dominant sets and game dynamics. In *Proc. IEEE Int. Conf. on Image Analysis and Processing*, pages 302–307, 2003.
7. J. W. Weibull. *Evolutionary Game Theory.* MIT Press, Cambridge, MA, 1995.

Color Image Segmentation Using Energy Minimization on a Quadtree Representation*

Adolfo Martínez-Usó, Filiberto Pla, and Pedro García-Sevilla

Dept. Lenguajes y Sistemas Informáticos, Jaume I Univerisity
Campus Riu Sec s/n 12071 Castellón, Spain
[auso,pla,pgarcia]@uji.es, http://www.vision.uji.es

Abstract. In this article we present the results of an unsupervised segmentation algorithm based on a multiresolution method. The algorithm uses color and edge information in an iterative minimization process of an energy function. The process has been applied to fruit images to distinguish the different areas of the fruit surface in fruit quality assessment applications. Due to the unsupervised nature of the procedure, it can adapt itself to the huge variability of colors and shapes of the regions in fruit inspection applications.

1 Introduction

Image segmentation is one of the primary steps in image analysis and visual pattern recognition. Most of image segmentation techniques are application-oriented and have been developed for specific purposes although that could be applied to a wide range of particular problems. Thus, the main motivation of the developed work has been to obtain a method able to segment images of fruits for their quality classification in visual inspection processes using a computationally efficient hierarchical representation. Particularly, the application problem that has motivated this work implies the following requirements:

1. An *unsupervised method* would be needed due to manifold variables which can arise in fruit images. Thus, any prior knowledge should be avoided for the segmentation procedure.
2. The segmentation method has to be mainly based on *color* and *edge criteria*, in order to define the segmented region boundaries as accurately as possible.

To meet the above mentioned requirements, a multiresolution Quadtree (QT) structure has been chosen to support the developed method due to its computational efficiency. The algorithm we present attachs great importance to an efficient strategy to solve the problem and the image segmentation will be conditioned and orientated by the image representation adopted, that is, by the QT and by the color and edge information as a particular and robust criterion.

* This work has been partly supported by grants DPI2001-2956-C02-02 from Spanish CICYT and IST-2001-37306 from the European Union

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 25–32, 2004.
© Springer-Verlag Berlin Heidelberg 2004

As part of any segmentation process, a criterion or condition is needed to know when the final segmentation has been reached. In our algorithm, the *ideal segmentation state* is achieved through a variational method that minimizes the segmentation energy.

The main contribution of the presented work is the proposed energy function that efficiently combines color (intra-region features) with edges (borders information) using a computationally efficient hierarchical representation, a QT, to guide the minimization process. The proposed framework yields satisfactory results, particulary in fruit inspection tasks.

2 Variational Image Segmentation

The goal of variational methods for image segmentation is to develop algorithms and their mathematical analysis to minimize the segmentation energy E represented by a real value. The segmentation energy measures how smooth the regions are, the similarity between the segmented image and the original one and the similarity between the obtained edges and the discontinuities of the original image.

The Mumford-Shah model [6] has been regarded as a general model within variational segmentation methods. This model looks for a piecewise smoothed image u with a set of discontinuities, edges of the original image g.

According to Mumford-Shah's conjecture, the minimal segmentation exists but it is not unique; for each image a set of minimal segmentations exists. Therefore, with the aim of minimizing the segmentation energy we can minimize the following equation, where K is the set of discontinuities in the image domain Ω representing the *edges* of g:

$$E(u, K) = \int_{\frac{\Omega}{K}} (|\nabla u(x)|^2 + (u - g)^2)dx + length(K) \tag{1}$$

Since Mumford-Shah's work, several approaches appeared that suggested modifications to the original scheme. Recent works change equation (1) in order to improve the results. In this sense, the boundary function, which is binary in the Mumford and Shah's formulation, was changed by a continuous one which obtains a clearly defined boundary in [5]. Furthermore, in [1] the authors analyze some possible generalizations of the Mumford-Shah functional for color images. They suggest that these changes accentuate different features in edge detection and restoration.

In general, formulating variational methods have several advantages:

1. A variational approach returns explicitly a measure of the quality of the segmentation. Therefore, we are able to know how good the segmentation is.
2. Many segmentation techniques can be formulated as a variational method.
3. A variational approach can be used as a quantitative criterion in order to measure the segmentation quality.
4. Finally, a variational approach provides a way to implement non-supervised processes by looking for a minimum in the segmentation energy.

3 Energy Minimization of the Quadtree Image Representation

In this work, a function to minimize the segmentation energy is proposed. With this assumption, it is important to point out that we cannot guarantee to find a global minimum but, the experimental results obtained show that the solutions are very satisfactory.

We have developed a statistically robust functional. That is, we use an energy function that takes into account any resolution or scale change producing the same segmentation results in each case.

Let u be a smoothed image and a set of discontinuities of the original image g, let R_i be a set, with $0 < i \leq \|\Omega\|$ and $R_i \subseteq \Omega$. R_i is a family of r regions in Ω such that $\bigcup_{i=1}^{r} R_i = \Omega$ and $R_i \bigcap R_j = \emptyset$ for $i \neq j$. Let B_i represent the border of region R_i, that is, $R_i' = R_i - B_i$ is the inner part of region R_i. Finally, let γ be certain very small value that avoids dividing by zero.

Thus, $\forall x \in R_i$ let us consider the function $E_i(R_i, B_i)$:

$$E_i(R_i, B_i) = \int_{R_i} (|u(x) - m_i|)\ dx + \frac{\int_{R_i'} |\nabla g(x)|}{\int_{B_i} |\nabla g(x)| + \gamma}\ dx \tag{2}$$

In the image u, $u(x)$ represents the color of an R_i element x, and m_i is a central measure for the color value of R_i. The final segmentation energy is expressed as $E(\Omega) = \sum_i E_i(R_i, B_i) + \|\Omega\|$.

The QT structure allows us to divide an image within a complete multiresolution tree representation including neighboring information. This spatial information can be further used by a clustering strategy which groups the QT leaves using color and edge information.

Let us see the following discrete version of (2) with the same nomenclature,

$$E_i(g) = k \cdot H_i + (1 - k) \cdot G_i + \lambda \cdot length(\Omega) \tag{3}$$

that returns the segmentation energy at each region, H_i and G_i are terms as follows:

$$H_i = \frac{\sum_{R_i} (D(u(x), m_i))}{\sigma_{image}} \qquad G_i = \frac{\sum_{R_i - B_i} |\nabla g(x)|}{\sum_{B_i} |\nabla g(x)| + \gamma} \tag{4}$$

Specifically, in the QT representation, R_i is the set of leaves of the QT belonging to region i and B_i represents the boundary leaves in R_i, $0 < i \leq r$, with r being the number of regions at each iteration. The function D calculates a distance between two colors (Euclidean, Manhattan, etc.). The value $|\nabla g(x)|$ returns the gradient magnitude at the pixel x. Note that the parameter $0 < k < 1$ allows to define the weight between the color and boundary information and λ allows the method to give more or less importance to the number of regions. Finally, the value σ_{image}, which is the sum of the standard deviations of each plane in the image, contributes to normalize the first term and makes the function statistically robust. Thus, in the energy functional (3) we can distinguish three components:

1. The first one takes into account the homogeneity of each region by means of a distance from each pixel to a central measure of its region.
2. The second component promotes that the gradient magnitude will be low in the interior leaves and high in the boundary ones.
3. Finally, the third term helps the function to punish a large number of regions such as the Mumford-Shah model or many other approaches about variational image segmentation [5][1].

4 The Algorithm

4.1 Use of Color Information

Any color representation could be used. Using perceptual spaces as L*a*b* or HSI or using other representations invariant to certain features, different segmentations results will be achieved. Therefore, the process here proposed can be used on any color representation although, obviously, the results obtained will depend on it.

Paying attention to color spaces, several ones have been used with different segmentations as a result. The final regions represent the most significant features of each color space. So, in the experiments carried out the following color representations have been tested in order to verify the effect of the methodology proposed using different color spaces:

- **RGB.** This space presents a high correlation among its planes but the final result have had a pleasing performance.
- **L*a*b* and L*u*v*.** Regarding to color, these are perceptual uniform spaces where the resulting regions would be proportional to the human perceptual difference between the color of each cluster.
- **Invariant features.** In order to meet specific features within a color space, we can use some of the invariant spaces described in the literature. For instance, we use HSI to take advantage of a perceptual space as well as H plane is less influenced by non-uniform illumination. We also make use of the invariant spaces described in [7] and develop a robust feature spaces discounting shadow, illumination highlights or noise.
- **Particular spaces.** We have tested several specific spaces in order to achieve the best performance when applying the method to images of fruits. All of them are transformations from the RGB input image.
  - From the results extracted of [2], we have developed an special plane $p1 = k \cdot \log(R + G - 2 \cdot B)$ which tries to find the ideal separation between the standard color of a defect and the standard color of a healthy fruit regarding oranges.
  - To have invariants for the dichromatic reflection model with white illumination a new color model $l_1 l_2 l_3$ is proposed in [4]. In our case, we only use l_2 and l_3 planes in order to make a $2D$-space which avoids the highlights and represents the chroma:

$$l_2 = \frac{(R-B)^2}{(R-B)^2 + (R-G)^2 + (G-B)^2} \tag{5}$$

$$l_3 = \frac{(G-B)^2}{(R-B)^2+(R-G)^2+(G-B)^2} \tag{6}$$

The functional described in the previous section has a first component H_i which represents our specific constraint to each region with the color information as a criterion. This component calculates the distance between a central measure (usually the median) and each pixel that belongs to this region. So, the smaller this result, the more homogeneous the cluster is. Note that this color term has an statistically robust behavior.

4.2 Use of Edge Information

The use of edge information in the segmentation process tries to avoid the merging of regions when the color criterion is satisfied but there exists an edge between the regions. In this sense, gradient information is used to develop a boundary map which is checked as the edge criterion. This boundary map is made from the maximum gradient value found in the R, G or B planes.

As with the color information, the functional (3) has a component G_i to try to find the boundaries as accurate as possible. This component promotes regions with no edges inside, because it promotes that the gradient magnitude will be low in the inner leaves of the region and high in the boundary ones.

4.3 The Segmentation Process

The QT structure allows us to divide an image within a complete multiresolution tree representation including neighboring information. This spatial information can be further used by a clustering strategy which joins the QT leaves using color and edge information. The multiresolution process allows us to analyze the image with a coarse-to-fine sequence described as follows:

1. We construct a hierarchical structure level by level. It is important to clarify that, talking about a QT, the meaning of *level* is a set of leaves with the same size. Thus, the first level will be the first four children leaves that descend from the whole image and so on. Therefore, while the previous named QT is created, each level is revised by the functional (3) in order to revise the clustering at that resolution. Each cluster created in any level will be taken into account in the next levels. Finally, when we finish the QT construction, the *salient regions* have been detected in a coarse way.
2. Focusing the attention on the *salient regions* (the coarse ones that have been labelled), they will be taken as the most significant groupings of the image. So, we continue the process expanding each cluster by means of a region growing method where each cluster aplies the functional (3) to its neighboring regions. This second step will take care of shaping the edges of each region by color and edge criteria.

Note that we use the functional (3) described in Sect. 3 in both of the previous steps but, whereas in the first one this application is guided by a hierarchical structure in order to develop each resolution level, in the second one the application of the functional follows a region growing strategy to achieve the final regions in a more accurate way.

Before summarizing the segmentation process, it is important to point out what are the main ideas the proposed method is based on:

1. We look for different features according to the color space used. The algorithm find groups that match regions in any color space we select, however, these groups will have properties according to the *salient features* of the color space used.
2. To guide the segmentation process, the following questions have to be solved:
   a) *the way to take to continue the segmentation process.*
   b) *how long the process have to continue.*

   The first question is solved by means of a multiresolution analysis of the image with a QT structure. Multiresolution is able to decompose the image in several resolution levels developing a coarse-to-fine process from the *salient regions* to the final shapes of each region. On the other hand, the question (*b*) will be determined by the functional (3) described in Sect. 3. It will be minimized in a progressive way until the functional energy stops decreasing.

The whole segmentation process is summarized in the following algorithm:

1. From RGB input image, make an edge map and create the reference color image with the selected color space.
2. Construct an oversegmented representation of the image, that is, expand the QT until every square region have all pixels with the same color. After this, create an ordered list according to region sizes.
3. The algorithm computes the functional (3) for each region and its neighbor regions in an iterative sequence that may be seen as a coarse-to-fine segmentation process.
4. If the whole image energy has decreased, reorder the list of regions by size and repeat the previous step.
5. Regroup small regions.

The previous algorithm shows the steps of the segmentation strategy used, which is basically an iterative process. Each cluster is compared to all its neighboring clusters and it will be merged when the segmentation criterion is satisfied (see Sect. 3). This clustering process stops when no other merging can be performed without increasing the energy of the segmentation. Finally, small regions are ignored and merged to their most similar neighbors giving more importance to the biggest ones.

It is important to point out that regions are arranged according to their size, giving more importance to bigger regions. This represents the spatial constraint in the merging process and facilitates merging small regions with big ones.

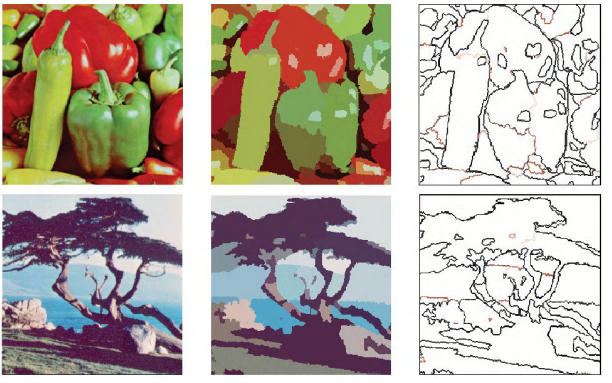


Fig. 1. Real images segmentation results.

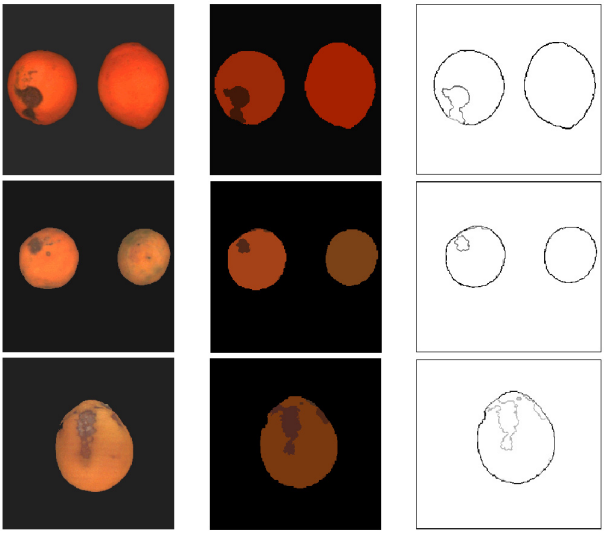


Fig. 2. Fruit images segmentation results.

5 Results

Results obtained with classical images (Fig.1) and fruit images (Fig.2) are presented. Columns represent the original images, segmented images, and the edge results, where the darker the edge line is, the greater difference between the color of neighboring regions. To show these results we have selected a perceptual color space like L*a*b* for the classical images and the transformation of equation (5) to segment the images of oranges.

Fruit image segmentation is used as input in a further process to characterize and classify the fruit surface. These visual inspection applications identify and

detect different types of defects and parts of the fruit. In this sense, images in Fig. 2 show some examples of the segmentation results on different fruits and situations to be characterized. Note how the segmentation process has adapted to the regions of each image due to its unsupervised nature. For instance, the first column shows examples of fruits with various stains produced by the effect of rot and how the segmentation obtained has found the different variations of the stains of the rotten zone. This will allow the extraction of region descriptors for their classification.

Finally, it is important to point out that the algorithm has been compared with the segmentation algorithm presented in [3] which is unsupervised and employs a perceptual color space. In comparison with this algorithm, our algorithm yields similar results when tested on classical images, and outperforms it on fruit images.

6 Conclusions

In this paper, color and edge information combined with a QT representation and the minimization function has been presented. The results obtained show how the algorithm can adapt to the different situations and variability of color regions, being able to segment areas and locating the borders due to the use of gradient information during the segmentation process. Thus, this unsupervised segmentation strategy can locate color regions, and find their contours satisfactorily.

The QT representation not only guides the minimization process but also allows the segmentation at different resolution levels improving the efficiency.

References

1. A. Brook, R. Kimmel, and N.A. Sochen. Variational restoration and edge detection for color images. *Journal of Mathematical Imaging and Vision*, 18(3):247–268, 2003.
2. Y-R. Chen, K. Chao, and Moon S. Kim. Machine vision technology for agricultural applications. *Computers and Elect. in Agriculture*, (36):173–191, November 2002.
3. D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. *IEEE Conf. Computer Vision and Pattern Recognition*, pages 750–755, 1997.
4. Theo Gevers and Arnold W.M. Smeulders. Color based object recognition. *Pattern Recognition*, (32):453–464, March 1999.
5. G.A. Hewer, C. Kenney, and B.S. Manjunath. Variational image segmentation using boundary functions. *IEEE Transactions on Image Processing*, 7(9):1269–1282, 1998.
6. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *CPAM*, 42(4), 1989.
7. J-M Geusebroek Rein van den Boomgaard, Arnold W.M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on PAMI*, 23(12):1338–1350, December 2001.

Segmentation Using Saturation Thresholding and Its Application in Content-Based Retrieval of Images

A. Vadivel[1], M. Mohan[1], Shamik Sural[2], and A.K.Majumdar[1]

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India
{vadi@cc, mmohan@cse, akmj@cse}@iitkgp.ernet.in
[2] School of Information Technology, Indian Institute of Technology, Kharagpur 721302, India
shamik@cse.iitkgp.ernet.in

Abstract. We analyze some of the visual properties of the HSV (Hue, Saturation and Value) color space and develop an image segmentation technique using the results of our analysis. In our method, features are extracted either by choosing the hue or the intensity as the dominant property based on the saturation value of a pixel. We perform content-based image retrieval by object-level matching of segmented images. A freely usable web-enabled application has been developed for demonstrating our work and for performing user queries.

1 Introduction

Segmentation is done to decompose an image into meaningful parts for further analysis, resulting in a higher-level representation of image pixels like the foreground objects and the background. In content-based image retrieval (CBIR) applications, segmentation is essential for identifying objects present in a query image and each of the database images. Wang et al [12] use the LUV values of a group of 4X4 pixels along with three features obtained by wavelet transform of the L component for determining regions of interest. Segmentation-based retrieval has also been used in the NeTra system [5] and the Blobworld system [1]. Some researchers have considered image segmentation as a stand-alone problem in which various color, texture and shape information has been used [2,3,8]. Over the last few years, a number of CBIR systems have been proposed. This includes QBIC [6], NeTra [5], Blobworld [1], MARS [7], SIMPLICity [12] and VisualSeek [10]. A tutorial survey of work in this field of research can be found in [9].

We segment color images using features extracted from the HSV space as a step in the object-level matching approach to CBIR. The HSV color space is fundamentally different from the widely known RGB color space since it separates out intensity (luminance) from the color information (chromaticity). Again, of the two chromaticity axes, a difference in hue of a pixel is found to be visually more prominent compared to that of saturation. For each pixel we, therefore, choose either its hue or the intensity as the dominant feature based on its saturation. We then segment the image by grouping pixels with similar features using the K-means clustering algorithm [4].

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 33–40, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Post-processing is done after initial clustering for merging small clusters into larger clusters. This includes connected component analysis and threshold-based merging for accurate object recognition. Segmentation information from each of the database images is stored as indexed files. During retrieval, a query image is segmented and the segmented image is matched with all the database images using a suitable distance metric. Finally, images that are ranked higher by the distance metric are displayed to the user.

The main contributions of this paper are as follows:

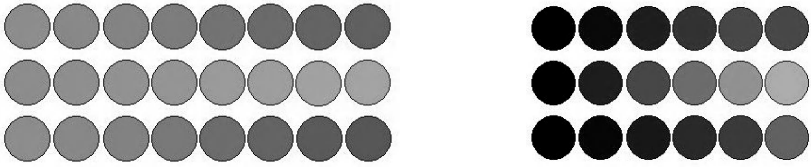
- Detailed analysis of the visual properties of the HSV color space.
- A new approach to image segmentation using the HSV color space properties.
- Development of a web-based image retrieval system using segmented images.

In the next section, we analyze the visual properties of the HSV color space. In section 3, we explain our HSV-based method for feature extraction and image segmentation. We describe the web-based image retrieval system in section 4. Experimental results are included in section 5 and we draw conclusions in the last section.

2 Analysis of the HSV Color Space

A three dimensional representation of the HSV color space is a hexacone, with the central vertical axis representing intensity [11]. Hue is defined as an angle in the range $[0,2\pi]$ relative to the red axis with red at angle 0, green at $2\pi/3$, blue at $4\pi/3$ and red again at 2π. Saturation is the depth or purity of color and is measured as a radial distance from the central axis with values between 0 at the center to 1 at the outer surface. For S=0, as one moves higher along the intensity axis, one goes from black to white through various shades of gray. On the other hand, for a given intensity and hue, if the saturation is changed from 0 to 1, the perceived color changes from a shade of gray to the most pure form of the color represented by its hue. Looked from a different angle, any color in the HSV space can be transformed to a shade of gray by sufficiently lowering the saturation. The value of intensity determines the particular gray shade to which this transformation converges. When saturation is near 0, all pixels, even with different hues, look alike and as we increase the saturation towards 1, they tend to get separated out and are visually perceived as the true colors represented by their hues. This is shown in Fig. 1(a). It is seen that the two leftmost circles in each row give similar impression of color to our eyes even though their hue values are quite different. This is due to low values of their saturation.

For low saturation, a color can be approximated by a gray value specified by the intensity level while for higher saturation, the color can be approximated by its hue. The saturation threshold that determines this transition is once again dependent on the intensity. For low intensities, even for a high saturation, a color is close to the gray value and vice versa as shown in Fig. 1(b). In this figure, it is seen that although the saturation is 1.0 for each of the circles and their hue values are quite different, the leftmost circles in each row give similar impression of color to our eyes. This is due to low values of their intensity.



1(a) 1(b)

Fig. 1. Variation of color perception with (a) saturation (Decreasing from 1 to 0 right to left) for a fixed value of intensity and (b) intensity (Decreasing from 255 to 0 right to left) for a fixed value of saturation

Saturation gives an idea about the depth of color and human eye is less sensitive to its variation compared to variation in hue or intensity. We, therefore, use the saturation of a pixel to determine whether the hue or the intensity is more pertinent to human visual perception of the color of that pixel and ignore the actual value of the saturation. It is observed that for higher values of intensity, a saturation of about 0.2 differentiates between hue and intensity dominance. Assuming the maximum intensity value to be 255, we use the following threshold function to determine if a pixel should be represented by its hue or its intensity as its dominant feature.

$$th_{sat}(V) = 1.0 - \frac{0.8V}{255} \quad . \tag{1}$$

In the above equation, we see that for V=0, $th_{sat}(V) = 1.0$, meaning that all the colors are approximated as black whatever be their hue or saturation. On the other hand, with increasing values of intensity, saturation threshold that separates hue dominance from intensity dominance goes down.

Thus, we treat each pixel in an image either as a "true color" pixel – a pixel whose saturation is greater than th_{sat} and hence, its hue is the dominant component or as a "gray color" pixel – a pixel whose saturation is less than th_{sat} and hence, its intensity is the dominant component.

3 Segmentation Using Saturation Thresholding

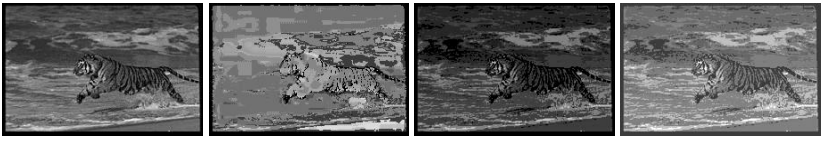
3.1 Feature Extraction

We effectively use visual properties of the HSV color space as described in the last section for color image segmentation. Each image can be represented as a collection of its pixel features as follows:

$$I \equiv \{(pos, [t|g], val)\} \; . \tag{2}$$

Here each pixel is a triplet where *pos* denotes the position of the pixel, [t|g] denotes whether the pixel is a "true color" pixel or a "gray color" pixel and *val* denotes the "true color" value or the "gray color" value. Thus, $val \in [0,2\pi]$ if [t|g] takes a value of *t* and $val \in [0,255]$ if [t|g] takes a value of *g*. Essentially, we approximate each pixel

either as a "true color" pixel or a "gray color" pixel with corresponding true/gray color values and then group similar "true color" or "gray color" values together to be represented by an average value for the group. In this approach, the feature of a pixel is the pair ([t|g], val) – whether it is a "true color" pixel or a "gray color" pixel and the corresponding hue or intensity value. Fig. 2(a) shows an original image and Fig. 2(b) shows the same image using the approximated pixels after saturation thresholding using Eq. (1). Pixels with sub-threshold saturation have been represented by their gray values while the other pixels have been represented by their hues. The feature generation method used by us makes an approximation of the color of each pixel in the form of thresholding. On the other hand, features generated from the RGB color space approximate by considering a few higher order bits only. In Figs. 2(c) - (d) we show the same image approximated with the six lower-order bits all set to 0 and all set to 1, respectively.



(a) (b) (c) (d)

Fig. 2. (a) Original Image (b) HSV Approximation (c) RGB approximation with all low order bits set to 0 and (d) RGB approximation with all low order bits set to 1

It is seen that the approximation done by the RGB features blurs the distinction between two visually separable colors by changing the brightness. On the other hand, the proposed HSV-based approximation can determine the intensity and shade variations near the edges of an object, thereby sharpening the boundaries and retaining the color information of each pixel.

3.2 Pixel Grouping by K-means Clustering Algorithm

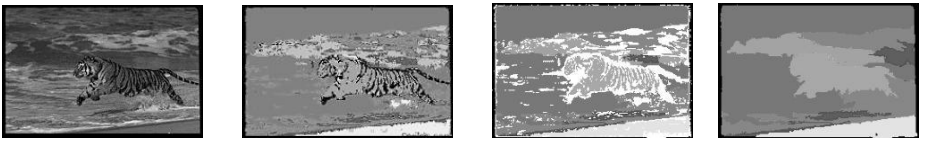
Once we have extracted each pixel feature in the form of ([t|g],val), a clustering algorithm is used to group similar feature values. The clustering problem is to represent the image as a set of *n* non-overlapping partitions as follows:

$$I \equiv \{O_1|O_2|O_3|\ldots.|O_n\} \quad . \tag{3}$$

Here each $O_i \equiv$([t|g], val, {pos}), i.e., each partition represents either a "true color" value or a "gray color" value and it consists of the positions of all the image pixels that have colors close to *val*. We use K-Means clustering for pixel grouping. In the K-Means clustering algorithm, we start with K=2 and adaptively increase the number of clusters till the improvement in error falls below a threshold or a maximum number of clusters is reached. We set the maximum number of clusters to 12 and an error improvement threshold over number of clusters to 5 %.

3.3 Post Processing

After initial K-Means clustering of image pixels, we get different color cluster centers and the image pixels that belong to these clusters. In Fig. 3(a), we show a natural scene image. In Fig. 3(b), we show the transformed image after feature extraction and K-Means clustering. It is observed that the clustering algorithm has determined five "true color" clusters, namely, Blue, Green, Orange, Yellow and Red for this particular image and three gray clusters – Black and two other shades of gray.



(a) (b) (c) (d)

Fig. 3. Different Stages of Image Segmentation. (a) Original image (b) Image after clustering (c) Image after connected component analysis and (d) Final segmented image

However, these clustered pixels do not yet contain sufficient information about the various objects in the image. For example, it is not yet known if all the pixels that belong to the same cluster are actually part of the same object or not. To ascertain this, we next perform a connected component analysis [11] of the pixels belonging to each cluster. Connected component analysis is done separately for pixels belonging to each of the "true color" clusters and each of the "gray color" clusters. At the end of the connected component analysis step, we get the different objects of each color. During this process, we also identify the connected components whose size is less than a certain percentage (typically 1%) of the size of the image. These small regions are to be merged with the surrounding clusters in the next step. Such regions which are candidates for merger are shown in white in Fig. 3(c). In the last post-processing step, the small regions are merged with their surrounding regions with which they have maximum overlap. The image at the end of this step is shown in Fig. 3(d). It is seen that the various foreground and background objects of the image have been clearly segmented.

4 Web Based Image Retrieval Application

We have developed a web-based CBIR application that matches images after segmenting them using the proposed method (www.imagedb.iitkgp.ernet.in/seg). A query in the application is specified by an example image. Initially, a random set of 20 images is displayed. Retrieval is done using the proposed feature extraction and segmentation approach with a suitable distance metric. The nearest neighbor result set is retrieved from the image database based on the query image and is displayed to the user. Users are often interested in retrieving images similar to their own query image. To facilitate this, we provide a utility to upload an external image file and use the image as a query on the database. We plan to enhance our application by displaying

the segmented image corresponding to the uploaded image as an extension of our work.

5 Results

In this section, we show results of applying our segmentation method on different types of images. Figs. 4(a)-(c) show a number of original images, segmentation results using the proposed method and also the corresponding results of segmentation using the RGB color space.

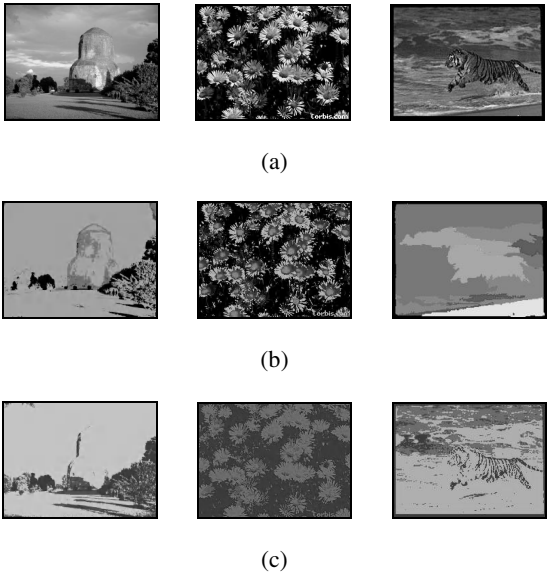


Fig. 4. (a) Original Images (b) Segmentation using HSV features and (c) Segmentation using RGB features

For RGB, we consider the higher order 2 bits to generate the feature vectors. In the images, we have painted the different regions using the color represented by the centroid of the clusters to give an idea about the differentiation capabilities of the two color spaces. Although exact segmentation of unconstrained color images is still a difficult problem, we see that the object boundaries can be identified in a way more similar to human perception of the same. The RGB features, on the other hand, fail to determine the color and intensity variations and come up with clusters that put neighboring pixels with similar color but small difference in shade to different clusters. Often, two distinct colors are merged together. In the HSV-based approach, better clustering was achieved in all the cases with proper segmentation. Fig. 5 shows some more examples of segmentation using the proposed approach that are considered difficult to segment using traditional methods.

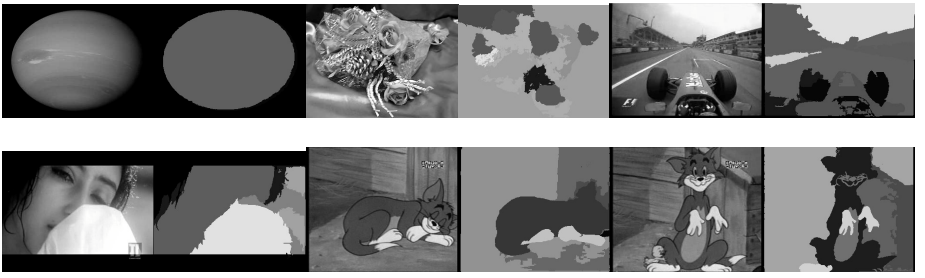
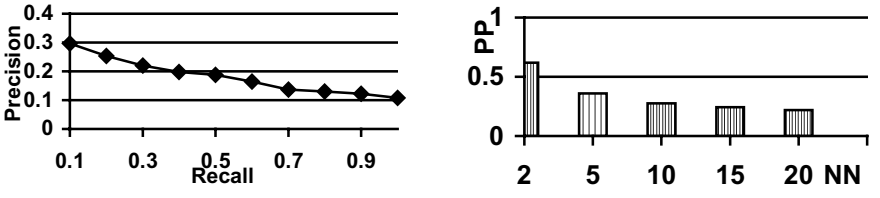


Fig. 5. Segmentation results in the proposed system. The first image in each pair is the original image and the second is the segmented image



6(a) 6(b)

Fig. 6. (a) Precision vs. recall on a controlled database of 2,015 images. (b) Perceived precision variation on a large un-controlled database of 28,168 images

We first show the recall and precision of retrieval in our CBIR application on a controlled database of 2,015 images in Fig. 6(a). The database has various image categories, each containing between 20-150 images. Any image belonging to the same category as a query image is assumed to be a member of the relevant set.

It should, however, be noted that the performance comparison of large content-based image retrieval systems is a non-trivial task since it is very difficult to find the relevant sets for an uncontrolled database of general-purpose images. One way of presenting performance for such databases is through the use of a modified definition of precision. Even though we do not exactly know the relevant set, an observer's perception of relevant images in the retrieved set is what can be used as a measure of precision. Thus, we re-define precision as "Perceived Precision" (PP) which is the percentage of retrieved images that are perceived as relevant in terms of content by the person running a query.

By measuring PP of a large number of users and taking their mean, we get a meaningful representation of the performance of a CBIR system. In our experiments, we have calculated perceived precision for 50 randomly selected images of different contents and taken their average. Our database currently contains 28,168 images downloaded from the. PP is shown for the first 2, 5, 10, 15 and 20 nearest neighbors (NN) in Fig. 6(b). It is seen that the perceived precision stays almost constant from five to twenty nearest neighbors which implies that the number of false positives does not rise significantly as a larger number of nearest neighbors are considered.

6 Conclusions

We have studied some of the important visual properties of the HSV color space and developed a framework for extracting features that can be used for effective image segmentation. Our approach makes use of the saturation value of a pixel to determine if the hue or the intensity of the pixel is more close to human perception of color that pixel represents. K-Means clustering of features is used to combine pixels with similar color for segmentation of the image into objects. A post-processing step filters out small extraneous clusters to identify correct object boundaries in the image. An image retrieval system has been developed in which database images are ranked based on their distance from a query image. Promising retrieval results are obtained even for a large database of about 28,000 images. We plan to increase the database size to about 80,000 images and compare our results with other segmentation-based retrieval systems.

Acknowledgement. The work done by Shamik Sural is supported by research grants from the Department of Science and Technology, India, under Grant No. SR/FTP/ETA-20/2003 and by a grant from IIT Kharagpur under ISIRD scheme No. IIT/SRIC/ISIRD/2002-2003.

References

1. Carson, C. et al: Blobworld: A System for Region-based Image Indexing and Retrieval. Third Int. Conf. on Visual Information Systems, June (1999)
2. Chen, J., Pappas, T.N., Mojsilovic, A., Rogowitz, B.: Adaptive Image Segmentation Based on Color and Texture. IEEE Conf. on Image Processing (2002)
3. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-texture Regions in Image and video. IEEE Trans. on PAMI, Vol. 23 (2001) 800-810
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York (1990)
5. Ma, W.Y., Manjunath, B.S.: NeTra: A Toolbox for Navigating Large Image Databases. IEEE Int. Conf. on Image Processing (1997) 568-571
6. Niblack W. et al: The QBIC Project: Querying Images by Content using Color Texture and Shape. SPIE Int. Soc. Opt. Eng., In Storage and Retrieval for Image and Video Databases, Vol. 1908, (1993) 173-187
7. Ortega, M. et al: Supporting Ranked Boolean Similarity Queries in MARS. IEEE Trans. on Knowledge and Data Engineering, Vol. 10 (1998) 905-925
8. Randen, T., Husoy, J.H.: Texture Segmentation using Filters with Optimized Energy Separation. IEEE Trans. on Image Processing, Vol. 8 (1999) 571–582
9. Smeulders, A.W.M. et al: Content Based Image Retrieval at the End of the Early Years. IEEE Trans. on PAMI, Vol. 22 (2000) 1-32
10. Smith, J.R., Chang, S.-F.: VisualSeek: A Fully Automated Content based Image Query System. ACM Multimedia Conf., Boston, MA (1996)
11. Stockman, G., Shapiro, L.: Computer Vision. Prentice Hall, New Jersey (2001)
12. Wang, J.Z., Li, J., Wiederhold, G.: SIMPLIcity: Semantics-sensitive Integrated Matching for Picture Libraries. IEEE Trans. on PAMI, Vol. 23 (2001).

A New Approach to Unsupervised Image Segmentation Based on Wavelet-Domain Hidden Markov Tree Models

Qiang Sun, Shuiping Gou, and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University
710071 Xi'an, China
qsun@mail.xidian.edu.cn

Abstract. In this paper, a new unsupervised image segmentation scheme is presented, which combines wavelet-domain hidden Markov tree (HMT) model and possibilistic C-means (PCM) clustering algorithm. As an efficient soft clustering algorithm, PCM is introduce into unsupervised image segmentation and used to cluster model likelihoods for different image blocks to identify corresponding image samples, on the basis of which the unsupervised segmentation problem is converted into a self-supervised segmentation one. The simulation results on synthetic mosaics, aerial photo and synthetic aperture radar (SAR) image show that the new unsupervised image segmentation technique can obtain much better image segmentation performance than the approach based on K-means clustering.

1 Introduction

Image segmentation schemes based on multiscale Bayesian strategy have gained more and more attention in image processing field. IN [1], a powerful statistical signal processing model, wavelet-domain hidden Markov tree (HMT) model, was proposed to capture inter-scale dependencies through a binary tree structure of wavelet coefficients of a 1-D signal, which provides a promising statistical signal modeling framework. Specially, the framework can be extended to 2-D signals, say images, for modeling the quadtree structures of their wavelet coefficients to implement different tasks. On the basis of the HMT model, a supervised multiscale image segmentation algorithm, HMTseg, was developed by Choi to further justify the effectiveness of HMT model [2]. An extension from supervised image segmentation to unsupervised segmentation using HMT-3S model and the JMCMS approach was studied in [3], where K-means clustering was used to identify the corresponding training samples for unknown textures based on the likelihood disparity of HMT-3S. As we all know, however, K-means clustering is a hard clustering approach, and can give distorted clustering results or even fail completely when noise is present in the data set. Hence, bad segmentation results can be got owing to the wrong identification of training samples. Here, we propose a new unsupervised image segmentation scheme, which

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 41–48, 2004.
© Springer-Verlag Berlin Heidelberg 2004

combines wavelet-domain HMT model [1] and possibilistic C-means (PCM) clustering algorithm [4], an efficient soft clustering approach. Due to the higher classification accuracy of PCM, better segmentation performance is obtained. The experimental results on synthetic mosaics, aerial photo and synthetic aperture radar (SAR) image demonstrate that the new unsupervised image segmentation scheme is more efficient.

The organization of this paper is as follows. In section 2, the multiscale HMT model is reviewed briefly. The supervised Bayesian image segmentation is then described in section 3. In section 4, PCM clustering is introduced into unsupervised image segmentation and used to cluster model likelihoods for different image blocks to identify corresponding image training samples. Simulation results on synthetic mosaics, aerial photo and synthetic aperture radar (SAR) image are given in section 5. Finally, a conclusion is drawn in section 6.

2 Multiscale Hidden Markov Tree Model

It is well known that the discrete wavelet transform (DWT) is an effective multiscale image analysis tool because of its intrinsic multiresolution analysis characteristics, which can represent different singularity contents of an image at different scales and subbands. In Fig. 1 (a), a quad-tree structure of wavelet coefficients is shown, which illustrates the parent-child dependencies of wavelet coefficients at three scales and three subbands.

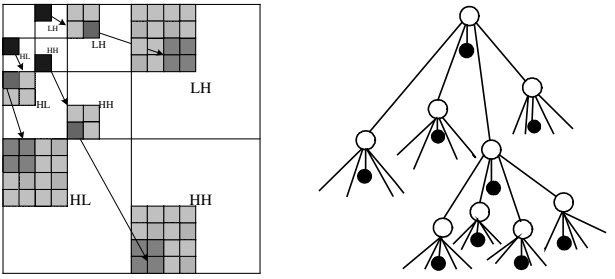


Fig. 1. (a) Quadtree structure of 2-D discrete wavelet transform. (b) A 2-D wavelet hidden Markov tree model for one subband. Each wavelet coefficient (black node) is modeled as a Gaussian mixture model controlled by a hidden state variable (white node) [1], the persistence across scales can be captured by connecting the hidden states vertically across scale in Markov-1 chains

As to multiscale singularity characterization above, one statistical model, hidden Markov tree (HMT) model, was proposed to model this structure [1]. The HMT associates with each wavelet coefficient a “hidden” state variable, which determines whether it is “large” or “small” (see Fig. 1 (b)). The marginal density of each coefficient is then modeled as a two-density Gaussian mixture: a large-variance Gaussian

for the large state and a small-variance Gaussian for the small one. As a result, this Gaussian mixture model can closely fit the non-Gaussian wavelet coefficient marginal statistics existing in most real-world images.

Grouping the HMT model parameters, i.e. state probabilities for root nodes of different quad-trees, state transition probabilities and Gaussian mixture variances, into a vector $\mathbf{\Theta}$, the HMT can be considered as a high-dimensional yet highly structured Gaussian mixture model $f(w|\mathbf{\Theta})$ that approximates the overall joint pdf of the wavelet coefficients *W*. Regarding each wavelet coefficient, the overall pdf $f(w)$ can be formulated as

$$f_W(w) = \sum_{m=1}^{M} p_S(m) f_{W|S}(w|S=m), \tag{1}$$

where *M* is the number of states and *S* is state variable. The HMT model parameters can be estimated using the iterative expectation-maximization (EM) algorithm according to maximum likelihood criterion.

To be noted, the HMT has a convenient nesting structure that matches the dyadic squares in an image [2]. Each subtree of the HMT is also an HMT, with the HMT subtree rooted at node *i* modeling the statistical characteristics of the wavelet coefficients corresponding to the dyadic square d_i in original image.

3 Supervised Image Segmentation

Image segmentation aims at addressing the problem of identifying different regions of homogeneous "textural" characteristics within an image. Supervised Bayesian image segmentation approach classifies an image using both image features and its prior knowledge. Usually maximum a posterior (MAP) estimation is involved [3], i.e.,

$$\hat{x} = \arg\max_{x} E[C_{MAP}(X,x)|Y=y], \tag{2}$$

where $C_{MAP}(X,x)$ is the cost function that assigns equal cost to any single erroneous estimation. The MAP estimation aims at maximizing the probability that all pixels are correctly classified. Later, Bouman et al [5] presented an alternative weighted cost function $C_{SMAP}(X,x)$ to overcome the expensive computation intrinsic by MAP estimator. Given that $Y^{(n)}$ is an image block at scale *n*, $X^{(n)}$ is its class label, $y(n)$ and $x(n)$ are the particular values of them, the SMAP estimator can be formulated as

$$\hat{x}^{(n)} = \arg\max_{x^{(n)}} \left\{ \log p_{y^{(n)}|x^{(n)}}(y|x^{(n)}) + \log p_{x^{(n)}|x^{(n+1)}}(x^{(n)}|\hat{x}^{(n+1)}) \right\}. \tag{3}$$

The two terms in (3) are the likelihood function of an image block and the contextual information from the next coarser scale, respectively. As to the second part of (3), a context-based Bayesian segmentation algorithm, HMTseg, was presented by

Choi et al in [1] where the contextual information is modeled as a context vector $v^{(n)}$. The contextual prior $p_{x(n)|v(n)}(c|u)$ is involved in the SMAP as the second part of (3). Given there are *N* different textures, SMAP estimation can be formulated as [6]

$$\hat{x}^{(n)} = \arg\max_{x^{(n)}} p_{x^{(n)}|v^{(n)},y^{(n)}}(x^{(n)}|\hat{v}^{(n)}, y^{(n)}), \tag{4}$$

where

$$p_{x^{(n)}|v^{(n)},y^{(n)}}(x^{(n)}|\hat{v}^{(n)}, y^{(n)}) \quad \frac{p_{x^{(n)}}(x^{(n)}) p_{v^{(n)}|x^{(n)}}(\hat{v}^{(n)}|x^{(n)}) f(y^{(n)}|x^{(n)})}{\sum_{c=1}^{N} p_{x^{(n)}}(c) p_{v^{(n)}|x^{(n)}}(\hat{v}^{(n)}|x^{(n)}=c) f(y^{(n)}|x^{(n)}=c)}, \tag{5}$$

and $p_{x^{(n)}}(c)$ is the probability mass function (PMF) of class *c* at scale *n* and $f(y^{(n)}|x^{(n)}=c)$ the likelihood function of image block $y^{(n)}$ with respect to class *c* that can be calculated with one upward sweep procedure in the EM algorithm [1].

In HMTseg, the HMT model was applied to characterize texture images, aiming at capturing inter-scale dependencies of wavelet coefficients with the assumption that the three subbands in the quad-tree structure of wavelet decomposition are independent statistically. Alternatively, an improved hidden Markov model, HMT-3S [7], was presented to characterize not only the dependencies of wavelet coefficients across different scales but those across three wavelet subbands to enhance the accuracy of characterizing image statistics. Meanwhile, JMCMS was combined with HMT-3S in [6] to capture more robust contextual information with multiple context models to improve the segmentation performance around boundaries. But segmentation methods above belong to supervised image segmentation, i.e., all image features have to be given in terms of HMT or HMT-3S models. In the following, we propose an new unsupervised image segmentation based on possibilistic C-means clustering, an efficient approach to clustering.

4 Unsupervised Image Segmentation Based on PCM Clustering

4.1 Possibilistic C-means (PCM) Clustering

In [3], a hard clustering algorithm, i.e., K-means clustering, was used to identify training samples for each class, where the goal of K-means clustering is to minimize the following objective function

$$J_e = \sum_{k=1}^{N} \sum_{l \in \Gamma_k} \left\| f(y_{k,l}^{(J)}|\theta) - f(y_k^{(J)}|\theta) \right\|^2, \tag{6}$$

where N is the number of textures in an image, J_e is the partition vatiance, $f(y_k^{(J)}|\theta)$ is the likelihood mean of class k at the coarsest scale J, and $f(y_{k,l}^{(J)}|\theta)$ is the likelihood of an image block l regarding the class k.

However, K-means is a hard clustering algorithm, and can give distorted results or even fail completely when noise is present in the data set. Aiming at this problem, Krishnapuram and Keller [4] presented the possibilistic family of clustering algorithms, which differs from the K-means and fuzzy C-means (FCM) algorithms in that the membership of a sample in a cluster is independent of all other clusters. The objective function of this algorithm is formulated as

$$J_m(L,U)=\sum_{k=1}^{N}\sum_{l\in\Gamma_k}(u_{ij})^m\left\|f(y_{k,l}^{(J)}|\theta)-f(y_k^{(J)}|\theta)\right\|^2+\sum_{k=1}^{N}\eta i\sum_{l\in\Gamma_k}(1-u_{ij})^m, \tag{7}$$

where $L=(\beta_1,\cdots,\beta_C)$ is a C-tuple of prototypes, $U=[u_{ij}]$ is the fuzzy C-partition matrix, N is the total number of textures, $m\in[1,\infty]$ is a weighting exponent called the fuzzifier (a value of 2 usually gives good results in practice), and η_i are suitable positive numbers, which determine the distance at which the membership value of a point in a cluster becomes 0.5. The updated equation of u_{ij} is

$$u_{ij}=\frac{1}{1+\left(\dfrac{\left\|f(y_{k,l}^{(J)}|\theta)-f(y_k^{(J)}|\theta)\right\|^2}{\eta_i}\right)^{\frac{1}{m-1}}}, \tag{8}$$

where η_i can be defined as

$$\eta_i=\frac{\sum_{j=1}^{N}u_{ij}^m\left\|f(y_{k,l}^{(J)}|\theta)-f(y_k^{(J)}|\theta)\right\|^2}{\sum_{j=1}^{N}u_{ij}^m}. \tag{9}$$

4.2 PCM Algorithm

A family of possibilistic clustering algorithm for image sample selection is specified as follows:

```
Set initialize the number of textures N and the fuzzi-
fier m :=2;
Set iteration counter t :=0; stop criterion ε :=1e-5;
Initialize the possibilistic N-partition U^(0) using FCM;
Estimate η_i using equation (9);
```

```
Begin
  Update the prototypes by U^(t) to obtain U^(t+1) by equa-
  tion (8);
  t=t+1;
  Until ||U^(t+1) − U^(t)|| < ε
End.
```

4.3 Image Sample Selection

After the possibilistic C-means clustering for the likelihoods of image blocks at the coarsest scale *J*, reliable training samples for different textures can be obtained. Thereafter, HMT model parameters for different textures can be obtained using the selected image sample blocks via the EM algorithm.

Finally, the segmentation step can be accomplished using the supervised segmentation, HMTseg algorithm, based on (3), (4) and (5).

5 Experimental Results

The simulation experiments are implemented on three sets of images. One set is composed of the synthetic mosaics with different kinds of homogeneous textures, the other sets are aerial photo and synthetic aperture radar (SAR) image respectively. The sizes of the images are all 256×256. Before segmentation, the three sets of images are firstly decomposed into 3 scales by Haar wavelet transform. Following [3], only the coarsest scale (with the size 32×32 at three subbands) is used to implement PCM clustering of the model likelihoods for different image blocks.

Table 1. Segmentatin Performance comparision (%)

Numerical Results	P_a		P_b		P_c	
Mosaics	K-means	PCM	K-means	PCM	K-means	PCM
Mosaic I	90.31	92.12	21.38	27.46	32.18	34.72
Mosaic II	85.31	90.26	19.36	26.90	31.21	33.56

In our work, two synthetic mosaics are used to perform image segmentation. According to [6], three numerical criteria are used to evaluate the performance of segmentation: P_a is the percentage of pixels that are correctly classified, P_b the percentage of boundaries that are consistent with the true ones and P_c the percentage of boundaries that can be detected. The numerical results of P_a, P_b and P_c are tabulated in Table 1 and the segmentation results shown in Fig. 2. As we can see from them, the performance of unsupervised image segmentation based on PCM clustering

is better than that of segmentation based on K-means clustering under the same condition.

Meanwhile, aerial photo and SAR images are also used in our experiment, and their segmentation results based on two clustering schemes are also shown in Fig. 2. Although the texture distributions of the two images are not-uniform, and even noisy, such as SAR image, the method proposed still obtain satisfactory results than that described in [3], this mainly owes to the higher classification accuracy and robustness to noise of PCM clustering than K-means clustering.

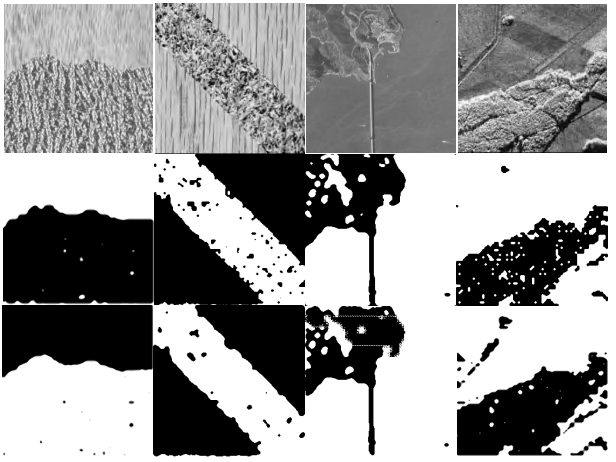


Fig. 2. Unsupervised image segmentation results. The first row is original images. From left to right: Synthetic mosaic I, Synthetic mosaic II, Aerial photo, SAR image. The second row is their segmentation results using the scheme in [3]. The third row is the corresponding segmentation results with the approach proposed

6 Conclusion and Discussion

In this paper, we combine wavelet-domain hidden Markov model and possibilistic C-means (PCM) clustering to implement unsupervised image segmentation. Much better segmentation performance is obtained due to the higher cluster accuracy and robustness against noise of PCM clustering algorithm. Simulation results justify the efficiency of this approach. Currently, we are investigating a more robust unsupervised image segmentation approach specific for SAR images to overcome the “speckle” effect inherent in them.

Acknowledgement. The authors would thank Professor Xinbo Gao for his helpful suggestion for the adoption of the possibilistic approach to clustering.

References

1. Crouse, M.S., Nowak, R.D., Baraniuk, R.G.: Wavelet-Based Signal Processing Using Hidden Markov Models. IEEE Trans. on Signal Processing. 46 (1998) 886–902
2. Choi, H., Baraniuk, R.G.: Multiscale Image Segmentation Using Wavelet-Domain Hidden Markov Models. IEEE Trans. on Image Processing. 10 (2001) 1309–1321
3. Song, X.M., Fan, G.L.: Unsupervised Bayesian Image Segmentation Using Wavelet-Domain Hidden Markov Models. In Proc. of IEEE International Conference on Image Processing. 2 (2003) 423–426
4. Krishnapuram, R., Killer, J.M.: A Possibilistic Approach to Clustering. IEEE Trans. on Fuzzy System. 1 (1993) 98–110
5. Bouman, C.A., Shapiro, M.: A Multiscale Random Field Model for Bayesian Image Segmentation. IEEE Trans. on Image Processing. 3 (1994) 162–177
6. Fan, G.L., Xia, X.G.: A Joint Multi-Context and Multiscale Approach to Bayesian Image Segmentation, IEEE Trans. on Geoscience and Remote Sensing. 39 (2001) 2680–2688
7. Fan, G.L., Xia, X.G.: Wavelet-Based Texture Analysis and Synthesis Using Hidden Markov Models. IEEE Trans. on Circuits and Systems. 50 (2003) 106–120

Spatial Discriminant Function with Minimum Error Rate for Image Segmentation

EunSang Bak

Electrical and Computer Engineering Department,
University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC 28223, U.S.A.
ebak@uncc.edu

Abstract. This paper describes how a normal discriminant function with minimum error rate can be applied to segment an image in a particular manner. Since the maximum likelihood method assigns pixels based on the underlying distributions in image, it is inevitable to make decision errors when there are overlapping areas between the underlying distributions. However, this overlapping area can be minimized by a conversion of distributions which is proposed in this paper. This method is derived by exploiting characteristics of a linear combination of random variables and its relation to the corresponding random vector. The suitable performance of the process is mathematically proved and the experimental results that support the effectiveness of the proposed method are provided.

1 Introduction

Segmentation in data space or equivalently classification in feature space is one of the most important issues in computer vision and pattern recognition. In general, segmentation or classification relies on a particular measure to segment the inhomogeneous data in the data space, or to classify the different feature data in the feature space, respectively. This measure must magnify dissimilarities between the classes [3], and also magnify the similarities within the class. The determination of a measure is the most important part of the both segmentation and classification process.

Such a measure can be obtained from a particular function called discriminant function. The discriminant function approach is one of the widely used methods in the classification and has been applied in a variety of applications [1,2,5]. In particular, distribution-based discriminant function for segmentation is focused in this paper. This approach links a segmentation task to the problem of distribution estimation, and the segmentation process is executed based on the estimated distributions. There are not as many applications of discriminant function for segmentation as for classification. In this paper, the two terms, classification and segmentation, are used interchangeably, since classification of pixels can be regarded as image segmentation. We shall address this issue in more detail.

The paper is organized as follows: In Section 2, a brief explanation of normal discriminant function is given and the motivation in deriving the proposed method is

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 49–56, 2004.
© Springer-Verlag Berlin Heidelberg 2004

suggested based on the discriminant function. The applicability of the normal discriminant function over a spatial domain is presented in Section 3, and the interpretation of the image data is described corresponding to the applicability of the discriminant function. In Section 4, a conversion of distributions scheme is proposed to achieve minimum error rate segmentation, which is mathematically proved in this section. The experimental results are presented in Section 5 and finally conclusions are given in Section 6.

2 Motivation

Suppose there are two classes having different underlying distributions. We label two classes with π_1 and π_2. Assuming that $\pi_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, we would like to assign a sample data $\mathbf{x}$, which is a k-dimensional random vector, to the most probable class between the π_1 and π_2. For this purpose, we use the discriminant function for the multivariate normal density p_i with the same covariance specified as follows.

$$S_1 - S_2 = \ln\left(\frac{p_2}{p_1}\right) \tag{1}$$

where

$$S_i = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}\mathbf{x} - 0.5\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\mu}_i \tag{2}$$

S_i is called a discriminant function for the i^{th} class. If $\mathbf{x}$ is a scalar, the discriminant function is simplified as

$$S_i = \frac{\mu_i}{\sigma^2}x - \frac{\mu_i^2}{2\sigma^2} \tag{3}$$

Assuming the normality for classes, we can produce an optimal decision rule with respect to the underlying distributions. This method is called maximum likelihood estimation, which assigns a pixel to the class in which the pixel has larger probability to belong to.

However, unless there is no overlapping area between the underlying distributions of the two classes, it is inevitable to make decision errors. In order to reduce the area of decision errors (ADE), we consider simple intuitive approaches to reduce the size of the overlapping area. There may be two approaches to reduce the size of this area: one is to make the distance between the means of two classes larger while maintaining their variances, the other is to make the variances of two classes smaller while maintaining their means. These two methods could be performed simultaneously. This is a motivation to develop our proposed method.

3 Discriminant Function in Spatial Domain

Consider the discriminant function for the multivariate normal case for which, in particular, all the components x_j of the vector $\mathbf{x}$ has the same variance in the i^{th} class.

The correspondence of the mean and the variance between x_j and $\mathbf{x}$ becomes as follows.

$$\begin{array}{ccc} x_j & \Rightarrow & \mathbf{x} \\ \begin{bmatrix} \sigma^2 \\ \mu_i \end{bmatrix} & \Rightarrow & \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{\mu}_i \end{bmatrix} = \begin{bmatrix} \sigma^2 \mathbf{C} \\ \mu_i \mathbf{1} \end{bmatrix} \end{array} \tag{4}$$

As can be seen in (4), the covariance of $\mathbf{x}$ is represented by the product of σ^2, which is a scalar, and the corresponding correlation coefficient matrix $\mathbf{C}$. The mean vector is also represented by the product of μ_i, which is a scalar, and the column vector $\mathbf{1}$ whose components are all ones. Therefore, the discriminant function S_i for $\mathbf{x}$ becomes

$$\begin{aligned} S_i &= \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}\boldsymbol{\mu}_i \\ &= \left(\frac{\mathbf{C}^{-1}}{\sigma^2}\mu_i\mathbf{1}\right)^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\left(\frac{\mathbf{C}^{-1}}{\sigma^2}\mu_i\mathbf{1}\right)^{\mathrm{T}}\mu_i\mathbf{1} \\ &= \frac{\mu_i}{\sigma^2}\left(\mathbf{C}^{-1}\mathbf{1}\right)^{\mathrm{T}}\mathbf{x} - \frac{\mu_i^2}{2\sigma^2}\left(\mathbf{C}^{-1}\mathbf{1}\right)^{\mathrm{T}}\mathbf{1} \\ &= \frac{\mu_i}{\sigma^2}\left(\sum_j \gamma_j x_j\right) - \frac{\mu_i^2}{2\sigma^2}\sum_j \gamma_j \end{aligned} \tag{5}$$

where $\sum \gamma_j = \mathbf{1}^T \mathbf{C}^{-1}\mathbf{1}$.

Even if (5) is the discriminant function for a multivariate $\mathbf{x}$, this can be seen from a different perspective. It can be interpreted as a discriminant function for a one dimensional random variable G that is a linear combination of the components x_j of the random vector $\mathbf{x}$ as in (6).

$$G = \sum \gamma_j x_j \tag{6}$$

This fact can be confirmed in the following way. When the expectation μ^* and the variance σ^{*2} of the new variable G are calculated, it will be noticed that the discriminant function of G is the same as that of the multivariate $\mathbf{x}$. Eq. (7) and (8) shows the expectation and the variance of the random variable G in the i^{th} class.

$$\mathrm{E}(G) = \mathrm{E}\left(\sum_j \gamma_j x_j\right) = \sum_j \gamma_j \mathrm{E}(x_j) = \sum_j \gamma_j \mu_i \tag{7}$$

$$\begin{aligned} \mathrm{Var}(G) &= \mathrm{Var}\left(\sum_j \gamma_j x_j\right) = \sum_p \sum_q \gamma_p \gamma_q \mathrm{Cov}(x_p, x_q) \\ &= \sigma^2 \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{C} \boldsymbol{\gamma} \\ &= \sigma^2 \mathbf{1}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{C} \mathbf{C}^{-1} \mathbf{1} \\ &= \sigma^2 \mathbf{1}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{1} \\ &= \sigma^2 \sum_j \gamma_j \end{aligned} \tag{8}$$

Since a linear combination of the normal random variables also follows a normal distribution, the distribution of the random variable G follows the normal distribution with the mean (7) and the variance (8) as in (9).

$$G \sim N(\mu^*, \sigma^{*2}) \qquad (9)$$
$$\mu_i^* = \mu_i \sum_j \gamma_j,\ \sigma^{*2} = \sigma^2 \sum_j \gamma_j$$

Considering G as a random variable, the discriminant function of G is simply described in (10) just as it was done in (3).

$$S_i = \frac{\mu_i^*}{\sigma^{*2}} G - \frac{\mu_i^{*2}}{2\sigma^{*2}} = \frac{\mu_i \sum \gamma_j}{\sigma^2 \sum \gamma_j} G - \frac{\mu_i \sum \gamma_j}{2\sigma^2 \sum \gamma_j} \left(\mu_i \sum \gamma_j\right) \qquad (10)$$
$$= \frac{\mu_i}{\sigma^2} \left(\sum \gamma_j x_j\right) - \frac{\mu_i}{2\sigma^2} \left(\mu_i \sum \gamma_j\right).$$

Comparing (10) with (5), we can see the equivalence of the discriminant functions between a multivariate random variable **x**, and a univariate random variable G that is a linear combination of the components of **x**.

This observation gives a clue to find a refined criterion to segment an image which contains regions of different classes. An image can be seen as a spatial collection of random variables. Each pixel is considered as a realization of its random variable that is supposed to be assigned to a particular class.

Let us define the image **X** where $\boldsymbol{X}_s$ is the intensity at a site s. Suppose that the **X** consists of two different classes, one for the objects and the other for background. The objects usually have rigid body so that the objects from one particular class are located in some particular regions and the other regions in the image are occupied by the class to which the background belongs.

In principle, before applying any algorithms to segment the objects against background, it is assumed that the neighboring pixels of a site s belong to the class to which $\boldsymbol{X}_s$ belongs with probability close to 1. This is called spatial continuity. This rule is violated only around the boundaries between different classes but still gives the value of probability close to 0.5 if there is a certain amount of uncertainty in choosing the class.

Local information around the neighborhood is the most frequently used information to circumvent the decision errors from distribution-based methods. We are also going to take neighboring pixels into consideration through the use of suitable discriminant functions. The main objective in using discriminant functions is reducing

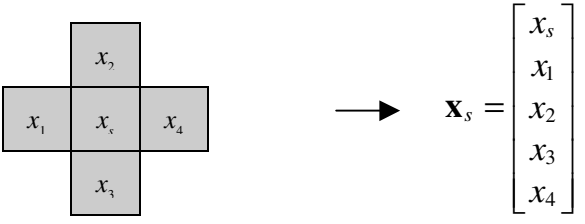


Fig. 1. Random vector at a site s.

the dimension of the random vector to one, and making the decision based on a scalar quantity from the discriminant function. Since each pixel is treated as a random variable, we make a *k*-dimensional random vector for each pixel combining the neighboring pixels. The dimensionality of the resulting random vector depends on the order of the neighborhood system. Given a pixel value x_s at a site *s*, we consider four neighboring pixels around the x_s, and the random vector for x_s is made as in Fig. 1.
Random vector **x**'s for the pixels in the same class build a sample space of the class and the statistical characteristics of the class can be estimated from this sample space. Due to the spatial continuity in the image, the variances of the components of the random vector in the same class are same and the means are the same as the mean of the class they belong to. Eqs. (4)-(10) shows the same case as we have in the image. Therefore, through the process of conversion of distributions, which will be explained in the next section, we will take advantage of the discriminant function of a combination of the random variables whose statistical features are equivalent.

4 Conversion of Distributions

Instead of gray level pixel values, we will generate another quantity *G* for each pixel by calculating (6). As a result, we shall have another image data consisting of *G* values which do not range in gray level. This image is referred to as **G**, and we call this process a conversion of distributions.
Let us look into the changes of the statistical characteristics of the classes in **G** in comparison with those in the original image **X**. For example, while the mean of the i^{th} class was μ_i in **X**, the mean in **G** is $\mu_i \sum \gamma_j$, and the variance changes to $\sigma^2 \sum \gamma_j$ from σ^2. As long as $k = \sum \gamma_j$ is greater than 1, the new distributions in image **G** have larger parameters than in **X**. More specifically, while the mean becomes *k* times larger, the standard deviation becomes $\sqrt{k}$ times larger than the standard deviation of **X**. This implies that the distance between the centers of the distributions of the classes gets larger than the amount of the extended spreads of the distributions. This fact results in

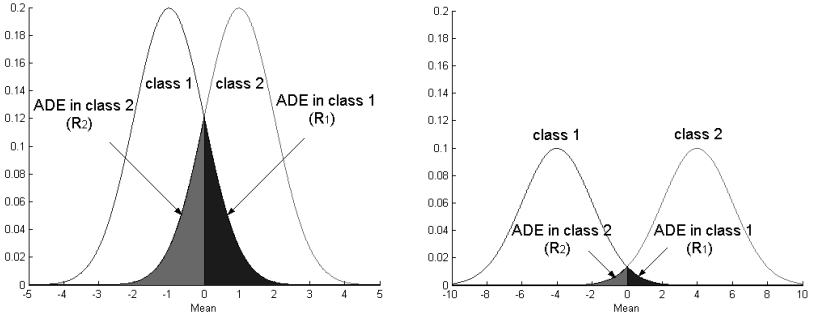


Fig. 2. Conversion of distributions.

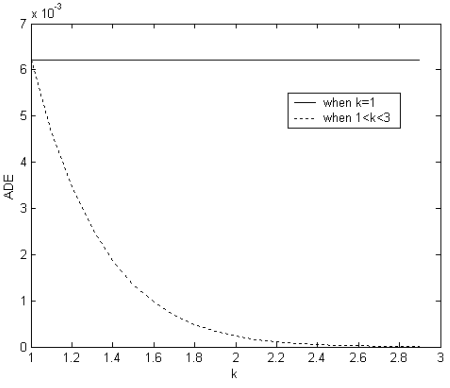


Fig. 3. Decreasing rate of ADE by increasing *k*.

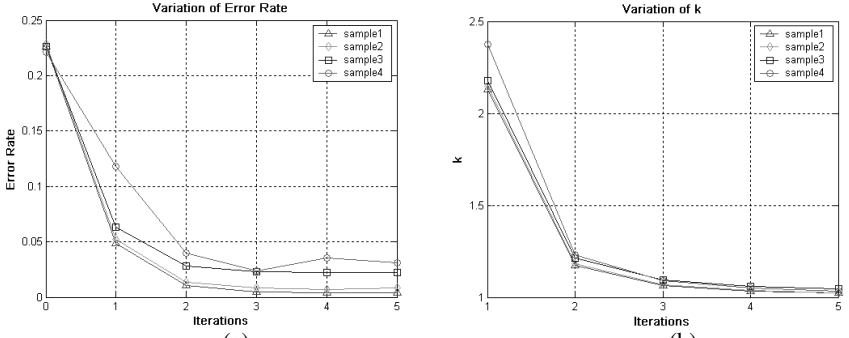


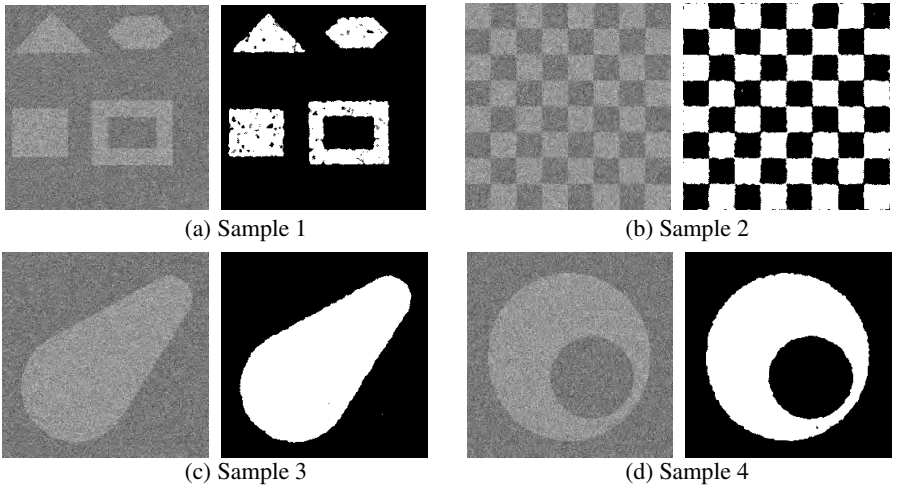
Fig. 4. Variation of error rates and *k*.

the decrease of the overlapping areas, which eventually makes less decision errors, by means of conversion of distributions. Fig. 2 shows the graphical explanation of conversion of distributions when *k* is 4.

Mathematically, it is proved that the conversion of distributions achieves less decision errors and gives better performances, however, the proof is omitted due to the lack of space. Fig. 3 shows the decrease of ADE by increasing *k*.

5 Experimental Results

The proposed method can be much efficiently implemented in an iterative manner. That is, if an image is converted by the process of conversion of distributions, the ADE gets smaller. Based on the image, **G** obtained by the previous conversion, it is converted again through the exactly same process. Theoretically, as long as *k* is greater than 1, it is supposed to give less ADE. Therefore, it will allow making a better decision to assign individual pixels in the image to the correct classes.



(a) Sample 1 (b) Sample 2

(c) Sample 3 (d) Sample 4

Fig. 5. Segmentation/classification results.

The critical value for better performance is k. This k is obtained from the inverse of correlation coefficient matrix $\mathbf{C}$. Since it is initially assumed that classes have a common variance, the means from classes in the image are required for pixel classification.

Fig. 4 shows the variation of the error rates and k's at iterations. For all the images, the proposed method decreases error rate significantly as expected and is compared with the conventional maximum a posteriori (MAP) method. Note that in Fig. 4(a), there is an iteration zero, which means that the error rate at iteration zero is not from the proposed method but from the conventional MAP method. Therefore, it shows how the proposed method overcomes the decision error made by MAP method and gives better results. Consequently, the proposed method can be considered as a new classification method with minimum error rate and the performance is superior to the MAP method.

The algorithm is terminated when either the difference of k's between consecutive iterations is less than 1% or the value itself is less than 1.1. Fig. 4(b) shows that the larger the value of k, the greater the performance. Fig. 5 illustrates the results of pixel classification from the various sample data. Such a pixel classification eventually separates the objects of difference classes in the image so that it can be seen as an image segmentation result. Even though the algorithm in the experiment uses a supervised classification scheme, it is possible for the proposed method to be implemented in an unsupervised way. The only difference is whether prior information is given or should be estimated by additional procedure.

In the light of the results, the proposed method which has been theoretically proved to give better results also gives better results experimentally and this fact indicates that this method will be an alternative to overcome the error rate that has been known as an optimal error rate in the MAP method.

6 Conclusions

This paper shows a new classification method applied to image segmentation with minimum error rate. The proposed method was proved mathematically and turned out to be a method with minimum error rate which outperformed the MAP method. The extension of the proposed method will be applications in multidimensional data space.

References

1. Chou, W.: Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition. Proc. IEEE. 88 (2000) 1201-1223
2. Hastie, T., Tibshirani, R.: Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Machine Intell. 18 (1996) 607-616
3. Kurita, T., Otsu, N., Abdelmalek, N.: Maximum Likelihood Thresholding Based on Population Mixture Models. Pattern Recognition. 25 (1992) 1231-1240
4. Mardia , K.V., Hainsworth, T.J.: A Spatial Thresholding Method for Image Segmentation. IEEE Trans. Pattern Anal. Machine Intell. 10 (1988) 919-927
5. Sakai, M., Yoneda, M., Hase, H.: A New Robust Quadratic Discriminant Function. Proc. Int'l Conf. Pattern Recognition (ICPR'98). 1 (1998) 99–102 1998.

Detecting Foreground Components in Grey Level Images for Shift Invariant and Topology Preserving Pyramids

Giuliana Ramella and Gabriella Sanniti di Baja

Istituto di Cibernetica E. Caianiello, CNR, Via Campi Flegrei 34,
80078, Pozzuoli (Naples), Italy
{g.ramella, g.sannitidibaja}@cib.na.cnr.it

Abstract. A method to single out foreground components in a grey level image and to build a shift invariant and topology preserving pyramid is presented. A single threshold is generally not enough to separate foreground components, perceived as individual entities. Our process is based on iterated identification and removal of pixels causing merging of foreground components with different grey levels. This is the first step to generate a pyramid which, within the limits of decreasing resolution, is shift invariant and topology preserving. Translation dependency is reduced by taking into account the four positions of the partition grid used to build lower resolutions. Topology preservation is favoured by identifying on the highest resolution pyramid level all foreground components and, then, by forcing their preservation, compatibly with the resolution, through lower resolution pyramid levels.

1 Introduction

Pyramids are convenient data structures for multiresolution analysis, as they provide successively condensed representations of the information in the input image. An advantage of pyramids is their ability to reduce the influence of noise, by eliminating the importance of details in lower resolutions. Moreover, one can work with a reduced data set, at low resolution, which still provides a reasonable representation of the most relevant regions of the pattern.

Pyramid representations have been employed for a number of tasks, e.g., line-drawing analysis or object contour extraction [1, 5] and segmentation [6, 10]. Pyramids are generally built by using a uniform subdivision rule that summarises fixed sized regions in the image, regardless of their contents. The representations tend to smooth out variations within regions when resolution decreases, resulting in the unavoidable loss of some information. Possible variations on this scheme concern: how lower resolution images are computed, how pixels are linked to each other, the size of the neighbourhood to be investigated in order pixels can find their *parent* (a pixel at lower resolution level), and so on [11]. Both continuos and discrete methods to build pyramids can be found in the literature. In this communication, a discrete method is presented and, hence, only discrete methods will be taken into account in the rest of the paper.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 57–64, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Two important features, not always satisfied by discrete pyramid generation methods, are shift invariance and topology preservation. Representations at low resolution can be severely distorted when the input is shifted. A different position of the sampling grid can, in fact, lead to a significantly modified pyramid structure. Moreover, the mechanism used to build the pyramid can significantly alter geometry and topology of the original single-scale image, so that the representative power of lower resolution images would become questionable.

Topology preservation is not considered in methods based on filters, [12,15], while is taken into account in generation methods of irregular pyramids even if, as a counterpart, the father-child relation is lost or altered, [16, 17]. Our aim is to obtain an almost shift invariant and topology preserving grey level pyramid, without altering or destroying the father-child relation, and based on a regular tessellation. In [18], a method for binary images was proposed to generate a pyramid which, within the limits of decreasing resolution, is shift invariant and topology preserving. Here, we borrow the ideas suggested in [18] and present a method for 256 grey level images to single out foreground components and to build almost shift invariant and topology preserving pyramids.

To reduce translation dependency, we combine the different results that would be originated by the decimation process when the partition grid is differently placed on the input image. Actually, decimation is performed only once, and the combination is obtained by using a suitable neighbourhood of each pixel and proper weights. To preserve topology through lower resolution levels, we first identify and single out in the input image all significant foreground components. Topology preservation is, then, obtained by opening *canals* to separate foreground parts that constituted individual entities at the highest resolution. In general, identification of foreground components in a grey level image cannot be achieved just by thresholding the image with a single threshold. Our simple and effective method to single out foreground components in the input grey level image is obtained by iterating at different grey levels a process consisting of two phases: i) non topological shrinking, and ii) topological expansion.

2 Foreground Components Identification

Let G be the input grey level image, and let θ be an a priori fixed threshold such that all pixels with grey level $g \leq \theta$ can, without any doubt, be interpreted correctly as background pixels. If G is segmented by assigning to the foreground all pixels with grey level $g > \theta$, and to the background all pixels with $g \leq \theta$, generally a number of components different from the expected one is obtained. The threshold θ should assume different values in different parts of the image, to allow correct identification of foreground components. Moreover, the effect of thresholding may significantly reduce the size of foreground components, as the threshold necessary to separate component equally affects all pixels in the component, including those far enough from the fusion area.

We iterate a process, based on non topological shrinking and topological expansion, which uses different values for θ. In particular, the initial value of θ can be set to the minimal grey level found in the image and the final value of θ can be set to the maximal grey level, decreased by 1. Of course, the initial and final values of θ can be

differently set, depending on user's needs. In general, the initial value should be such that all pixels with grey level smaller than or equal to θ can be interpreted without doubt as background pixels; the final value should be set in such a way to prevent foreground fragmentation into meaningless parts. Non topological shrinking and topological expansion are repeated for all the values of θ, which are obtained by automatically increasing the initial value by an increment δ. The increment δ can be set to any value greater than or equal to 1, depending on problem domain. In particular, a small increment is preferable for images where grey level distribution of the foreground is in a large range and also small variations of grey level are significant.

Let θ_1, θ_2, ..., θ_n be the n values of the threshold, where the threshold at the i-th iteration is $\theta_i = \theta_{i-1} + \delta$. The following subsections describe the process done at the i-th iteration, which are respectively i) non topological shrinking and ii) topological expansion.

2.1 Non Topological Shrinking

This phase is aimed at separating foreground components and it verifies whether, for the given threshold, fusions among components actually occurred.
At the i-th iteration, pixels candidate to removal are those with grey level g such that $\theta_1 < g \leq \theta_i + \delta$. These pixels are examined in increasing grey level order and are removed (i.e., their grey levels are lowered to the value θ_1) provided that they have at least an edge-neighbour with grey level equal to θ_1, regardless on whether connectedness is preserved. However, the number of connected components of pixels with grey level greater than θ_1 in the neighbourhood of any removed pixel is counted. If more than one component is found, this indicates that the removed pixel was a fusion point among foreground components. The removed pixels are orderly stored in a suitable list L, to keep track of their coordinates and grey levels. Removal (i.e., lowering to θ_1) is repeated on the image as far as removable pixels are found.

2.2 Topological Expansion

This phase is aimed to restore size and shape of foreground components, which have been modified by non topological shrinking, still maintaining separated the components that have been singled out. If no fusion occurred at θ_i (i.e., the number of connected components for all pixels in L was equal to 1), all pixels in L are newly assigned their initial grey level in the image G. In turn, if some fusion occurred, care should be taken when setting on G the pixels from L. To this purpose, pixels of L are examined in the opposite order with respect to their insertion order in L. For the current pixel p, the number of components of pixels with grey level greater than θ_1 is counted after p is restored in G. If only one component is found, p is just assigned on G its initial grey level, as no fusion occurred in p. Otherwise, p is also suitably marked on G to point out that p was a fusion point. Once all pixels of L have been processed and restored in G, the fusion marker is propagated onto the neighbours of the marked pixels that have been recovered from L. Marker propagation is repeatedly done to correctly identify all pixels in the fusion area. Actually, marker propagation

may include in the fusion area more pixels than those whose removal is necessary to separate components. Thus, a suitable process is done to reduce the thickness of the set of marked pixels, so as to better preserve the original contents of G. Finally, all pixels still marked as fusion points are set to θ_1, which results in the separation of foreground components at θ_i.

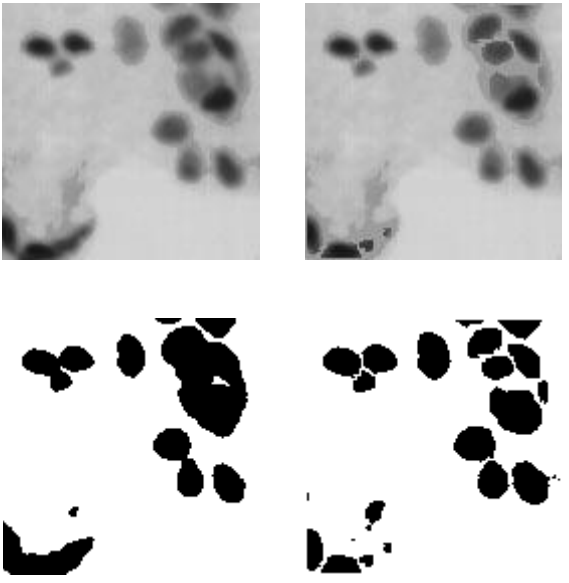


Fig. 1. Top: the input grey level image G, left, and the resulting image G_1, right. Bottom: the binarized images obtained from G, left, and G_1, right.

The threshold θ_i is incremented by δ and non topological shrinking and topological expansion are iterated to separate foreground parts at the various threshold values.
The effect of our process can be seen by comparing the two images in Fig.1 top, respectively showing the input grey level image G (left) and the image G_1 (right), resulting after non topological shrinking and topological expansion. In Fig. 1 bottom, the binarizations of G and G_1 are shown, where pixels with grey level equal to or smaller than θ_i are set to 0 and the remaining pixels are set to 1, to point out which are the foreground components singled out by the process, right, compared with the result of using a single threshold value, left.

3 Building the Pyramid

The pyramid is constructed based on recursive subdivision of the $2^n \times 2^n$ grey level image, G_1, into quadrants, until all pyramid elements represent individual pixels in the original image. However, resolution levels with less than 16 pixels in each dimen-

sion can be disregarded, as they would not be useful in shape analysis tasks. The base of the pyramid represents the original image G_1 at full resolution (128×128, in our case), the next level of the pyramid represents the image G_2 at a uniformly lower resolution (half the resolution in each dimension), and so on. The apex of the pyramid is, in our case, the 16×16 image G_4.

On G_1, foreground components singled out by the process described in Section 2, are identified by means of connected component labelling. G_2 is built by superimposing onto G_1 a grid partitioning it into non overlapping blocks of 2×2 pixels (cells). Each cell of the grid constitutes a single pixel in G_2. The four pixels in a cell are called the children, and the cell itself is called the parent pixel. The value of each parent pixel is computed by taking into account number, positions and grey levels of its four children. The label identifying the foreground component to which children belong is also transferred to the parent pixel. When children belong to different components, a majority rule is used to decide the label for the parent pixel.

Depending on the position of the partition grid over G_1, a pixel p in the image G_1 belongs to one out of four possible bocks. In turn, each of the 8 neighbours of the pixel p on G_1 belongs to a different number of blocks, depending on its position with respect to p. Each edge-neighbour of p belongs to two different 2×2 blocks including p, while each point-neighbour belongs to exactly one such 2×2 block. The decimation filter is a 3×3 mask of weights, with central value 4, edge-neighbours valued 2 and point-neighbours valued 1. This multiplicative mask is centred on each pixel p of G_1 with even coordinates (2j, 2k), and the value of the corresponding parent pixel p' with coordinates (j, k) in G_2, is computed.

The image G_2 is then re-scaled by linearly re-scaling its pixels to the range {0, 255} and the mask is applied to build G_3, and so on.

The basic foreground-background decisions are those of the OR-pyramid. Thus, no part of the pattern is lost, not even the very narrow ones. Moreover, since the multiplicative mask takes into account all possible 2×2 non overlapping blocks, the resulting pyramid will be almost shift invariant.

We point out that the result obtained by using our mask is equivalent to the result obtained by using a Gaussian filter based on 3×3 neighbourhood [12].

To understand how in general shape and topology are altered while building the pyramid, a landscape representation of the grey level image is useful. If the grey level of a pixel is interpreted as the height in a topographic map, the image G_1 can be interpreted as consisting of a number of islands, some of which may include lakes and craters. Islands of G_1 can merge into a unique component in G_2, lakes and craters may be transformed into less deep craters, and new lakes can appear, due to deformation of bays. To preserve topology as much as possible, we open suitable canals at lower resolutions to separate islands erroneously merged into a unique component. Moreover, we suitably treat spurious craters. In particular, we "open" rather than "fill" them. In this way, the effect produced by the modified value distribution better resembles that of the original image. This opening is done at each resolution level before computing the next pyramid level, so as to avoid chain reaction effects.

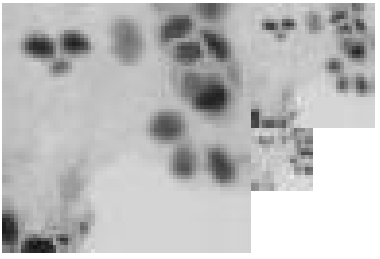


Fig. 2. The pyramid built starting from the grey level image shown in Fig.1. A zoom factor 2 is used.

Fig. 3. A document image where characters are partially merged (left), the image resulting after identification of foreground components (middle), and the binarized version (right).

Fig. 4. The pyramid built starting from the grey level image shown in Fig.3 middle. A zoom factor 2 is used.

To open canals, we scan G_2 and for each foreground pixel we check if it has a neighbour with different label than itself. If it has, the pixel with the lowest grey level (and, hence, the lowest relevance) in the adjoining pair is assigned to the background.

As for opening the craters, we note that a crater is a set of pixels with value strictly smaller than all values of the pixels on the 8-connected crater rim. Craters and lakes which do not correspond to craters and lakes in the full resolution image are interpreted as spurious entities, and are transformed into bays by iteratively identifying their bottom and by assigning to the rim of the bottom the grey level of the bottom.

The pyramid built for the grey level image shown in Fig.1, is illustrated in Fig.2. Only levels 64×64, 32×32, and 16×16 are shown, since level 128×128 is the image in

Fig. 1 top right. A zoom factor 2 has been used to make visible these resolution levels. We note that foreground components are preserved also at low resolution.

4 Conclusion

We have introduced a method to single out foreground components and to build an almost shift invariant and topology preserving pyramid, starting from a 256 grey level image.

Foreground components identification is done by using an iterated process, based on non topological shrinking and topological expansion. The pyramid is then built starting from this image. To limit translation dependency, the four different images that would be generated by shifting, in the four possible positions, the partition grid used to build lower resolutions are combined. Topology preservation is favoured by identifying on the highest resolution pyramid level all foreground components and by opening canals to separate, at lower resolution, foreground parts that constituted individual entities at the highest resolution level. Of course, topology is preserved compatibly with the resolution, through lower resolution pyramid levels.

The method we have developed was mainly intended to treat biological specimens, as the one shown in Fig.1 consisting in a cluster of epithelial cells. However, we think that the method could also be used in other fields, e.g., for document analysis. As an example, see Figs.3,4, showing a document image (Fig.3 left), and the pyramid (Fig.4) resulting after foreground components have been singled out (Fig.3 middle). In Fig. 3 right the binarized version is shown to make it visible the splitting effect caused by foreground components identification. In Fig. 4, only levels 64×64, 32×32, and 16×16 are shown, and a zoom factor 2 is used. More in general, the method is adequate for application areas dealing with grey level images in which foreground components can be distinguished from the background.

References

1. E. E. Stannard, D. Pycock, “Recognising 2-D shapes from incomplete boundaries”, Proceedings of IEE Colloquium on Applied Statistical Pattern Recognition, pp. 12/1 – 12/6; 1999.
2. S. B. Yacoub, J. –M. Jolion, “Hierarchical line extraction”, Proceedings of IEEE Conf. on Vision, Image and Signal Processing, v. 142, pp. 7-14, 1995.
3. C.L. Tan and S.K.K. Loh, “Efficient edge detection using hierarchical structures”, Pattern Recognition, vol. 26, no. 1, pp. 127-135; 1993.
4. G.N. Khan and D.F. Gillies, Extracting contours by perceptual grouping, Image and Vision Computing, vol. 10, no. 2, pp. 77-88; 1992.
5. P. Meer, C.A. Sher, and A. Rosenfeld, “The chain pyramid: Hierarchical contour processing”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 4, pp. 363-376; 1990.
6. J. Kim, H. Kim, “Multiresolution-based watersheds for efficient image segmentation”, Pattern Recognition Letters, 24, pp. 473-488, 2003.

7. J. A. Rodrìguez, C. Urdiales, A. Bandera, F. Sandoval, “A multiresolution spatiotemporal motion segmentation technique for video sequences based on pyramidal structures”, Pattern Recognition Letters, 23, pp. 1761-1769, 2002.
8. D. Prewer, L. Kitchen, “Soft image segmentation by weighted linked pyramid”, Pattern Recognition Letters, n. 22, pp. 123-132, 2001.
9. A. Rosenfeld, C. Sher, “Detecting image primitives using feature pyramids”, Journal of Information Sciences, 107, pp.127-147, 1998.
10. P.F.M. Nacken, “Image segmentation by connectivity preserving relinking in hierarchical graph structures”, Pattern Recognition, vol. 28, no. 6, pp. 907-920; 1995
11. M. Bister, J. Cornelis, and A. Rosenfeld, A critical view of pyramid segmentation algorithms, Pattern Recognition Letters, vol. 11, no. 9, 605-617; 1990.
12. P. J. Burt, “The Pyramid as a Structure for Efficient Computation”, in A. Rosenfeld, (ed.) “Multiresolution Image Processing and Analysis”, Springer-Verlag, Berlin; 1984.
13. H. Greenspan, S. Belongie, R. Goodman, P.Perona, S. Rakshit and C. H. Anderson, Overcomplete steerable pyramid filters and rotation invariance, Proceeding IEEE Computer Vision and Pattern Recognition, Seattle, Washington, pp. 222-228, 1994.
14. X. Kong and J. Goutsias, A study of pyramidal techniques for image representation and compression, Journal of Visual Communication and Image Representation , vol. 5, 190-203; 1994.
15. Wei Chen, S. T. Acton, Morphological pyramids for multiscale edge detection, Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 137-141; 1998.
16. F. Arman, J. A. Pearce, Unsupervised classification of cell images using pyramid node linking, IEEE Trans Biomedical Engineering, vol. 37, n. 6, pp. 647-50; 1990.
17. A. Montanvert, P. Meer and A. Rosenfeld, “Hierarchical Image Analysis Using Irregular Tessellations”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, n. 4, pp. 307-316; 1991.
18. G. Borgefors, G. Ramella, G.Sanniti di Baja, "Shape and topology preserving multi-valued image pyramids for multi-resolution skeletonization", Pattern Recognition Letters, vol 22, pp. 741-751, 2001.

Pulling, Pushing, and Grouping for Image Segmentation

Guoping Qiu[1] and Kin-Man Lam[2]

[1] School of Computer Science, The University of Nottingham
qiu@cs.nott.ac.uk
[2] Dept. of Electronic and Information Eng. The Hong Kong Polytechnic University
enkmlam@polyu.edu.hk

Abstract. This paper presents a novel computational visual grouping method, termed pulling, pushing and grouping, or PPG for short. Visual grouping is formulated as a functional optimisation process. Our computational function has three terms, the first pulls similar visual cues together, the second pushes different visual cues apart, and the third groups spatially adjacent visual cues without regarding their visual properties. An efficient numerical algorithm based on the Hopfield neural model is developed for solving the optimisation process. Experimental results on various intensity, colour and texture images demonstrate the effectiveness of the new method.

1 Introduction

Visual grouping is thought to be a crucial process in both human and computer vision systems. The volume of computer vision/image processing literature relating to the topic, often under different names, such as image segmentation, figure ground separation, and perceptual organization, is huge, and recent examples include [1-4]. Although a large amount of effort has been put into the study of the topic, many computational issues remain very difficult. In this paper, we present a new computational approach to visual grouping.

We consider visual grouping as a combined process of pulling, pushing and grouping, or PPG for short. Two visual cues that belong to the same visual group will pull each other together to become members of the same group. Conversely, two visual cues that belong to different visual groups will push each other apart to become members of the different groups. Simultaneously, local neighboring visual cues, regardless of their visual properties, will group together to become members of the same group. Similar to many other approaches in the literature, the PPG process is formulated as a functional optimization process. A PPG computational function, consisting of three terms, namely, the pulling, the pushing and the grouping terms, is constructed in such a way that the minimum of the function will correspond to a good solution to the visual grouping problem. An efficient numerical solution based on a neural computing model is developed to solve the optimization problem.

The organization of the paper is as follows. Section 2 presents the new PPG visual grouping method. Section 3 presents experimental results of the application of PPG to the segmentation of intensity, colour, and texture images. Section 4 very briefly discusses related prior work in the literature. Section 5 concludes the presentation.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 65–73, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Perceptual Organization Based on the Pulling, Pushing, and Grouping (PPG) Principle

Perceptual grouping is the process of organising visual elements into groups that manifest certain regularities. The Gestalt psychologists have suggested that visual grouping should follow some basic laws, including proximity, similarity, good continuation, amongst many others. To formulate computational models for the perceptual grouping process, a common practice is to optimise some cost functions which measure the quality of the grouping. Although the Gestalt grouping laws are intuitive, casting them into a computational function proves to be difficult. A common approach is to capture local interaction of visual cues in a global optimisation framework. Once a cost function is formulated, computing an optimal solution is again very challenging. In this section, we propose a pushing, pulling and grouping principle to formulate a perceptual grouping computational energy and develop a neural network based optimisation algorithm.

2.1 The PPG Computational Energy Function

We first consider the case where the image is to be partitioned into two groups (figure and ground). Let V_i, $i = 1, 2, ...N$, be the binary variables, $V_i = 0$ represents pixel i belong to group 0 and $V_i = 1$ represents pixel i belongs to group 1. If two pixels i and j belongs to the same group, we define a pulling energy, $A(i, j)$, between the pixels. If two pixels i and j belong to different groups, we define a pushing energy, $R(i, j)$, between the pixels. Then the PPG computational function is defined as

$$\begin{aligned} E = \lambda_1 \Bigg(& \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} V_i V_j A(i,j) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (1-V_i)(1-V_j) A(i,j) + \\ & \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} V_i (1-V_j) R(i,j) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (1-V_i) V_j R(i,j) \Bigg) + \\ & \lambda_2 \left(- \sum_{i=1}^{N} \sum_{j\in n_i, j\neq i} (2V_i - 1)(2V_j - 1) \right) \end{aligned} \tag{1}$$

where n_i is a local neighbourhood of i. λ_1, λ_2 are non-negative weighting constants. The values of V_i, $i = 1, 2, ...N$, that correspond to the minimum of E represent an effective partition of the image into two perceptual groups.

At a first glance, it may seem strange to write (1) in such a redundant way. As will become clear immediately, the current form of (1) helps an intuitive explanation of the meanings of each term. The first term measures the cost of putting two pixels into the same group (pulling). The second term measures the cost of putting two pixels into different groups (pushing). The third term measures the cost of putting a pixel and its neighbours into the same group regardless of the values of the original pixels (grouping). The relative importance of these terms is determined by their respective weighting constants. However, the importance of the first and second terms is the same, so the same weighting constant is used.

If two pixels have similar visual properties, such as similar intensity, similar colour, similar texture or similar edgels, it is likely that they belong to the same group. If such pixels are close to each other in the spatial domain, this likelihood (that they belong to the same group) will increase. On the other hand, if two pixels have dissimilar visual properties, different intensities, different colours, different textures or different edgels, it is likely that they will belong to different groups. If they are far apart in the spatial domain, this likelihood (that they belong to different groups) will increase. Based on these reasoning and the form of (1), we can now define the form of the pulling energy function $A(i, j)$ and the pushing energy function $R(i, j)$.

When two pixels have very similar visual cues or are very close to each other spatially, $A(i, j)$ should be small and $R(i, j)$ should be large such that the first term in (1) is favoured, effectively forcing a solution that V_i and V_j will have the same value, i.e., pulling pixels i and j into the same group. Conversely, when two pixels have very different visual cues and far apart spatially, $A(i, j)$ should be large and $R(i, j)$ should be small such that the second term in (1) is favoured, effectively forcing a solution that V_i and V_j will have different values, i.e., pushing pixels i and j to different groups. Let $F(i)$, $i = 1, ...N$, be the visual cue of pixel i, one possible definition of the pulling and pushing energy functions is (2) (other similar definitions are possible)

$$A_S(i,j)=\begin{cases}\dfrac{|i-j|}{\sigma_S} & \text{if } |i-j|\le\sigma_S\\ 0 & \text{Otherwise}\end{cases} \quad R_S(i,j)=\begin{cases}\dfrac{\sigma_S-|i-j|}{\sigma_S} & \text{if } |i-j|\le\sigma_S\\ 0 & \text{Otherwise}\end{cases} \tag{2}$$

$$A_F(i,j)=\exp\left(-\frac{\|F(i)-F(j)\|}{\sigma_F}\right) \quad R_F(i,j)=1-\exp\left(-\frac{\|F(i)-F(j)\|}{\sigma_F}\right)$$

$$A(i,j)=A_S(i,j)A_F(i,j) \qquad R(i,j)=R_S(i,j)R_F(i,j)$$

Fig.2 shows the shape of the visual components and the spatial components of the pulling and pushing energy. The shape is controlled by the free parameters σ_F and σ_S, which determine when the pulling power outweighs the pushing power for the pair of pixels, or vice versus. The pulling and pushing interactions amongst pixels are confined to local regions. The rationale is that when two pixels are too far apart, they will not affect one another.

2.2 A Numerical Algorithm

The Hopfield neural network models [5 -7] have been successfully used in solving optimization problems. Because a neural network is a naturally parallel structure, it has the potential for solving many complex optimization problems much more rapidly than other sophisticated and computationally intensive optimization techniques. Neurons can be highly and selectively interconnected so as to give rise to collective computational properties and create networks with good computational efficiency. Collective computational properties emerge from the existence of an energy function of the states of the neuron outputs, namely the computational energy. The computational energy function of the Hopfield network has the following form

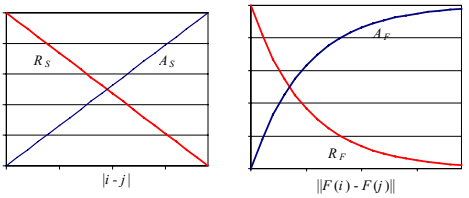


Fig. 1. The spatial and visual components of the pulling and pushing energy function.

$$E_H = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} V_i T_{ij} V_j - \sum_{i=1}^{N} I_i V_i \tag{3}$$

where V_i, i = 1, 2, ...N, is the outputs of the network, T_{ij} is the connection strength between neuron i and neuron j, I_i, i =1, 2, ...N, is the external input to neuron i. Hopfield has shown that if the connection matrix is symmetrical with zero diagonal elements (i.e., $T_{ij} = T_{ji}$, and $T_{ii} = 0$), the function of the neuron output in (4) is decreased by any state changes produced by the following asynchronous rule:

$$V_i = \begin{cases} 1 & \text{if } H_i \geq 0 \\ 0 & \text{if } H_i < 0 \end{cases} \quad \text{where } H_i = I_i + \sum_{j=1, j\neq i}^{N} T_{ij} V_j \tag{4}$$

For a given initial state, the network updates its output V_i according to equation (4) asynchronously, that is one neuron output at a time. When the network reaches its stable state in which further iterations do not change the output of V_i, the network outputs correspond to a local minimum of the computational energy function of the form (3). It follows that the interconnection strength between neurons, and the external inputs to the neurons, have the following relationship with the computational energy function:

$$T_{ij} = \frac{\partial^2 E_H}{\partial V_i \partial V_j} \qquad I_i = \left.\frac{\partial E_H}{\partial V_i}\right|_{V_j=0, \forall j} \tag{5}$$

Clearly, to use the collective computational properties of the Hopfield network to solve a problem, a cost function (computational energy) has to be defined in such a way that the lowest value represents an effective solution to the problem. For our current application, we can use the Hopfield neural model to optimize the PPG energy (1).

To find the optimum solution to (1), we find it numerically convenient to separate the visual feature dependent part and the visual feature independent part, and minimize each part in turn. We re-write (1) as

$$E_{PP} = \lambda_1 \left(\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} V_i V_j A(i,j) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (1-V_i)(1-V_j) A(i,j) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} V_i (1-V_j) R(i,j) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (1-V_i) V_j R(i,j) \right) \tag{6}$$

$$E_G = \lambda_2 \left(-\sum_{i=1}^{N} \sum_{j\in n_i, j\neq i} (2V_i - 1)(2V_j - 1) \right)$$

We can construct one Hopfield network to optimize E_{PP} and another Hopfield network to optimize E_G. For the E_{PP} network, we have the connection strengths and external inputs as

$$\begin{aligned} T_{ij}^{PP} &= -4\lambda_1 (A(i,j) - R(i,j)), j \neq i \\ I_i^{PP} &= 2\lambda_1 \sum_{j=1, j\neq i}^{N} (A(i,j) - R(i,j)) \end{aligned} \tag{7}$$

and for the E_G network, we have the connection strengths and external inputs as

$$T_{ij}^{G} = 4\lambda_2, \; j \neq i, j \in n_i \qquad I_i^{G} = -2(|n_i| - 1)\lambda_2 \tag{8}$$

Then the PPG visual grouping algorithm can be summarized:

The PPG Visual Grouping Algorithm

Step 1, Initialize V_i, $i = 1, 2, ...N$, to random values in the range (0, 1)
Step 2, For all $i = 1, 2, ...N$, update V_i, one at a time according to

$$V_i = \begin{cases} 1 & \text{if } H_i^{PP} \geq 0 \\ 0 & \text{if } H_i^{PP} < 0 \end{cases} \quad \text{where } H_i^{PP} = I_i^{PP} + \sum_{j=1, j\neq i}^{N} T_{ij}^{PP} V_j$$

Step 3, For all $i = 1, 2, ...N$, update V_i, one at a time according to

$$V_i = \begin{cases} 1 & \text{if } H_i^{G} \geq 0 \\ 0 & \text{if } H_i^{G} < 0 \end{cases} \quad \text{where } H_i^{G} = I_i^{G} + \sum_{j=1, j\neq i}^{N} T_{ij}^{G} V_j$$

Step 4, go to Step 2, until converge.

From (8), it is not difficult to see that Step 3 performs majority voting. A pixel will be assigned to the membership of that of the majority of its neighbours. It is also easy to see that the weighting parameter λ_2 for the grouping term becomes irrelevant in the decision making. The complexity of this Step is $O(mNn_i)$ binary operations, where N is the number of pixels, n_i the size of the local window and m the number of iterations. To perform Step 2, the connections and external inputs in (7) need only to be done once with a complexity of $O(N\sigma_S)$, where N is the number of pixels and σ_S is size of the local region within which the pulling and pushing are active. The complexity of Step 2 after computing the T_{ij} and I_i is $O(mN\sigma_S)$ addition operations. Again, the weighting parameter λ_1 becomes irrelevant in the decision making. We have observed that $n_i = 3 \sim 5$, $\sigma_S = 7 \sim 15$, $m = 5 \sim 10$ worked well in our experiments. An important

parameter in the algorithm is σ_F in (2). Typically, this should be set based on the statistics of the visual features used in the interaction. We found setting it according to the variance of the feature vectors worked well.

If it is required that the image be partitioned into more than two groups, the algorithm can be applied recursively to the resultant groups, until the desired number of groupings are generated.

3 Application to Scene Segmentation and Results

We have implemented the PPG algorithm based on intensity, colour and texture features. For the intensity images, the visual feature $F(i)$ in (2) takes the intensity of the pixel. For the colour image, we used the HSV colour space and each pixel was represented by a 3-d vector formulated as $F(i) = (v(i)_* s(i)_* \sin(h(i)), v(i)_* s(i)_* \cos(h(i)), v(i))$, where $v(i)$, $s(i)$, $h(i)$, are the Value, Saturation and Hue of pixel i. For texture images, we use Law's filtering approach to extract the texture features. The filter masks were t_1 = [1, 4, 6, 4, 1], t_2 = [-1, -2, 0, 4, 1], t_3 = [-1, 0, 2, 0, -1], t_4 = [-1, 2, 0, -2, 1], and t_5 = [1, -4, 6, -4, 1]. The resulting 25 filter outputs for each pixel position were used to form a 25-d feature vector for each pixel [4]. In each case, the variance of the feature, var_F, was calculated as

$$\mathrm{var}_F = \frac{1}{N}\sum_{i=1}^{N}\|F(i) - M_F\|, \qquad M_F = \frac{1}{N}\sum_{i=1}^{N}F(i) \tag{9}$$

The pulling and pushing energy parameter in (2) is set as $\sigma_F = \alpha \mathrm{var}_F$, where α is a constant, which is tuned manually. We found setting $0 < \alpha \leq 1$ worked well.

Fig. 2 shows results of grouping grey scale intensity images. It is seen that the results are very good. The two salient groups of visual pattern have been successfully separated.

Fig. 3 shows results of grouping images using colour features. It is again seen that the two salient groups of visual pattern have been successfully separated.

Fig. 4 shows results of grouping images using texture features. It is again seen that the two salient groups of visual pattern have been successfully separated. Notice that in the grouping process, only texture features extracted from the achromatic channel were used and colour information was not used. The images are shown in colour for visually more appealing presentation.

4 Related Work

The volume of literature in the area of image segmentation is huge, so we do not intend to list all possibly related prior work but merely to point out some recent work we believe are most related to our current work. We believe the present work is most closely related to recent work on graph based frameworks for image segmentation [1-3]. Whilst [1,3] proposed eigenvector or graph spectral theory based solutions, [2] used methods based on maximum likelihood graph clustering. Our method also can

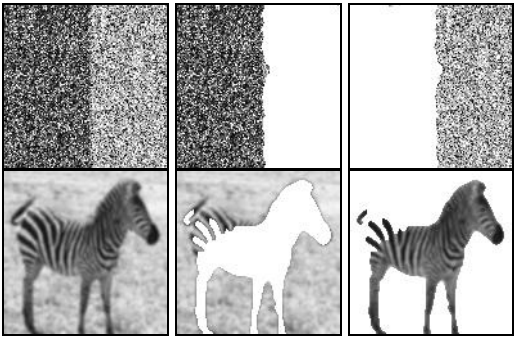


Fig. 2. Results of PPG on intensity images. Images in the 1st column are the original images. The two visual groups are shown in the 2nd and 3rd columns. All results are for parameters σ_s = 5, n_i =2, α =1 and after 5 iterations.

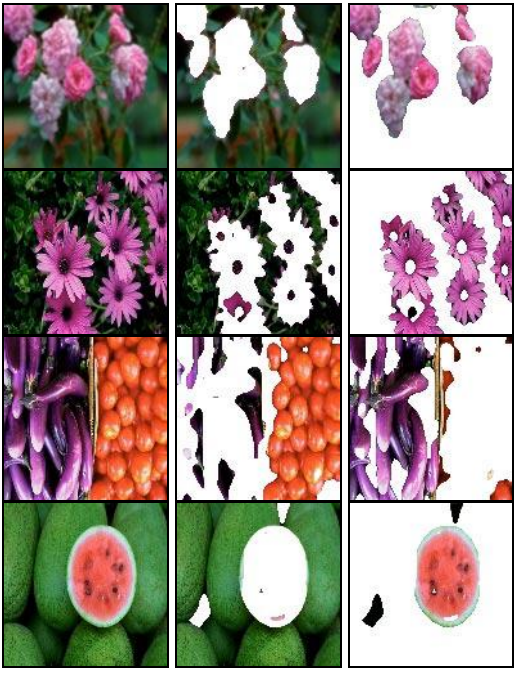


Fig. 3. Results of PPG on colour images. Images in the 1st column are the original images. The two visual groups are shown in the 2nd and 3rd columns. All results are for parameters σ_s = 11, n_i =2, α = 0.5 and after 15 iterations.

be regarded as representing the segmentation or grouping using a graph. However, the weights of the graph edges in our current work are defined differently, which not only measure the similarity between the nodes but also explicitly measure the dissimilarity of the nodes, i.e., the nodes not only pulling each other together, they also

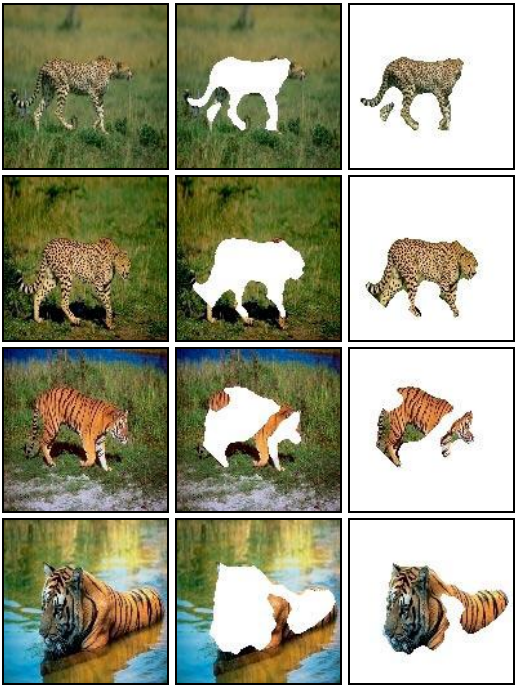


Fig. 4. Results of PPG based on texture features (notice that only achromatic signals were used in the grouping). Images in the 1st column are the original images. The two visual groups are shown in the 2nd and 3rd columns. All results are for parameters $\sigma_s = 11$, $n_i = 2$, $\alpha = 0.5$ and after 15 iterations.

pushing one another apart. We have formulated the energy function in such a way that it can be numerically solved by a Hopfield neural network model. Our solution can be considered very efficient numerically because only addition and thresholding operations are involved. We are investigating whether there are deeper relations between our current algorithm and those in the literature.

5 Concluding Remarks

We have presented a novel algorithm for visual grouping. Our pulling, pushing and grouping principle was proposed based on the rationale that, (i) if two pixels have similar visual properties, it is likely that they belong to the same group and this likelihood increases if they are close to each other spatially as well, on the other hand, (ii) if two pixels have dissimilar visual properties, it is likely that they will belong to different groups and this likelihood will increase if they are far away from each other spatially, and (ii) spatially close pixels are more likely belong to the same group regardless of their photometric values. We then cast this principle in a computational energy function and developed a neural network based solution. We have presented

experimental results to show that the new algorithms work very effectively and give very encouraging results.

As is the same with many other algorithms, there are several free parameters in our algorithm that needed to be decided by empirical means. We have provided some guidelines to select them and fortunately, we found that it was not difficult to find their values that would work well on a variety of data. We also found the algorithm converged very fast. However, one possible problem is that the algorithm may converge to a local minimum, and in some cases, the algorithm may perform unbalanced grouping resulting in one group being too large the other too small. Work is ongoing to test the algorithm using a variety of data to gain deeper understanding of its behaviour.

References

[1] J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Trans PAMI, vol. 22, no. 8, pp. 888 - 905, 2000

[2] A. Amir and M. Lindenbaum, "A generic grouping algorithm and its quantitative analysis", IEEE Trans PAMI, vol. 20, no. 2, pp. 168-185,1998

[3] Y. Weiss, "Segmentation using eigenvectors: a unifying view", ICCV 1999

[4] T. Randen and J. Husøy, " Filtering for Texture Classification: A Comparative Study", IEEE Trans PAMI, Vol 21, No. 4, pp. 291 - 310, 1999

[5] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proc. Natl. Acad. Sci., pp. 2554 - 2558, April 1982,

[6] J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons", Proc. Natl. Acad. Sci., pp. 3088 - 3092, April 1984,

[7] J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems", Biol. Cyern., pp. 141 - 152, 1985

Image Segmentation by a Robust Clustering Algorithm Using Gaussian Estimator

Lei Wang, Hongbing Ji, and Xinbo Gao

Lab. 202 of School of Electronic Engineering, Xidian University,
710071 Xi' an, China
leiwang@mail.xidian.edu.cn

Abstract. This paper presents a novel clustering-based image segmentation method, which incorporates the features of robust statistics. To overcome the sensitivity to noise and outliers in fuzzy clustering, a simple but efficient M-estimator, Gaussian estimator, has been introduced to clustering analysis as weight or membership function. When applied to image analysis, the proposed Robust Gaussian Clustering/Segmentation (RGCS) algorithm exhibits more reasonable pixel classification and noise suppression performance with respect to FCM. Moreover, by using a uniform resolution parameter scheme, this method avoids producing coincident clusters, which occurs often in possibilistic clustering segmentation. Experiments on both the synthetic data and real image demonstrate the validity and power of the proposed algorithm.

1 Introduction

Image segmentation is a fundamental component in many computer vision applications, and can be addressed as a clustering problem [1]. Many fuzzy clustering methods including Fuzzy C-means (FCM) algorithm have been successfully applied to image segmentation [2]. And they provide so-called soft segmentations of images, which allow regions or classes to overlap. However, the uniform probabilistic constraint on the membership of each feature point makes FCM be sensitive to noise and outliers, thus leading to obvious noise effect and unsatisfying margin description in segmented images.

Possibilistic C means (PCM) algorithm was proposed in [3] to address above drawbacks. It leads to higher noise immunity with respect to FCM, by relaxing above probabilistic constraint. Unfortunately, this interesting method has been reported to have a difficulty when applied to image segmentation, as it has undesirable tendency to produce coincident clusters even by starting from good initial partitions [4]. The risk of failure to recognize the whole structure underlying the data set restricts the further use of such method.

After noting the difficulties associated with FCM for locating centers in noisy data set, we introduce Gaussian estimator into clustering analysis and present a Robust Gaussian Clustering/Segmentation (RGCS) algorithm. This method exhibits high

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 74–81, 2004.
© Springer-Verlag Berlin Heidelberg 2004

immunity to noise and outliers in data sets and gives more reasonable results in image segmentation to reduce noise effect and margin blur, compared to FCM. Besides, the investigation on iterative formula of cluster center indicates a simple but efficient solution to the problem of coincident clusters in image segmentation by unconstrained clustering methods including PCM. Segmentation tests on the classical Lena image also verify the effeteness of such scheme.

2 Background

2.1 Fuzzy C-means Clustering

Fuzzy C-means (FCM) algorithm [2] is formulated to minimize the following objective function with respect to the membership function u_{ij} and the centroids C_i:

$$J_{FCM} = \sum_{j=1}^{N}\sum_{i=1}^{C}(u_{ij})^m d_{ij}^2 \tag{1}$$

where C is the number of clusters or classes, N is the total number of feature points or vectors., and $m \in (1,\infty)$ is a weighting exponent. u_{ij} denotes the grade that input vector x_j belongs to fuzzy cluster i, and should satisfy

$$0 \le u_{ij} \le 1 \qquad \forall i, j \tag{2a}$$

$$\sum_{i=1}^{C} u_{ij} = 1 \qquad \forall j \tag{2b}$$

$$0 < \sum_{j=1}^{N} u_{ij} < N \quad \forall i \tag{2c}$$

For each input vector x_j, using Lagrangian multiplier method, for $m > 1$, local minimum solutions of Eq. (1) was demonstrated if and only if

$$C_i = \frac{\sum_{j=1}^{N}(u_{ij})^m x_j}{\sum_{j=1}^{N}(u_{ij})^m} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}(d_{ij}/d_{kj})^{\frac{2}{m-1}}} \tag{4}$$

It can be seen from above formulation that FCM algorithm has considerable trouble of high sensitivity to noise and outliers. Since the memberships generated by the constraint (2b) are relative numbers, the noisy or outlying points may be also assigned relatively high membership values, thus affecting significantly the center estimation and pattern classification. It does not give a robust estimation of centers and appropriate membership assignment of noise and outliers in overlapping or contaminated data set. This may partially explain the noise effect and margin blur problem in segmented images, especially when the major peaks in histogram are very close.

2.2 Possibilistic C-means Clustering

To overcome the sensitivity to the noise and outliers, possibilistic C-means (PCM) algorithm was proposed in [3], which seeks a solution that minimize the following objective function with constraint (2b) being removed,

$$J_{PCM} = \sum_{i=1}^{C}\sum_{j=1}^{N}(u_{ij})^m d_{ij}^2 + \sum_{i=1}^{C}\eta_i \sum_{j=1}^{N}(1-u_{ij})^m \tag{5}$$

The membership value u_{ij} represents the typicality of the *j*th point in the *i*th class or the possibility of the *j*th point belonging to the *i*th class, and is given by

$$u_{ij} = \frac{1}{1+\left(d_{ik}/\eta_i\right)^{1/m-1}} \tag{6}$$

Here η_i is a measure of the radius of the *i*th cluster and is called the "bandwidth parameter" of the *i*th cluster. It may be estimated successively using the following rules:

$$\eta_i = K\frac{\sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m} \tag{7}$$

$$\eta_i = \frac{\sum_{x_j \in (\Pi_i)_\alpha} d_{ij}^2}{\left|(\Pi_i)_\alpha\right|} \tag{8}$$

Some unfavorable attributes of PCM that were reported in [4] include failure to recognize the overall structure of the data and generation of coincident clusters, especially in the applications to image segmentation.

3 The Proposed Algorithm

3.1 Robust Statistics Features

Huber [5] characterizes a robust procedure by the following features:

(1) it should have a reasonably good efficiency (accuracy) at the assumed model;
(2) small deviations from the model assumptions should impair the performance only by a small amount; and
(3) large deviations from the model assumptions should not cause a catastrophe.

The first requirement means that, when the data is "clean" and follows assumed model, a robust procedure must yield accurate estimates, whereas the third property is related to the concept of breakdown point [6] [7]. However, we should pay much attention to the second requirement in engineering practice, i.e., small deviations should not have significant negative effect on the performance, since such situations arise often in practice as noise effect and approximate modeling.

3.2 Robust Clustering Based on M-estimator

Dave investigated the relation between robust statistics and fuzzy clustering and summarized the following objective function for M-estimator based clustering [6],

$$J(\theta) = \sum_{j=1}^{N} \rho(r_j) \tag{9}$$

where θ is a parameter vector to be estimated, N is the number of observations, and r_j is defined as residual, i.e.,

$$r_j = x_j - \theta \tag{10}$$

If we let $\psi(r_j)$ represents the derivative of $\rho(r)$, then the estimate of satisfies the implicit equation,

$$\sum_{j=1}^{N} \psi(r_j) \frac{dr_j}{d\theta} = 0 \tag{11}$$

The problem defined by (9) or (11) is called the M- estimator. A general way to solve this problem is to reformulate the M-estimator and obtain the W-estimator [8] [9], as explained below.

Let $w(r)$ be defined as $w(r) = \psi(r)/r$. Substituting for $\psi(r_j)$ in (11) for location estimation, we obtain

$$\sum_{j=1}^{N} (x_j - \theta) w(x_j - \theta) = 0 \tag{12}$$

Rearranging, we get

$$\theta = \frac{\sum_{j=1}^{N} w(x_j - \theta) x_j}{\sum_{j=1}^{N} w(x_j - \theta)} \tag{13}$$

where $w(x_j - \theta)$ plays the role of the robust weight or possibilistic membership of a feature vector in the good class as opposed to the noise class. One may select different forms for $\psi(r)$ or $w(r)$ to achieve robust M- estimate of a particular parameter θ. For details of M-estimator based clustering, one may refer to [6].

3.3 Robust Gaussian Clustering/Segmentation

In clustering analysis, one needs to estimate C cluster centers, which may be implemented by a collection of C independent M-estimators working simultaneously. The key issue is how to choose an appropriate M-estimator to achieve center estimation with above robust statistics features. Here we introduce a simple but efficient M-estimator, i.e., Gaussian estimator, into clustering analysis. Its robust weight function or possibilistic membership is capable of reducing the influence of large residuals on the estimated fit and can be written as

$$w_{ij} = \exp\{-\beta \bullet d^2(x_j, c_i)\} \tag{14}$$

where $d(x_j, c_i)$ represents the distance between feature vector x_j and class or cluster c_i, and β is termed as *resolution parameter*. A greater resolution β indicates that w_{ij} has the tendency to descend fast, ensuring fewer outliers being included in each cluster, and the boundary shrinks for each compact cluster.

Substituting cluster center c_i for parameter θ in (13), we get

$$c_i = \frac{\sum_{j=1}^{N} w_{ij} x_j}{\sum_{j=1}^{N} w_{ij}} \tag{15}$$

Eq. (14) and (15) can be performed alternately to estimate each cluster center, thus forming a Robust Gaussian Clustering/Segmentation (RGCS) algorithm. Dave had suggested that in [6], with a particular objective function being used, PCM can also deduce such weight function as in (14). However, that objective function is designed passively, which lacks clear physical meaning and well-defined mathematical features, compared to the proposed one.

3.4 Choice of Resolution Parameter

In a basic data-clustering problem, an adaptive choice of resolution parameters can be used if these clusters vary in size. The following scheme can be used to detect clusters with different sizes in Multi-Resolution Robust Gaussian Clustering (MRRGC) algorithm,

$$\beta_i = K \times \frac{1}{N} \sum_{j=1}^{N} \frac{1}{d_{ij}^2} \tag{16}$$

where *K* is a constant to adjust the *resolution parameter*. A greater *K* is preferred to separate close clusters with a higher resolution, when some clusters overlap.

However, in pixel-intensity-based image segmentation, the actual objective function should be reformulated as

$$J(\theta) = \sum_{j=1}^{N} h_j \rho(r_j) \tag{17}$$

where h_j denotes the number of pixels whose densities are x_j in the 1-D histogram. Thus the resulting center estimate can be rewritten as

$$c_i = \frac{\sum_{j=1}^{N} h_j w_{ij} x_j}{\sum_{j=1}^{N} h_j w_{ij}} \tag{18}$$

It can be seen that histogram weight is an unconstrained weight and represents the absolute density, which plays a dominant role in the center estimate. On the other hand, cluster centers are often located in dense regions, the points round which usually have both big amplitudes h_j in the histogram and high robust weight w_{ij}, with

great consistency. So if (16) is used, the adjacent two centers will converge to the same dense region or middle place between clusters, thus resulting in coincident clusters like in PCM based segmentation. We suggest using uniform resolution parameter scheme to avoid over-weight in image segmentation. Besides, a great β is preferred to separate overlapping clusters, whose peaks are very close in histogram.

4 Results and Discussion

Let us begin with a synthetic data set, which contains two synthetic spherical Gaussian clusters as shown in Fig. 1(a). FCM, RGCS, MRRGC are tested separately. Parameter m is set to 2 in FCM, $\beta = 1.5$ in RGCS, and $K = 1.5$ in MRRGC. The circles enclose the points whose membership values are greater than 0.01.

It can be seen that the estimated centers deviate from the true positions by FCM, due to the effect of noise and outliers on estimation. However, RGCS and MRRGC can obtain robust estimation of cluster centers, with Gaussian estimator being used to reduce the effect of outlying points. Besides, by using the adaptive resolution scheme, MRRGC has more reasonable description of cluster volume than RGCS in general data clustering.

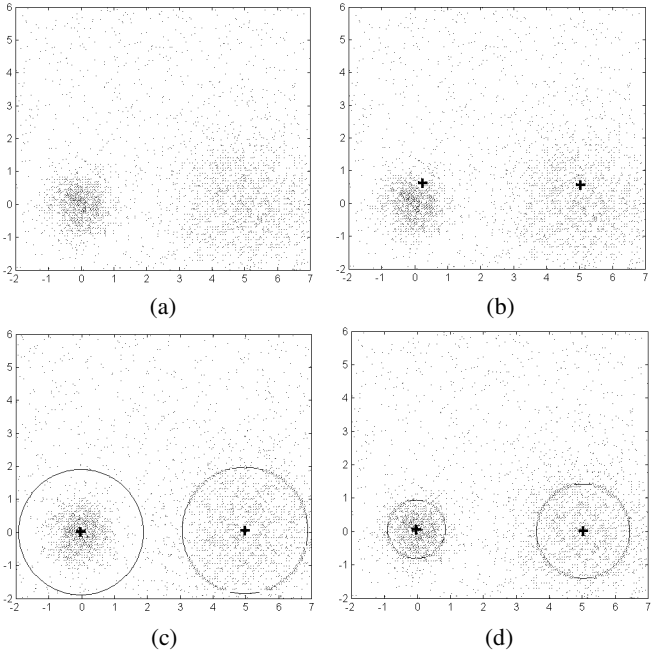


Fig. 1. Clustering results on a synthetic data set, (a) original data set; (b) result by FCM; (c) result by RGCS; (d) result by MRRGC.

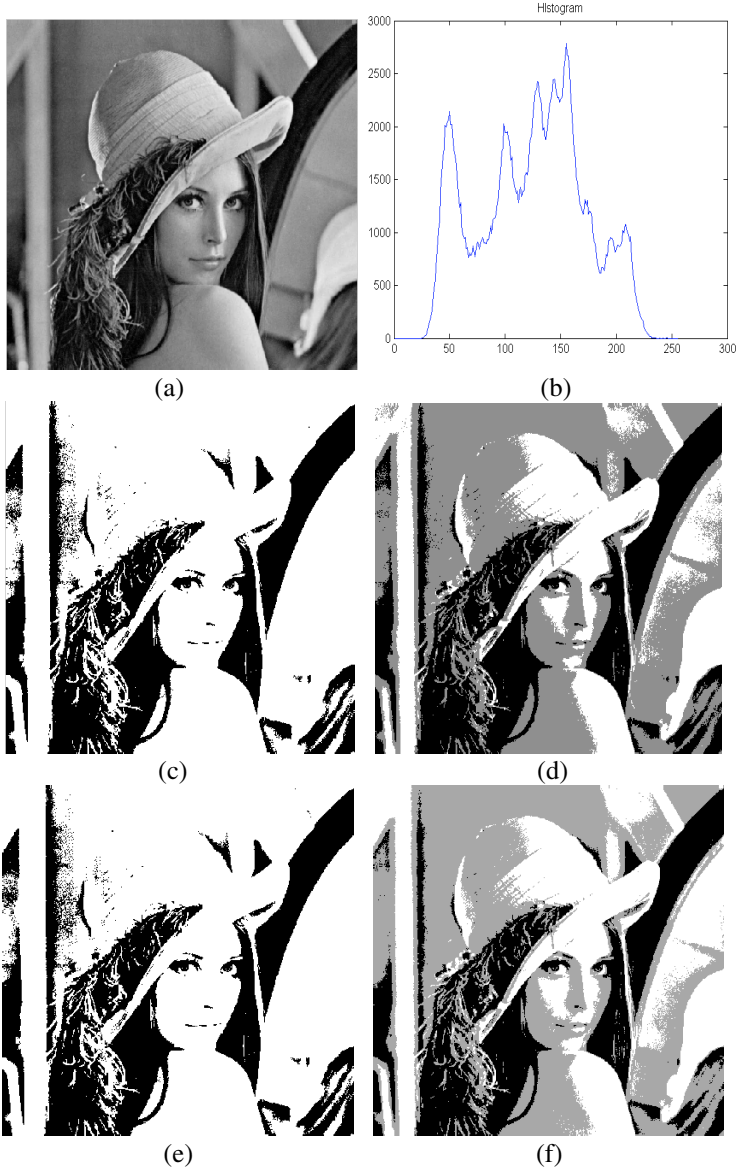


Fig. 2. Segmentation tests on the classic Lena image, (a) Lena image; (b) histogram of Lena image; (c) result by PCM; (d) result by FCM; (e) result by RGCS with $\beta = 0.2$; (f) result by RGCS with $\beta = 1.5$.

The second test is performed on a real data set, i.e., Lena image. With class being set to 3, PCM only finds two clusters and fails to recognize the whole structure. Such coincident clusters are resulted from over-weight by the unequal resolution (bandwidth) scheme. Though FCM finds three clusters, the segmented image exhibits obvi-

ous noise effect and margin blur problems, especially at the top right corner of the image. With a high resolution, the proposed algorithm obtains the best segmentation result, as shown in (f). Fig. 2(e) shows a low value of resolution for RGCS might also produced coincident clusters like that by PCM, since the clusters overlap each other and are difficult to separate, as shown in the histogram (b).

5 Conclusion Remarks

In this paper, we present a new clustering-based image segmentation method, which incorporates the features of robust statistics. Gaussian estimator is introduced into clustering analysis to achieve robust estimation of cluster centers and reduce the influence of large residuals on the estimated fit. Compared with FCM, the proposed algorithm exhibits more reasonable pixel classification and noise suppression performance.

We have proposed two choices of resolution parameter. With an adaptive scheme, MRRGC is capable of detecting the clusters with different sizes in a general data-clustering problem. However, in image segmentation, a uniform resolution scheme is suggested to avoid producing coincident clusters, which usually arises in PCM based segmentation. Such scheme can also be extended to other possibility or M-estimator based image segmentation. Besides, it is preferred to use a high resolution value in RGCS to separate close or overlapping clusters, which holds in most image segmentation practice.

References

1. A. Rosenfeld and A.C.Kak, Digital Picture processing. 2nd ed. Academic Press, Inc., New York, 1982.
2. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
3. R. Krishnapuram, and J. Keller, “A Possibilistic Approach to Clustering,” IEEE Trans. Fuzzy Systems, vol. 1, no. 2, pp. 98-110, May 1993.
4. M.Barni, et.al., Comments on “A Possibilistic Approach to Clustering”, IEEE Trans. Fuzzy Systems, vol. 4, no. 3, pp. 393-396, Aug. 1996.
5. P. J. Huber, Robust Statistics. New York: Wiley, 1981.
6. Davé, R.N. and Krishnapuram, R., Robust Clustering Methods: A Uniform View, IEEE Transactions on Fuzzy Systems, pp. 270-293, no 5,1997.
7. F. R. Hampel, E. M. Ponchotti, P. J. Rousseeuw, and W. A. Stahel, Robust Statistics: The Approach Based on Influence Functions. New York: Wiley, 1986.
8. C. Goodall, “M-estimator of location: An outline of the theory,” in Understanding Robust and Exploratory Data Analysis, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds. New York: 1983, pp. 339–403.
9. P. W. Holland and R. E. Welsch, “Robust regression using iteratively reweighted least squares,” Communication Statistics—Theory and Methods, vol. A6, no. 9, pp. 813–827, 1977.

A Multistage Image Segmentation and Denoising Method – Based on the Mumford and Shah Variational Approach

Song Gao and Tien D. Bui

Department of Computer Science, Concordia University, 1455 de Maisonneuve Blvd. W., Montreal, QC, H3G 1M8, Canada
{sgao, bui}@cs.concordia.ca

Abstract. A new multistage segmentation and smoothing method based on the active contour model and the level set numerical techniques is presented in this paper. Instead of simultaneous segmentation and smoothing as in [10], [11], the proposed method separates the segmentation and smoothing processes. We use the piecewise constant approximation for segmentation and the diffusion equation for denoising, therefore the new method speeds up the segmentation process significantly, and it can remove noise and protect edges for images with very large amount of noise. The effects of the model parameter ν are also systematically studied in this paper.

1 Introduction

Image segmentation and smoothing are two fundamental problems in computer vision and image processing. Variational methods have been applied successfully to image segmentation problems [1], [2]. The basic idea of variational methods is to model images through an energy functional then minimizes the functional to obtain the desired segmentation and smoothing results. There are two basic factors in most variational methods which are energy functional formulations and energy minimization. The energy functional was first introduced as a discrete energy by Geman and Geman [5] and later transposed by Mumford and Shah into a continuous domain [7]. Image segmentation and smoothing problem actually is an optimization problem for a given energy functional. To obtain the segmentations is equivalent to minimizing the energy functional.

Many energy minimization approaches have been proposed over the past decade that can be classified into two categories which are stochastic and deterministic. In stochastic approaches, a discrete energy functional is usually used. The simulated annealing [5] or mean field annealing techniques [9] were often employed to obtain the optimization solutions. Deterministic approaches are often the continuous version of the energy, and the Mumford-Shah energy is one of the most extensively studied models in this class. Some active contour models based on the Mumford-Shah energy have been proposed for solving the problems of image segmentation and smoothing [3], [11], [10]. Various advantages have been achieved by using the level set method for the numerical implementations [8]. Here we briefly present Geman-Geman's and Mumford-Shah's energies for image segmentation.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 82–89, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Let Ω be a bounded open set of R^2, the segmentation problem can be formulated as follows: Given an initial image u_0 (could be a noisy image), it is required to find a set of edges Γ and a final image u which is smooth (or stationary) in each region of Ω\Γ.

The segmentation problem now becomes the minimization of the energy $E[u, \Gamma|u_0]$. Geman and Geman modelled this energy [5] by:

$$E(u) = \sum_{i,j} [\lambda^2 |u(i+1, j) - u(i, j)|^2 (1 - v(i, j)) + \lambda^2 |u(i, j+1) - u(i, j)|^2 (1 - h(i, j)) + \alpha(h(i, j) + v(i, j)) + |u(i, j) - u_0(i, j)|^2] . \quad (1)$$

Where $h(i, j) \in \{0, 1\}$ and $v(i, j) \in \{0, 1\}$ are horizontal and vertical "line process" respectively. α and λ are model parameters. As noted by Mumford and Shah, the term $\alpha \sum_{i,j} (h(i, j) + v(i, j))$ could be an estimation of the length of the edges inside the image. They rewrote the total energy in the continuous form as follows [7]:

$$F(u, C) = \lambda \int_{\Omega} |u - u_0|^2 \, dxdy + \mu \int_{\Omega \setminus \Gamma} |\nabla u|^2 \, dxdy + \nu \cdot Length(\Gamma) . \quad (2)$$

Where λ, μ and ν are model parameters, x, y are coordinates in the domain R^2. If $\mu \to \infty$, then u must have zero gradients in each region. This leads to reducing the Mumford-Shah model to a piecewise constant approach [7]. Chan and Vese rediscovered this approach in terms of active contours in [3], [11]. In these works Chan and Vese employed partial differential equations and the level set techniques to implement a curve evolution on the plane. The curve stops on the boundaries of the objects within the image where the energy has its minimum.

The Chan-Vese piecewise smooth active contour model [11] and a similar approach developed by Tsai, Yezzi and Willsky [10] perform image smoothing and segmentation simultaneously. The smoothing process is governed by the damped Poisson equations. In this method, besides the computational cost problem, the model cannot give good denoising results for very noisy images. Very recently, we propose a hierarchical model for segmentation and smoothing based on the active contour model. A diffusion equation was used to smooth the noisy images [4].

There are two objectives in this paper. First, we want to study the piecewise constant model of the Mumford–Shah formulation in details with respect to the model parameter ν. We put special emphasis on ν since it controls the segmentation process. Second, we propose a new multistage smoothing method and use diffusion equation for image denoising because it can deal with very noisy images and preserve edges.

2 Segmentation Prior to Smoothing

The idea of our segmentation prior to denoising algorithm is to obtain different sub-regions within a given image by using the piecewise constant segmentation method first, then applying a diffusion equation to each sub-region of the original noisy image u_0 independently, but not across the edges. The final reconstruction of an image is obtained in terms of the combination of the smoothing results in all such sub-regions. This algorithm works in the following steps:

(1) Apply the piecewise constant segmentation method to the original image u_0, and partition the image into different regions.

The piecewise constant active contour model uses a level set function ϕ to represent the evolving curve C [3]. The region inside C is denoted by $\phi > 0$, outside C by $\phi < 0$ and on C by $\phi = 0$. An image is represented in two phases by two constants c_1 and c_2 which are the mean intensity values in regions $\{\phi > 0\}$ and $\{\phi < 0\}$ respectively. Under these approximations and using the *Heaviside* function $H(\phi) = \begin{cases} 1 & \text{if } \phi \geq 0 \\ 0 & \text{if } \phi < 0 \end{cases}$, the energy functional (2) can be represented in a level set formulation (for λ=1) as,

$$F(\phi, c_1, c_2) = \nu \int_\Omega |\nabla H| dxdy + \int_\Omega |u_0 - c_1|^2 \, H dxdy + \int_\Omega |u_0 - c_2|^2 \, (1 - H) dxdy. \qquad (3)$$

When minimizing this energy functional over the level set function ϕ, the curve evolution equation can be obtained as follows [3]:

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\left[\nu \, \nabla \cdot \left(\frac{\nabla \phi}{|\nabla \phi|}\right) - (u_0 - c_1)^2 + (u_0 - c_2)^2\right]. \qquad (4)$$

The segmented image therefore is a two-phase image, the curve C is the boundary between the sets $\{\phi > 0\}$ and $\{\phi < 0\}$. Hence the original image u_0 can be represented by u^{I} in region $\{\phi > 0\}$ and u^{II} in $\{\phi < 0\}$, where $u_0 = u^{\mathrm{I}} \cup u^{\mathrm{II}} \cup C$.

(2) Apply the diffusion equation to u^{I} and u^{II} separately. In order to solve the diffusion equations numerically in different regions, we need to extend u^{I} to the region $\{\phi < 0\}$ and u^{II} to the region $\{\phi > 0\}$. For instance, to extend u^{I} to the region $\{\phi < 0\}$, we can use the average of u_0 in region $\{\phi < 0\}$. Other extension methods can be found in [11]. The boundary conditions are $\partial u^{I} / \partial \vec{n} = 0$ or $\partial u^{II} / \partial \vec{n} = 0$ ($\vec{n}$ is the normal of the boundary C) when we extend u^{I} or u^{II} across the edges between regions. The diffusion does not cross the boundaries, and it is very similar to the idea of the *anisotropic diffusion* [12]. The general formula of a diffusion equation is,

$$\partial u / \partial t = \nabla \cdot (\mathrm{D} \cdot \nabla u), \qquad u(t = 0) = u_0. \qquad (5)$$

Where $\boldsymbol{D}$ is the diffusivity. For the simplest linear isotropic diffusion, $\boldsymbol{D}$ becomes a scalar constant.

We first demonstrate how the proposed algorithm works for images with two-phase representation through a *satellite* ice image in Fig.1. We also show the segmentation results related to the choice of the model parameter ν. When we fix λ=1, the step-sizes h = 1, and Δt = 0.1 in our finite difference (implicit) numerical implementation, ν is the only free parameter in the active contour segmentation model. If ν is small we can detect all detailed boundaries including small areas created by noise. As ν increases, larger and larger areas (or objects) can be detected but smaller objects like noisy points are removed. Systematic studies of the effects of changing the value of

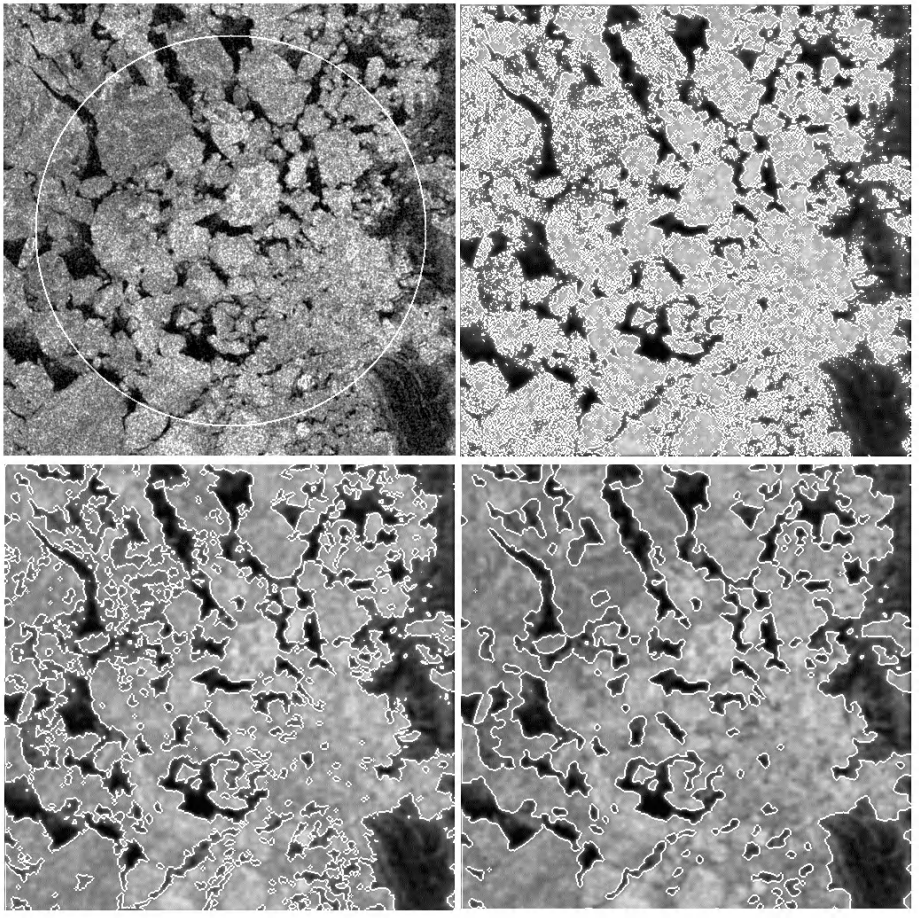


Fig. 1. Segmentation and smoothing of a satellite ice image. **Top left**: original noisy image with an initial level set curve, image size 512×512. **Top right**: segmented and smoothed image with ν=1, t = 9.3s. **Bottom left:** β = 1, t = 35.6s. **Bottom right:** β = 3, t = 106.6s.

the parameter ν do not exist in the literature. In this paper we present a systematic study of the parameter ν and propose a method to estimate ν.

We choose $\nu = \beta\sigma^2$, where σ^2 is the variance of the image u in a region Ω:

$$\sigma^2 = \int_\Omega (u-\overline{u})^2\, dxdy \Big/ \int_\Omega dxdy\,, \tag{6}$$

where $\overline{u}$ is the mean value of u in Ω, and β is a constant factor. We have determined β empirically to be between 0 and 5. The value of β depends on the amount of noise in the original image. From Fig.1 we can see that if we choose a very small value of ν then the image is *over-segmented* (see top right of Fig.1), because very small edges created by noise are also detected. When we increase the value of ν we obtain less

detailed segmentations (see bottom left of Fig.1). If we choose a large value of ν we obtain the final segmentations without the edges created due to noise (see bottom right of Fig.1).

For comparison, we provide the *CPU* time *t* in seconds for running our C++ program on Pentium IV 2.40GHz PC. The *CPU* times for the given images using different algorithms are detailed in the captions of the three figures.

3 Multistage Smoothing Algorithm

Since using smaller ν the active contour segmentation model can detect more edges, and the diffusion process which applies to different regions but not across the edges can preserve those edges. In order to remove noise effectively and also preserve as many 'good features' like edges of objects within the image as possible, we divide the segmentation and smoothing process into different sub-stages. This is characterized by choosing different values of the scale parameter $\nu = \beta\sigma^2$. When $\beta = \beta_{min}$ the edges are preserved. When $\beta = \beta_{max}$ the segmentation process can remove almost all noise, but it may also destroy some of the edges. Hence we propose the following multi-stage method.

At the first stage we use a small parameter ν_1 for segmentation and also apply the diffusion filter to each region, therefore most of the edges can be detected and preserved. However, some small "objects" may be formed by noise at the output of this stage.

The second stage takes the output of the first stage as input. However a larger scale parameter $\nu_1 < \nu_2 < \nu_{max}$ is used. Since we choose $\nu_2 > \nu_1$ some of the small "objects" formed at the previous stage and some very small edges are removed. The result obtained at this stage is coarser than the previous stage.

We can repeat this procedure until we reach the final segmentation and smoothing result. The proposed multistage algorithm is somewhat similar to region merging algorithm proposed by Koepfler *et al.* [6]. In practice one or two middle stages are good enough for most noisy images.

In Fig.2 we use a *knotted DNA* image to show the importance of our multistage method. The first row of Fig.2 shows results using one stage approach and the second row shows the results of our multistage approach. From the results in Fig.2 we can see that the result using our multistage approach is better than the result using only one stage segmentation and smoothing. The *DNA* knot chain is preserved in the multistage approach while it is broken in the one stage approach.

In Fig.3, we demonstrate how the proposed multistage method works for images with very large amount of Gaussian noise. We use a synthetic image and add the Gaussian noise to it. The amount of noise is measured by the signal-to-noise ratio (*SNR*) which is defined by:

$$SNR = 10\log_{10}\left(\|u\|^2 / \|u_0 - u\|^2\right),$$

where u_0 is the 'clear' image while u is the noisy image, and $\|\cdot\|$ denotes the Euclidean norm. *SNR* is usually measured in dB (*decibel*).

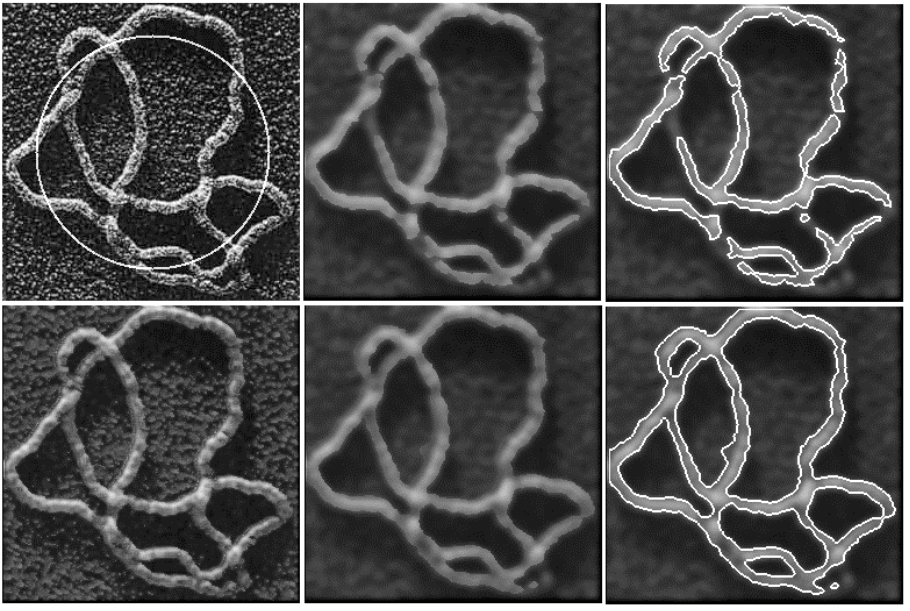


Fig. 2. Segmentation and smoothing results of a knotted DNA image with one-step and two-step smoothing methods. **Top left**: Original image with an initial level set circle, size: 256×256. **Top middle**: Result by one–step algorithm, $\nu = 8271.7$, $t = 19.8$s. **Top right**: Same result as in the top middle but the segmenting curves superimposed. **Bottom left**: the first stage result for the multistage algorithm, $\nu = 87.1$, $t = 3.5$s. **Bottom middle**: The final result after two-step segmentation and smoothing, $\nu = 3961.6$, $t = 16.2$s. **Bottom right**: Same result as in the bottom middle but the segmenting curves superimposed.

If we compare the result using one stage method (top right of Fig.3) with the result of our 3-stage method (bottom right of Fig.3), some image features are lost in the one-stage method while the multistage method preserves more features.

Using our multistage approach we can process very noisy images and still preserve the edges. This is useful in applications such as in medical image processing.

4 Conclusion

In this paper we have studied the effects of changing the scale parameter ν in the piecewise constant model. We also proposed and implemented new image segmentation and smoothing algorithms based on the Chan-Vese active contour model and PDE-based diffusion techniques. The level set methods are used in our numerical implementations. For denoising, this algorithm works in two steps: First the image is segmented by using the piecewise constant segmentation method, then PDE-based diffusion method is used to smooth and denoise each segmented region of the original image separately but not across the boundaries.

Fig. 3. Segmentation and smoothing results of a noisy synthetic image with multistage method. **Top left**: Original noisy image with an initial level set circle, *SNR* = 3.59 dB, image size: 512×512. **Top middle**: result by one-step algorithm, ν = 19632.2, *t* = 121.2s. **Top right**: Same result as in the top middle but the segmenting curves superimposed. **Bottom left:** The result of the first stage, ν = 109.1, *t* = 15.0s. **Bottom middle**: The images of the second stage, ν = 570.5, *t* = 14.8s. **Bottom right**: Final result from the 3-stage method with the segmenting curves superimposed, ν = 6478.5, *t* = 93.8s. The total *CPU* time for the 3-stage method is *t* = 123.6s.

If we need to deal with very noisy images, the proposed algorithm allows us to use multiple steps strategy by choosing different scale parameter ν incrementally from the first stage to the final stage as described in section 3. The most important advantage of this approach is more edges can be preserved while noise is removed.

Compared with the previous simultaneous segmentation and smoothing methods [11], [10], the proposed method is more efficient and flexible. It also improves the computational speed drastically. Furthermore, since we use the multistage smoothing approach by choosing different scale parameter ν for high levels of noise, we can obtain better segmentations for very noisy images. The proposed method allows us to apply different smoothing algorithms in different regions in an image, so it is very convenient when the applications need to highlight some special regions in an image. For example, the inverse diffusion technique can be implemented in our method for edge enhancement. Like anisotropic diffusion methods, the proposed algorithm only smoothes the image within the homogeneous regions but not across the boundaries, thus edges are preserved during the denoising process. The proposed method can process very noisy images with good performance.

In the present paper we use one level set function therefore we can segment an image into two regions. In general case multiple level set functions should be used for images with complicated features to obtain multiphase segmentations using two or

more level set functions [11]. The proposed multistage segmentation and denoising method can be extended to multiphase segmentations.

Acknowledgment. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. We would like to thank Dr. Langis Gagnon of the Computer Research Institute of Montreal (CRIM), Canada for providing us the satellite ice image.

References

1. Aubert, G., Kornprobst**,** P.: Mathematical Problems in Image Processing: Partial differential equations and the Calculus of Variations, Applied Mathematical Sciences, vol.147, Springer-Verlag, Berlin Heidelberg New York (2002)
2. Chambolle, A.: Image Segmentation by Variational Methods: Mumford and Shah Functional and the Discrete Approximations. SIAM Jour. on Appl. Math. 55 (1995) 827–863
3. Chan, T.F., Vese, L.V.: Active Contours without edges. IEEE Tran. Image Proces. 10 (2001) 266–277
4. Gao, S., Bui, T.D.: A New Image Segmentation And Smoothing Model. IEEE International Symposium on Biomedical Imaging, April (2004) 137–140
5. Geman, S., Geman, D.: Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. IEEE Trans. on PAMI. 6 (1984) 721–741
6. Koepfler G., Lopez, C., Morel, J.M.: A Multiscale Algorithm for Image Segmentation by Variational Method. SIAM J. Numer. Anal. 33 (1994) 282–299
7. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. 42 (1989) 577–685
8. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University Press (1999)
9. Snyder, W., Logenthiran, A., Santago, P., Link, K. Bilbro, G., Rajala, S.: Segmentation of Magnetic Resonance Images using Mean Field Annealing. Image and Vision Comput. 10 (1992) 361–368
10. Tsai, A., Yezzi, A., Willsky, A.S.: Curve Evolution Implementation of the Mumford–Shah Functional for Image Segmentation, Denoising, Interpolation, and Magnification. IEEE Tran. on Image Proces. 10 (2001) 1169–1186.
11. Vese, L.V., Chan, T.F.: A multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model. International Journal of Computer Vision 50(2002) 271–293
12. Weickert, J.: *Anisotropic diffusion in Image Processing*, Teubner, Stuttgart (1998)

A Multiresolution Threshold Selection Method Based on Training

J.R. Martinez-de Dios and A. Ollero

Grupo de Robótica, Visión y Control.
Departamento de Ingeniería de Sistemas y Automática.
Escuela Superior de Ingenieros. Universidad de Sevilla.
Camino de los Descubrimientos, 41092 Sevilla (Spain).
Phone: +34-954487357; Fax:+34-954487340.
{jdedios, aollero}@cartuja.us.es

Abstract. This paper presents a new training-based threshold selection method for grey level images. One of the main limitations of existing threshold selection methods is the lack of capacity of adaptation to specific vision applications. The proposed method represents a procedure to adapt threshold selection methods to specific applications. The proposed method is based on the analysis of multiresolution decompositions of the image histogram, which is supervised by fuzzy systems in which the particularities of the specific applications were introduced. The method has been extensively applied in various computer vision applications, one of which is described in this paper.

1 Introduction

Image segmentation is an essential process in image analysis. Threshold selection methods can be classified according to the information on which they rely for object/background classification [8]. Some methods rely on grey level information and ignore spatial dependence such as those based on maximization of entropy functions [6] and [1]; maximization of class separability [11] and minimization of misclassification error [7]. Some other thresholding methods use spatial information.

These methods are based on general object/background separability criteria and are not capable of adapting to specific applications. Thresholding problem is highly dependent on the vision application. The method uses knowledge extracted from training images of an application to supervise the threshold selection. This method is a procedure to design threshold selection methods adapted to specific problems.

2 Multiresolution Scheme for Threshold Selection

Most histogram-based threshold selection methods assume that pixels of the same object have similar intensities. Thus, objects in the images are represented as histogram modes. Some methods aim to identify the mode or modes corresponding to the object of interest -*object modes*-. The method presented in this paper divides the

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 90–97, 2004.
© Springer-Verlag Berlin Heidelberg 2004

identification and selection of object modes in L+1 steps. The method presented is based on the analysis of approximations of the image histogram at different levels of resolution l. Let $h(z)$ be the histogram of an image with NL intensity levels. Let $CA^l(z)$ be the multiresolution approximations of $h(z)$ at resolution $l \in \{L, L-1,...,0\} \subseteq \mathbb{N}$. $CA^l(z)$ is computed from $h(z)$ by applying Mallat's approximation decomposition algorithm [10] at level l. The wavelet decomposition uses Haar functions due to its efficient implementation [5]. The low-pass filtering effect increases with l. This paper also uses the concept of histogram region that will be defined as a set $\{z : a \le z \le b\}$, where a and b are two intensity levels.

Fig. 1 shows the general scheme of the presented method. The first two steps are the computation of $h(z)$ and the wavelet approximation of $h(z)$ at level L, $CA^L(z)$. Then is the iterative application of *Mode Selection System* from l=L to l=0. At each level l, *Mode Selection System* carries out the function f: $\{IHROI^l\} \rightarrow \{SHROI^l\}$. *Mode Selection System* selects the modes in $\{IHROI^l\}$ that correspond to the object of interest. With the selected modes it computes the output histogram region $\{SHROI^l\}$ at level l. The selection of the object modes will be described in Section 2.1. The histogram region with the selected modes at level l, $\{SHROI^l\}$, will be analyzed with more resolution at level l-1: assuming dyadic decompositions, at l-1 $\{SHROI^l\}$= $\{z: a \le z \le b\}$ is transformed to $\{IHROI^{l-1}\}$=$\{z: 2a \le z \le 2b\}$. The application of *Mode Selection System* from l=L to l=0 analyzes the histogram at increasing resolutions and perform an iterative restriction of the histogram region of interest. The resulting histogram region, $\{SHROI^0\}$, contains the modes selected as object modes in $CA^0(z) = h(z)$. Finally, the threshold is computed as the lower value of $\{SHROI^0\}$.

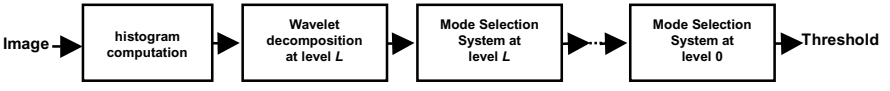


Fig. 1. General scheme of the multiresolution method for threshold selection.

2.1 Mode Selection System

Mode Selection System at level l analyzes $CA^l(z)$, $\{IHROI^l\}$, selects the modes corresponding to the object and, computes $\{SHROI^l\}$ from the selected modes. Assuming a probabilistic (Bayesian) approach commonly considered in the literature, histograms can be modeled as a mixture of Gaussian probability density functions. Histogram decomposition in Gaussian modes can be considered as an unsupervised problem, which can be solved by the method described in [2].

Assume S is the set of modes selected as corresponding to the object, $S = \{\text{selected } \omega_i\}$. The histogram can be divided in two components: one with the selected modes and one with the non selected modes, respectively

$h_s(z) = \sum_{\omega_i \in S} P(\omega_i) p(z|\omega_i)$ and $h_u(z) = \sum_{\omega_i \notin S} P(\omega_i) p(z|\omega_i)$. The aim of *Mode Selection System* is to select the modes such that $h_s(z)$ contains the modes of the object.

Histogram modes interpretation is not an easy problem. The *Fuzzy Supervising System*, also denoted by *FSS*, is responsible for selecting the histogram modes according to knowledge of the application. The aim of the *FSS* is to recognize mode features in order to classify a mode as corresponding to the object or to the background. Let $\{descriptor_d\}$, $d \in \{1, 2, ..., D\}$ be a set of features able to describe mode ω_u. The knowledge of the application was introduced in the *FSS* during the *FSS Training* (Section 3). The selection of $\{descriptor_d\}$ is detailed in Section 4.

Fig. 2 shows the scheme of the *Mode Selection System* at level *l*. *j* is the number of modes in which $CA^l(z)$ is decomposed. At each iteration the selection of one mode, denoted by ω_u^l, is analyzed. If a mode is selected, $CA_s^l(z)$ and $CA_u^l(z)$ are updated.

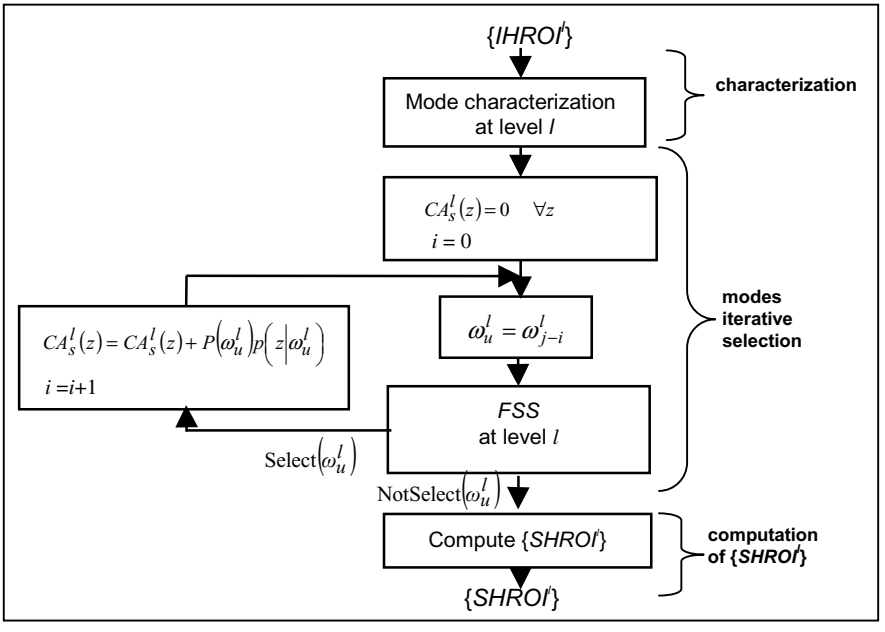


Fig. 2. Scheme of the *Mode Selection System* at level *l*.

To generalize the expressions for level *l*, their formulation should be transformed by substituting ω_i by ω_i^l, S by $S^l = \{\text{selected } \omega_i^l\}$, and $h_s(z)$ and $h_u(z)$ by $CA_s^l(z)$ and $CA_u^l(z)$. The proposed method assumes that the object corresponds to the modes with higher intensity levels in the histogram. In order to select all the modes corresponding to the object, the iterations continue in descending order (from ω_j^l to ω_0^l) until a mode is not selected. Once the iterative selection has finished, $\{SHROI^l\}$ is

computed as the histogram region $\left\{z : TH_s^l \le z \le \left(NL/2^l\right)-1\right\}$, which is lower bounded by TH_s^l, the value that optimally distinguishes $CA_s^l(z)$ from $CA_u^l(z)$, i.e. $CA_s^l\left(TH_s^l\right) = CA_u^l\left(TH_s^l\right)$. The computation of TH_s^l consists in solving a simple 2nd order equation (see [3]) by applying simple expressions.

2.2 Fuzzy Supervising System

Fig. 3 depicts the diagram block of the *FSS*. The input of the *FSS* is $\{descriptor_d\}$, $d \in \{1, 2, ..., D\}$. The output, $y \in [0,1]$, represents a possibility value to select ω_u^l as part of the object. ω_u^l is selected if $y \ge \alpha$, where $\alpha \in [0,1] \subset \Re$, will be called the *FSS decision threshold*. If the output (y) is higher than α, the *FSS decision threshold* will consider ω_u^l part of the object. Otherwise ω_u^l will be considered as corresponding to the background

The main advantage of using fuzzy systems for *FSS* is that the knowledge is expressed in form of explicit rules, which can be easier to understand by system designers. Besides, knowledge can be introduced in fuzzy systems via training methods. The design of *FSS* based on handcrafted rules can originate long test&error processes. In this paper a training method is applied to incorporate this knowledge.

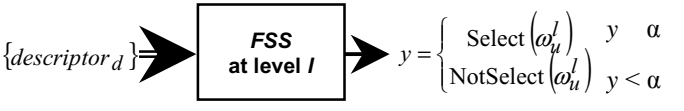


Fig. 3. Diagram block of the *Fuzzy Supervising System* (*FSS*).

3 FSS Training

The aim of the *FSS Training* is to extract knowledge from a set of training images of a vision application and, to introduce it in the *FSS* by means of a training process. The *FSS Training* can be divided in *Pair Set Generation* and *Fuzzy Identification*.

3.1 Pair Set Generation

Pair Set Generation is divided in two tasks: *Selection of Training Images* and *Pair Set Extraction*. The first one consists in selecting the training images and assigning their corresponding desired threshold values. The training images should be carefully chosen to consider a rich variety of conditions of the computer vision application. The threshold values for each training image should be carefully selected to best fit the application specifications.

The aim of *Pair Set Extraction* is to extract a set of data pairs that contain the knowledge of the application. These input-output data pairs express the desired input-output behavior of the *FSS* in order to minimize the error with the desired threshold th_n. *Pair Set Extraction* simulates the operation of the multiresolution method to identify the situations in which the *FSS* decides whether ω_u^l should be selected. Once these situations are identified, *Pair Set Extraction* computes the decisions in order to minimize the number of misclassified pixels. Each decision situation k generates an input-output data pair (I/O pair) of the form $dpairk=\left(\left\{descriptor_d\right\}_k, dy_k\right)$, where $\left\{descriptor_d\right\}_k$ is the point in the space described by $\left\{descriptor_d\right\}$ that represents the conditions of decision situation k, and dy_k is the desired output for decision situation k. In case that the desired decision is to select ω_u^l, the output should be $dy_k=\alpha+\beta>\alpha$, where α is the *FSS decision threshold* and $\beta\in[0,\ 1]\in\Re$ will be named as the protection band width. If the decision is not to select ω_u^l, $dy_k=\alpha-\beta<\alpha$.

Under the assumption that the object corresponds to the highest modes of the histogram, th_n, the threshold value manually chosen for $image_n$, is the lower bound of $\{DHROI_n^0\}$, $\{DHROI_n^l\}=\{z : th_n \le z \le NL-1\}$, which is the desired histogram region at level l=0 for $image_n$. $\{DHROI_n^l\}$, the dyadic representation of $\{DHROI_n^0\}$ at level l, is expressed by $\{DHROI_n^l\}=\left\{z : th_n/2^l \le z \le NL/2^l-1\right\}$.

Let $\{SHROI_n^l\}$ be the selected region of interest at level l for training image n. The criterion for mode selection is based on the minimization of the number of misclassified pixels. Misclassified pixels can be classified in: *detection error pixels* (denoted by pDE_n), those in $\{DHROI_n^l\}$ but not in $\{SHROI_n^l\}$, i.e. which intensity levels are $z\in\left\{\{DHROI_n^l\}/\{SHROI_n^l\}\right\}$; and *false positive pixels* (pFA_n), those in $\{SHROI_n^l\}$ but not present in $\{DHROI_n^l\}$, i.e. which intensity levels are $z\in\left\{\{SHROI_n^l\}/\{DHROI_n^l\}\right\}$. This criterion leads to the minimization of J:

$$J=\sum_{n=1}^{N} j_n\ ,\qquad j_n=(\tau)\,pDE_n+(1-\tau)pFA_n \tag{1}$$

The value of $\tau\in[0,1]$ establishes a compromise common in many applications.

3.2 Fuzzy Identification

Fuzzy Identification consists in training the *FSS* to approximate the input-output pairs that contain the knowledge of the specific application. In this paper *Fuzzy Identification* is performed by applying *ANFIS* method. *ANFIS* [4] uses a hybrid learning algorithm to identify the parameters of Sugeno-type fuzzy inference systems. It applies a combination of the least-squares method and the backpropagation gradient descent

method for training fuzzy parameters to fit a data set. *ANFIS* receives as inputs a set of *NP* filtered pairs, and trains the *FSS* by minimizing the following error:

$$MSE = \frac{1}{NP_f} \sum_{k=1}^{NP} (e_k)^2 , \qquad e_k = dy_k - FSS(\{ descriptor_d \}_k), \tag{2}$$

where $FSS(\{ descriptor_d \}_k)$ is the output of the *FSS* to input $\{ descriptor_d \}_k$. The *FSS* will be considered to be successfully trained if it classifies all pairs without error.

4 Selection of Descriptors

The set of descriptors represent the features that will be used for mode interpretation in the mode selection. In these examples the descriptors used are selected to provide information on **illumination conditions** and **mode similarity**. Changes in illumination conditions have severe impact on the performance of most thresholding methods. The effect of illumination on ω_u^l can be parameterized in terms of its mean and standard deviation. Several segmentation techniques have been developed basing on theoretical and empirical definitions of mode similarity measures such as Kullback-Leibler distance or Weighted-Mean-Variance distance [9]. In this work, mode similarity mainly refers to the notion of "proximity" and is estimated by a simplified definition of the Weighted-Mean-Variance distance.

5 Results and Discussions

The presented method has been applied in an automatic infrared detection system in outdoor urban scenarios installed on the roof of the building of the School of Engineering of Seville. The images in Fig. 4a-c show some of the training images. The desired threshold values for images in Fig. 4a-c (see Table 1a) were selected to fit the application specifications by using an image processing tool. The *FSS Training* is executed with 43 training images. The parameters of the method were L=4, α=0.5, β=0.25 and τ =0.8. Table 1b shows the threshold values resulting of applying the trained method to the images shown in Fig. 4a-c. Notice the low difference between the desired and resulting thresholds shown in Table 1a and Table 1b. The mean number of misclassified pixels per image is 4, which has no practical relevance (see images in Fig. 4d-f).

Fig. 5 shows one test image and its corresponding satisfactorily thresholded image. Fig. 6 shows the images resulting of applying the threshold values computed by Otsu's [11], Kittler-Illingworth's [7] and Ridler-Calvard's [12] methods to the image shown in Fig. 5a. They contain a high number of background pixels classified as object pixels. The proposed method bases on object/background separability criteria learned from application images while existing techniques rely on generic criteria.

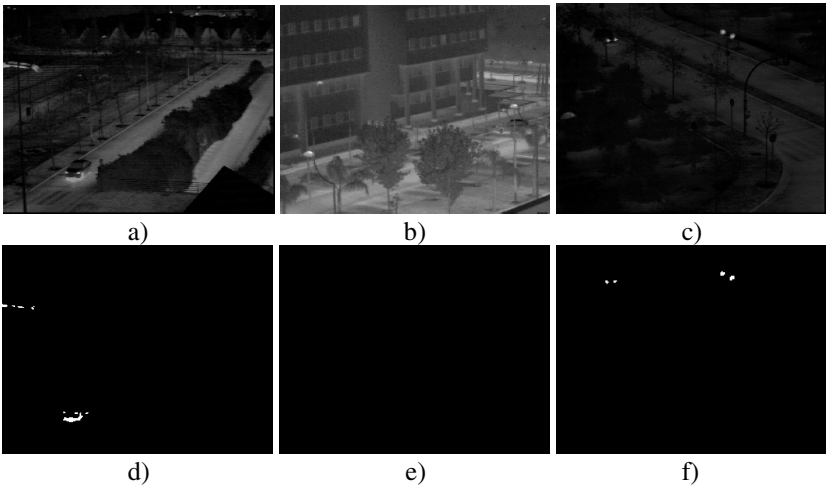


Fig. 4. a-c) Three training images of an infrared detection system in outdoor urban scenarios; d-f) thresholded images provided with the trained *FSS* corresponding to the images in Fig. 4a-c.

Table 1. Desired threshold values for the training images in Fig. 4a-c and resulting threshold values provided by the method with trained *FSS*.

image	a)	b)	c)
desired th_n	126	105	118

image	a)	b)	c)
resulting th_n	128	102	118

Fig. 5. Test images and thresholded images by the method with the trained *FSS*.

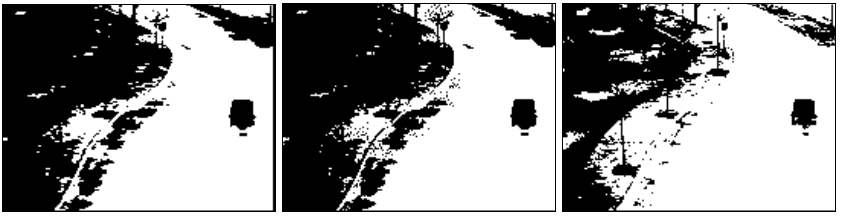


Fig. 6. Resulting thresholded images by applying: Otsu's, Kittler-Illingworth's and Ridler-Calvard's methods to the image shown in Fig 5a.

6 Conclusions

This paper presents a new training-based threshold selection method for grey-level images. It represents a new procedure to design threshold selection methods particularized to specific applications. The application-adaptation capabilities of the method bases on the utilization of knowledge of the specific application extracted from a set of training images. The training process extracts knowledge from the training images and incorporates it in a supervising system via *ANFIS* method. The main contribution of the presented method with respect to existing techniques is that it relies on object/background separability criteria learned for the specific application while existing techniques rely on generic criteria. The method has been extensively applied in various computer vision problems. Its performance exhibits considerable robustness to illumination conditions and noise in the images.

Acknowledgements. The work described in this paper has been developed in the frame of the following projects SPREAD (EVG1-CT-2001-00043), COMETS (IST-2001-34304) and EURFIRELAB (EVR1-CT-2002-40028), funded by the European Commission.

References

1. Abutaleb A. S., "Automatic Thresholding of gray level pictures using two-dimensional entropy", Computer Vision, Graphics and Image Processing, vol. 47, (1989). 22-32.
2. Duda R.O, P.E. Hart and D.G. Stork, "Pattern Classification", John Wiley and Sons, (2001).
3. Gonzalez R.C. and R.E. Woods, "Digital Image Processing", Addison-Wesley (1992).
4. Jang J.-S. R., "ANFIS: Adaptive-Network-based Fuzzy Inference Systems", IEEE Transactions on Systems, Man and Cybernetics, Vol. 23, no. 3, (1993). 665-685.
5. Kaiser G., "The Fast Haar Transform", IEEE Potentials, May/April, (1998). 34-37.
6. Kapur J.N., P.K. Sahoo, and A.K.C. Wong, "A new method for gray-level picture thresholding using the entropy if the histogram", Comput. Vision, Graphics, Image Processing, vol. 29, (1985) 273-285.
7. Kittler J. and J Illingworth, Minimum error thresholding, Pattern Recognition, 19, (1986).41-47.
8. Lee S. U., Y.S. Chung, and R. H. Park, "A comparative performance study of several global thresholding techniques for segmentation", Computer Vision, Graphics, and Image Processing, 52(2), (1990). 171-190.
9. Ma W. Y. and B. S. Manjunath, "Texture Features and Learning Similarity", In Proc. IEEE Conference on Computer Vision and Pattern Recognition, (1996), 425-430.
10. Mallat S., "A Theory for Multi-resolution Signal Decomposition: the Wavelet Representation", IEEE Trans. PAMI, vol PAMI-11, (1989). 674-693.
11. Otsu N., "A threshold selection method from grey-level histograms", IEEE Trans. Syst., Man, Cybern., vol. SMC-9, no. 1, (1979). 62-66.
12. Ridler T.W. and S Calvard, "Picture Thresholding using an Iterative Selection Method", IEEE Transations on Systems, Man and Cybernetics. Vol. 8. (1978). 630-632.

Segmentation Based Environment Modeling Using a Single Image*

Seung Taek Ryoo

Department of Software, HanShin University, Korea
stryoo@hs.ac.kr

Abstract. This paper introduces segmentation based environment modeling method for constructing the real-time image-based view-rendering. The image(or environment map) is segmented according to the objects that form the environment and the depth value is set by the characteristics of the classified objects for the segmentation based modeling. This method can easily be implemented on an environment map and makes the environment modeling easier through extracting the depth value by the image segmentation. It becomes possible to develop an environment modeling system with a full-view through this method.

1 Introduction

The image-based modeling method is the one that is being studied the most due to its efficiency. This method enables real-time rendering because it extracts the geometric information from the image that represents the environment in the pre-processing stage. The depth-image acquisition method is mostly used for the geometric recovery. The image registration method is used extract the depth value through corresponding point method using the stereo-image, but different methods are required for application on the environment map. Methods for representing the object or environment using range images are being studied recently. The rangefinder is sometimes used to acquire the range image and realistic image concurrently, but a high-resolution image cannot be obtained through this method. To acquire an accurate 3D object, the studies that are registered and integrated the range data and environment images are being done on this area as well as on acquiring 3D model mapped on high-resolution images. The objective of this paper is on the generation of the environment model as realistically as possible. A method for acquiring the depth-image through image segmentation is suggested to construct the environment model. This paper is organized as follows: The related works on 3D reconstruction is surveyed in chapter 2. The depth-image acquisition method and environment modeling/mapping method is discussed in detail in chapter 3. The results of segmented based environment modeling implementation are presented in the chapter 4. Finally, the conclusion of this paper and further areas to study are illustrated.

* This work was supported by HanShin University Research Grant in 2004.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 98–105, 2004.
© Springer-Verlag Berlin Heidelberg 2004

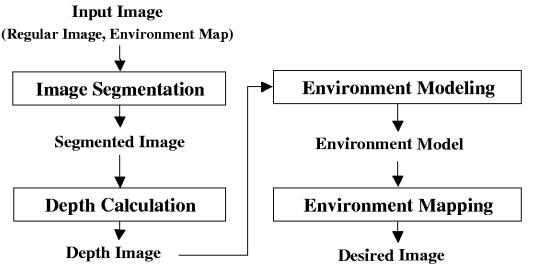


Fig. 1. The steps of segmentation based environment modeling

2 Related Work

If the information about the environment can be extracted, environment model can be made from a single image. Method of reconstructing the environment information using the vanishing point and vanishing line [1][2][3] and the image editing method through interaction [4][5] are some examples of this method. The first method performs the modeling of the environment using the vanishing point based on the acquired plane, which makes it difficult to acquire an accurate environment model. The second method enables a more accurate modeling but requires user interaction using various tools. A reference plane has been used to set the depth value from the image primitives [1][2][4]. Horry [1] used a plane constructed using spider mesh to set the depth value and Criminisi [2] took into account the relationship between the plane parallel to the reference plane and the vanishing point in calculating the distance. Oh [4] used the floor parallel tool and the perpendicular tool to set the location on the reference plane. In this paper, the method using the parallel/perpendicular of the plane has been used for generating an accurate model. The input image is divided into a floor, ceiling and surrounding objects and the segmented floor(ceiling) is set as the reference plane, to calculate the depth of the surrounding objects.

In this paper, method for reconstructing a 3D environment model using panoramic images has also been developed. The method using vanishing point [6] and the method using a multiple environment maps [7] are some of the methods used for panoramic images. The first method uses a vanishing point for the environment modeling which makes it difficult to acquire accuracy. The second method requires a multiple number of environment maps and requires a high cost for the relationship. In this paper, a segmented based environment modeling method without relying on the number or the type or the features of images has been suggested through the segmentation of the environment map. The characteristics of this method are shown in the next chapter.

3 Segmentation Based Environment Modeling

The model should be constructed using the depth value to freely represent the environment. The depth value should be extracted from the image or the en-

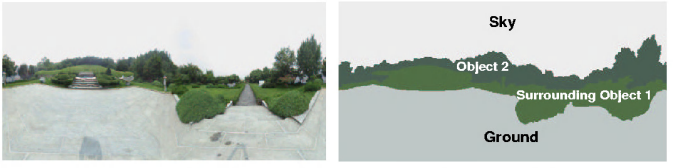


Fig. 2. Region-based image segmentation

vironment map acquired to construct an environment model. We propose an environment modeling method based on depth image acquired through image segmentation. Figure 1 shows the process of image based environment modeling. First of all, the input image is segmented according to the characteristics of the environment. The depth image is acquired from the segmented image. Then an environment model is constructed through a rectangular or triangular grid using the depth image and then the real image is mapped onto the environment model for the desired scene. These methods are shown in the following section.

3.1 Image Segmentation

An indoor virtual environment consists of a floor and ceiling and the outdoor of ground and sky. From such characteristics, the environment map is divided into a floor(ground), ceiling(sky), and the surrounding objects. A method for image segmentation is required to divide the image or environment map into these features. Image segmentation is a process of extracting one or more connected regions satisfying homogeneity criterion which is based on features derived from spectral components. The image segmentation method [8] can be divided into a pixel-based, region-based, an edge-based method and the hybrid method. The edge-based approach classifies pixels using a numerical test for a property such as image gradient or curvature. On the other hand, region based methods take the basic approach of dividing the image into regions and classifying pixels as inside, outside, or on the boundary of a structure based its location and the surrounding 2D regions. Because the region based segmentation is more robust, we used the region method based on the characteristics of the image in this paper. These approaches, including region growing, region splitting, region merging and their combination, attempt to group pixels into homogeneous regions. The region-based segmentation method [9] explained above has been applied to segment a regular image and the environment map into floor, ceiling and surrounding objects. The depth calculation method in the next paragraph is applied to the segmented image to acquire the depth image. Figure 2 shows the image after the environment map [10] has been divided into a ground, sky and surrounding objects.

3.2 The Extraction of the Depth Value

- The image plane is perpendicular to the XZ plane in the real-world coordinate.
- The floor(ground) and the ceiling consist of a plane that is parallel to the XZ plane in the real-world coordinate.
- In case of outdoor scene, the sky consists of a hemisphere.
- The surrounding objects are perpendicular to the ground or ceiling

Two stages are required to compute the depth value using these conditions. In first stage, the depth value of the floor or the ceiling is calculated. In second stage, the depth value of the surrounding objects is calculated by using the depth value from the first stage. The methods of each stage are shown in the following paragraph.

The depth calculation of the floor(ceiling). Figure 3 shows the process of calculating the depth value of the floor using a regular image. Figure 3-a shows the segmented image and the environment model through image coordinate system and Figure 3-b shows the relation between the image plane and the floor through world coordinate system. o is the origin of the real-world coordinate, b the distance between the central point(m) and the pixel, d the distance between the origin and the floor, and r represents the distance between the origin and the image plane. The central point(m) represents the average value between the maximum y value from the floor and the minimum y value from the ceiling. If images are tilted to one direction, a revision process that divides the floor and the ceiling is required. Using the similar triangle equation, as shown in Figure 3, the depth value of the floor can be acquired as follows.

$$Z(Depth) = r \times \frac{d}{b} \tag{1}$$

The depth value of the floor and the ceiling from the segmented image and environment map can be acquired using the above equation. The depth value of the floor shows large variation in the regions near the center of the image and show little change in areas far away, which can be classified as being non-linear. Therefore, a revision process is required to make the floor show a linear change.

The depth calculation of the surrounding objects. The floor and ceiling are assumed to be perpendicular to the surrounding objects. Therefore, it is necessary to find out the plane close to the surrounding objects to acquire the depth value of the surrounding objects. Such adjacent relationship can be acquired through the segmented image. To acquire the depth value of the surrounding objects, the image is scanned from the center of the image to vertical direction. As for the surrounding objects near the floor, the maximum depth value of the floor extracted from vertical direction is obtained. The objects near

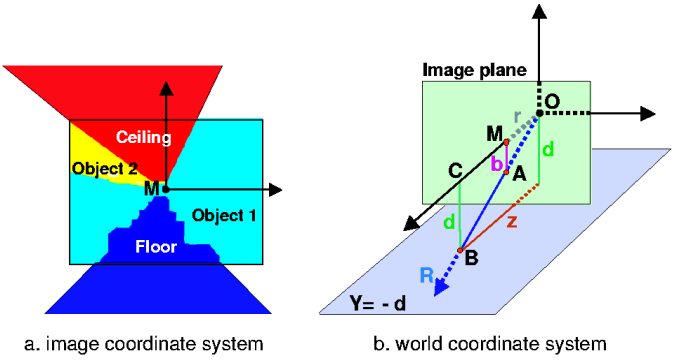


Fig. 3. The calculation of depth value(floor)

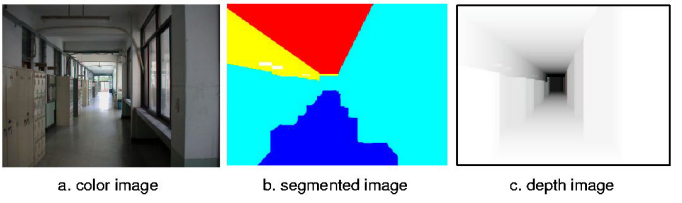


Fig. 4. The process of the depth calculation(hallway)

the ceiling are set as the maximum depth value of the ceiling. To set the next depth value, it is incremented to horizontal direction by the interval between pixels and the previous step is repeated. The whole depth values can be acquired through this process. The depth value cannot be acquired precisely with the assumption alone. For more precision, the surrounding objects should be segmented more thoroughly. Therefore, some other assumptions should be made.

- the relationship between the surrounding objects.
- the relationship between the floor(ceiling) and the surrounding objects.
- the relationship between the image plane and the surrounding objects.

A more specific depth value can be obtained by taking the surroundings into account. Figure 4 shows the process of extracting the depth value through the segmentation of the image. The image is segmented into a floor, ceiling and surrounding objects as shown in Figure 4-a. In this image, the surrounding objects are divided into the object(magenta) adjacent to the floor(blue), the object(yellow) adjacent to the ceiling(red), the object(white) perpendicular to the image plane(Figure 4-b). Figure 4-c shows the depth image by using the above method from the segmented image.

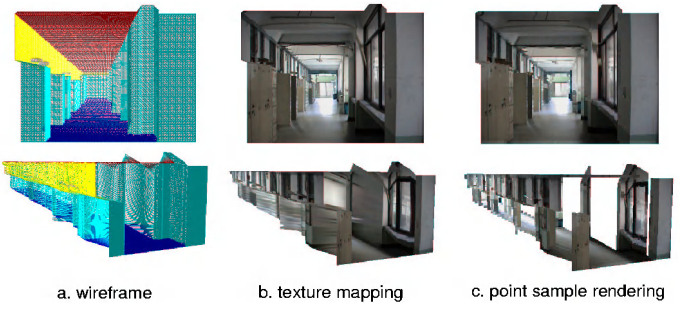


Fig. 5. The environment mapping of the environment model

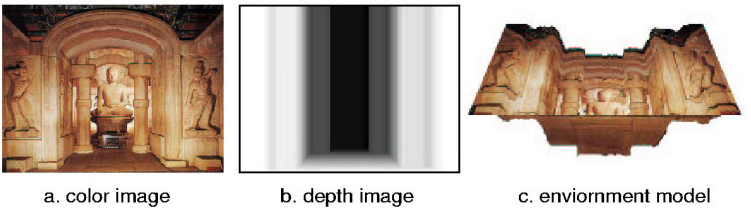


Fig. 6. Segmentation based environment model from a regular image

3.3 Environment Mapping

Environment mapping method is required for a more realistic representation. Ray casting(texel sampling) method, texture mapping method and point sampling methods have been applied for the environment mapping. Ray casting method indicates the method of directly sampling a pixel value from the environment map to represent the scene. The point sampling method and the bilinear interpolation method have been used in this paper. The texture mapping method selects the texels from the regular image or environment map for the environment mapping. We have used OpenGL for the mapping of the environment map onto the 3D environment model. Finally, the point sample rendering method is a method of using points as the most primitive unit for the rendering. This method does not triangulate the points in the environment but rather represent the points directly. Figure 5 is an image that represented the environment model using the texture mapping method and the point sample rendering method. Stretching effect of the texture occurs when parts of the object that was hidden due to the occlusion appears as shown in Figure 5-b. Also, in the point sample rendering method, the hole appears on the image due to the lack of data(Figure 5-c). Such problems can be overcome by through registration of the environment models and integrating them using multiple images.

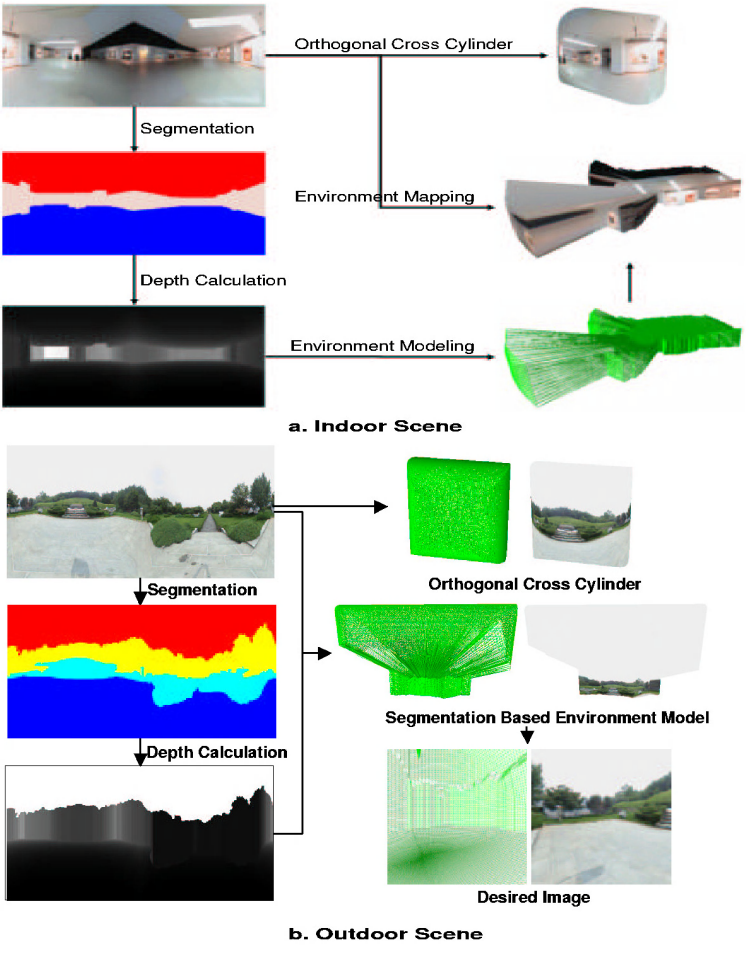


Fig. 7. Segmentation based environment model from an environment map

4 Conclusion and Future Works

This paper suggests segmentation based environment modeling method using the parallel/perpendicular of the plane for constructing the real-time image-based view-rendering. This method can easily be implemented on an environment map and makes the environment modeling easier through extracting the depth value by the image segmentation. It becomes possible to develop an environment modeling system with a full-view through this method.

Figure 6 is an image that represented the environment model through a regular image. Figure 6-c represent the model of the temple(seokkulam) using the color(Figure 6-a) and depth image(Figure 6-b). Figure 7 shows the process of reconstructing the environment model using a single environment map. Figure

7-a is an image that represented the environment model of the art gallery for indoor navigation. In case of outdoor scene(Figure 7-b), the depth value of the sky can acquire from the assumption (the sky consists of a hemisphere). As shown in these figures, the 3-dimensional environment model can easily be constructed from the image segmentation method.

However, an environment model that is constructed using a single image has a problem of the blurring-image effect caused by the fixed resolution and the stretching effect of the 3D model caused when an information that does not exist on the environment map occurs by the occlusion. To resolve this problem, a 3D environment model based on a multiple image is required. Also, segmentation based environment modeling method has an advantage in that it is easy to calculate the depth value of the image by segmenting the image into a floor, ceiling and surrounding objects, but also a disadvantage in that the only images whose viewpoint and the floor are perpendicular may be used. To apply any type of general image, the input images should be reconstructed by the relationship between the camera and the floor. Also, it must be made possible to acquire the depth value for a plane which is not perpendicular or parallel to the established plane.

References

1. Youichi Horry, Ken-Ichi Anjyo, Kiyoshi Arai, "Tour Into the Picture: Using a Spidery Mesh Inter-face to Make Animation from a Single Image", SIGGRAPH '97, pp 225-232, 1997
2. A. Criminisi, I. Reid, A. Zisserman, "Single View Metrology", Int. J. of Computer Vision, v.40, n.2, pp 123-148, 2000
3. P. Sturm and S. Maybank. "A method for interactive 3d reconstruction of piecewise planar objects from single images", In British Machine Vision, pp 265-274, 1999.
4. Byong Mok Oh, Max Chen, Julie Dorsey, Fredo Durand, "Image-based modeling and photo editing", SIGGRAPH 2001: 433-442
5. L. Zhang, G. Dugas-Phocion, J. S. Samson, and S. M. Seitz, "Single View Modeling of Free-Form Scenes", Proc. CVPR 2001
6. Heung-Yeung Shum, Mei Han, Richard Szeliski. "Interactive Construction of 3D Models from Panoramic Mosaics", CVPR 1998: 427-433.
7. Sing Bing Kang and R. Szeliski, "3-D scene data recovery using omnidirectional multibaseline stereo", CVPR 1996: 364-370.
8. W. Skarbek and A. Koschan, "Colour Image Segmentation - A Survey", Technical Report 94-32, Technical University of Berlin, 1994
9. E. N. Mortensen, L. J. Reese, and W. A. Barrett, "Intelligent Selection Tools", CVPR 2000: 776-777.
10. SeungTaek Ryoo. KyungHyun Yoon, "FullView Panoramic Navigation Using Orthogonal Cross Cylinder", Journal of WSCG, pp. 381-388, Feb 2002

Unsupervised Color-Texture Segmentation

Yuzhong Wang, Jie Yang, and Yue Zhou

Institute of Image Processing and Pattern Recognition,
Shanghai Jiaotong University, Shanghai 200030, P.R. China
{oliverwang, jieyang, zhouyue}@sjtu.edu.cn

Abstract. An improved approach for JSEG is presented for unsupervised color image segmentation. Instead of color quantization, an automatic classification method based on adaptive mean shift (AMS) based clustering is used for nonparametric clustering of image data set. The clustering results are used to construct Gaussian mixture modelling of image data set for the calculation of soft *J* value. The region growing algorithm used in JSEG is then applied in segmenting the image based on the multiscale soft *J*-images. Experiments show that the synergism of JSEG and the soft classification based on GMM overcomes the limitations of JSEG successfully and is more robust.

1 Introduction

Color image segmentation is useful in many applications. From the segmentation results, it is possible to identify regions of interest and objects in the scene. A variety of techniques have been proposed, for example: stochastic model based approaches [1], [4], [9], morphological watershed based region growing [11], energy diffusion [10], and graph partitioning [7]. However, due to the difficult nature of the problem, there are few automatic algorithms that can work well on a large variety of data.

The problem of segmentation is difficult because of image texture. In the reality, natural scenes are rich in color and texture. It is difficult to identify image regions containing color-texture patterns. Yining Deng [2] proposed a new approach called JSEG which can be used to segment images into homogeneous color-texture regions.

The basic idea of the JSEG method is to separate the segmentation process into two stages, color quantization and spatial segmentation. In the first stage, colors in the image are quantized to several representative classes that can be used to differentiate regions in the image. This quantization is performed in the color space without considering the spatial distributions of the colors. Then, the image pixel values are replaced by their corresponding color class labels, thus forming a class-map of the image. The class-map can be viewed as a special kind of texture composition. In the second stage, spatial segmentation is performed directly on this class-map without considering the corresponding pixel color similarity.

A criterion for "good" segmentation using spatial data points in class-map is proposed in JSEG. Let Z be the set of all N data points in a class-map. Let $z = (x, y), z \in Z$, and m be the mean,

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 106–113, 2004.
© Springer-Verlag Berlin Heidelberg 2004

$$m = \frac{1}{N}\sum_{z\in Z} z \tag{1}$$

suppose Z is classified into C classes, $Z_i, i = 1,...,C.$ Let m_i be the mean of the N_i data points of class Z_i,

$$m_i = \frac{1}{N_i}\sum_{z\in Z_i} z \tag{2}$$

Let

$$S_T = \sum_{z\in Z} \|z - m\|^2 \tag{3}$$

and

$$S_W = \sum_{i=1}^{C} S_i = \sum_{i=1}^{C}\sum_{z\in Z_i} \|z - m_i\|^2 \tag{4}$$

S_W is the total variance of points belonging to the same class. Define

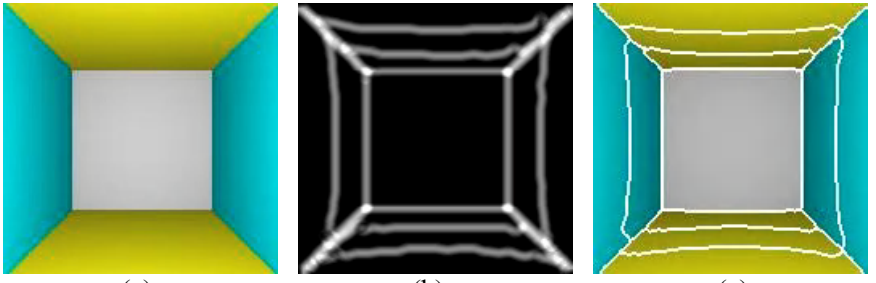
$$J = (S_T - S_W)/S_W \tag{5}$$

However, JSEG has two limitations which will affect the segmentation results. One is caused by color quantization parameter which determines the minimum distance between two quantized colors, and the quantization results directly influences the segmentation results. A good parameter value yields the minimum number of colors necessary to separate two regions. However, it's very difficult to select a good parameter. Another is caused by the varying shades due to the illumination. The problem is difficult to handle because, in many cases, not only the illuminant component, but also the color components of a pixel, change their values due to the spatially varying illumination [2] and this problem usually cause oversegmentation.

A synthetic image, its *J*-image at scale 2 and segmentation result are shown in Fig.1 (quantization parameter is 200, scale threshold is 2 and region merge threshold is 0.4). In the synthetic image, there is phenomenon of colors smooth transition in yellow and blue regions. Therefore, it's impossible to avoid oversegmentation no matter what quantization parameter is selected. This simple example well shows the limitations of JSEG.

In this paper, a new approach is presented to improve JSEG. First, we use (AMS) based clustering to finish color classification instead of using original color quantization algorithm. By this classification method, image data can be divided into appropriate clusters automatically, so the adaptability of JSEG without quantization parameter is improved. Second, enlightened from segmentations based on fuzzy theories, we make an assumption that colors distributions in the image obey Gaussian mixture modeling (GMM), and calculate soft *J* values to construct soft *J*- image using GMM. This can effectively restrain the oversegmentation in those regions with color

smooth transition. Experiments show that the synergism of JSEG and the soft classification based on GMM is successful.



(a) (b) (c)

Fig. 1. (a) A synthetic image. (b) *J*-image at scale 2. (c) Result after segmentation at scale 2.

2 The Improved Method Based on GMM

To overcome the first limitation of JSEG, a nonparametric clustering based on AMS is used for colors quantization. While to overcome the second limitation, GMM of the image data set, instead of the hard classification, is constructed by using the results of AMS clustering for labeling every pixel.

2.1 Adaptive Mean Shift-Based Clustering

Here we only review some of the results described in [3] which should be consulted for the details.

Assume that each data point $x_i \in R^d, i = 1,...,n$ is associated with a bandwidth value $h_i > 0$. The sample point estimator

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^d}k\left(\left\|\frac{x-x_i}{h_i}\right\|^2\right) \tag{6}$$

based on a spherically symmetric kernel K with bounded support satisfying

$$K(x) = c_{k,d}k\left(\|x\|^2\right) > 0 \qquad \|x\| \le 1 \tag{7}$$

is an adaptive nonparametric estimator of the density at location x in the feature space. The function $k(x), 0 \le x \le 1$ is called the profile of the kernel, and the normalization constant $c_{k,d}$ assures that $K(x)$ integrates to one. The function $g(x) = k'(x)$ can

always be defined when the derivative of the kernel profile $k(x)$ exists. Using $g(x)$ as the profile, the kernel $G(x)$ is defined as $G(x) = c_{g,d}\, g\left(\|x\|^2\right)$.

By taking the gradient of (6) the following property can be proven

$$m_G(x) = C\,\frac{\hat{\nabla} f_K(x)}{\hat{f}_G(x)} \tag{8}$$

Where *C* is a positive constant and

$$m_G(x) = \frac{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}}\, x_i\, g\left(\left\|\frac{x - x_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}}\, g\left(\left\|\frac{x - x_i}{h_i}\right\|^2\right)} \tag{9}$$

is called the mean shift vector. The expression (8) shows that at location *x* the weighted mean of the data points selected with kernel *G* is proportional to the normalized density gradient estimate obtained with kernel *K*. The mean shift vector thus points toward the direction of maximum increase in the density. The implication of the mean shift property is that the iterative procedure

$$y_{j+1} = \frac{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}}\, x_i\, g\left(\left\|\frac{y_j - x_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}}\, g\left(\left\|\frac{y_j - x_i}{h_i}\right\|^2\right)} \qquad j = 1,2,... \tag{10}$$

is a hill climbing technique to the nearest stationary point of the density, i.e., a point in which the density gradient vanishes. The initial position of the kernel, the starting point of the procedure y_1 can be chosen as one of the data points x_i. Most often the points of convergence of the iterative procedure are the modes (local maxima) of the density.

There are numerous methods described in the statistical literature to define h_i, the bandwidth values associated with the data points, most of which use a pilot density estimate. For computational reasons, the simplest way to obtain the pilot density estimate is by nearest neighbors [8]. Let $x_{i,k}$ be the *k*-nearest neighbor of the point x_i. Then, we take $h_i = \left\|x_i - x_{i,k}\right\|_1$. In [5], an approximation technique, locality-sensitive hashing (LSH), was employed to reduce the computational complexity of AMS and we can call this fast algorithm as Fast Adaptive Mean Shift (FAMS) procedure, and the selection of *k* was proved flexible.

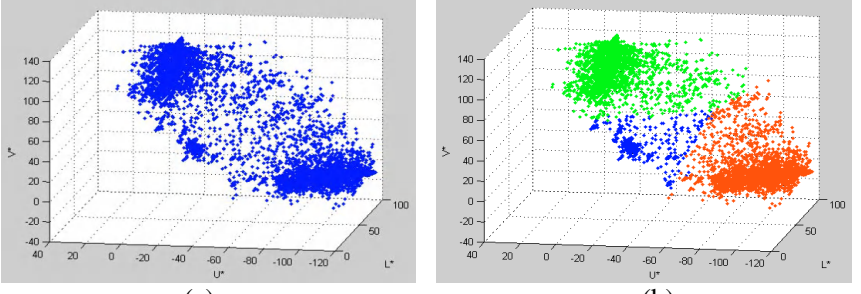
AMS clustering is employed to classify color image data. Images are usually stored and displayed in the *RGB* space. However, to ensure the isotropy of the feature space,

a uniform color space with the perceived color differences measured by Euclidean distances should be used. We have chosen the *L*U*V** space, whose coordinates are related to *RGB* values by nonlinear transformations，thus allowing the use of spherical windows [6]. We assume image data obey GMM in *L*U*V** space, so we employ the multivariate normal kernel

$$K(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|x\|^2\right) \tag{11}$$

in AMS procedure. In practical applications, we select *k* equal 500 and employ FAMS procedure. Convergence is declared when the magnitude of the shift becomes less than 0.1.

Fig.2 shows the colors distribution of synthetic image shown in Fig.1 in *L*U*V** color space and its colors classification result using FAMS clustering procedure. Visually the synthetic image should be classified into three color classes and it does be decomposed into three clusters with FAMS clustering procedure.



(a) (b)

Fig. 2. (a) Colors distribution of the synthetic image in *L*U*V** color space. (b) Corresponding clustering result using FAMS clustering procedure.

2.2 Soft *J* Value with GMM

Suppose $\{I_k\}, k = 1,..., N$ is the set of all pixels of the color image $I(x, y)$, and I_k obey Gaussian mixture distribution of C classifications. Mark sub-Gaussian distribution as $\omega_i, i = 1,..., C$. Then, the statistical distribution $p(I_k)$ of I_k can be approximately expressed with Gaussian mixture modelling of C classes, and the probability density function of every subsidiary Gaussian distribution ω_i can be expressed as following

$$p(I_k \mid \omega_i, \theta_i) = \frac{1}{(2\pi)^{\frac{3}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(I_k - \mu_i)^T \Sigma_i^{-1}(I_k - \mu_i)\right\} \quad i = 1,...,C \tag{12}$$

$\theta_i = (u_i, \Sigma_i)$ denotes the parameters of Gaussian mixture modelling, and μ_i is the mean and Σ_i is the covariance matrix; the prior probability of is $P(\omega_i)$. μ_i and Σ_i

can be calculated with the data belonged to the *i*th class and $P(\omega_i)$ is the ratio of the number of pixels of the *i*th class to total number of pixels.

Then we can calculate every pixel's membership ($\mu_{I_k,j}(k=1,...,N, i=1,...,C)$) of every class with Bayesian equation. After finishing calculation of pixel's membership, we redefine the calculation of *J* value, letting *Z* be the set of all *N* data points in a class-map and $z=(x,y), z \in Z$. Suppose image data set is classified into *C* classes. Equations (1) 、(3) and (5) needn't to be changed. Modify equation (2) as following

$$m_i = \frac{\sum_{z \in Z} z \cdot \mu_{z,i}}{\sum_{z \in Z} \mu_{z,i}} \qquad i = 1,..., C \tag{13}$$

and modify equation (4) as following

$$S_W = \sum_{i=1}^{C} S_i = \sum_{i=1}^{C} \sum_{z \in Z} \left(\mu_{z,i} \left\| z - m_i \right\|^2 \right) \tag{14}$$

(a) (b)

Fig. 3. (a) Soft J-image at scale 2 of the synthetic image. (b) Corresponding segmentation result.

Then, the *J* value calculated with new rules is called soft *J* value, and the new *J*-image constructed by soft *J* values is called soft *J*-image. The second limitation can be overcome by using region growing in soft *J*-image.

Soft *J*-image of the synthetic image and corresponding segmentation result are shown in Fig.3. The experimental results prove that the improved method overcomes the limitations of JSEG successfully.

3 Experimental Results

The improved algorithm is tested on a variety of images. Generally speaking, the new method looks more robust than JSEG.

Fig.4 shows three examples. The parameters used in JSEG are the same as those used in the simple example shown in Fig.1 and the scale threshold and the region merging threshold used in our method also adopt the same values. The results of Fig.4 (a) and (d) obtained from our method are obviously better than those obtained from JSEG. However, the result of Fig.4 (g) obtained from JSEG is similar to the result from our method. This can be explained as that the set of parameters is right suitable for Fig.4 (g); in another word, it indicates that our method has outstanding adaptability.

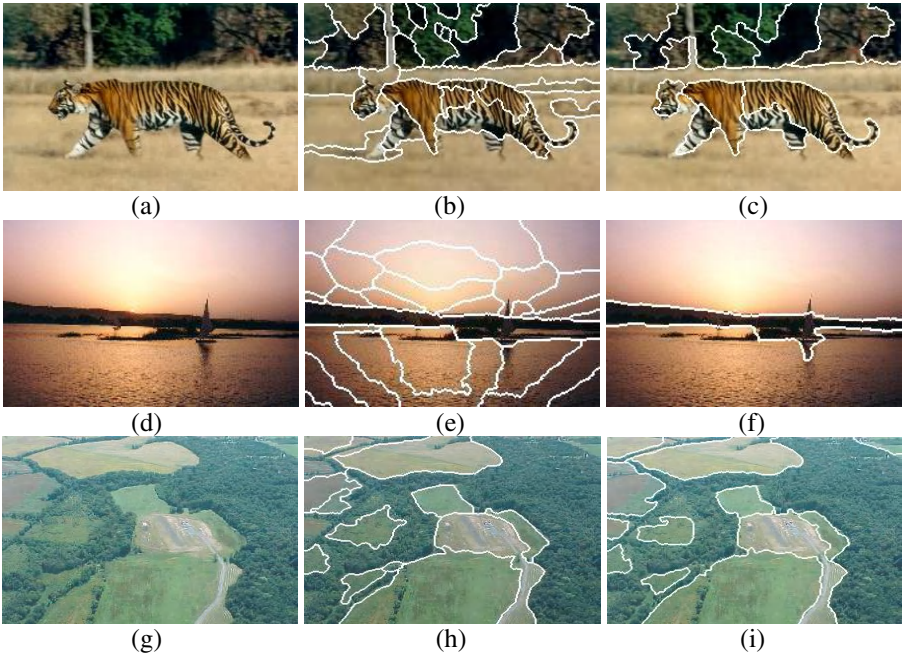


Fig. 4. The original images are shown in the left column while the results from JSEG in the middle column, and the results from our method in the right column.

4 Conclusions

In this work, an improved approach for JSEG is presented for the fully unsupervised segmentation color-texture regions in color images. An automatic classification method based on FAMS clustering is used for nonparametric clustering of image data set. GMM of image data constructed with classifications achieved by FAMS clustering procedure is applied in the calculation of soft J value.

If we want to get good results by JSEG, the parameters used in JSEG must be adjusted repeatedly. Fortunately, the influence of scale threshold and region merging threshold are much less than quantization threshold. Therefore, the selection of quantization threshold degrades efficiency in practical application to a great extent. Repeated selecting quantization threshold will exhaust users and is forbidden in

automatic systems. In the traditional clustering techniques, we know, the feature space is usually modeled as a mixture of multivariate normal distributions, which can introduce severe artifacts due to the elliptical shape imposed over the clusters or due to an error in determining their number. However, the AMS based nonparametric feature space analysis eliminates these artifacts. Therefore, GMM constructed from the results obtained by AMS based clustering method is consequentially more exact.

Experiments show the new method overcomes the limitations of JSEG successfully and is more robust. Excellent adaptability and flexibility of the improved method make it more applicable in practical systems.

References

1. Belongie, S. Carson, C. et. al.: Color- and texture-based image segmentation using EM and its application to content-based image retrieval. Proc. of ICCV. (1998) 675-682
2. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-texture Regions In Images and Video. IEEE Trans. PAMI. 8 (2001) 800-810
3. Comaniciu, D.: An Algorithm for Data-Driven Bandwidth Selection. IEEE Trans. PAMI. 2 (2003) 281-288
4. Delignon, Y., Marzouki, A. et. al.: Estimation of generalized mixtures and its application in image segmentation. IEEE Trans. Image Processing. 6 (1997) 1364-1376
5. Georgescu, B., Shimshoni, I., Meer, P.: Mean Shift Based Clustering in High Dimensions: A Texture Classification example. Proc ninth Int'l Conf. Computer Vision. (2003) 456-463
6. D. Comaniciu, P. Meer: Robust Analysis of Feature Spaces: Color Image Segmentation. IEEE Proc. CVPR. (1997) 750-755
7. Shi, J. Malik, J.: Normalized cuts and image segmentation. Proc. of CVPR. (1997) 731-737
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2001)
9. Wang, J.-P.: Stochastic relaxation on partitions with connected components and its application to image segmentation. IEEE Trans. PAMI. 6 (1998) 619-636
10. Ma, W.Y., Manjunath, B.S: Edge flow: a framework of boundary detection and image segmentation. Proc. of CVPR. (1997) 744-749
11. Shafarenko, L., Petrou, M., Kittler, J.: Automatic watershed segmentation of randomly textured color images. IEEE Trans. Image Processing. 11 (1997) 1530-1544

Hierarchical MCMC Sampling

Paul Fieguth

Department of Systems Design Engineering
University of Waterloo
Waterloo, Ontario, Canada
pfieguth@uwaterloo.ca

Abstract. We maintain that the analysis and synthesis of random fields is much faster in a hierarchical setting. In particular, complicated long-range interactions at a fine scale become progressively more local (and therefore more efficient) at coarser levels. The key to effective coarse-scale activity is the proper model definition at those scales. This can be difficult for locally-coupled models such as Ising, but is inherent and easy for those models, commonly used in porous media, which express constraints in terms of lengths and areas.
Whereas past methods, using hierarchical random fields for image estimation and segmentation, saw only limited improvements, we find reductions in computational complexity of two or more orders of magnitude, enabling the investigation of models at much greater sizes and resolutions.

Keywords: Posterior sampling, MCMC methods, Hierarchical sampling, Porous media, Ising, Random Fields, Energy minimization

1 Introduction

The cure to arthritis and collapsing buildings lies in the fast random sampling of large images!

A trifle optimistic, to be sure, however drug delivery in cartilage and the cracking of concrete both rely on a detailed understanding of *porous media* [8, 10], and thus a corresponding need to model, generate, and manipulate large stochastic 2D images and 3D volumes.

As motivated by Figure 1, we seek hierarchical approaches to modelling and computation, specifically for two reasons: first, for those media which are inherently multi-scale (concrete, for example, has pore sizes ranging from sub-micron to millimetre) and, secondly, to more effectively model those non-local relationships on the finest scale, but which become progressively more local (and simple) on coarser levels.

To be sure, a variety of hierarchical [1,2,5,6,7] and region-based [9] methods exist, however they differ from our current context in a few ways:

– Most methods, certainly among those in the image processing literature, are designed for estimation [1,2,6,7] (thus guided/driven by measurements), not random sampling (purely model based).

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 114–121, 2004.
© Springer-Verlag Berlin Heidelberg 2004

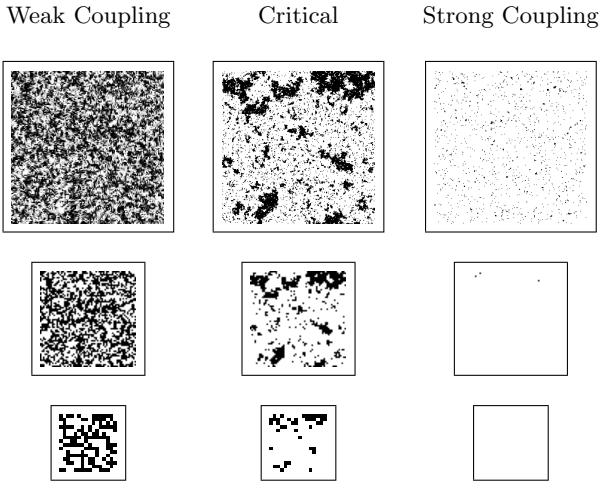


Fig. 1. The analysis and synthesis of random fields is easier and faster in a hierarchical setting. For Ising models, weakly and strongly coupled models become progressively random and uniform, respectively, at coarser scales [11]. Although critically-coupled structures do not simplify with scale, there are computational benefits: the long-range interactions implied by a large cluster at a fine scale become progressively more local (and therefore more efficient).

– In most cases there is a certain ambiguity or arbitrariness in the selection of coarse-scale models.

In many cases, the coarse scales served only as a weak regularizer for a densely-measured, well-conditioned fine-scale estimation problem, and led to only marginal improvements. We are interested in problems involving sparse or no measurements, in which case the finest scale is very poorly conditioned, and where the coarse scales, if appropriately defined, have a great deal to offer. In the following three sections will examine three different models which, although greatly simplified, are representative of models used in studies of porous media. A hierarchical approach will, in each case, lead to computational improvements at or exceeding two orders of magnitude.

2 Ising Model

We first look at the well-known Ising model [3] — very widely studied — in which the elements of a binary field x interact with their immediate neighbours:

$$E = \beta \sum_{ij} x_{i,j} x_{i+1,j} + \beta \sum_{ij} x_{i,j} x_{i,j+1}. \tag{1}$$

The coupling β controls the degree to which adjacent x_{ij} should be the same, and as β increases so does the inter-pixel correlation (Figure 3).

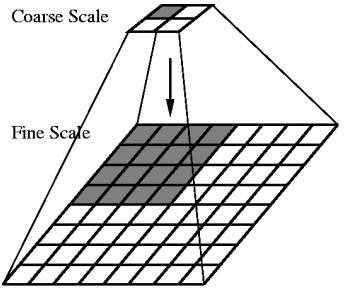


Fig. 2. A standard model of hierarchical random fields: to evaluate the energy of a coarse-scale field, we project it to the finest scale and evaluate the energy there.

For small β the sampled field is essentially random, and all samplers, whether flat or hierarchical, converge quickly. However as β increases the structures present in x grow larger, and the longer-range relationships become increasingly difficult to deduce from a local model, thus hierarchical methods begin to outperform (Figure 4). The hierarchical samplers proceed from coarse to fine, in a single pass, sampling for some iterations at each scale.

The key challenge relates to the definition of coarse-scale models. The coarsification of (1) is not obvious [5], and is most often circumvented by defining coarse-scale models implicitly in terms of the finest-scale by projection [6] (Figure 2). In the Ising model this implies the widely-used model $\beta_s = 2^s\beta$ at s scales above the finest.

The problem is that this is wrong. Figure 3 makes it clear that for small β the coupling should *decrease* with scale. Using $\beta_s = 2^s\beta$ leads to stiff, large-scale structures created on coarse scales which then need to be undone at finer scales. If a properly-renormalized model (here derived experimentally) is used, with the correct value of β at each scale, then the sampler at each scale needs only to insert the details unresolvable at the coarser scale, a *much* easier task than undoing incorrect structure, and thus converging *far* faster, as seen in Figure 4.

The key is the production of properly renormalized coarse-scale models.

3 Correlation Model

Although effective as an illustration, the familiar Ising model is arguably not a good representation of cartilage, concrete, or other porous media, nor was the Ising coarse-scale model easily found.

The key difficulty was to infer, from a local, interpixel model such as Ising, the interpixel relationships on coarser scales. Instead, common stochastic porous media models [10] often involve concepts of correlation, chord-length, or area — all of which are *nonlocal* constraints on the finest scale, and thus rescale almost trivially, since it is relatively easy to express a correlation, for example, between coarse pixels, on the basis of a stipulated correlation model.

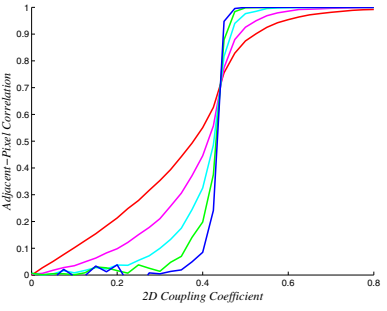


Fig. 3. The correlation between adjacent pixels as a function of the coupling β and scale. Coarser scales (blue) are progressively random or uniform, depending on the value of β relative to criticality (at $\beta \approx 0.44$).

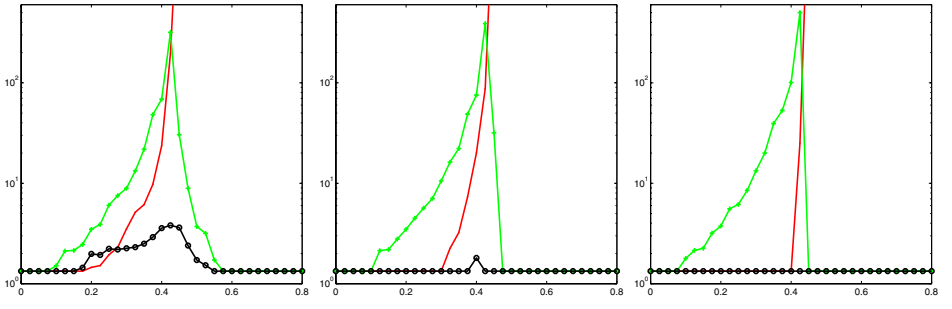


Fig. 4. The convergence of a flat (red), standard hierarchy (green), and renormalized hierarchical model (black) as a function of coupling, for three different scales. Around criticality, where structures are most complex, the properly renormalized model converges more than 100 times faster.

Consider, for example, the correlation structure of Figure 5: we seek regions on the order of 10 to 20 pixels in size. We can sample from such a model by computing the empirical correlation of a sample x, and accepting/rejecting a pixel change on the basis of the degree of improvement in the empirical correlation toward the ideal. The model rescales inherently, since a coarsification of x is simply equivalent to rescaling the horizontal axis of the correlation plot in Figure 5.

Since the chosen correlation of Figure 5 is arbitrary, the pictures which result are not of any particular importance, however the convergence results of Figure 6 are significant, where the degree of fit between a desired correlation $c(d)$ and an average, empirical correlation $\hat{c}(d)$ is chosen to be

$$\|c - \hat{c}\| = \sum_{d} \frac{|c(d) - \hat{c}(d)|}{d} \tag{2}$$

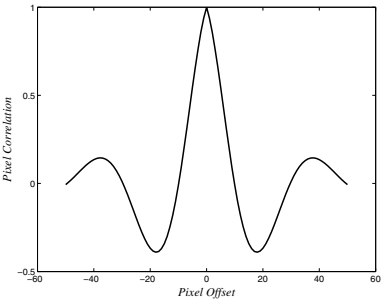


Fig. 5. The asserted correlation function: we seek random fields which are correlated out to a typical size of 10 pixels, and negatively correlated beyond that, out to roughly 30 pixels. The model rescales trivially, since this involves only an axis rescaling.

where d is the spatial offset, measured in pixels, and where division by d de-emphasizes correlation at long ranges.

When the coupling between sample field x and the desired correlation is weak, then x is essentially random and local, devoid of large-scale structure, and is easily sampled by a flat model. However as the coupling increases, large-scale structures begin to appear, and the flat-model convergence time increases rapidly. This slowing down is well understood; in the Ising case this is a critical phenomenon, related to the random-walk nature of information propagation in the pixellated lattice, and where the number of iterations to produce large regions grows very rapidly with region size. In this correlation model, the walk is not quite random, however the problem is analogous: the criterion at the pixel level does not provide strong guidance in the formation of large-scale regions, and so a random sampler wastes a great deal of time on trial-and-error before finding improved solutions.

With strong coupling, since the sample fields consist of relatively large regions, much of the desired structure can be realized at coarse scales where there are fewer pixels, where the regions are smaller, and where iterations proceed much more rapidly. As before, only the details of the region boundaries remain to be refined at finer scales, rather than the induction of entire regions at the finest scale of a flat model, leading to one or two orders of magnitude improvement.

4 Multi-scale Porous Media

The previous section introduced a correlation model, which scales inherently, however most of the structure lives on a single scale (roughly 10-20 pixels in size). A persuasive example of a porous medium needs to be truly multi-scale in nature, containing both large and small structure.

A final example is based on the criteria of Figure 7: we permit tiny regions (0% – 1% of domain), embedded in larger regions (5% – 20% of domain), in

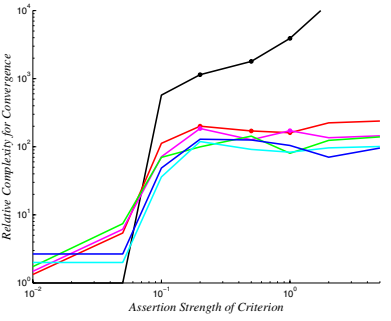


Fig. 6. Convergence time as a function of coupling strength. For weakly-coupled (nearly random) models, left, a flat sampler is ideal, however synthesizing larger structures becomes much more difficult. Asterisks (*) denote extrapolated convergence time for cases which failed to converge. Colour indicates change in iterations with scale (Red: fewer iterations, Blue: more iterations, Black: flat model, no iterations at coarse scales).

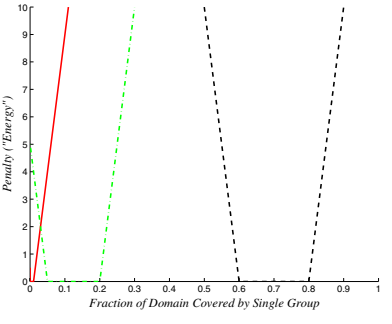


Fig. 7. Criteria for a true, multi-scale porous-media model: the lines plot penalty as a function of region size. The model consists of small pores (red) embedded in large regions (green) embedded in a background medium.

turn embedded in a background. To prevent the growth of long, tendril-like regions, there is a constraint limiting region shape (by penalizing the ratio of region perimeter to square-root of area). The model is trivially renormalized across scale, since all parameters are expressed in terms of lengths and areas. The model is also binary: the tiny pores and the background are of the same material.

The random sampling results, shown in Figure 8, follow a pattern by now familiar. For weakly-constrained problems (Figure 8 left), where the sample is mostly random, a flat sampler performs well. However as the constraints increase (right), and larger regions begin to appear the flat sampler fails completely, whereas the hierarchical sampler begins to produce very interesting, very credible samples, involving a wide range of structural scales.

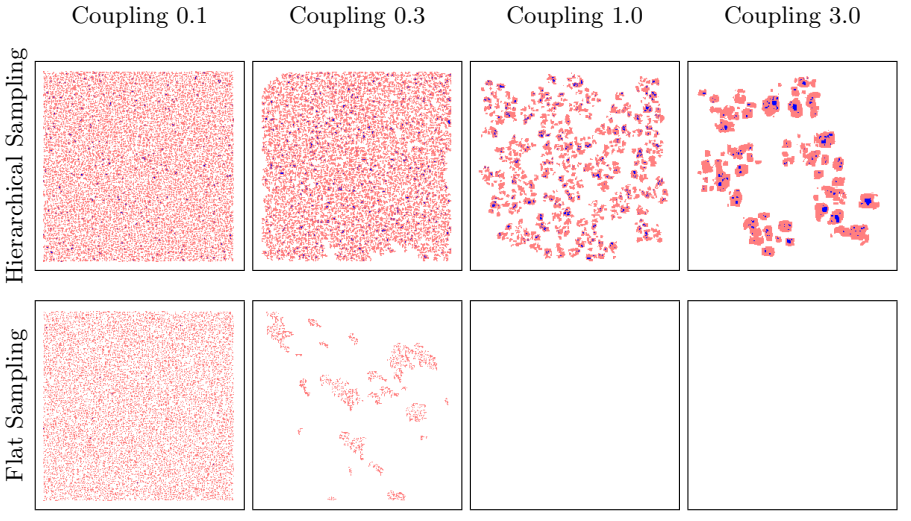


Fig. 8. Random samples from hierarchical and flat approaches, for four coupling strengths. Consistent with other experiments, the differences are minimal for weakly-coupled models (left), but are striking in strongly-coupled cases (right): the flat model fails to be able to synthesize large-scale structure, because of an inhibition on tiny regions. All cases were initialized with a flat, white field. Initializing with a random field had no effect on the hierarchical case, and produced very different, although no better, results for flat sampling.

In this case the computational benefit of the hierarchical approach is unmeasurable (that is, almost infinite), since the time constant to convergence for the flat sampler diverges, and may be infinite. The failure of flat sampling is easy to diagnose: there are strong barriers inhibiting the creation of small foreground regions within a uniform background. Initializing with a white field does not allow the production of local regions; initializing with random pixels does not allow a background to form. Only by initializing with a credible solution (small regions within larger regions on a background) can a flat sampler converge. A critic could charge that a redesign of the criteria in Figure 7 could solve this problem, however this leaves us with a flat sampler, sensitive to minor perturbations in the energy function, and still converging orders of magnitude slower than a hierarchical sampler.

5 Conclusions

It may be argued that flat samplers are not meant to perform well on strongly-coupled fields, that such problems are meant to be solved by annealing[4]. Indeed, our research in hierarchical sampling is driven by a long term interest in hierarchical annealing. However, we maintain that the best annealer is the one built around the best, fastest, and most robust sampler.

Although hierarchical annealing and sampling is not new, flat sampling and annealing methods are still widely practiced, certainly for porous media. In this paper we have clearly shown the degree of improvement available with hierarchical approaches, with only very modest algorithmic changes, and the importance of properly renormalizable or rescalable models. The improvement in computational complexity is very clearly seen to be the synthesis of large-scale structures at coarse scales, with only local details remaining to be refined at finer scales.

Motivated by the clear successes in this paper, our research foci are twofold: one, the development of less arbitrary or contrived models, more physically meaningful for a particular porous-media context, and secondly the development of annealing techniques and appropriate temperature schedules built around hierarchical samplers.

References

1. Charles Bouman and Bede Liu, *Multiple resolution segmentation of textured images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), no. 2, 99–113.
2. C. Bouman and M. Shapiro, *A multiscale random field model for Bayesian image segmentation*, IEEE Image Processing **3** (1994), no. 2, 162–177.
3. D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, 1987.
4. S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **6** (1984), 721–741.
5. Basilis Gidas, *A renormalization group approach to image processing problems*, IEEE Trans. PAMI **11** (1989), no. 2, 164–180.
6. Z. Kato, M. Berthod, and J. Zerubia, *A hierarchical Markov random field model* ..., GMIP **58** (1996)
7. Jan Puzicha and Joachim M. Buhmann, *Multiscale annealing for grouping and unsupervised texture segmentation*, CVIU **76** (1999), no. 3, 213–230.
8. Dietrich Stoyan and Helga Stoyan, *Fractals, random shapes and point fields*, J. Wiley, 1994.
9. R. H. Swendson and J. S. Wang, *Nonuniversal critical dynamics in Monte Carlo simulations*, Physical Review Letters **58** (1987), 86–88.
10. M. S. Talukdar, O. Torsaeter, and M. A. Ionnidis, *Stochastic reconstruction of particulate media from two-dimensional images*, Journal of Colloid and Interface Science **248** (2002), 419–428.
11. K. Wilson and J. Kogut, *The renormalization group and the ϵ-expansion*, Phys. Rep. **C12** (1974), 75–200.

Registration and Fusion of Blurred Images

Filip Sroubek and Jan Flusser

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 182 08, Praha 8
{sroubekf, fluser}@utia.cas.cz

Abstract. We present a maximum a posteriori solution to problems of accurate registration of blurred images and recovery of an original undegraded image. Our algorithm has the advantage that both tasks are performed simultaneously. An efficient implementation scheme of alternating minimizations is presented. A simulation and a real-data experiment demonstrate the superb performance of the algorithm.

1 Introduction

Imaging sensors and other devices have their physical limits and imperfections, therefore, an acquired image represents only a degraded version of the original scene. Two main categories of degradations are recognized: color (or brightness) degradations and geometric degradations. The former degradations are caused by such factors as wrong focus, motion of the scene, media turbulence, noise, and limited spatial and spectral resolution of the sensor; they usually result in image blurring. The latter degradations originate from the fact that each image is a 2-D projection of 3-D world. They cause deformations of object shapes and other spatial distortions of the image.

Since the geometric and color degradations are in principle inevitable in real applications, analysis and interpretation of degraded images represents the key problem. *Image fusion* provides a solution to this problem and consists of two steps. First the geometric deformations are removed by means of image registration, and second, the color (intensity) information is combined. If we can model the color deformation by convolution, the second step corresponds to a multichannel blind deconvolution (MBD) problem. In this paper, we address the problem of registration of blurred images (channels) from the perspective of image fusion.

Image registration is a process of transforming two or more images into a geometrically equivalent form. It eliminates degradation effects caused by the geometric distortion. For images which are not blurred, the registration has been extensively studied in the recent literature (see [1] for a survey). However, blurred images require special registration techniques. They can be, as well as the general-purpose registration methods, divided into two groups – global and landmark-based techniques. Regardless of the particular technique, all feature extraction methods, similarity measures, and matching algorithms used in the registration process must be insensitive to image blurring.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 122–129, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Global methods do not search for particular landmarks in the images. They try to estimate directly the between-channel translation and rotation. Myles and Lobo [2] proposed an iterative method working well if a good initial estimate of transformation parameters is available. Zhang *et al.* [3] proposed to estimate the registration parameters by bringing the channels into a canonical form. Since blur-invariant moments were used to define the normalization constraints, neither the type nor the level of the blur influences the parameter estimation. Kubota *et al.* [4] proposed a two-stage registration method based on hierarchical matching, where the amount of blur is considered as another parameter of the search space. Zhang and Blum [5] proposed iterative multiscale registration based on optical flow estimation in each scale, claiming that optical flow estimation is robust to image blurring. All global methods require a considerable (or even complete) spatial overlap of the channels to yield reliable results, which is their major drawback.

Landmark-based blur-invariant registration methods have appeared very recently, just after the first paper on the moment-based blur-invariant features [6]. Originally, these features could only be used for registration of mutually shifted images. The proposal of their rotational-invariant version [7] in combination with a robust detector of salient points [8] led to registration methods that are able to handle blurred, shifted and rotated images [9].

Although the above-cited registration methods are very sophisticated and can be applied almost to all types of images, the result tends to be rarely perfect. The registration error is usually few pixels for blurred images. However, the second step of image fusion (MBD in our case) requires perfectly aligned channels. Current MBD methods, see e.g. [10,11,12,13], are not sufficiently robust to handle the registration error. In the field of super-resolution image reconstruction, a few techniques, such as [14], were proposed that simultaneously estimate motion vectors and reconstruct the image, but they are not blind or assume the same parametrized blur in the channels. The first step towards more robust MBD was given in [15], which deals with blind deconvolution of translated channels degraded by different but simple motion blurs.

In this paper, we propose a novel technique that can be applied after removing large between-channel misregistrations and which then performs image fusion in one step, i.e. fine registration and blind deconvolution simultaneously. Image blurring can be arbitrary and is unknown, while the geometric misregistrations are supposed to be also unknown but limited to "small" translations. In the next section, we formulate the solution as a maximum *a posteriori* estimator (MAP) and use an alternating minimization (AM) algorithm to find the solution. We derive *a priori* probabilities of the original image and the blurs from properties of bounded variation functions and the multichannel framework, respectively. Experimental results are given in Section 3. Finally, Section 4 concludes the paper.

2 MAP Analysis

Let us assume that the k-th acquired image (channel) z_k can be modelled by blurring the "ideal" image u and shifting the result by few pixels

$$z_k(x + a_k, y + b_k) = (u * h_k)(x, y) + n_k(x, y), \tag{1}$$

where vector $(a_k, b_k) = t_k$ represents the unknown translation of the k–th channel, h_k is the unknown blur mask with a characteristic of a low-pass filter, and n_k denotes additive noise. In the discrete domain, this degradation model takes the form:

$$\mathbf{z}_k = \mathbf{T}_k \mathbf{H}_k \mathbf{u} + \mathbf{n}_k \,, k = 1, \ldots, K,$$

where $\mathbf{z}_k$, $\mathbf{u}$, and $\mathbf{n}_k$ are discrete lexicographically ordered equivalents of the image functions z_k, u, and n_k, respectively. $\mathbf{T}_k$ is a translation operator shifting the image by t_k pixels, i.e. a linear filter with the delta function at the position t_k. One can readily see that the matrix product $\mathbf{T}_k \mathbf{H}_k = \mathbf{G}_k$ defines convolution with a mask $\mathbf{g}_k$ that is a shifted version of a mask $\mathbf{h}_k$ (discrete representation of h_k). This degradation model closely resembles the model used in super resolution except that a subsampling operator is not present in our case. By concatenating the channels, the previous equation can be rewritten in two equivalent forms

$$\mathbf{z} = \mathbf{G}\mathbf{u} + \mathbf{n} = \mathbf{U}\mathbf{g} + \mathbf{n} \,, \tag{2}$$

where $\mathbf{z} \equiv [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T$, $\mathbf{G} \equiv [\mathbf{G}_1^T, \ldots, \mathbf{G}_K^T]^T$, $\mathbf{n} \equiv [\mathbf{n}_1^T, \ldots, \mathbf{n}_K^T]^T$, $\mathbf{g} \equiv [\mathbf{g}_1^T, \ldots, \mathbf{g}_K^T]^T$, and $\mathbf{U}$ is a block-diagonal matrix with K blocks each performing convolution with the image $\mathbf{u}$.

We adopt a stochastic approach and follow the MAP formulation proposed in our previous work [16]. The conditional pdf $p(\mathbf{z}|\mathbf{u}, \mathbf{g})$ follows from (2) and from our assumption of white Gaussian noise, i.e.

$$p(\mathbf{z}|\mathbf{u}, \mathbf{g}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{z} - \mathbf{G}\mathbf{u})^T \Sigma^{-1} (\mathbf{z} - \mathbf{G}\mathbf{u}) \right\},$$

where Σ is the noise diagonal covariance matrix with $\{\sigma_k^2\}_{k=1}^K$ on the corresponding positions on the main diagonal. If the same noise variance σ^2 is assumed in each channel, Σ^{-1} reduces to a scalar σ^{-2}.

A general model for the prior distribution $p(u)$ is a Markov random field which is characterized by its Gibbs distribution given by $p(u) \propto \exp(-F(u)/\lambda)$, where λ is a constant and F is called the *energy function*. One can find various forms of the energy function in the literature, however, the most promising results have been achieved for variational integrals. The energy function then takes the form

$$F(u) = \int \phi(|\nabla u|) \,, \tag{3}$$

where ϕ is strictly convex, nondecreasing function that grows at most linearly. Examples of $\phi(s)$ are s (total variation), $\sqrt{1 + s^2} - 1$ (hypersurface minimal

function) or $\log(\cosh(s))$. The energy function based on the variational integral is highly nonlinear and to overcome this difficulty we follow a half-quadratic scheme described in [17]. In addition, we confine the distribution to an amplitude constraint set $C_u \equiv \{u|\alpha \leq u \leq \beta\}$ with amplitude bounds derived from the input images, typically $\alpha = 0$ and $\beta = 255$. The prior distribution then takes the form

$$p(\mathbf{u}) = \begin{cases} \frac{1}{Z} \exp\left\{-\frac{1}{2\sigma_u^2}\mathbf{u}^T\mathbf{L}(v)\mathbf{u}\right\} & \text{if } \mathbf{u} \in C_u\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where Z is the partition function, σ_u^2 denotes the image variance, $\mathbf{u}^T\mathbf{L}(v)\mathbf{u}$ represents the discretization of (3) and v is the auxiliary variable introduced by the half-quadratic scheme, which is calculated as $v(x,y) = \phi'(|\nabla u(x,y)|)/|\nabla u(x,y)|$.

The shape of the prior distribution $p(\mathbf{g})$ can be derived from a fundamental multichannel constraint stated in [10]. Let $\mathbf{Z}_k$ denote the convolution matrix with the degraded image $\mathbf{z}_k$. If noise $\mathbf{n}_k$ is zero and the original channel masks $\{\mathbf{h}_k\}$ are *weakly coprime*, i.e. their only common factor is a scalar, then the blurs $\{\mathbf{g}_k\}$ satisfy

$$\mathbf{Z}_i\mathbf{g}_j - \mathbf{Z}_j\mathbf{g}_i = \mathbf{0}\,, \quad 1 \leq i < j \leq K \tag{4}$$

provided that the blur size is large enough to accommodate the maximum original blur size and the maximum shift between the channels. After stacking the system of equations (4), we obtain

$$\mathcal{Z}\mathbf{g} = \mathbf{0}\,, \tag{5}$$

where $\mathbf{g} \equiv [\mathbf{g}_1^T, \ldots, \mathbf{g}_K^T]^T$. If the noise term $\mathbf{n}_k$ is present, it follows from (2) that the left-hand side of (5) equals a realization of the Gaussian process of zero mean and covariance $\mathcal{C} = \mathcal{G}\Sigma\mathcal{G}^T$, where $\mathcal{G}$ has the form of $\mathcal{Z}$ in (5) with $\mathbf{Z}_i$ replaced by $\mathbf{G}_i$.

It is desirable to include also other prior knowledge about the blurs, such as positivity or constant energy. We therefore define a set of admissible solutions as $C_g \equiv \{\mathbf{g}|g_k(x,y) \geq 0 \wedge \sum_{x,y} g_k(x,y) = 1,\ k = 1, \ldots, K\}$ and propose the following prior distribution:

$$p(\mathbf{g}) = \begin{cases} \frac{1}{Z} \exp\left\{-\frac{1}{2}\mathbf{g}^T\mathcal{Z}^T\mathcal{C}^{-1}\mathcal{Z}\mathbf{g}\right\} & \text{if } \mathbf{g} \in C_g\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

The matrix $\mathcal{C}$ is constructed by the blurs $\mathbf{g}$, that are to be estimated, and its inversion is not trivial. To overcome this difficulty, we approximate $\mathcal{C}$ by a diagonal matrix $\mathcal{D}$ such that $\mathrm{diag}(\mathcal{D}) = \mathrm{diag}(\mathcal{C})$, where $\mathrm{diag}(\cdot)$ denotes the main diagonal of a matrix. The elements of $\mathcal{D}$ take the form $\sigma_i^2\|\mathbf{g}_j\|^2 + \sigma_j^2\|\mathbf{g}_i\|^2$ for $1 \leq i < j \leq K$. The value of $\|\mathbf{g}_i\|^2$ is not known in advance, but a good initial approximation can be calculated. Since $\mathbf{g} \in C_g$, $\frac{1}{\prod S_g} \leq \|\mathbf{g}_i\|^2 \leq 1$ and we use the bottom limit for $\|\mathbf{g}_i\|^2$.

The MAP estimation is then given by

$$\{\hat{\mathbf{u}}, \hat{\mathbf{g}}\} = \arg \min_{\substack{\mathbf{u}\in C_u \\ \mathbf{g}\in C_g}} \left\{ (\mathbf{z}-\mathbf{G}\mathbf{u})^T \Sigma^{-1} (\mathbf{z}-\mathbf{G}\mathbf{u}) + \frac{1}{\sigma_u^2} \mathbf{u}^T \mathbf{L}(v)\mathbf{u} + \mathbf{g}^T \mathcal{Z}^T \mathcal{D}^{-1} \mathcal{Z}\mathbf{g} \right\}. \tag{6}$$

Such problems can be solved by means of genetic algorithms or simulated annealing. In this paper we have adopted an approach of alternating minimizations (AM-MAP) over $\mathbf{u}$ and $\mathbf{g}$. The advantage of this scheme lies in its simplicity. Each term in (6) is convex, the derivatives w.r.t. $\mathbf{u}$ and $\mathbf{g}$ can be easily calculated and we alternate between two simple linear equations.

One must supply the blur size to the algorithm. An important feature, which makes the proposed method particulary appealing, is that an accurate estimation is not necessary; we must only guarantee that the blur size is not underestimated. In addition, the noise covariance Σ and the image variance σ_u^2 are obligatory in the algorithm. If noise has the same variance σ_n^2 in every channel, the MAP expression (6) is simplified and only the signal to noise ratio σ_u^2/σ_n^2 is required in the algorithm.

3 Experimental Results

To illustrate the performance of the AM-MAP registration and fusion method, we conducted two experiments: one with synthetic data and one with real data.

In the first experiment, we created two degraded images depicted in Fig. 1(a). This was done by convolving the original image with two different 5×5 blurs, adding noise of SNR = 50 dB and shifting one of the blurred images by 5×5 pixels. The AM-MAP algorithm was initialized with the overestimated blur size 12×12. The fused image and the estimated blur masks are shown in Fig. 1(b). Reconstructed blurs are properly shifted and *de facto* perform registration. The fused image is by visual comparison much sharper than the input channels and no artifacts are visible.

In real applications one cannot expect that translation is the only geometric deformation that may occur. As mentioned in the introduction, the registration methods suitable for blurred images must be first applied in such cases. However, precise registration is seldom achieved. If our fusion algorithm is used, one can reduce the effort to reach accurate registration. Fast registration that removes only the major geometric degradation will suffice. To test this issue and applicability of the proposed algorithm, we conducted the experiment with real data.

We photographed a house facade (relatively planar scene) under low-light conditions with a standard 3 Mp digital camera. Three photos (see Fig. 2) were taken from the same spot with the camera held in hand and set to ISO 100. The first two have a long exposure time of 1/4 s which was necessary for the given light conditions. Both images exhibit strong blurring due to the hand motion. We slightly tilted the camera before the second image was taken. The third photo was

(a) (b)

Fig. 1. AM-MAP image fusion: (a) two blurred and mutually shifted images of size 230×260 pixels; (b) fused image and estimated translated blur masks.

acquired with a shorter exposure time of 1/100s to avoid the blurring effect but at the cost of low contrast. Due to the finite quantization step of the digital camera, some details in the image (compare for example sphinx's face) are irrecoverably lost. This image was used only for comparison purposes. Image fusion proceeded as follows. The two degraded images in Figs. 2(a) and 2(b) were first registered with a landmark-based method. Rough registration was performed by means of three control points and an affine mapping function. We could do better and decrease, but never completely cancel, a relatively large registration error (≈ 10 pixels), if more control points and/or a more complex mapping were used. Since the goal was to test the robustness of our method, inaccurate registration was sufficient. The roughly registered images then served as inputs to the AM-MAP fusion algorithm. We estimated the blur size to be 15×15 pixels, which was done by first inspecting the blurred letters (see Fig. 3) in the photos and then multiplying the size by two to accommodate possible misregistration. The algorithm was restarted with different signal to noise ratios and provided the best results according to our visual assessment for $\sigma_u^2/\sigma_n^2 = 2000$, which corresponds to SNR between 30 and 40 dB. The fused image and the estimated blurs after 20 iterations are shown in Figs. 2(d). On closer inspection, details in Fig. 3 reveal that the fused image is not as sharp as the short-exposure photo but the overall impression of the image is very pleasing.

4 Conclusions

We have developed an iterative fusion algorithm that performs simultaneously registration and restoration of translated blurred images. The fusion problem is formulated as the MAP estimation with the prior probabilities derived from the variational integral and from the mutual relation of coprime channels. We envisage that possible future extensions of the proposed method are to include space-variant deconvolution (the blur mask may depend on the position in the image) and resolution enhancement of the fused image (super-resolution).

Fig. 2. Real data experiment: A house facade was photographed with a standard digital camera under low-light conditions. All photos were taken with the hand-held camera and downscaled by four. (a)-(b) Two images acquired with a correct but long exposure time of 1/4 s. The images exhibit high contrast but severe blurring is evident due to the hand motion. (c) Image acquired with a short exposure time of 1/100 s. This image is sharp but too dark. The two top images were roughly registered and then fused with the AM-MAP algorithm. (d) Result of image fusion.

Fig. 3. Close-ups: (top) detail of the blurred image in Fig. 2(a); (middle) detail of the sharp image in Fig. 2(c); (bottom) detail of the fused image in Fig. 2(d).

Acknowledgments. Financial support of this research was provided mainly by the Grant Agency of the Czech Republic under the project No. 102/04/0155 and partially by the bilateral project 2003CZ0009 CSIC-Academy of Sciences of the Czech Republic.

References

1. Zitová, B., Flusser, J.: Image registration methods: A survey. Image and Vision Computing **21** (2003) 977–1000
2. Myles, Z., Lobo, N.V.: Recovering affine motion and defocus blur simultaneously. IEEE Trans. Pattern Analysis and Machine Intelligence **20** (1998) 652–658
3. Zhang, Y., Wen, C., Zhang, Y., Soh, Y.C.: Determination of blur and affine combined invariants by normalization. Pattern Recognition **35** (2002) 211–221
4. Kubota, A., Kodama, K., Aizawa, K.: Registration and blur estimation methods for multiple differently focused images. In: Proceedings International Conference on Image Processing. Volume II. (1999) 447–451
5. Zhang, Z., Blum, R.: A hybrid image registration technique for a digital camera image fusion application. Information Fusion **2** (2001) 135–149
6. Flusser, J., Suk, T., Saic, S.: Recognition of blurred images by the method of moments. IEEE Trans. Image Processing **5** (1996) 533–538
7. Flusser, J., Zitová, B.: Combined invariants to linear filtering and rotation. Intl. J. Pattern Recognition Art. Intell. **13** (1999) 1123–1136
8. Zitová, B., Kautsky, J., Peters, G., Flusser, J.: Robust detection of significant points in multiframe images. Pattern Recognition Letters **20** (1999) 199–206
9. Flusser, J., Zitová, B., Suk, T.: Invariant-based registration of rotated and blurred images. In Tammy, I.S., ed.: Proceedings IEEE 1999 International Geoscience and Remote Sensing Symposium, Los Alamitos, IEEE Computer Society (1999) 1262–1264
10. Harikumar, G., Bresler, Y.: Perfect blind restoration of images blurred by multiple filters: Theory and efficient algorithms. IEEE Trans. Image Processing **8** (1999) 202–219
11. Pai, H.T., Bovik, A.: On eigenstructure-based direct multichannel blind image restoration. IEEE Trans. Image Processing **10** (2001) 1434–1446
12. Panci, G., Campisi, P., Colonnese, S., Scarano, G.: Multichannel blind image deconvolution using the bussgang algorithm: Spatial and multiresolution approaches. IEEE Trans. Image Processing **12** (2003) 1324–1337
13. Šroubek, F., Flusser, J.: Multichannel blind iterative image restoration. IEEE Trans. Image Processing **12** (2003) 1094–1106
14. Woods, N., Galatsanos, N., Katsaggelos, A.: EM-based simultaneous registration, restoration, and interpolation of super-resolved images. In: Image Processing, 2003. Proceedings. Volume 2. (2003) 303–306
15. Rav-Acha, A., Peleg, S.: Restoration of multiple images with motion blur in different directions. In: IEEE Workshop on Applications of Computer Vision (WACV). (2000) 22–27
16. Šroubek, F., Flusser, J.: Shift-invariant multichannel blind restoration. In: Proceedings of the 3rd Int'l Symposium on Image and Signal Processing and Analysis, ISPA'03, Rome, IEEE (2003)
17. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing. Springer Verlag, New York (2002)

A New Numerical Scheme for Anisotropic Diffusion

Hongwen Yi[1] and Peter H. Gregson[2]

[1] Postdoctoral follow, iDLab, Dalhousie University, Halifax, NS, Canada
[2] NSERC Chair in Design Innovation, Director of iDLab
Dalhousie University, Halifax, NS, Canada

Abstract. Automatically stopping the diffusion process is a challenging task in anisotropic diffusion (AD). Without a preset number of iterations, over-smoothing of semantically meaningful features occurs very easily with current discrete version of AD (DAD). We address this problem by considering the difference in the behavior of DAD and its continuous counterpart. A new numerical scheme is proposed in this paper in which the non-negative part of the derivative of flux is employed for the first time to control the smoothing strength. Our proposed algorithm implements the desired AD operation with over-smoothing prevented.

1 Introduction

There has been a considerable amount of research into AD theory and applications since AD was introduced by Perona and Malik (PM) in 1987 [8]. AD is a nonlinear selective smoothing technique that should remove noise and trivial parts of images while preserving or enhancing semantically meaningful edges. The AD algorithm is suitable for parallel implementation in hardware, thereby making feasible implementing real time image processing at low cost. Also, AD theory suggests that AD can be used for unsupervised image processing tasks such as robotics and surveillance in unconstrained environments. We seek results that do not depend on a predefined or manually determined iteration count and that are easy to segment in later processing. They should have a "cartoon-like" quality.

Automatically stopping the diffusion process is one of the most critical problems in AD research. Current AD techniques are strongly image-dependent and person-dependent because oversmoothing of meaningful edges occurs easily and "good" results exist only within an image-dependent number of diffusion iterations. While stopping the iterations manually before over-blurring becomes apparent is popular, creating a criterion to automatically stop AD at "optimal" results is attractive [12]. Considerable work has been carried out on the conduction function (CF) that controls the diffusion process. It has been proposed that

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 130–138, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the CF be calculated from a Gaussian-filtered image [2] or from a morphologically filtered image [11]. Modified and new CFs have been proposed [5,13,18]. Time-dependent AD parameters have been proposed [13], but are problematic to implement. Tukey's biweight function forces the diffusion effect to zero [1], but how AD process evolves remains obscure. Oversmoothing can be mitigated by a fidelity term λ that controls the similarity between the result image I and the original image I_0 [5,9,14]. But it is not effective at removing noise and trivial details.

$$I_t = div(c(x, y, t)\nabla I) - \lambda(I - I_0) \tag{1}$$

Though the continuous AD (CAD) predicts that less, zero or backward smoothing is implemented at edges, PM DAD cannot preserve edges if the number of iterations is not artificially limited. The effectiveness of DAD has not enjoyed enough research since the introduction of AD. Among the large relevant literature, very few papers deal with AD's numerical aspects [14,15]. PM DAD has been explained under various conditions such as CFs with unit sum [10], or with a discretized space step of one [6]. Semidiscrete PM DAD has been studied in [4]. You et al. [17] showed that the continuous and discrete AD schemes behave differently in backward diffusion. CAD is ill-posed [18] while the PM DAD is well-posed except for "staircasing" [15]. This difference has been either accepted or ignored in AD research.

We seek a DAD algorithm that has the same qualitative performance as CAD. In [16] we show that the derivative of flux (DF) with restricted negative values should be used as the diffusion coefficient to control the diffusion process. We show in [16] that there are three different processes in AD based on the DF: a mathematically well-posed process, an ill-posed process and a process of no action. This paper addresses the effectiveness of PM DAD, and shows that the DF should be employed in DAD to control the smoothing strength. Our proposed DAD prevents the problem of oversmoothing with its solution converging to that of CAD consistently and stably.

Section 2 presents the background and a new interpretation of AD. In Sect. 3, we present a DAD that converges to the CAD stably and consistently. Section 4 presents our experiments and results. Conclusions are in Sect. 5.

2 Background

In their pioneer work, Perona and Malik (PM) proposed AD as [7]:

$$I_t = div(c(x, y, t)\nabla I) = c(x, y, t)\Delta I + \nabla c \cdot \nabla I \ . \tag{2}$$

where I is image intensity. div, ∇ and Δ are the divergence, gradient and Laplacian operators respectively. The CF $c(x, y, t)$, normally chosen to be a non-negative, monotonously decreasing function of the gradient magnitude starting

at 1 with $\lim_{x \to \infty} c(x) = 0$, controls the diffusion process. PM proposed two CFs:

$$c(x, y, t) = e^{-(\frac{\|\nabla I\|}{K})^2} \quad . \tag{3}$$

and

$$c(x, y, t) = \frac{1}{1 + (\frac{\|\nabla I\|}{K})^2} \quad . \tag{4}$$

where $\|\nabla I\|$ is the gradient magnitude and threshold K controls the diffusion. PM proposed a 4-nearest-neighbor operator to implement AD:

$$I_{i,j}^{t+1} = I_{i,j}^{t} + \lambda \sum_{D \in N,S,E,W} c_{D_{i,j}}^{t} \cdot \nabla_D I_{i,j}^{t} \quad . \tag{5}$$

where i, j is the pixel position; t is the discrete time step; scalar λ controls the rate of diffusion; and D denotes the relationship between pixel (i, j) and its neighbors in north (N), south (S), east (E) and west (W) respectively. The c_D are approximated by ∇_D, the differences between pixel (i, j) and its neighbors.

Denoting flux by $\phi(I_x) = c(I_x, t)I_x$, we rewrite (2) in the 1D case as:

$$\frac{\partial I}{\partial t} = \frac{\partial}{\partial x}(c(I_x, t)I_x) = \frac{\partial}{\partial x}(\phi(I_x)) = \phi_x(I_x)I_{xx} \quad . \tag{6}$$

Now the DF $\phi_x(\cdot)$ is considered one function and is used for the diffusion coefficient. Let us consider the DF's behavior.

$$\phi_x = \frac{\partial}{\partial I_x}(I_x \cdot c(I_x, t)) = c(I_x, t) + I_x \cdot c_x(I_x, t) \quad . \tag{7}$$

Note that only the $I_x c_x(I_x, t)$ remains unknown. Assume that both $c(I_x, t)$ and $c_x(I_x, t)$ are continuous functions. The boundary values of $I_x c_x(I_x, t)$ are:

$$I_x \cdot c_x(I_x, t)|_{I_x=0} = 0 \quad . \tag{8}$$

$$\lim_{I_x \to \infty} I_x \cdot c_x(I_x, t) = \lim_{I_x \to \infty} \frac{c(I_x, t)}{lnI_x} = 0 \quad . \tag{9}$$

$c_x(I_x, t)$ is negative. There is a number $M > 0$ such that: $|I_x \cdot c_x(I_x, t)| \leq M$. If we set $M = |min\{I_x \cdot c_x(I_x, t)\}|$, M is $|-\frac{2}{e}|$ and $|-\frac{1}{2}|$ for (3) and (4) respectively. Thus $I_x c_x(I_x, t)$ is negative except at its boundaries. The DF is a function starting at 1 and ending at $\lim_{I_x \to \infty} \phi_x(I_x) = 0$, as in Fig. 1. Negative values of DF are not employed due to their side-effects. It can be shown that this DF implements noise reduction and edge preservation.

3 Discrete Implementation of Anisotropic Diffusion

For simplicity, we consider an 1-D continuous image $I(x, t)$ sufficiently smooth in $\mathcal{R} : \{x \in [\ 0 \quad l\ ],\ 0 \leq t\ )\ \}$ with initial condition $I(x, t)|_{t=0} = u(x)$. Dividing

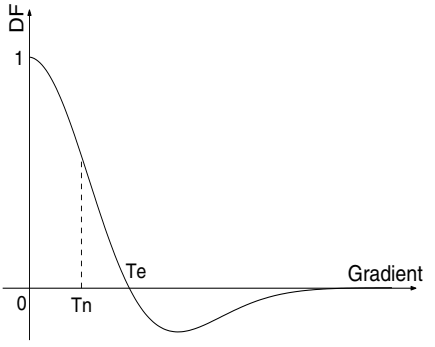


Fig. 1. A typical DF.

$I(x,t)$ in space evenly with $\Delta x = \frac{l}{N}$ and numerically approximating it by time step Δt, the discrete value of $I(x,t)$ at position i and time k is:

$$I_i^k = I(x,t)|_{x=i\Delta x, t=k\Delta t} \ . \tag{10}$$

Using Taylor's series, we get a discrete representation of (6):

$$\frac{I_i^{k+1} - I_i^k}{\Delta t} - \phi_x(I_i^k)\frac{I_{i+1}^k - 2I_i^k + I_{i-1}^k}{(\Delta x)^2} = \frac{\Delta t}{2}\frac{\partial^2 I(x_i,\widehat{t})}{\partial t^2} - \phi_x(I_i^k)\frac{(\Delta x)^2}{12}\frac{\partial^4 I(\widehat{x},t_k)}{\partial x^4}$$
$$t_k \le \widehat{t} \le t_{k+1} \text{ and } |\widehat{x} - x_i| \le \Delta x \ . \tag{11}$$

which is satisfied by the initial conditions:

$$I_i^0 = u(i\Delta x) \qquad (i = 0, 1, 2, \cdots, N-1) \ . \tag{12}$$

Denote the right hand side of (11), the cut-off error due to discretization, as ϵ_i^k. With the assumption that ϵ_i^k is zero, AD equation (6) is approximated by:

$$\frac{\dot{I}_i^{k+1} - \dot{I}_i^k}{\Delta t} - \phi_x(\dot{I}_i^k)\frac{\dot{I}_{i+1}^k - 2\dot{I}_i^k + \dot{I}_{i-1}^k}{(\Delta x)^2} = 0 \ . \tag{13}$$

with

$$\dot{I}_i^0 = u(i\Delta x) \qquad (i = 0, 1, 2, \cdots, N-1) \ . \tag{14}$$

We use symbol $\dot{I}(x,t)$ to distinguish between (11) and (13) because (13) is based on the assumption that ϵ_i^k is zero. With:

$$\lambda = \phi_x(\dot{I}_i^k)\frac{\Delta t}{(\Delta x)^2} \tag{15}$$

the numerical approximation of AD (13) is rewritten as:

$$\dot{I}_i^{k+1} = \lambda\dot{I}_{i+1}^k + (1-2\lambda)\dot{I}_i^k + \lambda\dot{I}_{i-1}^k \ . \tag{16}$$

If (16) is a valid approximation to CAD (6), its solution should converge to the solution of (6) as $\Delta x \to 0$ and $\Delta t \to 0$.

Theorem 1. *Assume that a continuous solution of (6) exists in* $R : \{x \in [\ 0\ l\ ],\ t \in [0\ T]\}$*, and that continuous partial derivatives* $\frac{\partial^2 I}{\partial t^2}, \frac{\partial^4 I}{\partial x^4}$ *exist, then the solution approximated by (16) converges to that of (6) if* $0 \leq \lambda \leq \frac{1}{2}$

The difference between the solutions of (11) and (13) is:

$$V_i^k = I_i^k - \dot{I}_i^k \qquad (i = 0, 1, 2, \cdots, N-1)\ . \tag{17}$$

From (11) - (17), there is:

$$\frac{V_i^{k+1} - V_i^k}{\Delta t} - \phi_x(I_i^k)\frac{V_{i+1}^k - 2V_i^k + V_{i-1}^k}{(\Delta x)^2} = \epsilon_i^k\ . \tag{18}$$

which meets

$$V_i^0 = 0 \qquad \text{for } i = 0, 1, 2, \cdots, N-1\ . \tag{19}$$

Denote its maximum value at time $k\Delta t$ as:

$$V^k = max|V_i^k| \quad (i = 0, 1, 2, \cdots, N-1)\ . \tag{20}$$

Introduce two constants for our discussion:

$$M_1 = max\left|\frac{1}{2}\frac{\partial^2 I(x,t)}{\partial t^2}\right|\ . \tag{21}$$

$$M_2 = max\left|\frac{\phi_x(I_i^k)}{12}\frac{\partial^4 I(x,t)}{\partial x^4}\right|\ . \tag{22}$$

and

$$M = M_1\Delta t + M_2(\Delta x)^2\ . \tag{23}$$

It is clear that:

$$|\epsilon_i^k| \leq M \qquad \text{for } i = 0, 1, 2, \cdots, N-1\ . \tag{24}$$

From (18)-(24) and the restriction for λ, there is:

$$\begin{aligned} |V_i^{k+1}| &= |\lambda V_{i+1}^k + (1-2\lambda)V_i^k + \lambda V_{i-1}^k + \Delta t\epsilon_i^k| \\ &\leq |\lambda V^k| + |(1-2\lambda)V^k| + |\lambda V^k| + |M\Delta t| \\ &\leq V^k + M\Delta t \qquad (i = 0, 1, 2, \cdots, N-1)\ . \end{aligned} \tag{25}$$

With (19), (20), (25) and $(k+1)\Delta t \leq T$, we get:

$$\begin{aligned} V^{k+1} &\leq V^k + M\Delta t\ \leq V^{k-1} + 2M\Delta t\ \leq \cdots \\ &\leq V^0 + (k+1)M\Delta t\ \leq MT\ \leq \left(M_1 + \frac{M_2}{\lambda}\right)T\Delta t\ . \end{aligned} \tag{26}$$

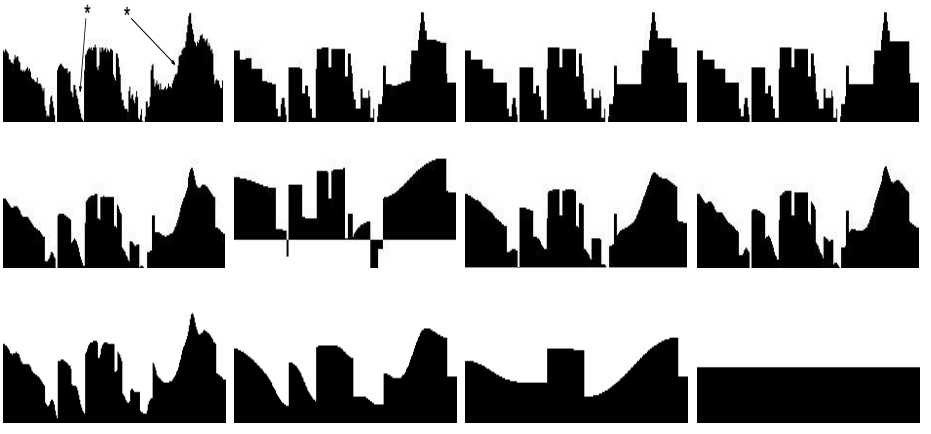


Fig. 2. 1D results: Top (from left): test image, NDAD result at t=10^2, NDAD final result with "frac" CF, NDAD final result with other CFs. Middle (from left): FAB results without a fidelity term (t=10 and 10^3) and with fidelity terms ($\lambda = 0.01$ and ($\lambda = 0.05$). Bottom (from left): PMDAD results (t=10, 10^2, 10^3 and 88710).

$(M_1 + \frac{M_2}{\lambda})T$ is independent of k, thus the difference between the solution of DAD and that of CAD converges to zero as $\Delta t \to 0$. Now we consider the stability of (16) since the ϵ_i^k at every step affects the solution of the next step and its accumulation could result in a poor solution or even an uncontrollable process. Approximating the DF with piecewise constants, the Lax equivalence theorem [3] tells us that the solution of (16) converges to the solution of CAD stably and consistently. Thus the DF should be used to form the diffusion coefficient to control the diffusion process. We propose a numerical AD scheme as:

$$I_{i,j}^{t+1} = I_{i,j}^t + \lambda \sum_{D \in N,S,E,W} d_{D_{i,j}}^t \cdot \nabla_D I_{i,j}^t \tag{27}$$

where $d(\cdot)$ is approximated by the non-negative part of the DF; and other parameters have the same meaning as those in (5).

Equations (27) and (5) share the same form but with a significant difference. The performance of (27) is determined by the DF, Δx and Δt. When (27) is used to approximate its continuous counterpart, it gives convergent, stable and consistent results if $\lambda \leq \frac{1}{2}$ in the 1D case. However, flux in (5) is controlled by the CF itself directly. It is difficult for edges to survive a diffusion process controlled by a positive CF over the whole range of gradients [16].

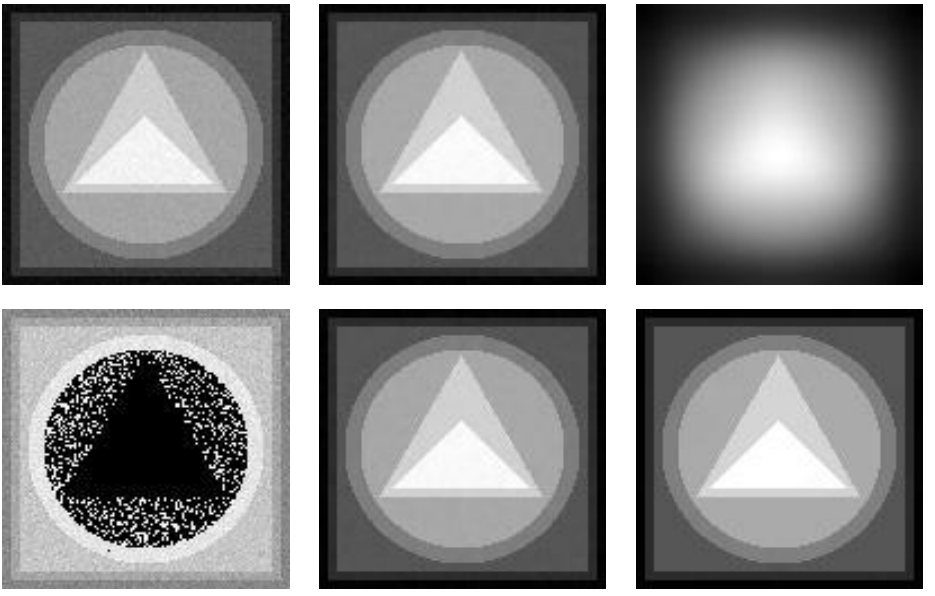


Fig. 3. 2D results on synthetical image. Top (from left): test image, PMDAD results (t=10 and 10^3). Bottom (from left): FAB result ($\lambda = 0.9$, t=40), NDAD result (t=10) and final NDAD result (t=372).

4 Experiments

To compare our proposed new DAD (NDAD) to PM DAD and the forward-and-backward diffusion (FAB), a number CFs were employed that include Tukey's biweight function [1], (1), (3), (4), the CF for implementing FAB proposed in [5]:

$$c(x) = \frac{1}{1 + (\frac{x}{K_1})^n} - \frac{\alpha}{1 + (\frac{x - K_2}{w})^{2m}} \tag{28}$$

where K_1 controls the forward "force", K_2 and w control the backward "force", and α sets the balance between them; and a "ramp" CF defined as follows:

$$c(x) = \begin{cases} 1 - \frac{|x|}{K_e} & \text{for} \quad 0 \le |x| \le K_e \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

For convenience, (3), (4), Tukey's biweight function, (1), (28) and (29) are called as "exp", "frac", "Tukey", "fidel", "FAB" and "ramp" respectively.

Experiment results are presented in Fig. 2 (1D) and Figs. 3 (synthetic 2D) and 4 (real 2D). All the results show that the behavior for 2D images is similar to the 1D behavior, but convergence takes more iterations. PM DAD can produce acceptable results partway through the diffusion process. But without

Fig. 4. 2D results on real image. Top (from left): test image, PMDAD results (t=10^2 and 10^4). Bottom (from left): FAB result ($\lambda = 0.1$, t=10), NDAD results (t=10^2 and 10^4).

a preset number of iterations, diffusion continues until meaningless results are produced. FAB, though it sharpens edges, creates distortion. Though a fidelity term mitigates the distortion, noise or trivial details are then maintained. Due to result uncertainty, close supervision is necessary for both PM DAD and FAB. With NDAD, edge preservation is performed early. Further processing makes areas of noise and trivial details smooth and image results reach a stable state without limiting iterations. Noise and trivial details are removed and meaningful edges are enhanced or preserved throughout the diffusion process.

5 Conclusion

A new numerical scheme is proposed in this paper for approximating AD. The problem of oversmoothing is prevented. Instead of using the CF directly to conduct the diffusion process as in most current DAD approaches, the non-negative part of the DF is used to control the smoothing strength. This approach agrees with AD theory, with a desirable combination of forward smoothing preformed in noise regions and zero/backward smoothing carried out at edges. Side-effects of using a negative diffusion coefficient are avoided. Our discrete scheme keeps the semantically meaningful features throughout the diffusion process, thereby making feasible the implementation of unsupervised computer vision systems. The effectiveness of our proposed NDAD is illustrated by experiments.

References

1. Black, M., Sapiro, G., Marimont, D., Heeger, D.: Robust anisotropic diffusion. IEEE Trans. IP. (**7**) (1998) 421–432
2. Catte, F., Lions, P., Morel, J., Coil, T.: Image selective smoothing and edge detection by nonlinear diffusion. SIAM J. Num. Anal. (**29**) (1992) 182–193
3. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology. (**6**) II, Springer-Verlag, Berlin. (1988)
4. Esedoglu, S.: An analysis of Perona-Malik scheme. Comm. Pure Appl. Math. (2001) 1442–1487
5. Gilboa, G., Sochen, N., Zeevi, Y.Y.: Forward-and-backward diffusion processes for adaptive image enhancement and denoising. IEEE Trans. IP. (**11**) (2002) 689–703
6. Jin, J.S., Wang, Y., Hiller, J.: An adaptive nonlinear diffusion algorithm for filtering medical images. IEEE Trans. Inform. Technol. Biomed. (**4**) (2000) 298–305.
7. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. PAMI. (1990) 629–639
8. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Computer Society Workshop on Computer Vision - Miami. (1987) 16–22
9. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D. (**60**) (1992) 259–268
10. Saint-Marc, P., Chen, J.S., Medioni, G: Adaptive smoothing: a general tool for early vision. PAMI. (**13**) (1991) 514–529
11. Segall, C.A., Acton, S.T.: Morphological anisotropic diffusion. Int. Conf. on IP. (**3**) (1997) 348–351
12. Solo, V.: A fast automatic stopping criterion for anisotropic diffusion. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. (2002) 1661–1664
13. Torkamani-Azar F., Tait, K.E.: Image recovery using the anisotropic diffusion equation. IEEE Trans. IP. (**5**) (1996) 1573–1578
14. Weickert, J.A.: Applications of nonlinear diffusion in image processing and computer vision. Acta Mathematica Universitatis Comenianae. (**70**) (2001) 33–50
15. Weickert, J.A., Romeny, B.H., Florack, L., Koenderink, J., Viergever: A review of nonlinear diffusion filtering. Invited paper. Springer, Berlin. (1997) 3–28
16. Yi, H., Gregson, P.H.: Behavioral analysis of anisotropic diffusion for image processing", submitted to IEEE Trans. Image Processing. (2004)
17. You, Y., Kaveh, M.: Differences in the behaviors of continuous and discrete anisotropic diffusion equations for image processing. ICIP98. (1998) 249–253
18. You, Y., Xu, W., Tannenbaum, A., Kaveh, M.: Behavioral analysis of anisotropic diffusion in image processing. IEEE Trans. IP. (**5**) (1996) 1539–1553

An Effective Detail Preserving Filter for Impulse Noise Removal

Naif Alajlan and Ed Jernigan

PAMI Lab, E & CE, UW, Waterloo, ON, N2L 3G1, Canada.
naif@pami.uwaterloo.ca

Abstract. Impulsive noise appears as a sprinkle of dark and bright spots. Linear filters fail to suppress impulsive noise. Thus, non-linear filters have been proposed. The median filter works on all image pixels and thus destroys fine details. Alternatively, the peak-and-valley filter identifies noisy pixels and then replaces their values with the minimum or maximum value of their neighbors depending on the noise (dark or bright). Its main disadvantage is that the estimated value is unrealistic. In this work, a variation of the peak-and-valley filter based on a recursive minimum-maximum method is proposed. This method preserves constant and edge areas even under high impulse noise probability and outperforms both the peak-and-valley and the median filters.

1 Introduction

Filtering a digital image to attenuate noise while preserving the image detail is an essential part of image processing. For example, in many applications where operators based on computing image derivatives are applied, any noise in the image can result in serious errors. Noise can appear in images from a variety of sources during the acquisition process, due to quality and resolution of cameras, and illumination variations. For most typical applications, image noise can be modeled with either Gaussian, uniform, or impulse distributions. Gaussian noise can be analytically described and has the characteristic bell shape. With uniform noise, the gray level values of the noise are evenly distributed across a specific range. Impulse noise generates pixels with gray level values not consistent with their local neighbors. It appears in the image as a sprinkle of dark and light spots. Transmission errors, malfunctioning pixel elements in the camera sensors, or faulty memory locations can cause impulse noise.

Linear filters, which consist of convolving the image with a constant matrix, fail to deal with impulse noise although they are effective in reducing Gaussian and uniform noise distributions. They usually produce blur and incomplete impulse noise suppression [1]. To overcome these difficulties, nonlinear filters have been proposed. The most popular nonlinear filter is the median filter. When considering a small neighborhood, it is highly efficient in removing impulse noise. The main disadvantage of the median filter is that it is applied on all the points of the image regardless if they are noisy or not, which results in the loss of fine

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 139–146, 2004.
© Springer-Verlag Berlin Heidelberg 2004

image detail and produces streaks and blotches in the restored image [2]. Finding a method that is efficient in both noise reduction and detail preservation is an active area of research.

Various forms of non-linear techniques have been introduced to solve the problem based on the average performance of the median filter. Examples of those techniques are the weighted median filter [3], the adaptive trimmed mean filter [4], the center weighted median filter [5], the switching-based median filter [6], the mask median filter [7], and the minimum-maximum method [8]. These approaches involve a preliminary identification of corrupted pixels in an effort to prevent alteration of true pixels. The recursive minimum-maximum filter [2] performs better than other filters including the standard median filter. It is good at preserving fine details, but its main disadvantage is that it requires thresholding to detect noisy pixels, which may require several iterations to achieve its best results since each image region has different properties. Consequently, the efficiency is reduced. To overcome the thresholding problem, the peak-and-valley filter [9] offers a fast and non-iterative method to detect noisy pixels and then it replaces their values with the minimum or maximum of the neighbor's values.

In this work, an efficient and detail preserving filter for impulse noise removal is proposed. It takes the advantages of the filters of [9,2] and works in two stages. First, it detects noisy pixels by examining the surrounding pixels as in the peak-and-valley filter. Then, it replaces the noisy pixel values using the recursive minimum-maximum method. The remaining of the paper is organized as follows. Sections 2 and 3 give explanations of the median and peak-and-valley filters, respectively. Section 4 introduces our proposed filter followed by comparative studies of its performance with the median and peak-and-valley filters in section 5. Finally, we conclude our work in section 6.

2 The Median Filter

The median filter is the most popular example of non-linear filters based on order statistics. Consider a 3×3 window shown in Fig. 1, the output of an order statistic filter is given by:

$$y = \sum_{i=1}^{9} \alpha_i d_i^k. \tag{1}$$

Where d_i^k are the order statistics of the nine inputs. The constants α_i may be chosen for a particular application. The median filter is a particular case of (1) with the coefficients $\alpha_i = 0$ except $\alpha_5 = 1$. We can also define the local mean filter by taking $\alpha_i = 1/9$.

Bovik *et al.* [10] showed that the optimal order statistic filter tends toward the median filter, as the noise becomes more impulsive, based on the minimum mean squared error between the original noise-free and noisy filtered images. The median filter is effective when the noise spatial extent is less than half the window size.

d_1	d_2	d_3
d_4	d_9	d_5
d_6	d_7	d_8

Fig. 1. Window used to detect and process impulse noisy pixels.

3 The Peak-and-Valley Filter

The peak-and-valley filter [9] is a non-linear non-iterative filter for impulse noise reduction based on order statistics and a minimal use of the background information. It consists of applying two conditional rules. The noisy pixels are identified and replaced in a single step. The replacement gray value is taken from the neighbors' gray levels.

To understand how the peak-and-valley filter works, consider the 1-D case where it takes the following shape:

$$y_i = \begin{cases} \min(d_{i-1}, d_{i+1}) & if \quad d_i < \min(d_{i-1}, d_{i+1}) \\ \max(d_{i-1}, d_{i+1}) & if \quad d_i > \max(d_{i-1}, d_{i+1}) \\ d_i & else \end{cases} \tag{2}$$

The peak-and-valley filter eliminates all the "peaks" and "valleys" which are thinner than two pixels and fills them following a sequence of cutting/filling then filling/cutting operations, while displacing all along the rows and columns of the image. For the cutting operation, if the middle pixel has a gray level higher than its two neighbors, its gray level value is replaced by the maximum of the other two. For the filling operation, if the middle pixel is smaller than the other two, its gray level value is replaced by the smallest value among its neighbors. All these operations are recursively applied to assure that no peaks and/or valleys remain in the filtered image.

The expression of the filter for the 2-D case, considering 3×3 window shown in Fig. 1 and $i \in [1:8]$, is:

$$y = \begin{cases} \min(d_i) & if \quad d_9 < \min(d_i) \\ \max(d_i) & if \quad d_9 > \max(d_i) \\ d_9 & else \end{cases} \tag{3}$$

4 The Proposed Filter

The proposed filter is a non-linear, non-iterative filter that is based on order statistics to remove impulse noise from an image. It operates in two steps. First, the noisy pixels are detected in the same manner as in the peak-and-valley filter. Then, the corrupted pixels' gray level values are estimated using the recursive minimum maximum method [2]. The motivation behind this work is, unlike

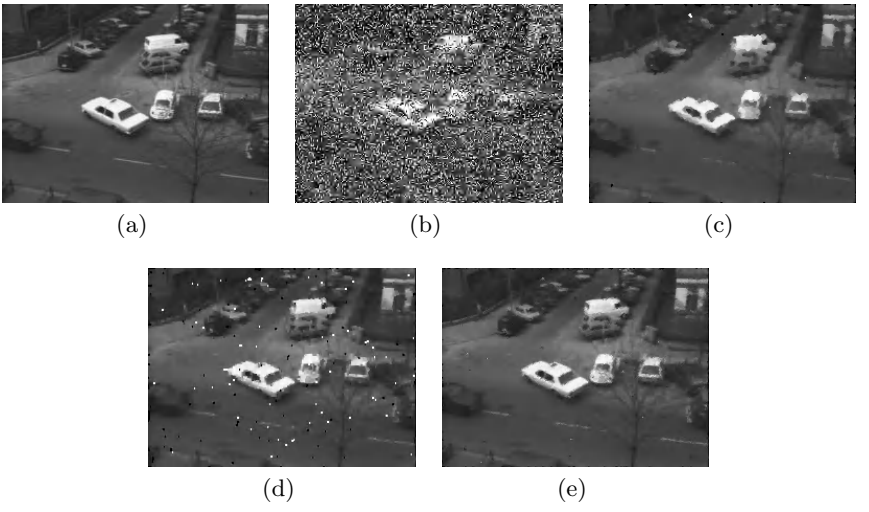


Fig. 2. Hamburg taxi images and filtering results: (a) original,(b) 30% corrupted, (c) median, (d) peak-and-valley, and (e) proposed.

Fig. 3. Lena images and filtering results: (a) original, (b) 30 % corrupted, (c) median, (d) peak-and-valley, and (e) proposed.

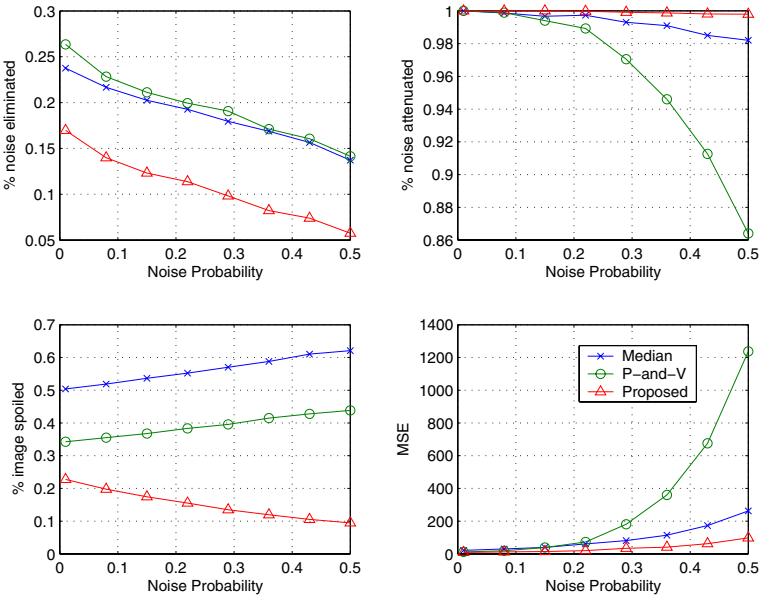


Fig. 4. Objective performances on the *Hamburg Taxi* image.

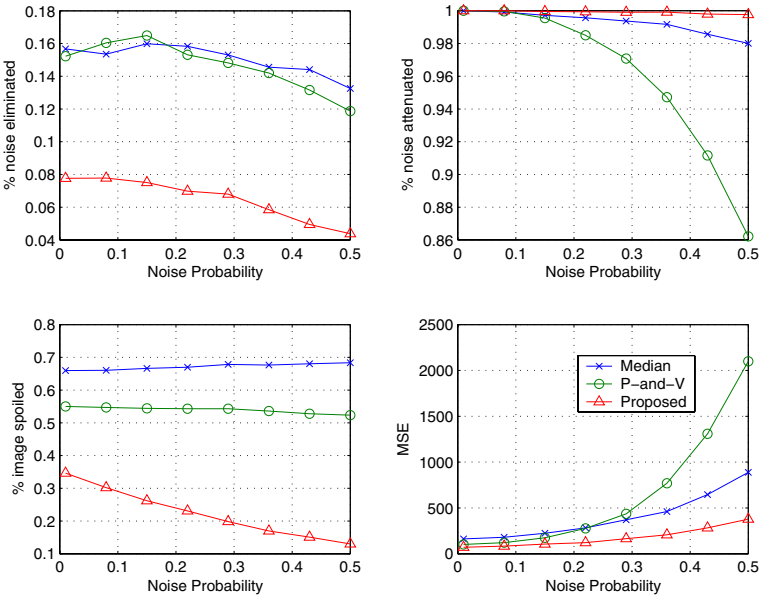


Fig. 5. Objective performances on the *Cameraman* image.

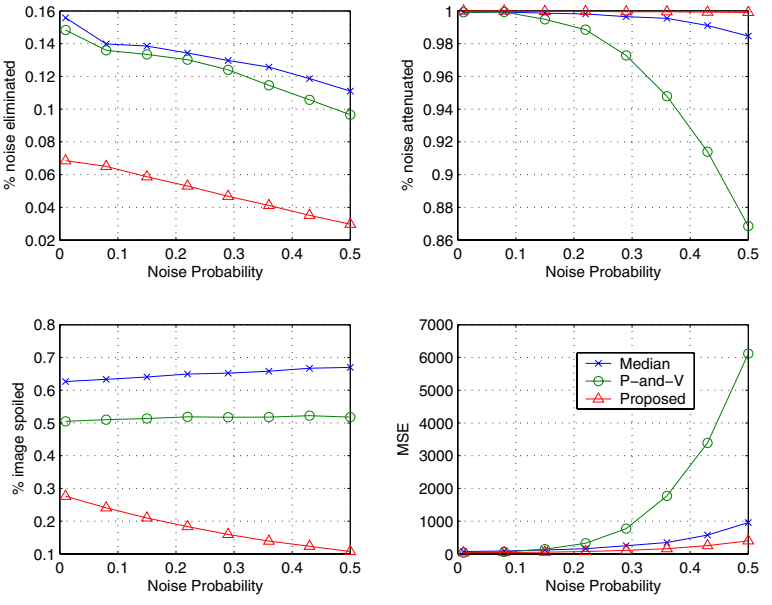


Fig. 6. Objective performances on the *Lena* image.

the median filter that modifies all pixels and destroys fine details, to have a detection approach that is simple and non-iterative. This enables the filter to be applicable to all image types. Afterwards, the recursive minimum maximum method provides an estimate of the corrupted pixels at constant signal as well as edges even when the noise probability is high. This estimation of the original pixel's value is more realistic than the estimation used in the peak-and-valley filter, which is just the minimum or maximum value of the surrounding pixels. The proposed algorithm for impulse noise filtering works as follows:

1. For a 3×3 window centered at the test pixel, as shown in Fig. 1.
2. If $d_9 \geq max(d_i)$ or $d_9 \leq min(d_i)$ where $1 \leq i \leq 8$, then d_9 is a noisy pixel and must be estimated, go to step 3 . Otherwise $y = d_9$.
3. When a noisy pixel is detected, its gray level is estimated as follows. For $1 \leq i \leq 4$, let $L_i = max(d_i, d_{9-i})$ and $E_i = min(d_i, d_{9-i})$. Set $P_{min} = min(L_1, .., L_4)$ and $P_{max} = max(E_1, .., E_4)$. Then $y = (P_{min} + P_{max})/2$.

Note that if there are three identical noisy pixels along one direction within the window, then the output of the filter is largely influenced by the noisy pixels. In this case, either P_{max} or P_{min} is equal to the level of the noisy pixel. However, (d_1, d_2, d_3, d_4) in Fig. 1 are in practice the previous outputs of the filter, instead of the original degraded image data. Thus, the output of the filter is derived recursively from the last four outputs and the present five inputs in the window.

5 Comparative Studies

We implemented the median, the peak-and-valley, and the proposed filters to compare their performances. To provide consistent comparison, only the recursive versions of these filters are considered. The peak-and-valley filter is implemented as a pair of 1D filters, applied in the horizontal then in the vertical directions because this version provides the best performance [9]. We tested the performance of these filters on three standard images used by the image processing research community. The first one was the first frame of a public domain twelve-frame sequence, known as *Hamburg taxi* (190×256 pixels), shown in Fig. 2(a). The second was the *cameraman* image (256×256 pixels). The third image was the well-known *Lena* image (512×512 pixels) shown in Fig. 3(a). The images contain a nice mixture of detail, flat regions, shading, and texture that do a good job of testing various image processing algorithms. We restricted our tests using a 3×3 window size to reduce the computational complexity of the algorithms.

The outcomes of the median, peak-and-valley, and proposed filters applied to the *Hamburg taxi*, and *Lena* images, at impulse noise probability of 30%, are shown in Figs. 2, and 3, respectively. In addition to the quality of the visual appearance, four performance measures are used to compare the filters [9]: the number of the noisy pixels replaced by the true values, the number of noisy pixels attenuated, the number of true pixels modified, and the mean squared error between the original noise-free and filtered images. All images were corrupted with impulse noise probability ranging from 1 % to 50 %. The four performance measures are plotted versus the impulse noise probability, as shown in Figs. 4, 5, and 6.

For all images, the proposed filter impulse noise attenuation rate is near 100 % even when the noise probability is high. The peak-and-valley filter noise attenuation rate reduces dramatically as the noise probability increases. The median filter is the best in terms of estimating the actual value of a noisy pixel, but it tends to change the values of more than 50 % of true pixels, which results in destroying fine details in the image. Interestingly, the proposed filter modifies fewer true pixels as the noise probability increases, which results in high detail preservation. Finally, the proposed filter outperforms other filters in the minimum mean squared error sense. From these results, the proposed filter outperforms other filters in the overall performance.

6 Conclusion

In this work, we proposed a non-linear, non-iterative filter for impulse noise attenuation. Unlike thresholding techniques, it detects noisy pixels non-iteratively using the surrounding pixel values, which makes it suitable for all image types. Then, it uses the recursive minimum-maximum method to estimate the value of corrupted pixels. This estimation provides an accurate estimation even when the noise probability is high. The performance of the proposed filter is compared

with two other filters, the median and the peak-and-valley. The proposed filter outperformed other filters in terms of noise suppression and detail preservation. In conclusion, the proposed filter represents an interesting replacement for the median filter, which is used for preliminary processing in most of the state-of-the-art impulse noise filters.

References

1. Moreno, H.G., Bascon, S.M., Manso, M.U., Martin, P.M.: Elimination of impulsive noise in images by means of the use of support vector machines. XVI National Symposium of URSI (2001)
2. Xu, Y., Lae, E.M.: Restoration of images contaminated by mixed gaussian and impulse noise using a recursive minimum-maximum method. Vision, Image and Signal Processing, IEE Proc. **145** (1998) 264–270
3. Brownrigg, D.: The weighted median filter. Communications of the ACM **27** (1984) 807–818
4. Restrepo, A., Bovik, A.C.: Adaptive trimmed mean filters for image restoration. IEEE Transactions on Acoustics, Speech, and Signal Processing **36** (1988) 1326–1337
5. Ko, S.J., H., L.Y.: Center weighted median filters and their applications to image enhancement. IEEE Transactions on Circuits and Systems **38** (1991) 984–993
6. Sun, T., Neuvo, Y.: Detail preserving median based filters in image processing. Pattern Recognition Letters **15** (1994) 341–347
7. Cabrera, L., Escanmilla, P.: Two pixel preselection methods for median type filtering. Vision, Image and Signal Processing, IEE Proc. **145** (1998) 30–40
8. Imme, M.: A noise peak elimination filter. CVGIP: Graph. Models Image Process. **53** (1991) 204–211
9. Windyga, P.S.: Fast impulsive noise removal. IEEE Transaction on Image Processing **10** (2001) 173–179
10. Bovik, A.C., Huang, T., Munson, D.: A generalization of median filtering using linear combinations of order statistics. IEEE Trans. Acous., Speech, and Signal Process **31** (1983) 1342–1350

A Quantum-Inspired Genetic Algorithm for Multi-source Affine Image Registration

Hichem Talbi[1], Mohamed Batouche[2], and Amer Draa[3]

[1]USI Emir Abdelkader, Constantine, Algeria, hichem_talbi@hotmail.com
[2,3]Lire Laboratory, Mentouri University, Constantine, Algeria
batouche@wissal.dz, draa_amer@yahoo.fr

Abstract. In this paper we propose a new algorithm for image registration which is a key stage in almost every computer vision system. The algorithm is inspired from both genetic algorithms and quantum computing fields and uses the mutual information as a measure of similarity. The proposed approach is based on some concepts and principles of quantum computing such as quantum bit and states superposition. So, the definitions of the basic genetic operations have been adapted to use the new concepts. The evaluation of each solution is performed by the computation of mutual information between the reference image and the resulting image. The process aims to maximize this mutual information in order to get the best affine transformation parameters which allow the alignment of the two images.

1 Introduction

The alignment of images is a central task in most of vision systems. It is required in different applications such as objects recognition, 3D reconstructions and data fusion. Basically, image registration can be defined as the process which consists in finding the best geometric transformation that allows the alignment of the common parts of two images. To solve this problem, which is a combinatorial optimization one, many approaches have been proposed. All of them aim to reduce the computing complexity and at the same time avoid local optimums. Among the proposed methods we can mention those based on artificial neural networks, simulated annealing, taboo search, genetic algorithms [1], ants colonies, and artificial immune systems.

Quantum computing is a new field in computer science which has induced intensive investigations and researches during the last decade. It takes its origins from the foundations of the quantum physics. The parallelism that the quantum computing provides reduces obviously the algorithmic complexity. Such an ability of parallel processing can be used to solve combinatorial optimization problems which require the exploration of large solutions spaces. So, the quantum computing allows the design of more powerful algorithms that should change significantly our view about solving hard problems. However, the quantum machines that these algorithms require to be efficiently executed are not available yet. By the time when a powerful quantum machine would be constructed, some ideas such as simulating quantum algorithms on conventional computers or combining them to existing methods have been suggested

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 147–154, 2004.
© Springer-Verlag Berlin Heidelberg 2004

to get benefit from this science [2]. Within this perspective, we are interested in the combination of genetic algorithms and quantum computing for image registration with the use of the mutual information as a measure of similarity.

Consequently, the rest of the paper is organized as follows. The section 2 gives some concepts about genetic algorithms, mutual information and quantum computing. The proposed approach is described in the section 3. The section 4 illustrates some experimental results. Finally, conclusion and some perspectives are drawn up.

2 Basic Concepts

2.1 Genetic Algorithms

Genetic algorithms derive from the evolution theory. They were introduced in 1975 by John Holland and his team as a highly parallel search algorithm. Later, they have been mainly used as optimization device.

According to the evolution theory, within a population only the individuals well adapted to their environment can survive and transmit some of their characters to their descendants. In genetic algorithms, this principle is traduced into the problem of finding the best individuals represented by chromosomes. So, each chromosome encodes a possible solution for the given problem and, starting from a population of chromosomes, the evolution process performs a parallel search through the solutions' space. The fitness is measured for each individual by a function related to the objective function of the problem to be solved.

Basically, a genetic algorithm consists of three major operations: selection, crossover, and mutation. The selection evaluates each individual and keeps only the fittest ones in the population. In addition to those fittest individuals, some less fit ones could be selected according to a small probability. The others are removed from the current population. The crossover recombines two individuals to have new ones which might be better. The mutation operator induces changes in a small number of chromosomes units. Its purpose is to maintain the population diversified enough during the optimization process.

2.2 Entropy Based Measures and Mutual Information

The entropy is a statistical measure defined by Shannon in 1948. It summarizes the randomness of a given variable. The more random a variable is, the larger entropy it will have.

Given a random variable represented by a probability distribution X, i.e. a set of couples (x_i, p_i) where p_i is the probability to have the value x_i. The entropy of X is given by:

$$H(X) = -\sum p_i \log_2 p_i \quad (1)$$

Intuitively, entropy measures the average information provided by a given distribution.

When dealing with two random variables represented by two probability distributions X and Y, we are interested by answering the question: "How likely the two distributions are functionally dependant?" In total dependence case, a measurement of one distribution discards any randomness about the other. As a consequence, quantifying the independence is equivalent to quantifying the randomness. The joint entropy is given by:

$$H(X,Y) = -\sum \sum p(x,y)\ log_2\ p(x,y) \tag{2}$$

In the case of total independence between X and Y, the joint distribution is the product of the marginal distributions.

$$P(X,Y)=P(X).P(Y) \tag{3}$$

In terms of entropy, this leads to:

$$H(X,Y)=H(X)+H(Y) \tag{4}$$

The mutual information is a measure of the reduction on the entropy of Y given X and is then given by:

$$MI(X,Y)=H(X)+H(Y)-H(X,Y) \tag{5}$$

The mutual information is maximized when the two variables are totally dependant.

2.3 Quantum Computing

In early 80, Richard Feynman's observed that some quantum mechanical effects cannot be simulated efficiently on a computer. His observation led to speculation that computation in general could be done more efficiently if it used this quantum effects. This speculation proved justified in 1994 when Peter Shor described a polynomial time quantum algorithm for factoring numbers.

In quantum systems, the computational space increases exponentially with the size of the system which enables exponential parallelism. This parallelism could lead to exponentially faster quantum algorithms than possible classically [3].

The *quantum bit* (*qubit*) is the elementary information unit. Unlike the classical bit, the qubit does not represent only the value 0 or 1 but a superposition of the two. Its state can be given by:

$$\Psi = \alpha\ |0\rangle+\beta|1\rangle \tag{6}$$

where $|0\rangle$ and $|1\rangle$ represent respectively the classical bit values 0 and 1; α and β are complex numbers such that

$$|\alpha|^2 + |\beta|^2 = 1 \tag{7}$$

If a superposition is measured with respect to the basis $\{|0\rangle, |1\rangle\}$, the probability that the *measured* value is $|0\rangle$ is $|\alpha|^2$ and the probability that the measured value is $|1\rangle$ is $|\beta|^2$.

In classical computing, the possible states of a system of n bits form a vector space of n dimensions, i.e. we have 2^n possible states. However, in a quantum system of n qubits the resulting state space has 2^n dimensions. It is this exponential growth of the state space with the number of particles that suggests a possible exponential speed-up of computation on quantum computers over classical computers. Each quantum

operation will deal with all the states present within the superposition in parallel. The basis of the state space of a quantum system of n qubits is: {|00...0⟩, |00...1⟩… |11...1⟩}.

The *measurement* of a single qubit projects the quantum state onto one of the basis states associated with the measuring device. The result of a measurement is probabilistic and the process of measurement changes the state to that measured. Multi-qubit measurement can be treated as a series of single-qubit measurements in the standard basis.

The dynamics of a quantum system are governed by *Schrödinger*'s equation. The *quantum gates* that perform transformations must preserve orthogonality. For a complex vector space, linear transformations that preserve orthogonality are unitary transformations, defined as follows. Any linear transformation on a complex vector space can be described by a matrix. A matrix M is unitary if $M.M'=I$. Any unitary transformation of a quantum state space is a legitimate quantum transformation and vice-versa. Rotations constitute one among the unitary transformations types.

One important consequence of the fact that quantum transformations are unitary is that they are reversible. Thus quantum gates, which can be represented by unitary matrices, must be reversible. It has been shown that all classical computations can be done reversibly.

3 The Proposed Algorithm

Having two images I_1 and I_2 obtained from either similar or different sensors, the proposed algorithm allows the estimating of the affine geometric transformation which overlays the two images. A similar work that concerns only rigid transformation class can be found in [4].

As in genetic algorithms, initial solutions are encoded in N chromosomes representing the initial population. The difference in our algorithm is that each chromosome is represented using quantum bits.

The geometric transformation that aligns the image I_2 on the image I_1 is affine. Affine transformations form the most commonly used type of spatial transformations for registration. A chromosome encodes the six parameters of the affine transformation. Having such parameters, the position of each pixel in the resulting image (x',y') can be calculated from the original position in the second image (x_2,y_2) as follows:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} dx \\ dy \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \tag{8}$$

This transformation does not have the properties associated with the orthogonal rotation matrix. Angles and lengths are no longer preserved, but parallel lines remain parallel. More general spatial distortions such as skew and changes in aspect ratio can be represented within this formulation.

Each parameter is encoded using a binary representation. A bit in a chromosome does not represent only the value 0 or 1 but a superposition of the two. In this way, all the possible solutions are represented in each chromosome and only one solution

among them can be measured at each time according to the probabilities $|\alpha_i|^2$ and $|\beta_i|^2$. A chromosome is then represented by:

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \ldots\ldots & \alpha_N \\ \beta_1 & \beta_2 & \ldots\ldots & \beta_N \end{pmatrix}$$

where each column represents a single qubit. In our algorithm α_i and β_i are real values only.

Initially we generate randomly 4 chromosomes. Each one is composed of N=48 qubits, 8 qubits for each parameter. *dx* and *dy* are the 2D translation parameters and belong to the interval *[-127,+127]*. The other parameters belong to the interval [*-2, +2*]. (the interval is subdivided into 2^8=256 values).

During the whole process we keep in memory the global best solution.

The algorithm consists on applying cyclically 4 quantum genetic operations (Figure 1):

The first operation is a quantum interference which allows a shift of each qubit in the direction of the corresponding bit value in the best solution. That is performed by applying a unitary quantum operator which achieves a rotation whose angle is function of α_i , β_i and the value of the corresponding bit in the best solution.

δθ has been chosen experimentally equal to π/45 and its direction is function of α, β and the bit's value in the best solution (Table 1).

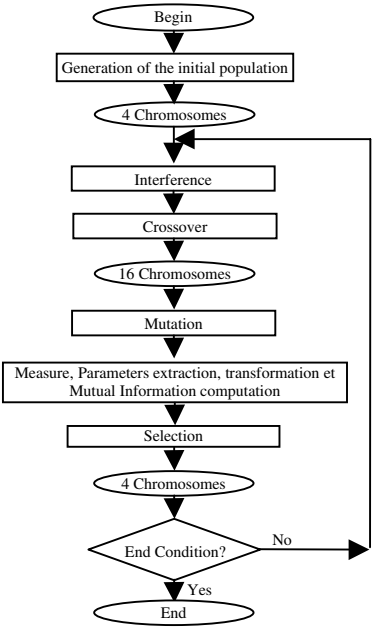


Fig. 1. The proposed algorithm

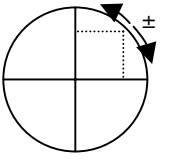


Fig. 2. Quantum Interference

Table 1. Lookup table of the rotation angle

α	>0	>0	>0	>0	<0	<0	<0	<0
β	>0	>0	<0	<0	>0	>0	<0	<0
Reference bit value	1	0	1	0	1	0	1	0
Angle	+δθ	-δθ	-δθ	+δθ	-δθ	+δθ	+δθ	-δθ

The second operation is a crossover performed between each pair of chromosomes at a random position. The figure 3 shows an example of a crossover between two chromosomes.

At each iteration, we obtain from the 4 initial chromosomes 12 new ones. The population becomes composed of 16 chromosomes.

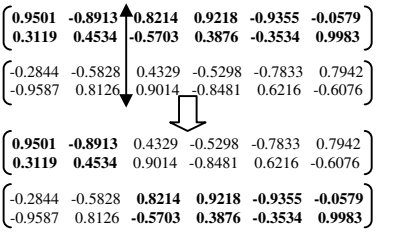


Fig. 3. Quantum Crossover

The third operation consists on a quantum mutation which will perform for some qubits, according to a probability, a permutation between their values $_i$ and $_i$. That will invert the probabilities of having the values 0 and 1 when applying a measurement (Figure 4).

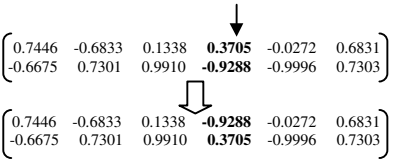


Fig. 4. Quantum Mutation

Finally, we perform a selection of 4 chromosomes among the 16. For this, we apply first a measurement on each chromosome to have from it one solution among all those present in superposition. But unlike pure quantum systems, the measurement here does not destroy the states' superposition. Since our algorithm operates on a conventional computer and does not require the presence of a quantum machine, it is possible and in our interest to keep all the possible solutions in the superposition for the next iterations. For each measurement result, we extract the transformation parameters and use them to transform the second image and have a calculated image. To evaluate the quality of an obtained solution, we compute the mutual information between the first image (reference image) and the calculated image. Greater the mutual information is, better the solution will be considered [5]. Afterwards, we select the 3 chromosomes from which derive the 3 best results and we select also randomly one chromosome from the others (in order to maintain a good diversity). So we have all in all 4 chromosomes which form the new population. The global best solution is then updated if a better one is found and the whole process is repeated until having satisfaction of a stopping criterion (maximum number of iterations in our experiments).

4 Experimental Results

We have applied the proposed algorithm on many pairs of images. The present section illustrates some obtained results when applying the registration algorithm.

The figure 5 shows at the left side two images belonging to the same modality (the second image is the result of the application of a geometric transformation on the first one) on which we have applied the proposed quantum genetic algorithm. The third image is calculated from the second one according the obtained parameters values.

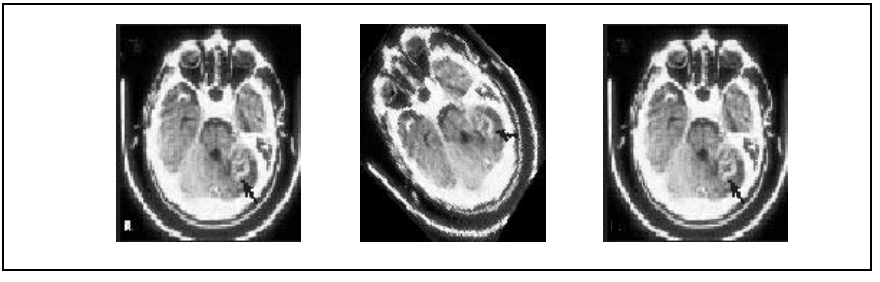


Fig. 5. Test images and registration result

The obtained transformation parameters values are:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} +5 \\ +1 \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} +0.4922 & +0.7422 \\ -0.8984 & +0.7578 \end{pmatrix}$$

The final mutual information is 1.6229

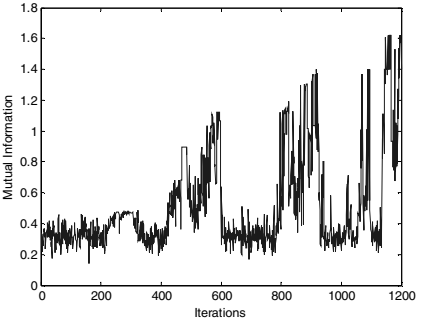


Fig. 6. Best mutual information evolution

The figure 6 illustrates the evolution of the best value of the mutual information through time.

The results below concern two images obtained from different sensors. The first is a Magnetic Resonance Image (MRI) and the second is an X-ray Computed Tomography Image (CTI).

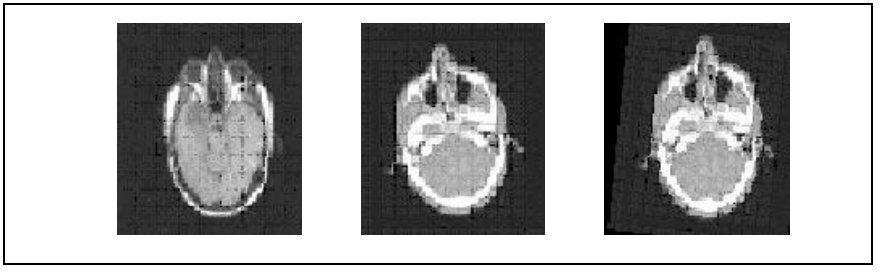


Fig. 7. Test images and registration result

The obtained transformation parameters values are:

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} -1 \\ +6 \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} +1.0469 & +0.1094 \\ -0.1094 & +0.9531 \end{pmatrix}$$

The final mutual information is 1.0015

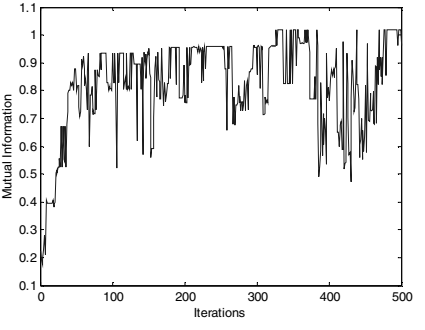


Fig. 8. Best mutual information evolution

The figure 6 illustrates the evolution of the best value of the mutual information through time.

We should notice that a good solution is obtained after less than 1500 iterations in both cases.

Classical genetic algorithms used to perform the same task [1] require an obviously greater number of chromosomes (usually more than 50), and dozens of thousands of iterations are needed to have acceptable solutions.

5 Conclusion

Our algorithm, compared to genetic algorithms, provides the advantage of giving a greater diversity by using quantum coding of solutions, i.e. all the solutions exist within each chromosome and what change are the probabilities to have one of them as a result of a measurement. Therefore, the size of the population does not need to be great. So, we have chosen to have only 4 chromosomes at the origin of each generation. Another advantage is that the interference provides in some way a guide for the population individuals and reinforces therefore the algorithm convergence. The use of the mutual information has allowed the registration of images coming from different sensors. That will provide the possibility of data fusion.

The obtained results are usually good. In comparison to classical genetic algorithms, the proposed quantum genetic algorithm has the advantage of using a small population size and the number of necessary iterations to have acceptable solution is also relatively smaller.

As ongoing work, we have first to search for the best choices concerning the population size, the mutation rate and the interference degree in order to warrant an optimal alignment for every pair of images. We would also use similar algorithms to solve efficiently other optimization problems.

References

[1] S. Meshoul, M. Batouche and K. Belhadj-moustefa. “An evolutionary framework for image data fusion based on the maximization of mutual information.” *Proceeding of the International Symposium on Software and Systems (I3S'2001)*, February 2001.

[2] K. Han and J. Kim, “Quantum-inspired evolutionary algorithm for a class of combinatorial optimization.” *IEEE transactions on evolutionary computation, vol. 6, no. 6*, December 2002.

[3] E. Rieffel and W. Polak, “An introduction to quantum computing for non-physicists”, *arxive.org*, quant-ph/9809016 v2, January 2000.

[4] H. Talbi, A. Draa and M. Batouche, “A quantum genetic algorithm for image registration.” *Proceeding of the 14th International Conference on Computer Theory and Applications (ICTTA’2004)*, April 2004, *IEEE Press, ISBN: 0-7803-8482-2/04*

[5] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, “Multimodality image registration by maximization of mutual information”, IEEE transactions on medical imaging, 16(2):187-198, 1997.

Nonparametric Impulsive Noise Removal

Bogdan Smolka[1⋆] and Rastislav Lukac[2]

[1] Polish-Japanese Institute of Information Technology,
Koszykowa 86 Str, 02-008 Warsaw, Poland
[2] The Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
University of Toronto, 10 King's College Road, Toronto ON, M5S 3G4, Canada

Abstract. In this paper a novel class of filters designed for the removal of impulsive noise in color images is presented. The proposed filter class is based on the nonparametric estimation of the density probability function in a sliding filter window. The obtained results show good noise removal capabilities and excellent structure preserving properties of the new impulsive noise removal technique.

1 Introduction

The majority of the nonlinear, multichannel filters are based on the ordering of vectors in a sliding filter window. The output of these filters is defined as the lowest ranked vector according to a specific vector ordering technique.

Let the color images be represented in the commonly used RGB color space and let $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$ be N samples from the sliding filter window W. Each of the $\mathbf{x}_i$ is an m-dimensional multichannel vector, (in our case $m = 3$). The goal of the vector ordering is to arrange the set of N vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ belonging to W using some sorting criterion.

In [1,2] the ordering based on the cumulative distance function $R(\mathbf{x}_i)$ has been proposed, $R(\mathbf{x}_i) = \sum_{j=1}^{N} \rho(\mathbf{x}_i, \mathbf{x}_j)$, where $\rho(\mathbf{x}_i, \mathbf{x}_j)$ is a function of the distance among $\mathbf{x}_i$ and $\mathbf{x}_j$. The ordering of the scalar quantities according to $R(\mathbf{x}_i)$ generates the ordered set of vectors. The most commonly used measure to quantify distance between two multichannel signals is the Minkowski norm $\rho_\gamma(\mathbf{x}_i, \mathbf{x}_j) = [\sum_{k=1}^{m} |x_{ik} - x_{jk}|^\gamma]^{1/\gamma}$. The Minkowski metric includes the city-block distance ($\gamma = 1$), Euclidean distance ($\gamma = 2$) and chess-board distance ($\gamma = \infty$) as the special cases.

One of the most important noise reduction filter is the vector median. In the scalar case of gray scale images, given a set W containing N samples, the median of the set is defined as $x_{(1)} \in W$ such that $\sum_j \left|x_{(1)} - x_j\right| < \sum_j |x_i - x_j|$, $\forall\, x_i, x_j \in W$. Median filters exhibit good noise reduction capabilities, (especially when long tailed noise is involved) and outperform simple nonadaptive linear filters in preserving signal discontinuities. As in many applications the signal is multidimensional, in [4] the *Vector Median Filter* (VMF) was introduced, by

⋆ This research has been supported by a grant No *PJ/B/01/2004* from the Polish - Japanese Institute of Information Technology

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 155–162, 2004.
© Springer-Verlag Berlin Heidelberg 2004

generalizing the scalar definition using a suitable vector norm. Given a set W of N vectors, the vector median of the set is defined as $\mathbf{x}_{(1)} \in W$ satisfying

$$\sum_j \left\|\mathbf{x}_{(1)} - x_j\right\| < \sum_j \left\|\mathbf{x}_i - \mathbf{x}_j\right\|, \quad \forall\ \mathbf{x}_i, \mathbf{x}_j \in W\,. \tag{1}$$

The orientation difference between two vectors can also be used as their distance measure. This so-called vector angle criterion is used by the *Vector Directional Filters* (VDF), to remove vectors with atypical directions, [3]. To improve the efficiency of the directional filters, another method called *Directional-Distance Filter* (DDF) was proposed. This filter retains the structure of the BVDF, but utilizes the combined distance criterions to order the vectors inside the processing window, [3,5].

2 Nonparametric Estimation

Nonparametric Density Estimation is based on placing a kernel function on every sample and on the summation of the values of all kernel function values at each point in the sample space, [6,7]. The nonparametric approach to estimating multichannel densities can be introduced by assuming that the color space occupied by the multichannel image pixels is divided into m-dimensional hypercubes. If h_N is the length of an edge of a hypercube, then its volume is given by $V_N = h_N^m$. If we are interested in estimating the number of pixels falling in the hypercube of volume V_N, then we can define the window function $\phi(\mathbf{x_i}) = 1$, if $|x_{ij}| \leq 1/2$, $j = 1, \ldots, m$ and 0 otherwise, which defines a unit hypercube centered in the origin.

The function $\phi\left(\|\mathbf{x} - \mathbf{x}_i\| / h_N\right)$ is equal to unity if the pixel $\mathbf{x}_i$ falls within the hypercube V_N centered at $\mathbf{x}$ and is zero otherwise. The number of pixels in the hypercube with the length of edges equal to h_N is then $k_N = \sum_{i=1}^{N} \phi\left(\|\mathbf{x} - \mathbf{x}_i\|/h_N\right)$ and the estimate of the probability that a sample $\mathbf{x}$ is within the hypercube is $p_N = k_N/NV_N$, which gives

$$p_N(\mathbf{x}) = (NV_N)^{-1} \sum_{i=1}^{N} \phi\left(\|\mathbf{x} - \mathbf{x}_i\|/h_N\right). \tag{2}$$

This estimate can be generalized by using a smooth kernel function K in place of $\phi(\cdot)$ and the width parameter h_N satisfying: $K(\mathbf{x}) = K(-\mathbf{x})$, $K(\mathbf{x}) \geq 0$, $\int K(\mathbf{x})\, d\mathbf{x} = 1$ and $\lim_{N\to\infty} h_N = 0$, $\lim_{N\to\infty} h_N^m = \infty$.

The multivariate estimator in the m-dimensional case is defined as

$$p_N^*(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_1 \ldots h_q} \mathcal{K}\left(\frac{|x_1 - x_{i1}|}{h_1}, \cdots, \frac{|x_q - x_{iq}|}{h_m}\right), \tag{3}$$

with $\mathcal{K}$ denoting a multidimensional kernel function $\mathcal{K}\colon \mathbb{R}^m \to \mathbb{R}$, $h_1, \ldots, h_m$ denoting bandwidths for each dimension and N being the number of samples in W. A common approach to build multidimensional kernel functions is to use a *product kernel* $\mathcal{K}(u_1, \ldots, u_m) = \prod_{i=1}^{m} K(u_i)$, where K is a one-dimensional kernel function

$$p_N^*(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{m}\left(\frac{|x_{ij}-x_j|}{h_i}\right). \tag{4}$$

The shape of the approximated density function depends heavily on the bandwidth chosen for the density estimation. Small values of h lead to spiky density estimates showing spurious features. On the other hand too big values of h produce over-smoothed estimates that hide structural features.

If we chose the Gaussian kernel, then the density estimate of the unknown probability density function at $\mathbf{x}$ is obtained as a sum of kernel functions placed at each sample $\mathbf{x}_i$

$$p_N(\mathbf{x},h) = \frac{1}{N\left(h\sqrt{2\pi}\right)^m}\sum_{i=1}^{N}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2h^2}\right). \tag{5}$$

The smoothing parameter h depends on the local density estimate of the sample data. The form of the data dependent smoothing parameter is of great importance for the non-parametric estimator.

Choosing the Gaussian kernel function for $\mathcal{K}$, the optimal bandwidth is

$$h^* = (4/(m+2))^{-\frac{1}{m+4}}\,\hat{\sigma}\,N^{-\frac{1}{m+4}}, \tag{6}$$

where σ denotes the approximation of the standard deviation of the samples. In one dimensional case (6) reduces to the well known, 'rule of thumb', $h^* = 1.06N^{-\frac{1}{5}}\hat{\sigma}$, [6,7]. A version which is more robust against outliers in the sample set can be constructed if the interquartile range is used as a measure of spread instead of the variance, [6]. This modified estimator is $h^* = 0.79\varrho N^{-\frac{1}{5}}\hat{\sigma}$, where ϱ is the inter-quartile range. Another robust estimate of the optimal bandwidth is $h^* = 0.9AN^{-\frac{1}{5}}\hat{\sigma}$ with $A = \min(\hat{\sigma}, \varrho/1.34)$.

Generally the simplified rule of choosing the optimal bandwidth h can be written as

$$h_1^* = C\;\hat{\sigma}\,N^{-\frac{1}{m+4}}, \tag{7}$$

where C is an appropriate weighting coefficient.

From the maximum likelihood principle and assuming independence of the samples, one can write the likelihood of drawing the complete dataset as the product of the densities of one sample

$$\mathcal{L}(h) = \prod_{j=1}^{N} p_N(\mathbf{x}_j,h) = \prod_{j=1}^{N}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\left(h\sqrt{2\pi}\right)^m}\exp\left(-\frac{\|\mathbf{x}_j-\mathbf{x}_i\|^2}{2h^2}\right). \tag{8}$$

As this likelihood function has a global maximum for h=0, in [8] a modified approach has been proposed

$$\mathcal{L}^*(h) = \left[\prod_{j=1}^{N}\frac{1}{N}\sum_{i=1,\,i\neq j}^{N}\frac{1}{\left(h\sqrt{2\pi}\right)^m}\exp\left(-\frac{\|\mathbf{x}_j-\mathbf{x}_i\|^2}{2h^2}\right)\right]^{\frac{1}{m}}. \tag{9}$$

This function has one maximum for h, which can be found by setting to 0 the derivative of the logarithm of $\mathcal{L}^*(h)$ with respect to h. A crude but rather fast

way to obtain an approximate solution is to assume that the density estimate of Eq. (4) on a certain location $\mathbf{x}$ in the feature space is determined by the nearest kernel only, [8]. In this case

$$\frac{\partial \log(\mathcal{L}^*(h))}{\partial h} = \frac{1}{N} \sum\nolimits_{j=1}^{n} \frac{\|\tilde{\mathbf{x}}_j - \mathbf{x}_i\|^2}{h^3} = \frac{m}{h}\,. \tag{10}$$

In this paper we use the optimal h derived from (10) defined as

$$h_2^* = C \left((mN)^{-1} \sum\nolimits_{j=1}^{N} \|\tilde{\mathbf{x}}_j - \mathbf{x}_j\|^2 \right)^{\frac{1}{2}}\,, \tag{11}$$

where $\tilde{\mathbf{x}}_i$ represents the nearest neighbor of the sample $\mathbf{x}_i$, and C is a tuning parameter.

3 Proposed Algorithm

Let us assume a filtering window W containing N image pixels, $\{x_1, \ldots, x_N\}$ and let us define the similarity function $\mu : [0; \infty) \to \mathbb{R}$ which is non-ascending and convex in $[0; \infty)$ and satisfies $\mu(0) = 1$, $\mu(\infty) = 0$. The similarity between two pixels of the same intensity should be 1, and the similarity between pixels with minimal and maximal gray scale values should be very close to 0. The function $\mu(x_i, x_j)$ defined as $\mu(x_i, x_j) = \exp\{-[(x_i - x_j)/h]^2\}$, where h is the bandwidth of the Gaussian kernel, defined by (7) or (11), satisfies the required conditions.

Let us additionally define the cumulated sum M of similarities between a given pixel and all other pixels belonging to window W. For the central pixel x_1 we introduce M_1 and for the neighbors of x_1 we define M_k as

$$M_1 = \sum_{j=2}^{N} \mu(x_1, x_j), \quad M_k = \sum_{j=2,\ j\neq k}^{N} \mu(x_k, x_j),\ k > 1, \tag{12}$$

which means that for x_k, which are neighbors of x_1, we do not take into account the similarity between x_k and x_1, which is the main idea of this algorithm. The omission of the similarity $\mu(x_k, x_1)$ when calculating M_k, privileges the central pixel, as in the calculation of M_1 we have $N-1$ similarities $\mu(x_1, x_k)$, $k > 2$ and for M_k, $k > 1$ we have only $N - 2$ similarity values, as the central pixel x_1 is excluded from the calculation of M_k, [9,10], (see Fig. 1).

In the construction of the new filter, the reference pixel x_1 in the window W is replaced by one of its neighbors if $M_1 < M_k,\ k = 2, \ldots, N$. If this is the case, then x_1 is replaced by that x_{k^*} for which $k^* = \arg\max M_k,\ k = 2, \ldots, N$. In other words x_1 is detected as being corrupted if $M_1 < M_k,\ k = 2, \ldots, N$ and is replaced by its neighbors x_k which maximizes the sum of similarities M between all the pixels from W excluding the central pixel.

The basic assumption is that a new pixel must be taken from the window W (introducing pixels, that do not occur in the image is prohibited like in the VMF). For this purpose μ must be convex, which means that in order to find a maximum of the sum of similarity functions M it is sufficient to calculate the values of M only in points $x_1, x_2, \ldots, x_N$.

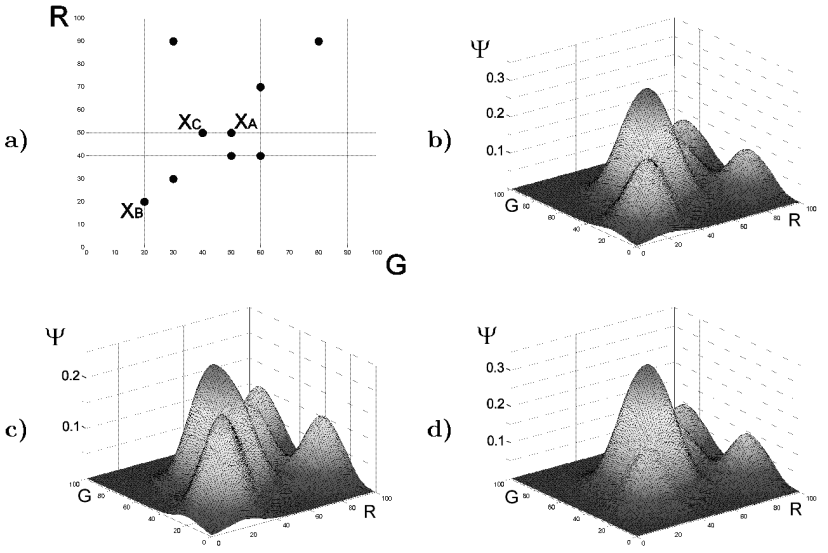


Fig. 1. Impulsive noise removal technique in the 2D case. Fig. **a**) depicts the arrangement of pixels in W and Fig. **b**) their nonparametric probability density estimation. Figs. **c**) and **d**) present the density plots for the cases when the central pixels x_A and x_B are removed from W. It can be seen that in the first case **c**) the pixel $x_1 = x_A$ will be retained and in the second case **d**) the pixel $x_1 = x_B$ will be replaced by x_A. The pixel x_A will be preserved, as in Fig. **c**) the plot attains its maximum at x_C, but this maximum is less than the maximum for x_A in Fig. **b**). Regarding sample x_B, its rejection causes that the maximum is attained at x_A and this pixel will replace the central pixel x_B.

The presented approach can be applied in a straightforward way to multichannel images using the similarity function defined as $\mu(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-[\|\mathbf{x_i} - \mathbf{x_j}\|/h]^2\}$, where $\|\cdot\|$ denotes the specific vector norm and h denotes the bandwidth. Now in exactly the same way we can maximize the total similarity function M for the vector case.

4 Results

The performance of the proposed impulsive noise reduction filters was evaluated using the widely used PSNR quality measure. Figure 2a) shows the dependence of the noise attenuation capability of the proposed filter class on the bandwidth type h_1^* and h_2^* defined by (7) and (11). Clearly the filter based on the h_2^* outperforms the technique based on the h_1 bandwidth for the whole range of used contamination probabilities p, ($p = 0.01$ - 0.1).

Figure 2b) presents the dependence of the PSNR restoration quality measure on the kind of the Minkowski norm. Surprisingly, the L_∞ norm yields significantly better results than the L_1 or L_2 norms. This is the result of the construction of the h_2^* bandwidth, which depends on the nearest neighbor in the sliding

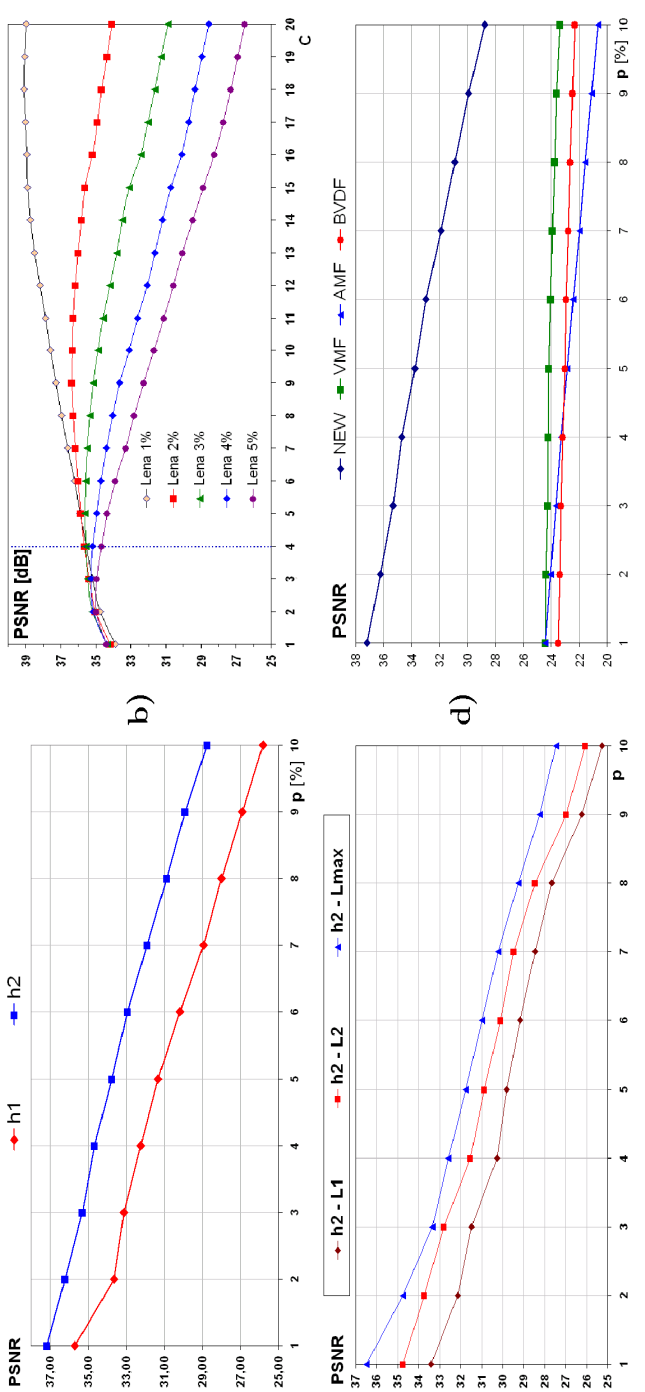


Fig. 2. Dependence of the efficiency of the proposed filtering scheme on the bandwidth h_1^* (7) and h_2^* (11) - (**a**), below the dependence of the PSNR value on the value of the tuning parameter C in (11) - (**b**) and the dependence on the kind of Minkowski norm for the bandwidth h_2* - (**c**). At the bottom (**d**) the comparison of results obtained using the h_2^* bandwidth, L_∞ norm and $C = 4$ with the standard multichannel filters VMF and BVDF, (test were performed on the color image *LENA*); p denotes the probability of a pixel corruption - to RGB channels random, uniformly distributed values from the interval [0,255] were assigned.

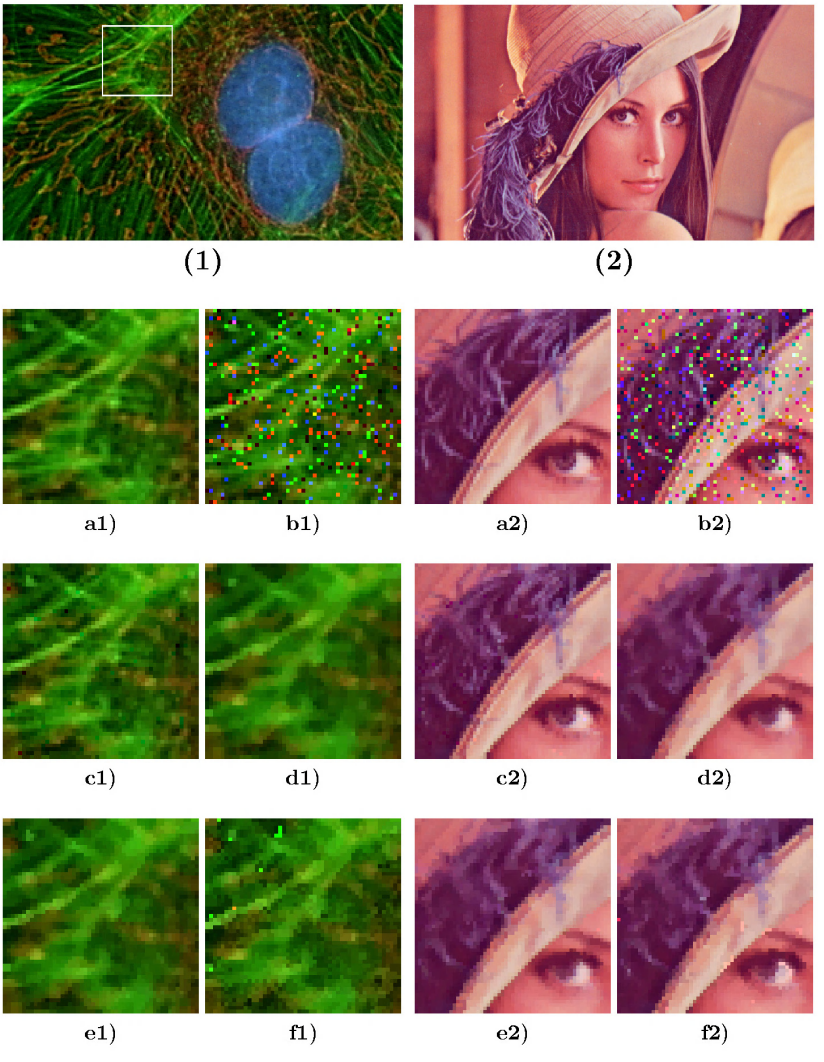


Fig. 3. Illustrative example of the efficiency of the proposed algorithm: **a**) zoomed part of a LENA color test image, **b**) image corrupted by 3% of impulsive noise, **c**) image after filtering with the proposed filter, **d**) VMF output, **e**) DDF output, **f**) BVDF output.

filter window. This behavior is advantageous, as the calculation of the L_∞ norm is much faster than the evaluation of distances determined by L_1, L_2 norms.

The efficiency of the filters based on adaptive h_1^* and h_2^* bandwidths are dependent, (especially for very small noise contamination) on the coefficient C in (7) and (11). Figure 2c) shows the dependence of PSNR for the filter based on h_2^* as a function of C in (11). For low noise intensity the parameter C should be

significantly larger than for the case of images corrupted by heavy noise process. However, setting C to 4 is an acceptable trade-off, as can be seen in Fig. 2 d), which depicts the efficiency of the proposed filter in comparison with VMF, AMF and BVDF. It can be observed that although the $C = 4$ is not an optimal setting for the whole range of tested noise intensities, nevertheless the described filter yields much better results than the traditional techniques.

This is also testified by Fig. 3, which compares the filtering results obtained by the filter based on adaptive h_2^* bandwidth with the performance of the *reference* VMF, BVDF, DDF filter. As can be observed the new filtering has much better detail preserving properties than VMF, BVDF and DDF.

5 Conclusions

In this paper a new nonparametric technique of impulsive noise removal in multichannel images has been proposed. The described filter class is based on the estimation of the kernel bandwidth using the technique proposed in [8]. The experiments revealed, that the proposed algorithm yields the best results when applying the L_∞ norm, which makes the filter computationally very attractive. The obtained results show that the proposed technique excels significantly over the standard techniques like VMF, BVDF and DDF.

References

1. Pitas, I., Tsakalides, P.: Multivariate ordering in color image processing. IEEE Trans. on Circuits and Systems for Video Technology, 1, 3, (1991) 247-256
2. Tang, K., Astola, J., Neuovo, Y.: Nonlinear multivariate image filtering techniques. IEEE Trans. on Image Processing, 4, 6, (1995) 788-797
3. Trahanias, P.E., Venetsanopoulos, A.N.: Vector directional filters: a new class of multichannel image processing filters. IEEE Trans. on Image Processing, 2, 4, (1993) 528-534
4. Astola, J., Haavisto, P., Neuvo, Y.: Vector median filters. Proceedings of the IEEE, 78, (1990) 678-689
5. Plataniotis, K.N., Venetsanopoulos, A.N.: Color Image Processing and Applications. Springer Verlag (2000)
6. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. London, Chapman and Hall (1986)
7. Scott, D.W.: Multivariate Density Estimation. New York, John Wiley (1992)
8. Kraaijveld, M.A.: A Parzen classifier with an improved robustness against deviations between training and test data. Pattern Recognition Letters, 17, (1996) 679-689
9. Smolka, B., Plataniotis, K.N., Chydzinski, A., Szczepanski, M., Venetsanopulos, A.N., Wojciechowski, K.: Self-adaptive algorithm of impulsive noise reduction in color images. Pattern Recognition, 35, (2002) 1771-1784
10. Smolka, B., Lukac, R., Chydzinski, A., Plataniotis, K.N., Wojciechowski, K.: Fast adaptive similarity based impulsive noise reduction filter. Real Time Imaging, 9, (2003) 261-276

BayesShrink Ridgelets for Image Denoising

Nezamoddin Nezamoddini-Kachouie, Paul Fieguth, and Edward Jernigan

University of Waterloo, Department of Systems Design Engineering, Waterloo, Ontario, N2L 3G1, Canada
{nnezamod, pfieguth, jernigan}@uwaterloo.ca
ocho.uwaterloo.ca

Abstract. The wavelet transform has been employed as an efficient method in image denoising via wavelet thresholding and shrinkage. The ridgelet transform was recently introduced as an alternative to the wavelet representation of two dimensional signals and image data. In this paper, a BayesShrink ridgelet denoising technique is proposed and its denoising performance is compared with a previous VisuShrink ridgelet method. To derive the results, different wavelet bases such as Daubechies, symlets and biorthogonal are used. Experimental results show that BayesShrink ridgelet denoising yields superior image quality and higher SNR than VisuShrink.

1 Introduction

Data obtained from the real world in the form of signals do not exist without noise. This noise might decrease to some negligible levels under ideal conditions such that denoising is not necessary, but usually to recover the signal the corrupting noise must be removed for practical purposes. For this reason noise elimination is a main concern in computer vision and image processing. Noise undesirably corrupts the image by perturbations which are not related to the scene under study and ambiguates the underlying signal relative to its observed form.

The goal of denoising is to remove the noise and to retain the important signal features as much as possible. To achieve this goal, traditional approaches use linear processing such as Wiener filtering. In the presence of additive noise, linear filters, which consist of convolving the image with a constant matrix to obtain a linear combination of neighborhood values, can produce a blurred and smoothed image with poor feature localization and incomplete noise suppression. To overcome this shortcoming, nonlinear filters have been proposed. Much research has focused recently on signal denoising using nonlinear techniques; of which one of the most important is wavelet based denoising. The wavelet transform generally separates signal and noise; as a result it can be used to remove the noise while preserving the signal characteristics. Researchers have employed various approaches to nonlinear wavelet-based denoising: In one approach, wavelet thresholding, a hard threshold function keeps a coefficient if it is larger than a threshold and sets it to zero otherwise; in another,

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 163–170, 2004.
© Springer-Verlag Berlin Heidelberg 2004

wavelet shrinkage takes the coefficient and shrinks it toward zero by some threshold. Both approaches are nonlinear and operate on one wavelet coefficient at a time.

Recently the Ridgelet and Curvelet transforms were developed to reduce the limitations of wavelet-based image processing. The two-dimensional wavelet image transform produces large coefficients along important edges even at fine scales. Hence, edges of an image appear as large wavelet coefficients repeatedly at fine scales, so to properly reconstruct the edges of the image many wavelet coefficients are required. The estimation of so many coefficients makes wavelet denoising techniques complex.

On the other hand, wavelet transforms can catch the point singularities of one-dimensional (1-D) signals, thus they have a good performance for one-dimensional smooth functions. To discover 1-D singularities in two-dimensional (2-D) signals, the wavelet transform faces some difficulties. Images that contain 2-D smooth signals have 1-D singularities across the edges which separate the smooth regions. Although edges are generally smooth curves, as borders of two smooth regions they are discontinues. Since the 2-D wavelet transform is the product of 1-D wavelets, it discovers the singularities across the edges but it doesn't recognize the smoothness along the edges. To compensate for this weakness of the wavelet transform in higher dimensions, ridgelet and curvelet transforms were recently introduced by Candes and Donoho [1-4].

Different denoising methods have been proposed for signal denoising via wavelet. On the other hand VisuShrink ridgelet thresholding has been recently introduced [5] as an alternative to the wavelet denoising and performs better than wavelet for the images with straight lines. In this paper BayesShrink ridgelet image denoising is proposed and the results are compared with those of VisuShrink ridgelet method. The following Section explains rigelet image denoising. In Section three ridgelet thresholding techniques are described. The proposed method is presented in Section four. In Section five the results of the proposed method and the previous VisuShrink technique are compared and the conclusions are presented in Section 6.

2 Ridgelet Image Denoising

The ridgelet transform was proposed as an alternative to the wavelet transform to represent 2-D signals. Since a sparse representation of smooth functions and straight edges is provided by the ridgelet transform, this new expansion can accurately represent both smooth functions and edges with a few nonzero coefficients and achieves a lower mean square error (MSE) than the wavelet transform.

2.1 Ridgelet Transform

The ridgelet transform effectively represents line singularities of 2-D signals. It maps the line singularities into point singularities in the Radon domain by employing the embedded Radon transform. Therefore, the wavelet transform can efficiently be ap-

plied to discover the point singularities in this new domain. Having the ability to approximate singularities along a line, several terms with common ridge lines can effectively be superposed by the ridgelet transform. The bivariate ridgelet transform in R^2 is defined by:

$$\Re_{\alpha,\beta,\theta}(\kappa) = \alpha^{-1/2}\omega((\kappa_1 \cos\theta + \kappa_2 \sin\theta - \beta)\ /\alpha) \tag{1}$$

where ω is a univariate wavelet function on $R \rightarrow R$. $\alpha > 0$, β and θ are scale, location and orientation parameters respectively. Along the ridgelet lines $\kappa_1 \cos\theta + \kappa_2 \sin\theta$, ridgelets are constant and they are equal to the wavelets in the orthogonal direction. ridgelet coefficients of a bivariate function $I(\kappa)$ in R^2 are given by:

$$\Re_I(\alpha,\beta,\theta) = \int \Re_{\alpha,\beta,\theta}\ (\kappa) I(\kappa) d\kappa \tag{2}$$

The reconstruction formula is given by:

$$I(\kappa) = \int_0^{2\pi} \int_{-\infty}^{+\infty} \int_0^{\infty} \Re_I(\alpha,\beta,\theta)\, \Re_{\alpha,\beta,\theta}\ (\kappa) \frac{d\alpha}{\alpha^3} d\beta \frac{d\theta}{4\pi} \tag{3}$$

and is valid for integrable (and square integrable) functions. Like Fourier and wavelet transforms, any arbitrary function can be represented by continuous superposition of ridgelets. Considering the 2-D ridgelet transform as a 1-D wavelet transform in the Radon domain, the ridgelet coefficients of function $I(\kappa)$ can be defined as:

$$\Re_I(\alpha,\beta,\theta) = \int \Re_t(\theta,\tau)\alpha^{-1/2}\omega((\tau-\beta)/\alpha) d\tau \tag{4}$$

where $R_t(\theta,t)$ is the Radon transform of function $I(\kappa)$ and is given based on the Dirac distribution (δ) as:

$$\Re_t(\theta,\tau) = \int I(\kappa_1,\kappa_2)\delta(\kappa_1 \cos\theta + \kappa_2 \sin\theta - \tau) d\kappa_1 d\kappa_2 \tag{5}$$

2.2 Ridgelet Denoising Concept

To explain the ridgelet denoising procedure, assume *I[i,j]* to be the original *M* by *M* image, where *i and j = 1, 2,…, M*, and *S[i,j] = I[i,j] + n[i,j]* is the image corrupted by additive noise *n[i,j]* which is identically distributed and independent of *I[i,j]*. In the first step of ridgelet denoising, the observed image *S* is transformed into the ridgelet domain. Then the ridgelet coefficients are thresholded and finally the denoised coefficients are transformed back to reconstruct the image. Let R_D and R_R be the forward ridgelet decomposition and inverse ridgelet reconstruction transforms. Assume *T* and τ to be the thresholding operator and threshold respectively. The ridgelet thresholding can be summarized as:

(i) $I_R = R_D(I)$, *(ii)* $I_T = T(I_R, \tau)$, *and (iii)* $\hat{I} = R_R I_\tau$

The choice of the threshold and the method which is used to calculate the threshold, determine how efficient the denoising technique would be. Although selecting a small threshold may produce an output image close to the input, the recovered image may still be noisy. On the other hand, a choice of a large threshold may yield a blurred image by setting most of the wavelet coefficients to zero. Two different thresholding techniques, VisuShrink and BayesShrink, are explained in the following Section.

3 Thresholding Techniques

The ridgelet denoising is used to recover the original signal from the noisy one by removing the noise. In contrast with denoising methods that simply smooth the signal by preserving the low frequency content and removing the high frequency components, the frequency contents and characteristics of the signal would be preserved during ridgelet denoising.

3.1 VisuShrink

VisuShrink, proposed by Donoho and Johnstone [6,9], uses the universal threshold given by:

$$U_t = \sigma_n \sqrt{2 \log M} \tag{6}$$

where σ_n and M are the noise variance and the number of image pixels respectively. Donoho and Jonstone have proved [6-8] that the maximum of any M independent and identically distributed (iid) values with high probability is less than the universal threshold U_t. As M is increased the probability will be closer to one, so with a high probability pure noise signals are set to zero. Since the universal threshold is obtained by considering the constraint that the noise is less than the threshold with high probability as M increases, it tends to be high for large values of M and as a result will shrink many noisy ridgelet coefficients to zero and produce smoothed estimated images.

3.2 BayesShrink

As it was suggested for the subband wavelet coefficients by [10,11], the subband ridgelet coefficients of a natural image can also be described by the Generalized Gaussian Distribution (GGD) as:

$$GG_{\sigma_I,\gamma}(I) = P(\sigma_I, \gamma) \exp\{-[\delta(\sigma_I, \gamma) \mid I \mid]^{\gamma}\} \tag{7}$$

where $-\infty < I < +\infty$, $\gamma > 0$ and,

$$\delta(\sigma_I, \gamma) = \sigma_I^{-1}\left[\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}\right]^{\frac{1}{2}} \tag{8}$$

and,

$$P(\sigma_I, \gamma) = \frac{\gamma \bullet \delta(\sigma_I, \gamma)}{2\Gamma\left(\frac{1}{\gamma}\right)} \tag{9}$$

σ_I is the standard deviation of subband ridgelet coefficients, γ is the shape parameter and Γ is Gamma function. For most natural images the distribution of the ridgelet coefficients in a subband can be described with a shape parameter γ in the range of [0.5,1]. Considering such a distribution for the ridgelet coefficients and estimating γ and σ_I for each subband, the soft threshold T_S which minimizes the Bayesian Risk [10,11], can be obtained by:

$$\Re(T_S) = E(\hat{I} - I)^2 = E_I E_{J|I}(\hat{I} - I)^2 \tag{10}$$

where $\hat{I}$ is $T_S(J)$, $J|I$ is $N(I,\sigma)$ and I is $GG_{\sigma_I,\gamma}$. Then the optimal threshold T_S^* is given by:

$$\mathrm{T_S^*}(\sigma_\mathrm{I}, \gamma) = \arg\min_{T_S} \ \Re(T_S) \tag{11}$$

Numerical calculation is used to find T_S^* since it does not have a closed form solution. A proper estimation of the value T_S^* is concluded by setting the threshold as [10,11]:

$$\hat{T}\ (\hat{\sigma}_I) = \frac{\hat{\sigma}_n}{\hat{\sigma}_I} \tag{12}$$

4 Calculating the BayesShrink Threshold by the Proposed Method

Subband dependent threshold is used to calculate BayesShrink ridgelet threshold. The estimated threshold is given by (12) where σ_n and σ_I are noise and signal standard deviations respectively. The 1-D ridgelet coefficients corresponding to different directions are depicted in Fig. 1. In this figure each column corresponds to a specific direction, hence the number of columns determines the number of directions and each column contains subband detail coefficients for L different decomposition levels.

To estimate the noise variance σ_n^2 from the subband details, the median estimator is used on the 1-D subband coefficients:

$$\hat{\sigma}_n = median(|Details|)/0.6745 \tag{13}$$

Signal standard deviation is calculated for each direction in each subband detail individually. Thus having *N* directions and *L* subband, *NxL* different σ_I must be estimated corresponding to *NxL* subband-directions coefficients. Note that in BayesShrink wavelet denoising, σ_I is estimated on 2-D dyadic subbands [10,11]. Thus having *L* decomposition levels, *3xL* different σ_I must be estimated to calculate the thresholds for the different subbands. To estimate the signal standard deviation *(σ_I),* the observed signal *S* is considered to be $S = I + n$ and signal *(I)* and noise *(n)* are assumed to be independent. Therefore,

$$\sigma_S^2 = \sigma_I^2 + \sigma_n^2 \tag{14}$$

where σ_S^2 is the variance of the observed signal. So $\hat{\sigma}_I$ is estimated by:

$$\hat{\sigma}_I = \sqrt{\max((\hat{\sigma}_S^2 - \hat{\sigma}_n^2),0)} \tag{15}$$

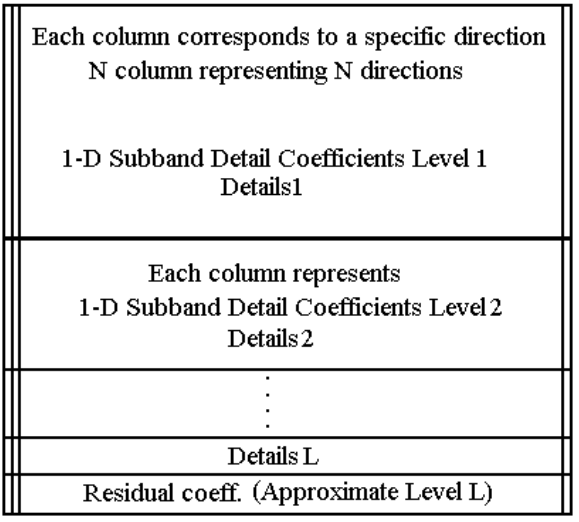


Fig. 1. Subband ridgelet coefficients: *N* directions (columns) and *L* levels which conclude *NxL* subband-direction coefficients

5 Results

In this section the proposed ridgelet denoising technique is used to recover the noisy images which are corrupted with additive white noise. BayesShrink and VisuShrink ridgelet image denoising methods are implemented and based on different wavelet bases the results are compared. Since the ridgelet transform performs better on images with straight lines, the test image in the following experiments, as depicted in Fig. 2, is an image with perfectly straight lines which has been used in [5]. Denoised images depicted in Fig. 2(c1)-2(e1) and 2(c2)-2(e2) are derived using the BayesShrink and VisuShrink thresholding methods respectively. The results are obtained based on

three different wavelet bases including Daubechies, Symlets and Biorthogonal. As we can observe according to the SNR measurements, the results obtained by BayesShrink ridgelet method are better than those obtained by VisuShrink ridgelet method using different wavelet bases. On the other hand based on image quality BayesShrink provides superior results than VisuShrink. Therefore, regardless of the wavelet bases BayesShrink ridgelet provides better performance than VisuShrink ridgelet denoising.

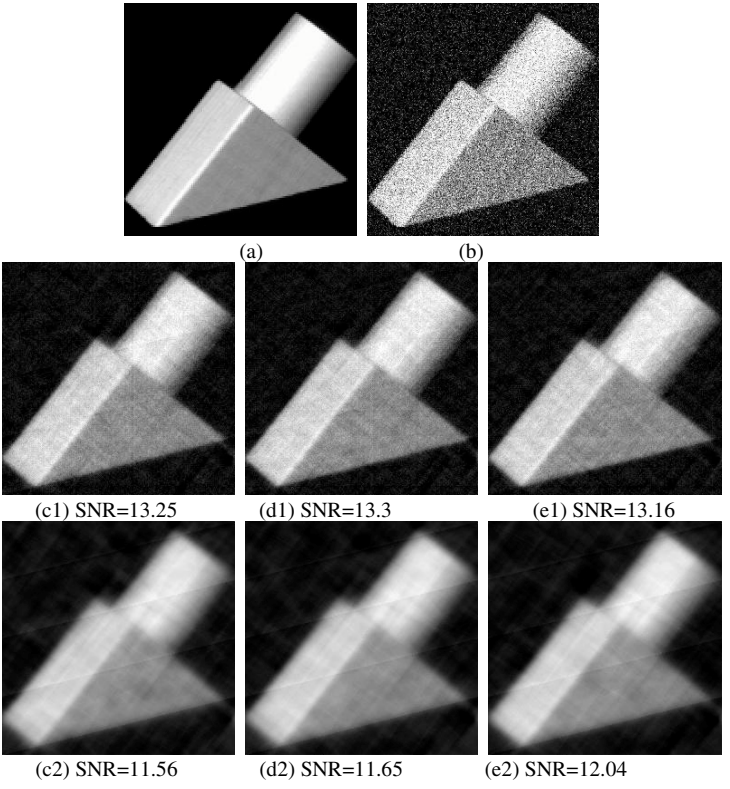


Fig. 2. (a) Original Image. (b) Noisy Image with SNR = 7.22. BayesShrink Ridgelet Denoising: (c1) db4. (d1) sym8. (e1) bior3.9. VisuShrink Ridgelet Denoising: (c2) db4. (d2) sym8. (e2) bior3.9.

6 Conclusions

In this paper the ridgelet transform for image denoising was addressed. BayesShrink ridgelet denoising was proposed. The proposed method was applied on test images with perfectly straight lines. The denoising performance of the results was compared with that of the VisuShrink ridgelet image denoising method. The experimental results by the proposed method showed the superiority of the image quality and its higher SNR in comparison with VisuShrink ridgelet technique. Furthermore we ob-

served that regardless of the selected wavelet basis, BayesShrink ridgelet performs better than VisuShrink ridgelet denoising method. However, the choice of the wavelet bases might affect the performance of both methods.

Future work is needed to improve the performance of this method. The BayesShrink curvelet denoising would also be compared with BayesShrink ridgelet denoisisng method. Moreover, the effect of the wavelet bases and the number of the decomposition levels on the performance of the denoised images would be investigated based on wavelet, ridgelet and curvelet methods.

References

1. Candes, E. J.: Ridgelets: Theory and Applications, Ph.D. thesis, Department of Statistics, Stanford University (1998)
2. Candes, E. J., Donoho, D. L.: Ridgelets: a key to higher dimensional intermittency?, Phil. Trans. R. Soc. Lond. A. (1999) 2495-2509
3. Donoho, D. L., Duncan, M. R.: Digital Curvelet Transform: Strategy, Implementation and Experiments, Proc.SPIE, Vol. 4056 (2000) 12-29
4. Starck, J. L., Candes, E. J., Donoho, D. L.: The Curvelet Transform for Image Denoising, IEEE Tran on Image Processing, Vol. 11, No. 6 (2002) 670-684
5. Do, M. N., Vetterli, M.: The Finite Ridgelet Transform for Image Representa tion, IEEE Tran. on Image Processing, Vol. 12, No.1 (Jan. 2003) 16 – 28
6. Donoho, D. L., Johnstone, I. M.: Ideal Spatial Adaptation via wavelet Shrinkage, Biometrika, Vol. 81 (Sept. 1994) 425-455
7. Donoho, D. L., Johnstone, I. M.: Adapting to Unknown Smoothness via Wavelet Shrinkage, Biometrika, Vol. 81 (Sept 1994) 425-455
8. Donoho, D. L.: Denoising by Soft Thresholding, IEEE Tran. on Inf. Theory, Vol. 41 (May1997) 613-627
9. Taswell, C.: The What, How, and Why of Wavelet Shrinkage Denoising, IEEE Journal Computing in Science and Engineering, Vol. 2, No. 3. (May-June 2000) 12-17
10. Chang, S. G., Yu, B., Vetterli, M.: Adaptive Wavelet Thresholding for Image Denoising and Compression, IEEE Trans. on Image Processing, Vol. 9, No. 9 (2000) 1532-1546
11. Chang, S. G., Yu, B., Vetterli, M.: Spatially Adaptive Wavelet Thresholding with Context Modeling for Image Designing, IEEE Tran on Image Processing, Vol. 9, No. 9 (2000) 1522-1531

Image Salt-Pepper Noise Elimination by Detecting Edges and Isolated Noise Points

Gang Li and Binheng Song

School of Software, Tsinghua University,
100084, Beijing, P.R. China
lig02@mails.tsinghua.edu.cn
bsong@math.tsinghua.edu.cn

Abstract. It deals an algorithm for removing the impulse noise, which is also called salt-pepper noise, in this paper. By evaluating the absolute differences of intensity between each point and its neighbors, one can detect the edges, the isolated noise points and blocks. It needs to set up a set of simple rules to determine the corrupted pixels in a corrupted image. By successfully identifying the corrupted and uncorrupted pixels, especially for the pixels nearing the edges of a given image, one can eliminate random-valued impulse noise while preserving the detail of the image and its information of the edges. It shows, in the testing experiments, that it has a better performance for the algorithm than the other's mentioned in the literatures.

1 Introduction

There are two kinds of image filters: linear and nonlinear. The linear and nonlinear filters are suitable for removing additive and Gaussian noise [1] while the nonlinear ones play well in case of the impulse noise, a kind of black-white spots usually called salt-pepper noise [2]. It has proposed a method for removing impulse noise in the paper. It is sensitive to nonstationarity, which is prevalent to images and blurring of image edges and structures, for linear filters [3], so it only studies the nonlinear filters here in this paper.

The median filter, which appeared as a tool for time series analysis [4] and was first applied in image processing by Pratt [5] and Frieden [6], has the abilities of suppressing impulse and preserving image edges and details. But its performance is not satisfactory in case of the heavy noise density. It has developed many nonlinear filters to improve the performances based on the median filter. They are the center weighted median filter (CWMF)[7], the multistage median filter (MMF)[8], the multi-state median filter (MSMF)[3], the nonlinear adaptive filter (NLAF)[9], the improved rank conditioned median filter (IRCMF)[10], the improved length-self-adaptive filter [11], the local-signal-statistical-character-based filter [12], the RCRS filter [13], the ROM filter [14] and etc.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 171–178, 2004.
© Springer-Verlag Berlin Heidelberg 2004

A so-called isolated point-edge-detected weighted median filter (IPEDWMF), based on edges and isolated noise points detection, is well studied in this paper. It demonstrates its capabilities in preserving edges and details of images while suppressing impulse noise. Combined with MSMF and IPEDWMF, we get an algorithm having better performance than the others for removing impulse noise.

The rest of this paper is organized as following. It introduces a noise model in section 2. In section 3, the algorithm IPEDWMF and the one of synthesizing of MSMF and IPEDWMF are introduced. The performances comparisons for different algorithms are presented in Section 4, followed by the conclusions and the acknowledgement.

2 Impulse Noise Model

A noise model, which is close to the realistic situation proposed in [3], is adopted in this paper. Let *Sij* and *Xij* denote the intensity values of the pixels at location (*i*,*j*) for the original and corrupted image respectively. The corrupted image with noise ratio *p* is defined as following:

$$X_{ij} = \begin{cases} S_{ij}, \textit{with probability } 1\text{-}p \\ N_{ij}, \textit{with probability } p \end{cases} \tag{1}$$

where N_{ij} is a random variable with a uniform distribution between 0 and 255.

3 Filtering Algorithms

3.1 Isolated Point-Edge-Detected Weighted Median Filter (IPEDWMF)

According to [15], pixels are classified into four types: common uncorrupted pixels, edge pixels, isolated impulse noise pixels and non-isolated impulse noise pixels. And [15] also proposes a method for identifying the pixel types applied to relevant filters to process the image. But the method is effective only in case of the fixed-valued impulse noise. And more, its pixel's type identifying procedure is complicated.

Here is our method for identifying pixel's type, whose principle is illustrated in Figure (1). And four parameters used in the algorithm are defined as follows: weight of the central pixel (*Weight*, *W*), noise block threshold (*NoiseBlockThreshold*, *NBT*), noise threshold (*NoiseThreshold*, *NT*) and isolated factor (*IsolatedFactor*, *IF*). Calculate the absolute differences of gray-value between the central pixel and its neighbors for each 3×3 filter windows. Sort the eight values in an array in ascending order. Select a noise threshold (*NT*), which is decided by the following algorithm. Compare the values of the array with *NT*, if its maximum is smaller than *NT*, then the central pixel is considered to be an uncorrupted pixel ((a) in Figure (1)), where the green square means that its absolute difference with the center one is less than *NT* and black means it is not less than *NT*; If the minimum of the array is bigger than *NT*, the central

pixel is assumed to be an isolated noisy pixel, as shown in (b). The central pixel displayed in (c) could be the one in a non-isolated noise blocks, in (d) could be the one on an edge, and in (e) is likely to be a pixel in a noise block. In the noise identifying map (f), where the green square is regarded as common uncorrupted pixels and the black is regarded as the others, the long tilt black line could be regarded as an edge, so it does not be treated as noise points as others do. It is shown by experiments that the proposed method detects the noise types effectively.

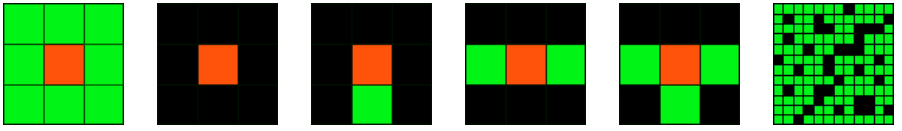


Fig. 1. The red square represents the central pixel. The green squares mean that its absolute gray value difference with the central pixel is smaller than *NT*. The black squares mean that its absolute gray value difference with the central pixel is bigger than or equal to *NT*. We give the isolated factor *IF* of corresponding (b), (c), (d) and (e) respectively to be 0, 1, 2 and 3. In (f) the green squares denote the uncorrupted pixels and the black denote the pixels identified as noise. A noise map like (f) will be generated for each combination of *IF* and *NT*.

IPEDWMF is based on the above principle. It removes impulse noise through three steps: i) compute the noise map (f) for an image; ii) determine the uncorrupted pixels including the edge pixels from the map; iii) eliminate noisy pixels. The range of *IF* is [0,3]. *NT* takes its value of 64, 48, 32 or 16. *W* (It's correlative with Med^{w} in formula (6), but not the same as *w*.) and *NBT* are determined by:

$$W = \begin{cases} 3, W > 0 \; or \; (W = 0 \; and \; IF \le 1) \\ 0, others \end{cases} \tag{2}$$

$$NBT = \begin{cases} 3, & IF = 1 \; and \; NT \le 16 \\ NT/8 - 2, & IF > 1 \; or \; NT > 16 \end{cases} \tag{3}$$

In order to record the detected noise information, two matrixes NIM_0 and NIM_1 with same size as the image are used to save the correlative information. Let D_0 denote the array of the current pixel's absolute intensity difference with its eight neighboring pixels (using 3×3 filter window) and D_1 be the array in ascending corresponding to D_0. NIM_0 is calculated by:

$$NIM_0[i,j] = \begin{cases} 1, & D_{1,ij}[IF] \ge NT \\ 0, & others \end{cases} \tag{4}$$

NIM_0 stores the information of the isolated noise pixels, noise blocks and edges, where it need to differ the edge pixels and the noise pixels. Define noise block diameter *d* as the maximum length of continuous '1's on *X*- or *Y*-axis in NIM_0. We can obtain the NIM_1:

$$NIM_1[i,j] = \begin{cases} NIM_0[i,j], d_{ij} < NBT, NIM_0[i,j] = 1 \\ 0\ , \qquad others \end{cases} \tag{5}$$

Let X, $\hat{X}$ and Y denote the corrupted image, the estimation of the uncorrupted image and the image filtered by a modified CWMF respectively, we get:

$$Y = Med^w(X[i,j]) \tag{6}$$

$$\hat{X}[i,j] = \begin{cases} Y[i,j]\ , & NIM_1[i\ ,j] = 1 \\ X[i,j]\ , & others \end{cases} \tag{7}$$

where $Med^w(X[i,j])$ denotes the median value in the window with the size of w at the location (i,j) in X, w is started at 3 increased by 2 if the corrupted pixel in the $w \times w$ window is over 20% till to 9.

To improve the performance, the filtering process is iterated by different values of NT and IF. NT takes every value from {64,48,32,16} and IF gets its value from {0,1,2,3,3,2,1,0}. So there are 32 iterations. Obviously the performance is improved at the cost of computational efforts.

3.2 Synthesis of MSMF and IPEDWMF

To further improve the filter performance, MSMF is used as the pre-processor, whose output is as the input of IPEDWMF. We define the corrupted ratio CR and use $\hat{CR}$ to denote the estimation of CR:

$$CR = \frac{NUM_c(X)}{NUM_t(X)} \times 100\% \tag{8}$$

where $NUM_t(X)$ and $NUM_c(X)$ are functions to count the total numbers of the pixels and the corrupted pixels in X respectively. Let I be the output of X processed by MSMF, w=3×3, T=15, w_{max}=5, iteration=4, and calculate $\hat{CR}$ by:

$$\hat{CR} = \frac{NUM_d(I,X)}{NUM_t(I)} \times 100\% \tag{9}$$

where $NUM_d(I,X)$ is a function to count the total number of pixels of different gray scale in I and X, then we can use $\hat{CR}$ to control the parameter of W in IPEDWMF:

$$W = \begin{cases} 3\ , & \hat{CR} < 25\ or\ (\hat{CR} \geq 25\ and\ IF \leq 1) \\ 0\ , & others \end{cases} \tag{10}$$

while keeping the other parameters unchanged.

4 Analysis and Comparison of Experiment Results

A set of images are obtained to test the algorithm by corrupting an original image of 320×320 with the proposed salt-pepper noise model of probability 0.05, 0.10, … , 0.50. Each image is processed by SMF (Standard Median Filter), CWMF, NLAF, IRCMF, MSMF, IPEDWMF and MSMF+IPEDWMF individually. By playing with different parameter settings, we get the best images output corresponding to each filter. They form the results set for comparison and analysis.

Before going to the experiment results, we first briefly introduce two criteria used in this paper to compare the performance of the filter. The first one is mean square error (*MSE*), which reflects the closeness between the filtered image $\hat{X}$ and the original image; the second one is the difference ratio (DR), which tells us how similar $\hat{X}$ and the original image are from another aspect. Their definitions are:

$$MSE = \sqrt{\frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (\hat{X}[i, j] - I[i, j])^2} \tag{11}$$

$$DR = \frac{NUM_d(\hat{X}, I)}{M \times N} \times 100\% \tag{12}$$

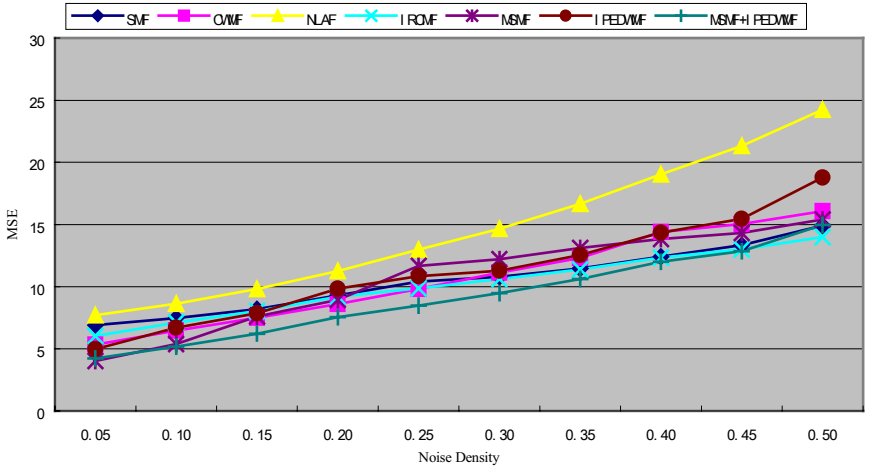


Fig. 2. MSE Graph

Figure (2) plots the *MSE* of different filters versus noise density. We can see from it that the performance of NLAF is the poorest. This is because the filter is designed assuming a fixed-valued salt-pepper model. In the fixed-valued model the intensity of corrupted pixels is either 255 or 0 (or within a small range around 255 or 0). This is not the actual situation. The performances of the rest are very close for different noise

density except MSMF+IPEDWMF. Though the quality of the images processed by IRCMF is better than that of the others when noise density is very high, the filter blurs the edges and details of the image and introduces many small black-white speckles. The performance of combined filter MSMF+IPEDWMF is the best among all algorithms for each noise density.

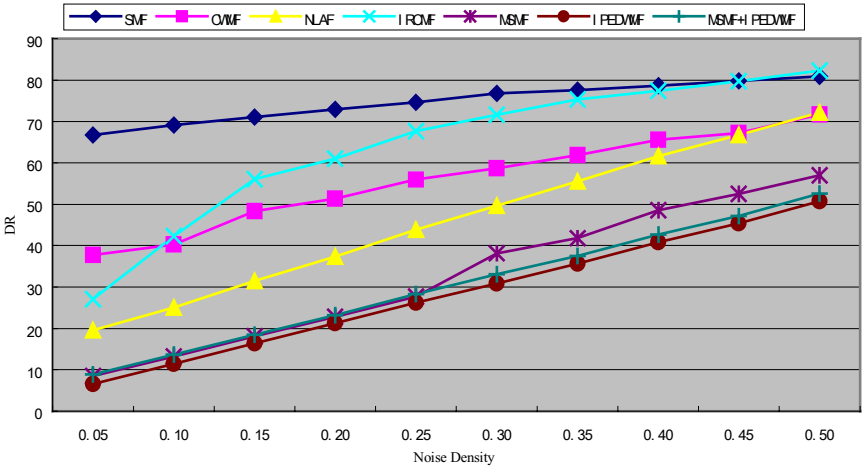


Fig. 3. Difference Ratio Graph

Figure (3) shows *DR* versus noise density for the processed images by different algorithms. We see that the performances of MSMF, IPEDWMF and MSMF + IPEDWMF in the *DR* sense are better and the differences among them are very small. This means that these three filters can work more efficiently and preserve the original gray scale of the pixels more effectively than others. Thus these filters have a better performance in preserving edges without sacrificing the noise elimination efficiency.

Figure (4) shows the original photograph, the photo after being corrupted by salt-pepper noise of intensity 0.2, and the final results after being processed by different filters. All the images are of size 320×320. We can see that the noise suppression performance of SMF is fine except for some exiguous spots, as shown in (c). However, SMF erodes the edges and blurs the details of the images greatly. (d) is the result by CWMF is clearer than (c), but there is still a slight blurring increases slightly, too. The filtering effect of NLAF is the poorest, as shown in (e). This means that NLAF is not suitable for the noise model. (f) shows that the photo processed by IRCMF still has many black-white speckles and its ability of preserving edges and details is not strong. The result of MSMF shown in (g) is very clear, although many black-white blocks appear in it. The photo obtained by IPEDWMF in (h) is not clearer than (g), but the black-white blocks are not apparent. The output image of MSMF+IPEDWMF (k) shows that the effect of the filter is the best and it has the advantages of both MSMF and IPEDWMF.

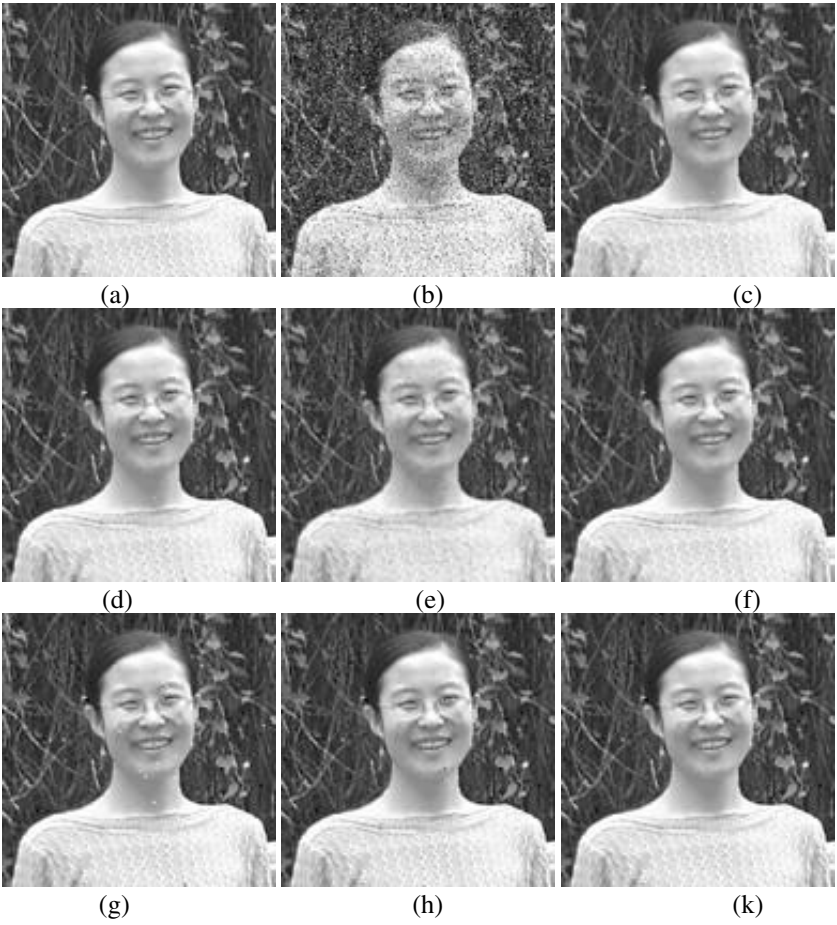


Fig. 4. The size of images is 320×320 pixels. (a) is the uncorrupted original image. (b) is the corrupted image with the probability of 0.2. The rest of images are obtained by using filters respectively as following: (c) SMF with 3×3 filter window and 1 iteration; (d) CWMF with 3×3 filter window, 2 iterations and W=3; (e) NLAF with T_1=36, T_2=1 and 4 iterations; (f) IRCMF with 3×3 filter window, 2 iterations, W=1 and T=40; (g) MSMF with 3×3 filter window, 4 iterations, W=5 and NT=15; (h) IPEDWMF with 1 iteration; (k) MSMF+IPEDWMF, for MSMF: 3×3 filter window, 4 iterations, W=5 and NT=15; for IPEDWMF: 1 iteration.

5 Conclusion

By differing the edge and isolated noise pixels in an images, a new algorithm IPEDWMF is introduced and performs well for removing the salt-pepper noise. Combined with MSMF, the IPEDWMF has an even better performance in salt-pepper noise suppression while preserving edges and details. Another advantage of the combination usage is that the filter parameters are fixed-valued or self-adaptive, which

facilitates the automation of image processing by computers. However, the algorithm requires more computation efforts, and the performance of suppressing big speckles generated in filtering processing needs to be improved.

Acknowledgments. The authors would like to thank Ms. Li Pingping for the usage of her photo.

References

1. M. Gabbouj, E. Coyle, N.C. Gallagher Jr., An overview of median and stack filtering, Circuit System Signal Processing, 11(1), 1992, 7–45.
2. Huang Xutao, Two-dimensional Digital Signal Processing II — Transformation and Median Filter (Beijing: Science Press, 1985).
3. Tao Chen, Hong Ren Wu, Space Variant Median Filters for the Restoration of Impulse Noise Corrupted Images, IEEE Transactions on, Circuits and Systems II, 8(48), 2001, 784–789.
4. J. W. Turky, Exploratory Data Analysis (Reading, MA: Addison-Wesley, 1971).
5. W. K. Pratt, Median filtering, Semianual Report, Image Proc. Institute, Univ. of Southern California, 1975, 116–123.
6. B. R. Frieden, A new restoring algorithm for the preferential enhancement of edge gradients, J. Opt. Soc. Amer. 66 (3), 1976, 280–283.
7. S.-J. Ko, Y.-H. Lee, Center weighted median filters and their applications to image enhancement, Circuits and Systems, IEEE Transactions on, 9(38), 1991, 984–993.
8. A. Nieminen, P. Heinonen, Y. Neuvo, A new class of detail-preserving filters for image processing, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(9), 1987, 74–90.
9. Li Shutao, Wang Yaonan, Non-Linear Adaptive Removal of Salt and Pepper Noise from Images, Journal of Image and Graphics. 12(5(a)), 2000.
10. Kh. Manglem Singh, Prabin K. Bora, Improved Rank Conditioned Median Filter for Removal of Impulse Noise from Images, TENCON '02. Proceedings, Volume: 1, 2002, 557–560.
11. Lin H M, Willson A N, Median filters with adaptive length, IEEE CAS1, 35(6), 1988, 675–690.
12. Florencio D A F, Schafer R W, Decision–based median filter using local signal statistics, Proc. SPIE Vol. 2308, 1994, 268–275.
13. Hardie R E, Barner K E, Rank conditioned rank selection filters for signal restoration, IEEE IP, 3(2), 1994, 192–206.
14. Abreu E, Lightone M, Mitra S Ketal, A new efficient approach for the removal of impulse noise from highly corrupted images, IEEE IP, 5(6), 1996, 1012–1025.
15. How-Lung Eng, Kai-Kuang Ma, Noise adaptive soft-switching median filter for image denoising, 2000 IEEE International Conference on , Volume: 6, 2000, 2175–2178.

Image De-noising via Overlapping Wavelet Atoms

V. Bruni and D. Vitulano

Istituto per le Applicazioni del Calcolo "M. Picone"
C. N. R.
Viale del Policlinico, 137 00161 Rome, Italy
{bruni, vitulano}@iac.rm.cnr.it

Abstract. This paper focuses on a novel approach for image denoising: *WISDOW* (Wavelet based Image and Signal De-noising via Overlapping Waves). It is based on approximating any singularity by means of a basic one in a wavelet domain. This approach allows us to reach some interesting mathematical properties along with good performances in terms of both subjective and objective quality. In fact, achieved results are comparable to the best wavelet approaches requiring a low computational effort and resulting completely automatic.

1 Introduction

Image de-noising is one of the most investigated topics of Computer Vision. Its difficulty stems from the fact that Fourier based approaches typically yield suppression of image high frequencies with an unavoidable blurring of its edges. From this point of view, wavelet transform is attractive for its good time-frequency localization [1]. Wavelet denoising approaches can be broadly split in two wide classes:

- attenuation based ones: wavelet coefficients are shrunk accounting for signal to noise ratio (see for instance chap. 10 of [1] and [2,3,4]);
- selection based ones: only coefficients trapping signal information, i.e. those over a noise dependent threshold, are retained (see for instance [5,6,7,8]).

All aforementioned approaches strongly rely on the hypothesis that wavelet bases are able to compact information of the original signal in a few coefficients. Nonetheless, only sub-optimal results can be achieved for real world signals leading researchers to take, again, different ways. *Matching pursuit* guarantees satisfying results but it is computationally very expensive ([1], chap. 9). New bases have also been proposed in literature (see for instance [9,10,11]) but they reach optimal approximation only for a restricted class of functions. Finally, the combination of different wavelet bases has been investigated, trying to optimize the intrinsic trade-off between quality and complexity [2,12,13].

The aim of the present work is to improve and exploit the signal representation in an arbitrary selected wavelet basis. The underlying idea consists of

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 179–186, 2004.
© Springer-Verlag Berlin Heidelberg 2004

approximating any signal singularity with the simplest isolated one, depicted in Fig. 1 (leftmost, solid line). In fact, in the wavelet domain more complicated singularities slightly differ from the basic waveform, whose amplitude is proportional to the change of the slope of the signal in correspondence to the singularity location. Therefore, any discontinuity waveform is approximated using the one depicted in Fig. 1 (rightmost, solid line), and its amplitude is estimated according to the real data. A classical least square method can be used for this estimation. Moreover, exploiting the linearity of the wavelet operator, *overlapping effects* principle helps us to simply manage interfering (i.e. non isolated) singularities. The proposed approach is attractive since it allows us to completely recover all wavelet coefficients representing a single singularity. More precisely, it is able to reconstruct coefficients both *over* and *under* noise threshold thanks to the adopted singularities representation. Some interesting mathematical properties have been found [14]; in particular, it has been proved that this approach theoretically outperforms hard thresholding under some constraints. With regard to experimental results, *WISDOW* improves results of existing techniques which use a pre-processing thresholding step. Moreover, it is able to reach results comparable to the best available adaptive models, requiring a lower computational time and avoiding a manual tuning of parameters.

2 De-noising in a Fixed Wavelet Basis

Signal de-noising can be mathematically written as follows:

$$z_i = x_i + \eta_i, \quad 1 \leq i \leq N \tag{1}$$

where $\{z_i\}_{0\leq i\leq N}$ is the noisy signal, $\{x_i\}_{0\leq i\leq N}$ the original one while $\{\eta_i\}_{0\leq i\leq N}$ is a zero-mean gaussian white noise.

As previously described, we investigate how to exploit at the best, signal representation in an *a priori* selected wavelet basis. The core of the model consists of studying the behaviour of the wavelet transform of a basic singularity, as depicted in Fig. 1 (leftmost, solid line), which represents the simplest singularity we can deal with. Using this basic waveform, we will describe any other discontinuity of the noisy signal: no matter what is its Lipschitz's order [1].

Let us first investigate the case concerning isolated singularities. Mathematically speaking, the basic signal of Fig. 1 (leftmost solid line) is defined as:

$$f(t) = \begin{cases} \alpha & t < n \\ \alpha_1 t + \beta_1 & t \geq n \end{cases} \tag{2}$$

with $\beta_1 = \alpha - \alpha_1 n$.

Its wavelet transform at time u and scale s is:

$$\begin{aligned} Wf(u,s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} f(t)\psi(\frac{t-u}{s})dt = \\ = \alpha_1 s\sqrt{s}\,(z\Psi(z) + \Phi(z)), \end{aligned} \tag{3}$$

where $z = \frac{n-u}{s}$, $\Psi(z) = \int_{-\infty}^{z} \psi(t)dt$, $\Phi(z) = \int_{z}^{+\infty} t\psi(t)dt$, while ψ is the a priori selected mother wavelet function.

It can be easily seen that (3) can be rewritten as:

$$Wf(z) = \alpha_1 \, C \, P(z), \tag{4}$$

where $C\,P(z) = s\sqrt{s}\,(z\Psi(z) + \Phi(z))$ and it is the same for all infinite ramps, except for α_1 that regulates slope.

It is clear that an approximation error occurs when (2) is used for representing any kind of singularity, as depicted in Fig. 1. On the other hand, it is intuitive that the higher the noise distortion, the more negligible the approximation error [14]. It can be proved that using the above approximation:

$$\exists\, I \in R :\ \forall\, \alpha_1 \ \in I \quad E_{LS} \leq E_D \tag{5}$$

where E_{LS} is the error relative to the proposed approach while E_D is the hard thresholding error, under the constraint:

$$\sum_{u \in O_T} \Upsilon^2(u) \geq \sum_{u \in O_T} P^2(u) - \frac{(\sum_{u \in O_T} P(u)K(u))^2}{\sum_{u \in O_T} K^2(u)} \tag{6}$$

where Υ, P and K are the wavelet transform respectively of noise, generic singularity and bisecting line. Sums are computed in O_T, i.e. for coefficients over threshold.

In particular, proposition (5) is verified if parameter α_1 is estimated from noisy data in the least square sense. A rigorous proof of aforementioned results can be found in [14].

WISDOW implementation is quite simple: it can be defined *estimate and replace*. In practice, using the proposed approximation, singularity waveform behaviour is known except for its amplitude, i.e. α_1 of (4). Hence, it is estimated for each maximum of the noisy wavelet details, since each singularity is detected by its most important maximum in the wavelet domain. Nonetheless, the noisy maximum location may have a displacement from the original one. From a theoretical point of view, displacement error does not drastically change the above mentioned results that can be easily generalized [14]. Mathematically speaking, (4) still holds under a bit more restrictive constraint. In practice, this error can be ignored without annoying effects in performances, as it will be shown in the experimental results.

The *overlapping effects principle* can be applied whenever two or more singularities interfere, bearing in mind that wavelet transform is a linear operator. So, the above algorithm can be still performed at each noisy maximum, but, in this case, the influence of the nearest singularities has to be accounted while estimating α_1 by least squares. Unfortunately, we do not know the contribution of each singularity in advance. Nevertheless, we can cope with this problem performing the algorithm from the greatest to the lowest maximum.

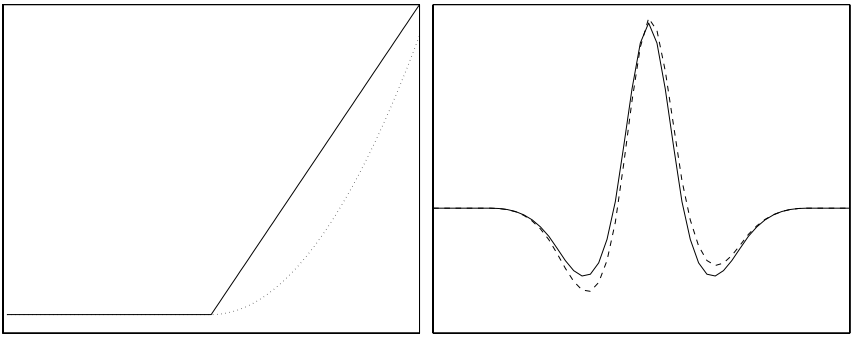


Fig. 1. Left): basic discontinuity (solid line) and a 2^{nd} degree piecewise polynomial one (dashed line). Right): The corresponding wavelet coefficients using *rbio2.4* basis (Matlab notation).

2.1 2-D Extension

The proposed approach can be simply extended in two dimensions. In fact a multi-resolution analysis of a function $f(x,y) \in L^2(R^2)$ corresponds to expanding it in a separable basis. This latter is generated by the functions

$$\phi_1(x,y) = \phi(x)\phi(y), \quad \psi_1(x,y) = \phi(x)\psi(y)$$

$$\psi_2(x,y) = \psi(x)\phi(y), \quad \psi_3(x,y) = \psi(x)\psi(y)$$

which respectively yield approximation, horizontal, vertical and diagonal components (see [1] chap. 7).

Then, at each scale level, each column of horizontal band and each row of vertical one can be processed by the 1-D algorithm. It is equivalent to recover the signal $f * \phi$, ignoring the averaging effect on the noise of the scaling function ϕ. As regards diagonal details, rows or columns can be processed, according to the order used for the decomposition.

As a variant of the proposed framework, a better recovering of the vertical and horizontal details can be achieved by exploiting, at a given scale level s, the corresponding approximation band for estimating α_1. In this case, we find the equivalent of (3) computing the convolution of the function in eq. (2) with the scaling function ϕ (associated to the mother wavelet ψ):

$$\begin{aligned} Sf(u,s) = \alpha\sqrt{s} \int_{-\infty}^{+\infty} \phi(t)dt + \\ +\alpha_1 s\sqrt{s}(I_1(z) - zI_2(z)) \end{aligned} \tag{7}$$

where $I_1(z) = \int_z^{+\infty} t\phi(t)dt$ and $I_2(z) = \int_z^{+\infty} \phi(t)dt$.

It can be noticed, again, that in the adopted approximation, once the wavelet function $\psi(t)$ is fixed, $Sf(u,s)$ is completely known except for the parameters α and α_1.

Fig. 2. Top) $512 \times 512 \times 8b$ original Lena (*left*) and Barbara (*right*) images; **Middle)** noisy Lena ($PSNR = 22.06db$) and Barbara ($PSNR = 22.15db$) images ; **Bottom)** denoised Lena ($PSNR = 31.50db$) and Barbara ($PSNR = 28.50db$) images using *WISDOW*.

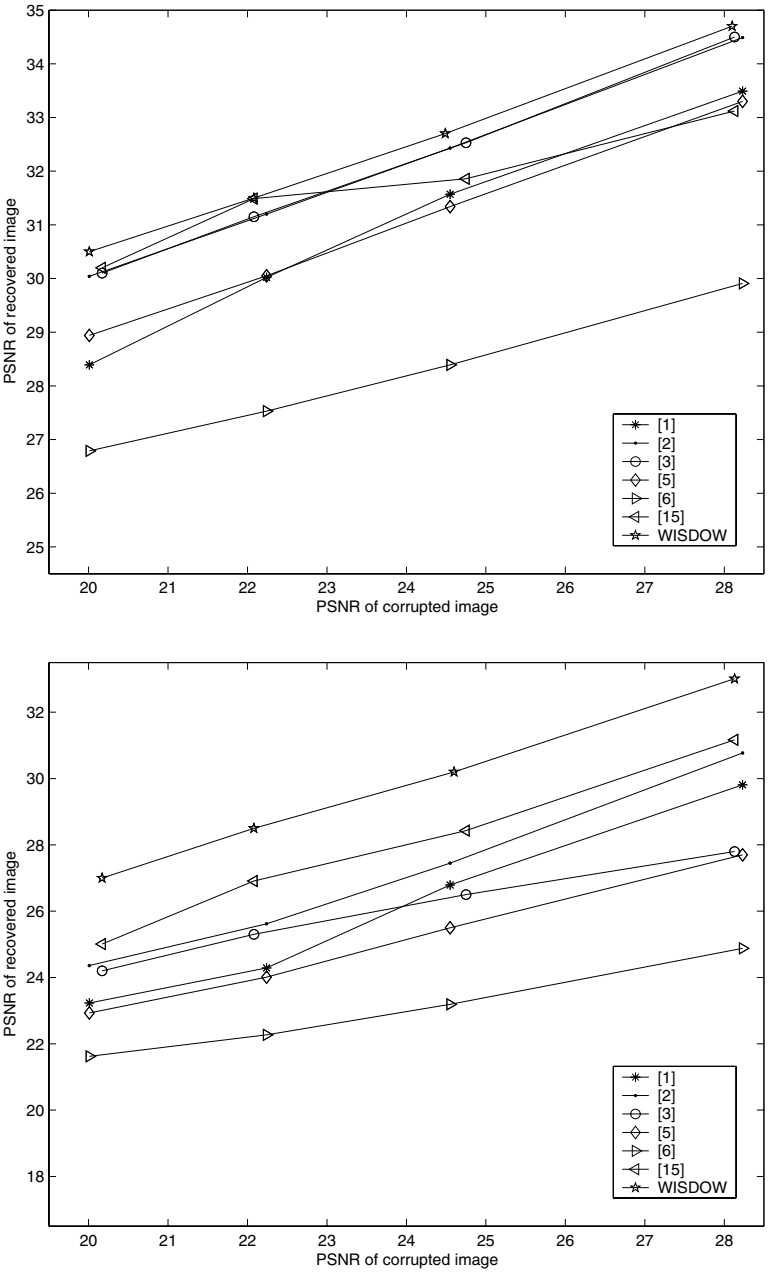


Fig. 3. PSNR values versus noise variance for [1] (Wiener filter), [2,3,5,6,15] and *WISDOW*: **Top)** *Lena* image; **Bottom)** *Barbara* image.

Therefore, for a selected maximum of the (horizontal or vertical) detail band, the parameter α_1 can be also estimated from the correspondent data in the approximation band (i.e. the data in the cone of influence of the analysed singularity). Then the final value of α_1 will be the weighted average of the two achieved estimations. In this case the estimation of α_1 is:

$$\overline{\alpha_1} = \frac{\omega_a \alpha_1^{(a)} + \omega_d \alpha_1^{(d)}}{\omega_a + \omega_d} \tag{8}$$

where indices a and d respectively indicate approximation and detail band, while ω_a and ω_d weight the two estimations such that $\omega_a + \omega_d = 1$. Accounting for the fact that noise becomes flatter and flatter along scale levels, we fix $\omega_d = 1/3$ and $\omega_a = 2/3$, since approximation band results cleaner than details one.

3 Experimental Results and Discussion

In this section we will present some experimental results achieved on some test images along with some comparisons with selected existing wavelet based approaches.

For the sake of clarity, it is necessary to shortly describe some implementation details. A biorthogonal wavelet (rbio2.4 in Matlab notation) associated to an over-complete multiresolution decomposition has been adopted. This wavelet admits an explicit and tractable mathematical function for ψ. Nevertheless, the proposed approach is still valid for any wavelet basis considering the corresponding impulse response. A three scale wavelet decomposition has been used in all tests.

In order to discard maxima originated by noise with high probability, we have hard-thresholded wavelet coefficients before performing our algorithm [5]. Universal threshold has been employed without corrections, even though some variants may sometimes greatly improve results. This allows us to avoid thresholds to be tuned. As regards the maxima location error, we ignore it, considering the computed one as correct.

For the lack of space we will just present results on $512 \times 512 \times 8$ bits Lena and Barbara images. It can be observed in Fig. 2 that *WISDOW* performs well in terms of both subjective and objective ($PSNR$, Peak Signal to Noise Ratio) quality, requiring a moderate computing time. In fact, $O(N)$ operations are required for hard thresholding, where N is the number of wavelet coefficients. Local extrema determination cost is $O(M)$, where M is the number of coefficients over threshold. Finally, least squares need $O(|J|)L$, where L is the number of local extrema, while J are coefficients associated to them.

As mentioned above, we have compared *WISDOW* with some of the most recent and effective wavelet based approaches for image denoising — see Fig. 3. In particular, we have selected those approaches having hard-thresholding as a pre-processing step, i.e. the approach proposed by Kazubek [3], Choi et al. [2], Ishwar et al. [15] along with classical hard [5] and soft [6] thresholding and Wiener filtering [1]. It can be noticed that our model outperforms the selected ones even

though in its "rough" version. Nonetheless, it has been observed that it may reach results of more sophisticated adaptive approaches like [4] and [7], suitably tuning universal threshold. Our work is only at its first stage of study and then some aspects must be further improved and investigated. In particular, better performances may be reached by investigating how to better manage interference between very close singularities and how to tie scale levels information. These two aspects constitute our future research.

References

1. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1998)
2. Choi, H., Baraniuk, R.: Analysis of wavelet - domain wiener filters. Proceedings of SPIE, San Diego (1997)
3. Kazubek, M.: Wavelet domain image denoising by thresholding and wiener filtering. IEEE Signal Processing Letters **10** (2003) 324–326
4. Portilla, J., Strela, V., Wainwright, M., Simoncelli, E.: Image denoising using scale mixtures of gaussians in the wavelet domain. IEEE Transactions on Image Processing **12** (2003) 1338–1351
5. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation via wavelet shrinkage. Biometrika **81** (1994) 425–455
6. Donoho, D.L.: De-noising by soft thresholding. IEEE Transactions on Information Theory **41** (1995) 613–627
7. Chang, S., Yu, B., Vetterli, M.: Spatially adaptive thresholding with context modeling for image denoising. IEEE Transactions on Image Processing **9** (2000) 1522–1531
8. Mihcak, M., Kozintsev, I., Ramchandran, K., Moulin, P.: Spatially adaptive thresholding with context modeling for image denoising. IEEE Signal Processing Letters **6** (1999) 300–303
9. Do, M.N., Vetterli, M.: Contourlets: a new directional multiresolution image representation. Proceedings of 36^{th} Asilomar Conference on Signals Systems and Computers **1** (2002) 497–501
10. Pennec, E.L., Mallat, S.: Non linear image approximation with bandelets. Tech. Rep. CMAP/ Ecole Polytechnique (2003)
11. Starck, J.L., Candes, E.J., Donoho, D.L.: The curvelet transform for image denoising. IEEE Transactions on Image Processing **11** (2002) 670–684
12. Bruni, V., Vitulano, D.: A wiener filter improvement combining wavelet domains. Proceedings of 12^{th} International Conference on Image Analysis and Processing (2003) 518–523
13. Ishwar, P., Ratakonda, K., Moulin, P., Ahuja, N.: Image denoising using multiple compaction domains. Proceedings of ICASSP 1998 **3** (1998) 1889–1892
14. Bruni, V., Vitulano, D.: Image and signal denoising in a fixed wavelet basis. IAC Report, CNR (2004)
15. Ishwar, P., Moulin, P.: Multiple domain image modeling and restoration. Proceedings of IEEE International Conference on Image Processing (1999) 362–366

Gradient Pile Up Algorithm for Edge Enhancement and Detection

Leticia Guimarães[1], André Soares[1,2], Viviane Cordeiro[1], and Altamiro Susin[1,2]

[1] Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Sul - UFRGS, Av. Osvaldo Aranha, 103, CEP. 90 035-190,Porto Alegre, RS, Brazil
leticia@eletro.ufrgs.br
http://www.lapsi.eletro.ufrgs.br
[2] Instituto de Informática, Universidade Federal do Rio Grande do Sul – UFRGS, Av. Bento Gonçalves, 9500, CEP. 91501-970, Porto Alegre, RS, Brazil

Abstract. Edge detection plays a fundamental role on image processing. The detected edges describe an object contour that greatly improves the pattern recognition process. Many edge detectors have been proposed. Most of them apply smooth filters to minimize the noise and the image derivative or gradient to enhance the edges. However, smooth filters produce ramp edges with the same gradient magnitude as those produced by noise. This work presents an algorithm that enhances the gradient correspondent to ramp edges without amplifying the noisy ones. Moreover, an efficient method for edge detection without set a threshold value is proposed. The experimental results show that the proposed algorithm enhances the gradient of ramp edges, improving the gradient magnitude without shifting the edge location. Further, we are testing the implementation of the proposed algorithm in hardware for real time vision applications.

1 Introduction

The estimation of the boundary location of an object is an important subject on image processing. Since edge detection have been studied, several techniques for edge enhancement were proposed. Most of the edge enhancement methods apply the derivative to enhance the edge of the objects on an image. Usually, the derivative of an image is produced by the convolution with a gradient operators [1], as Sobel, Roberts, Prewitt, for example.

However, there are three problems on the use of the derivative as edge detector: a) false edges produced by noise are enhanced; b) ramp edges produce lower derivative magnitude; c) the location of the ramp edges are not accurate. Several edge detection methods as proposed by Marr-Hildreth [2], Canny [3], Perona-Malik [4], have been proposed in order to provide tools to overcome the edge detection problems by applying the derivative to edge detection. Most of these methods apply a smooth gaussian operator in order to suppress the noisy edges, then, estimate the location of the edges by detection of the zero-crossing on the second derivative of an image. Edge detectors proposed by Petrou-Kitter [5] and Zang et. al. [6] are based on ramp edges model. Petrou-Kitter[5] propose a ramp edge model to an optimal edge

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 187–194, 2004.
© Springer-Verlag Berlin Heidelberg 2004

detector. Zang et. al. [6] applies correlation matching to enhance the gradient at ramp edge location. In [7] D. Wang proposed a multiscale gradient algorithm that enhances the morphological gradient of ramp edges, however, the edge of thin objects can be lost.

Our approach aims to enhance the gradient value of blurred edges without enhances noisy gradient. In order to achieve our objective the enhancement occurs over the two component vector of the gradient, the horizontal and vertical gradient map separately. First, a smooth filter is applied to the original image. Second, the horizontal and vertical components of the gradient of an image are calculated. Third, each component of the gradient generates a connected component map, the gradient enhancement map (GEM), that guide the gradient enhancement process. Finally, the horizontal and vertical gradient map are enhanced by a process similar to a pile up, guided by their respective gradient enhancement map (GEM). The experiments presented in this paper demonstrate the improvement of the performance on edge detection using the gradient map enhanced by the proposed method.

2 Gradient Map Calculation

There are two components of the gradient, the horizontal and vertical component. The two-dimensional function gh(x,y) represents the horizontal gradient map and gv(x,y) represents the vertical gradient map of an image f(x,y), respectively. The horizontal gradient map gh(x,y) is calculated by the convolution of the image f(x,y) with a horizontal gradient operator. The vertical gradient map gv(x,y) is calculated by the convolution of the image f(x,y) with a vertical gradient operator. The gradient magnitude map g(x,y) is the magnitude of the vector composed by two components. An example of horizontal and vertical operators are the Prewitt operators.

3 Gradient Enhancement Map Generation

The gradient enhancement map GEM guides the pile up process. The pile up process occurs on two gradient maps separately following the same rules. Therefore, two gradient enhancement maps are generated, $GEM_H(x,y)$ that guides the enhancement of the horizontal gradient gh(x,y) and $GEM_V(x,y)$ that guides the enhancement of the vertical gradient gv(x,y).

The GEM is based on connected components map proposed by Bieniek and Moga [8]. The GEM is composed by seven elements: a) zero (Z); b)right arrow (➞); c) left arrow (←); d) up arrow (↑); e) down arrow ↓ ; f) plateau and (P); g) maximum (M). In [8] the connected components map is used to locate the minima, in our approach we intend to locate the maxima. Therefore, the arrows on GEM point to the maximum neighbor in magnitude. Since the direction of the arrows are related to the gradient direction, GEM_H is composed only by up and down arrow and GEM_V is composed by only left and right arrow.

The GEM is generated in two steps: first, the initial map is generated, then, the final map is produced eliminating false maximum plateaus and detecting hidden zeros of the initial map. The false maximum plateaus are labeled with arrows and the true maximum plateaus with M.

The plateaus *P* are the candidates to maximum *M* and are present only on the initial map. Considering that the origin of an image coordinates is located at the upper left of an image, the labeling for the initial GEM is as follows:

$$gem_V(x,y)=\begin{cases} Z & if \quad gv(x,y)=0 \quad \vee \quad \left(|gv(x,y)|<|gv(x-1,y)| \ \wedge \ |gv(x,y)|<|gv(x+1,y)|\right), \\ \leftarrow & if \quad |gv(x,y)|<|gv(x-1,y)| \ \wedge |gv(x,y)|\geq|gv(x+1,y)|, \\ \rightarrow & if \quad |gv(x,y)|<|gv(x+1,y)| \ \wedge |gv(x,y)|\geq|gv(x-1,y)|, \\ P & otherwise. \end{cases} \tag{1}$$

$$gem_h(x,y)=\begin{cases} Z & if \quad gh(x,y)=0 \quad \vee \quad \left(|gh(x,y)|<|gh(x,y-1)| \ \wedge \ |gh(x,y)|<|gh(x,y+1)|\right), \\ \uparrow & if \quad |gh(x,y)|<|gh(x,y-1)| \ \wedge \ |gh(x,y)|\geq|gh(x,y+1)|, \\ \downarrow & if \quad |gh(x,y)|<|gh(x,y+1)| \ \wedge \ |gh(x,y)|\geq|gh(x,y-1)|, \\ P & otherwise. \end{cases} \tag{2}$$

where $gem_h(x,y)$ and $gem_v(x,y)$ represents the initial horizontal enhancement map and initial vertical enhancement map, respectively.

The final GEM is obtained by the detection of noisy zeros and false maximum plateaus detection and substitution is performed as follows:

$$GEM_H(x,y)=\begin{cases} \downarrow & if \quad gem_h(x,y)=P \ \wedge \ gem_h(x,y-1)=\downarrow \ \wedge \ gem_h(x,y+1)=P, \\ \uparrow & if \quad gem_h(x,y)=P \ \wedge \ gem_h(x,y+1)=\uparrow \ \wedge \ gem_h(x,y-1)=P, \\ M & if \quad gem_h(x,y)=P \ \wedge \ gem_h(x,y-1)=\downarrow \ \wedge \ gem_h(x,y+1)=\uparrow, \\ Z & if \quad \left(gem_h(x,y)=\uparrow \ \wedge gem_h(x,y-1)=\downarrow\right) \ \vee \ \left(gem_h(x,y)=\downarrow \wedge \ gem_h(x,y+1)=\uparrow\right). \end{cases} \tag{3}$$

$$GEM_V(x,y)=\begin{cases} \rightarrow & if \quad gem_v(x,y)=P \ \wedge \ gem_v(x-1,y)=\rightarrow \ \wedge \ gem_v(x+1,y)=P, \\ \leftarrow & if \quad gem_v(x,y)=P \ \wedge \ gem_v(x+1,y)=\leftarrow \ \wedge \ gem_v(x-1,y)=P, \\ M & if \quad gem_v(x,y)=P \ \wedge \ gem_v(x-1,y)=\rightarrow \ \wedge \ gem_v(x+1,y)=\leftarrow, \\ Z & if \quad \left(gem_v(x,y)=\rightarrow \wedge gem_v(x,y-1)=\rightarrow\right) \ \vee \ \left(gem_v(x,y)=\rightarrow \wedge \ gem_v(x,y+1)=\leftarrow\right). \end{cases} \tag{4}$$

where $GEM_H(x,y)$ and $GEM_V(x,y)$ represents the final horizontal enhancement map and final vertical enhancement map, respectively.

Fig. 1 shows an one-dimensional ramp edge, the corresponding gradient, the initial and final GEM. Notice the gradient value is only one, even though, the difference between **a** and **b** level is five. Furthermore, the gradient location is not evident, in this case.

4 Gradient Enhancement Process (Pile Up Process)

The gradient piled-up process is guided by the gradient enhancement maps GEM of the gradient maps, the horizontal and vertical map. The GEM indicate the direction, the start and the end points of the gradient pile up process. The arrows indicate the pile up direction. The start point of the pile up process is an arrow that the neighbor is zero (Z) and points to another arrow or maximum. The gradient value of a point at an

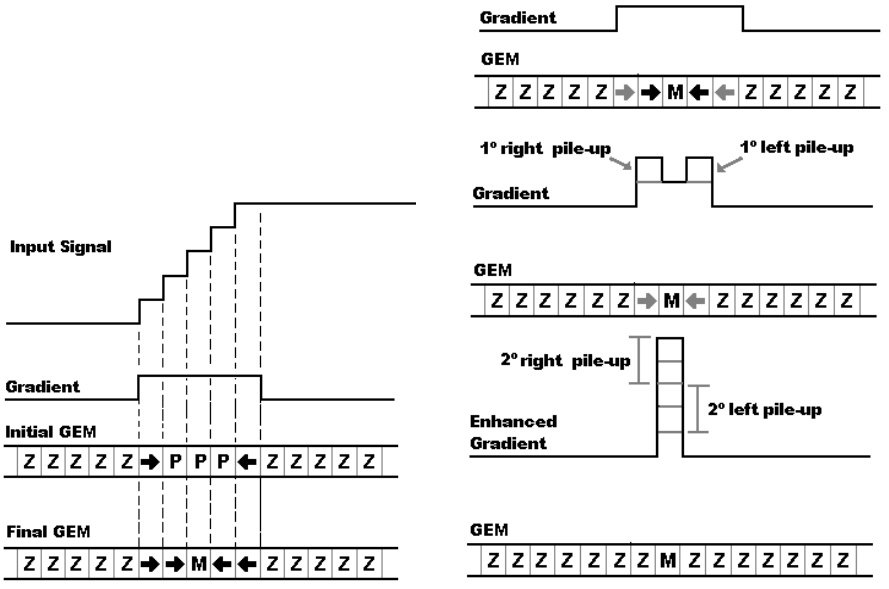


Fig. 1. Example of gradient enhancement process for one dimensional signal; (a) gradient enhancement map; (b) Gradient enhancement process (pile up algorithm) for the one dimensional ramp edge of the Fig. 1(a).

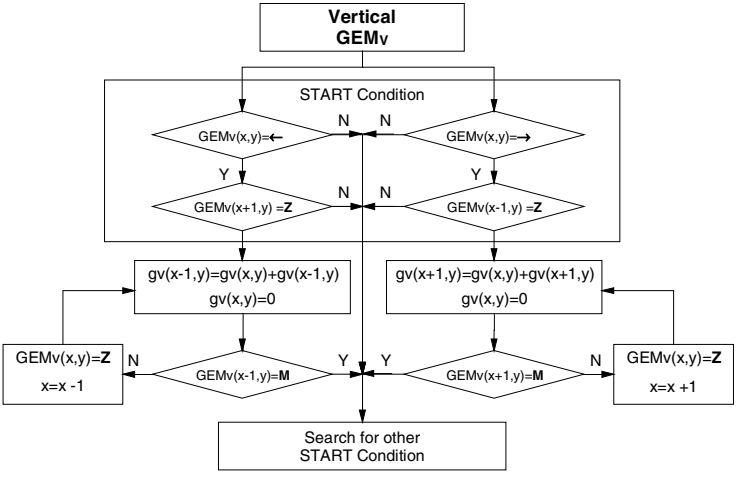


Fig. 2. Schema of the gradient enhancement process (pile up algorithm) for the vertical gradient enhancement .

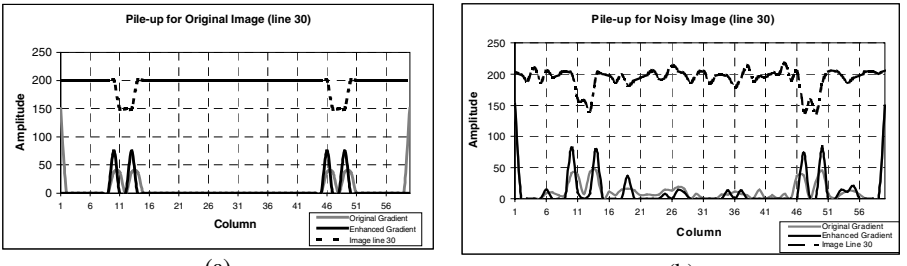
(a) (b) (c)

Fig. 3. (a) Synthetic test image (60X60); (b) GEM_H of the Prewitt gradient of synthetic image in (a); (c) GEM_V of the Prewitt gradient of the synthetic image in (a).

arrow is piled-up over the neighbor gradient value that presents an arrow of same direction or a maximum. The process ends at a maximum. Fig. 1, shows the gradient enhancement process (pile up process) for one-dimensional ramp edge. Observe, the arrows on final GEM in Fig.1(a) are substituted by zeros (Z) as the pile up occurs, in Fig. 1(b). At the end of the pile up process, the GEM presents only maxima (M) and zeros (Z) and the gradient is enhanced from one, the original gradient value, to five, the difference between **a** and **b** level in Fig. 1(a). Moreover, the edge location is evident at the maximum on GEM. Fig. 2 shows the flowchart for the left and right pile up process to the vertical gradient enhancement.

The GEM for a synthetic image is shown in Fig. 3. Fig. 4 shows the original Prewitt gradient and the enhanced gradient for a line of the gradient map.

The edge points of edge map are those which correspond to the maximum at the GEM after the pile up process. However, only the piled up maximum is considered edge point. We intend to implement in hardware the proposed pile up algorithm, in order to accelerate the edge detection on vision chips applications.



(a) (b)

Fig. 4. The pile up process for the line y=30 on synthetic image in Fig. 4. (a) line y=30 of the original image, Prewitt gradient and enhanced Prewitt gradient; (b) line y=30 of the noise image, Prewitt gradient and enhanced gradient.

5 Performance Evaluation

The performance of the gradient enhancement algorithm was proceeded by two methods: one applying the Pratt´s Figure Of Merit [9] FOM and other by mean absolute error MAE between the ideal edges and the detected edges.

(a) (b) (c) (d)

Fig. 5. Test images; (a) original image; (b) image corrupted by uniform gaussian noise with variance σ=0.001; (c) edge detected at the best FOM using the Prewitt (FOM=83.66% and MAE=30.88%); (d) edge detected at the best FOM using the enhanced Prewitt gradient (FOM= 83.84% and MAE=7.99%).

The Figure of merit FOM is given by:

$$FOM = \frac{1}{\max(Na, Ni)} \sum_{i=1}^{Na} \frac{1}{\left(1 - \alpha d_i^2\right)} \tag{5}$$

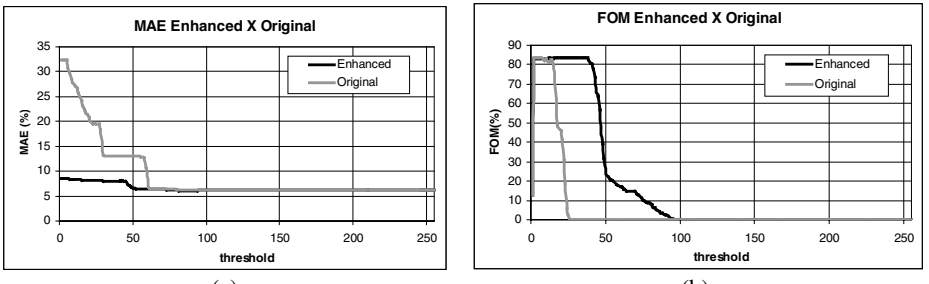
where *Na* is the number of detected edge points, *Ni* is the number of ideal edge points (ground truth), α a scaling factor (we adopt α =0.1), and d_i distance for the *ith* detected edge point.

The edge map was produced by thresholding the gradient magnitude map, however, for the original gradient the edges were thinned by the morphological skeleton operation. The FOM and MAE curve were calculated for each threshold edge map, shown in Fig. 6. The mean absolute error MAE is calculated by eq. (6).

$$MAE = \frac{1}{M \times N} \sum_{y=1}^{M} \sum_{x=1}^{N} \left| mi(y,x) - me(y,x) \right| \tag{6}$$

where *mi* is the ideal edge map, *me* is the resulting edge map, *M* and *N* are height and weight of the *mi* and *me*.

The MAE performance at the best FOM for the original Prewitt gradient and the enhance Prewitt gradient of image in Fig. 5 corrupted by gaussian noise, are shown in Fig. 7. Fig. 8 and 9 shows the results of edge detection applied to the *Lena* and *Boat* images. The ideal edge map of *Lena* in Fig. 8(b) was constructed manually in order to compare with the other edge detection techniques by calculating the FOM and MAE.



(a) (b)

Fig. 6. Performance of the original and enhanced gradient for the image in Fig. 6(b); (a) Figure Of Merit (FOM) x Threshold value ; (b) Mean Absolute Error (MAE) x Threshold value.

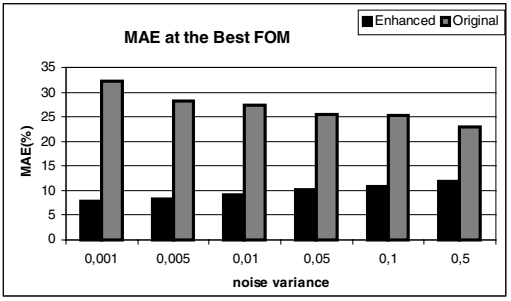


Fig. 7. Comparison of the Performance of Prewitt *vs*. Enhanced Prewitt for the image in Fig. 5.

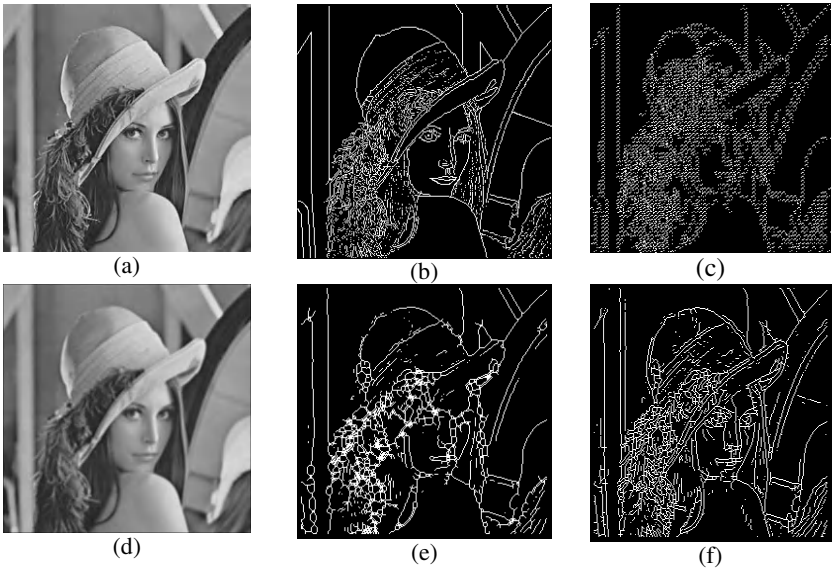


Fig. 8. Edge detection on Lena; (a) original image; (b) ideal edge map of Lena; (c) best FOM edge map from Canny with standard deviation $\sigma = 1$ (FOM=75.28% and MAE=16.37%); (d) smoothed Lena by mean filter 3×3; (e) best FOM edge map from the Prewitt gradient of Lena (FOM=79.18% and MAE=27.12%) and; (f) edge map from enhanced Prewitt (FOM=77.43% and MAE=14.72%).

Fig. 9. Edge map of the *Boat* image after pre-processed by a mean filter 3×3; (a) Original Boat image; (b) *Boat* image after the pre-processing; (c) the edge map generated by the proposed algorithm.

6 Conclusions

The experimental results for the synthetic image show that the best FOM for the enhanced gradient is near from the best FOM of the original gradient, while for the enhanced gradient the MAE decreased 50% of the original gradient, at the best FOM threshold. Although the best FOM for original Prewitt gradient of *Lena* in Fig. 8, (FOM=79.18%) had been greater than the best FOM of the proposed enhanced gradient (FOM=77.43%), the MAE of the edge map obtained applying the proposed pile up method had been reduced from a MAE=27.12%, for the original gradient, to MAE=14.72% at the best FOM. Therefore, the proposed gradient enhancement method is effective to enhance gradient edges and suppress noisy gradient.
Furthermore, the FOM for the edge map produced by the proposed method is higher than the edge map generated applying the Canny edge detector, in Fig. 8(c), without setting any parameter as threshold value.
The proposed pile up process preserves the edges even for thin objects in an image, as shown in Fig. 9(c).

Acknowledgements. This work was supported by Conselho Nacional de Pesquisa CNPq – BRAZIL.

References

1. Gonzalez, R., Woods, R., Digital Image Processing, Addison-Wesley, 1992.
2. Marr, D. and Hildreth, E. C., Theory of Edge Detection, Proc. of the Royal Society of London B207, pp. 187-217, 1980.
3. Canny, J., A Computational Approach to Edge Detection, PAMI, V.8, No.6, pp. 679-698, 1986.
4. Perona, P. and Malik, J., Scale-Space and Edge Detection Using Anisotropic Diffusion, PAMI ,V.12 , No. 7, 1990.
5. Petrou, M. and Kitter, J., Optimal Edge Detectors for Ramp Edges, PAMI, V. 13, No. 5, pp. 483-491, 1991.
6. Wang, Z., Rao, K. R. and Ben-Arie, J., Optimal Ramp Edge Detection Using Expansion Matching, PAMI, V. 18, No. 11, pp. 1092-1097, 1996.
7. Wang, D., A Multiscale Gradient Algorithm for Image Segmentation using Watersheds, Pattern Recognition, V. 30, No. 12, pp. 2043-2052, 1997.
8. Bieniek, A and Moga, A., An efficient watershed algorithm based on connected components. Pattern Recognition, V. 33 , No. 6 , pp. 907-916, 2000.
9. Ballard, D.H., Brown, C.M., Computer Vision, Prentice Hall Inc., 1982.

Co-histogram and Image Degradation Evaluation*

Pengwei Hao[1,2], Chao Zhang[1], and Anrong Dang[3]

[1] Center for Information Science, Peking University, Beijing, 100871, China
{phao, chzhang}@cis.pku.edu.cn
[2] Department of Computer Science, Queen Mary, University of London, E1 4NS, UK
phao@dcs.qmul.ac.uk
[3] Center for Science of Human Settlements, Tsinghua University, Beijing, 100084, China
danrong@tsinghua.edu.cn

Abstract. The tool for image degradation evaluation addressed in this paper is called co-histogram, which is a statistic graph generated by counting the corresponding pixel pairs of two images. The graph is a two-dimensional joint probability distribution of the two images. A co-histogram shows how the pixels are distributed among combinations of two image pixel values. By means of co-histogram, we can have a visual understanding of PSNR, and the symmetry of a co-histogram is also significant for objective evaluation of image degradation. Our experiments with image degradation models of image compression, convolution blurring and geometric distortion perform the importance of the co-histogram.

1 Introduction

Images may be corrupted by some degradation sources, which may arise during image capture or processing, such as blurring, geometric distortion, or compression. A lot of image degradation models have been proposed in literature and a great deal of effort has been made to assess image quality objectively but close to subjective evaluation [1-3]. However, image degradation evaluation is so difficult that only limited success has been achieved [7].

The subjective evaluation methods are human vision oriented, considering human visual system (HVS) characteristics [3]. Mean opinion score (MOS) has been used for a long time, but it is inconvenient, slow and expensive. The subjective metrics also suffer from the physiological, psychological and environmental impact on the viewers.

So far, the widely used objective metrics for compressed image quality assessment are peak signal-to-noise ratio (PSNR) and mean square error (MSE). If the peak value of an image is 255, the mathematical representations are

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \qquad MSE = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\big(f(x,y)-g(x,y)\big)^2 \tag{1}$$

where $f(x,y)$ and $g(x,y)$ are two images of size MxN.

* This work was supported by the Foundation for the Authors of National Excellent Doctoral Dissertation of China, under Grant 200038.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 195–203, 2004.
© Springer-Verlag Berlin Heidelberg 2004

However, they are considered not so satisfying. In many instances, it provides an inaccurate representation of image quality.

An ideal objective metric should be reliable, easily-computable and directly-intercomparable. There are two ways to overcome the shortcomings of PSNR: (i) to find another better objective metric to replace PSNR and to assess image degradation more effectively, and (ii) to find another objective metric as a complement to reinforce PSNR and to jointly assess image degradation more comprehensively. This paper is mainly based on the latter idea.

The tool we applied in this paper is called co-histogram [8]. A co-histogram is a statistic graph generated by counting the corresponding pixel pairs of two images. A co-histogram is also a two-dimensional joint probability distribution of the two images. A co-histogram shows how the pixels are distributed among combinations of two image pixel values, and visually it also gives an intuitive interpretation of PSNR.

For image degradation evaluation, a co-histogram can be easily obtained and then the metrics, the corresponding PSNR and its symmetry, are easily computable. Our experiments with the standard test images and using some degradation models, such as DCT-based JPEG and wavelet-based JPEG2000 compression methods, Gaussian blurring, and StirMark-like geometric distortion, perform the reliability of the co-histogram, its width and symmetry for image degradation evaluation.

2 Co-histogram and Its Properties

A histogram of an image is a statistical distribution of the image pixel values. It has found many applications such as image enhancement, thresholding, retrieval, classification and recognition.

The probability that pixel value p occurs in a digital image $f(x,y)$ of size M-by-N, $H_f(p)$, is counted as

$$H_f(p) = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}\delta\big(f(x,y),p\big) \tag{2}$$

where $\delta(f,p)$ is the Kronecker function: $\delta(f,p) = 1$ (if $f = p$) and 0 (if $f \neq p$).

For all possible p, $H_f(p)$ gives the histogram of image $f(x,y)$.

For two images of the same size MxN, $f(x,y)$ and $g(x,y)$, the joint probability that pixel value pair (p,q) occurs is

$$H(p,q) = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}\delta\big(f(x,y),p\big)\cdot\delta\big(g(x,y),q\big) \tag{3}$$

For all possible pixel value pair (p,q), $H(p,q)$ makes the co-histogram of image pair $f(x,y)$ and $g(x,y)$.

A co-histogram has following properties.

(1) Axial projection

$$\sum_{q}H(p,q) = \frac{1}{MN}\sum_{q}\sum_{y=1}^{M}\sum_{x=1}^{N}\delta\big(f(x,y),p\big)\cdot\delta\big(g(x,y),q\big) = H_f(p) \tag{4}$$

(Histogram of image $f(x,y)$)

$$\sum_{p} H(p,q) = \frac{1}{MN} \sum_{p} \sum_{y=1}^{M} \sum_{x=1}^{N} \delta\big(f(x,y), p\big) \cdot \delta\big(g(x,y), q\big) = H_g(q) \tag{5}$$

(Histogram of image $g(x,y)$)

(2) Mean pixel value

$$I_f = \sum_{p} pH_f(p) = \sum_{p,q} pH(p,q) \tag{6}$$

$$I_g = \sum_{q} qH_g(q) = \sum_{p,q} qH(p,q) \tag{7}$$

If $I_f = I_g$, we have $\sum_{p,q} (p-q)H(p,q) = 0.$

(3) Variance of pixel values

$$Var(f) = \sum_{p} (p - I_f)^2 H_f(p) = \sum_{p,q} p^2 H(p,q) - {I_f}^2 \tag{8}$$

$$Var(g) = \sum_{q} (q - I_g)^2 H_g(q) = \sum_{p,q} q^2 H(p,q) - {I_g}^2 \tag{9}$$

If $I_f = I_g$, we have $Var(f) - Var(g) = \sum_{p,q} (p^2 - q^2) H(p,q).$

(4) Diagonal projection

To project in the direction of $p=q$, or its parametric representation, $p(t) = t + r$ and $q(t) = t$, we have

$$\begin{aligned} \sum_{t} H\big(p(t), q(t)\big) &= \sum_{t} H\big(t+r, t\big) = \sum_{x,y} \sum_{t} \delta\big(f, t+r\big) \cdot \delta\big(g, t\big) \\ &= \sum_{x,y} \delta\big(f(x,y) - g(x,y), r\big) = H_{f-g}(r) \end{aligned} \tag{10}$$

In fact, it is the histogram of the difference image $d(x,y) = f(x,y) - g(x,y)$.

(5) Mean of the difference image

$$I_d = \sum_{r} rH_d(r) = \sum_{r} r \sum_{t} H(t+r,t) = \sum_{p,q} (p-q)H(p,q) = I_f - I_g \tag{11}$$

(6) Variance of the difference image

$$\begin{aligned} Var(d) &= \sum_{r} (r - I_d)^2 H_d(r) = \sum_{r} r^2 \sum_{t} H(t+r,t) - I_d^2 \\ &= \sum_{p,q} (p-q)^2 H(p,q) - (I_f - I_g)^2 = MSE - (I_f - I_g)^2 \end{aligned} \tag{12}$$

where MSE is the mean square error between two images.

3 PSNR and Symmetry of Co-histogram

For image degradation evaluation, we take the original image and the degraded image as above two images $f(x,y)$ and $g(x,y)$. Thus, for most image degradations, we generally have $I_f \approx I_g$. Accordingly, the degradation evaluation metrics PSNR and MSE correspond to the variance of the difference image, and then to the variance of the diagonal projection of the co-histogram, which is visually a "width" measure of the

co-histogram. Therefore, PSNR is a measure of the co-histogram width along the diagonal.

If $Var(d) = 0$, or the whole co-histogram is strictly on the diagonal, the two images must be identical, and no degradation occurs. If $Var(d) \neq 0$, or the off-diagonal distribution is not entirely zero, the image must have been degraded. Thus, the PSNR can be calculated directly from the co-histogram:

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \approx 48.1308 - 10\log_{10}\left(Var(d)\right) \tag{13}$$

The co-histogram symmetry (CHS) is defined as:

$$CHS = \frac{\alpha\sum_{p} H^2(p,p) + \sum_{p,q}(p-q)^2 H(p,q)H(q,p)}{\alpha\sum_{p} H^2(p,p) + \sum_{p,q}(p-q)^2 H^2(p,q)} \tag{14}$$

where α is a positive constant and less than 1, which plays a role of the weight of the unchanged pixel possibilities (distributed on the diagonal in the co-histogram). We use 1/4 in our experiments.

CHS reflects statistic information of pixel pairs and some similarity between the individual histograms of the two images. It is 1 if a co-histogram is exactly symmetric.

CHS is mathematically independent of the PSNR, and therefore is a valid complement to PSNR. Symmetry together with PSNR makes the depiction of co-histogram more comprehensive.

CHS also gives a good estimation of accuracy in image classification and other applications [8]. Actually, the CHS given in (14) is a weighted version of a co-histogram symmetry, which enhances the difference between two images.

4 Evaluation Methodology

For evaluation purposes, we employed some image degradation models, including pixel value degradation and geometric distortion.

Geometric distortion was usually ignored in image degradation evaluation. We use StirMark-like bending [4] and image rotation, cropping and scaling for comparison.

Pixel value variation methods include image blurring, and image compression.

The blurring of an image can be caused by many factors [2], such as out-of-focus, motion, and atmospheric turbulence. Blurs are generally modeled as convolution filters. In this paper, we use Gaussian low-pass filter to generate the blurred images.

The image compression methods we considered are two international standardized image-coding frameworks, DCT-based JPEG and wavelet-based JPEG2000. Due to the block artifacts of former JPEG and many other shortcomings, new standard named JPEG2000 was published in 2000, which provides many more flexibilities and functions besides better compression quality.

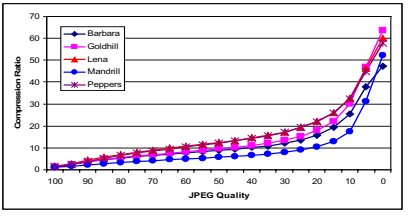
Test images we used are 5 widely-used standard test images, `Barbara`, `Goldhill`, `Lena`, `Mandrill` and `Peppers`. All of them are 512x512 8-bit gray-level images. We take each degraded version and its original image as a pair of images, and

make the co-histogram by statistics. Then, we use the co-histogram to find its "width" measure PSNR and its symmetry. Having collected all the PSNRs and the corresponding symmetries, we try to find out if there is any relationship between PSNR and CHS, and how much information of image degradation they manifest.

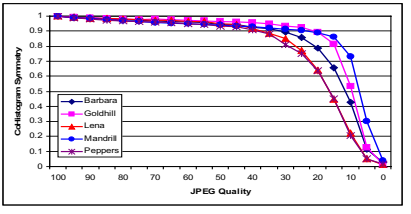
5 Experiments and Analysis

5.1 Image Compression

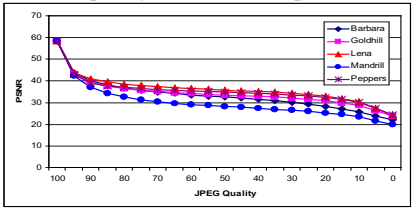
JPEG. JPEG compression is configured by image quality options [6]. All the 5 test images are compressed at image quality factor of 100 through 0 decreased by 5. The relation between the quality factors and the compression ratios are shown in Figure 1(a). JPEG quality factor vs. PSNR and co-histogram symmetry are in Figure 1 (b) and (c). Three co-histograms with Lena are shown in Figure 1(d).



(a) JPEG quality factor vs. compression ratio



(c) JPEG quality factor vs. CHS



(b) JPEG quality factor vs. PSNR (dB)

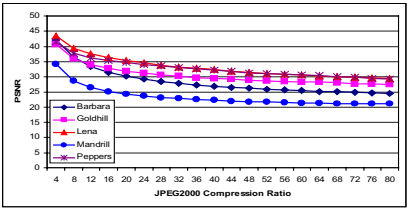
(d) Co-histograms at quality factors 30, 10, 0

Fig. 1. JPEG Compression

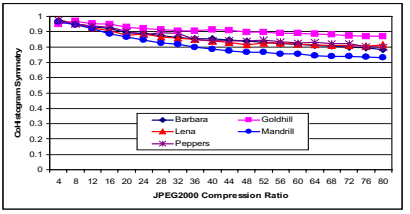
For JPEG compression, PSNRs are around 30dB or less for compression quality factor 80 or less, and CHS' are very low (even close to 0) when the JPEG quality is very low, and the compression ratio is between 30 and about 60.

JPEG2000. JPEG2000 compression can be controlled directly by bit rates or compression ratios [5]. We use the compression ratios from 4 to 80, increased by 4. The results are in Figure 2 (a) and (b). Three co-histograms of compression at ratio 32, 64 and 80 with Lena are shown in Figure 2 (c).

JPEG2000 compression gives higher PSNRs than JPEG at similar compression ratios, and the co-histogram symmetries are all larger than 0.7. It leads to a conclusion that JPEG2000 scheme is better than the old JPEG.



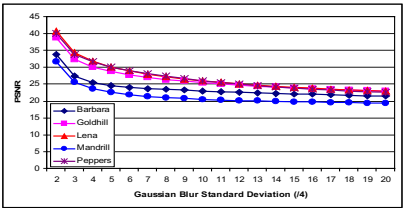
(a) JPEG2000 compression ratio vs. PSNR



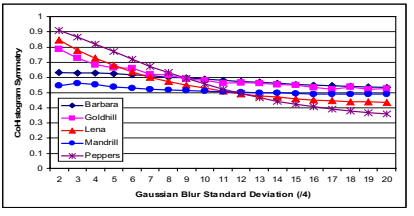
(b) JPEG2000 compression ratio vs. CHS

(c) Co-histograms at compression ratios 32, 64 and 80 (Lena)

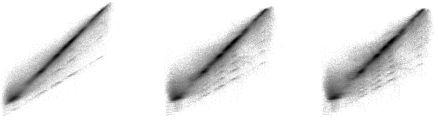
Fig. 2. JPEG2000 Compression



(a) Blur strength vs. PSNR



(b) Blur strength vs. CHS



(c) Co-histograms at standard deviation 8, 15 and 20 (Lena)

Fig. 3. Gaussian Image Blurring

5.2 Blurring

There are many models for image blurring [2]. In this paper, we use a widely-used typical model, convolution Gaussian blur. The model is defined as a 2D Gaussian filter:

$$h(x,y) = \frac{h_g(x,y)}{\sum_{x,y} h_g(x,y)} \qquad \text{where} \qquad h_g(x,y) = \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{15}$$

The filter is sized in 21 pixels and the standard deviation σ is tested from 1/4 to 20/4, increased by 1/4 as shown in Figure 3 (a) and (b). Three co-histograms at Gaussian blue standard deviation of 8, 15 and 20 with Lena are shown in Figure 3 (c).

PSNRs are almost under 20, and the co-histogram symmetries are mostly low, below 0.6.

Some other experiments with motion blur are very similar to above with Gaussian blur.

5.3 Geometric Distortion

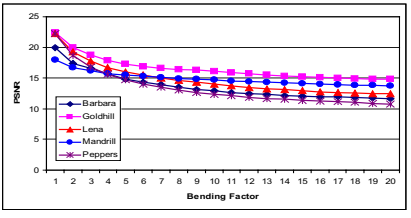
Punch-pinch bending. All the pixels are moved using a smooth sine function. Pixels at the corners are not displaced while the pixel in the center of the image is moved the most. This is similar to a "punch-pinch" effect. The degradation parameter is defined by the number of pixel displacement allowed for the center of the image. The displacement is defined as in StirMark software [4]:

$$\begin{cases} x' = x + b \cdot \sin x\pi / W \\ y' = y + b \cdot \sin y\pi / W \end{cases} \tag{16}$$

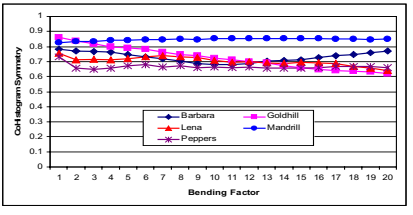
where the bending factor b varies from 1 to 20.

In our experiments, we use quadratic interpolation method and point sampling after geometric bending. Results and some co-histograms are shown in Figure 4.

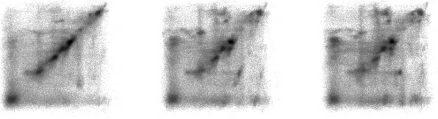
In these experiments, PSNRs are very low, while the co-histogram symmetries stay high.



(a) Bending factor vs. PSNR (dB)



(b) Bending factor vs. CHS

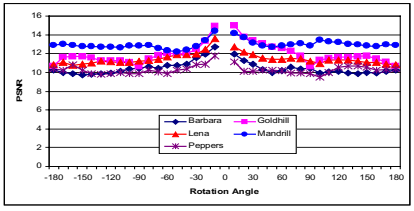


(c) Co-histograms at bending factors 8, 15 and 20 (Lena)

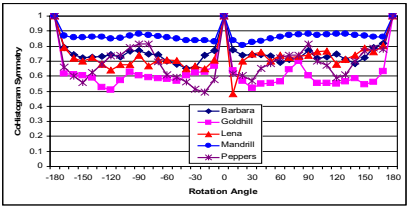
Fig. 4. Geometric Bending

Rotation. Rotation by an angle followed by centered cropping and rescaling to keep the original size of the image. Rotation, cropping and scaling are commonly used methods of geometric distortion, and the random distortion of an image can also be modeled as such distortion combinations for individual image blocks. We test the rotation angles from -180 to 180 degrees, increased by 5 degrees.

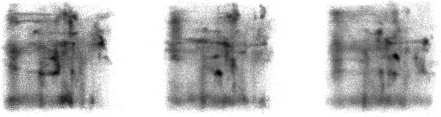
As shown in Figure 5, the results are similar to geometric bending. PSNRs are very low, while the co-histogram symmetries are high.



(a) Rotation angle vs. PSNR



(b) Rotation angle vs. CHS



(c) Co-histograms at rotation angles 30, 60 and 90 (Lena)

Fig. 5. Rotation, Cropping and Scaling

6 Conclusions

(1) Co-histogram is the joint probability distribution of two images. It gives the perfect statistic depiction of the pixel difference between two images.
(2) Co-histogram provides a visual perception of PSNR, which reflects the width of co-histogram along the diagonal.
(3) Co-histogram symmetry is independent of and a valid complement to PSNR, and it is also an objective statistic metric. Symmetry together with PSNR makes the depiction of a co-histogram more comprehensive. Its calculation is direct and as easy as that of PSNR.
(4) Co-histogram is significant in evaluation of image degradation.
(5) Some by-products of our experiments with PSNR and co-histogram symmetry: JPEG 2000 performs better than JPEG; geometric distortion changes images less than intensity distortion.

References

1. Baird, H.: The State of the Art of Document Image Degradation Modeling. In Proc. of 4 th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro, Brazil (2000) 1-16
2. Lee, H.-C.: Review of image-blur models in a photographic system using principles of optics. Optical Engineering, Vol. 29, No. 5 (1990) 405-421
3. Pappas, T.N., Safranek, R.J.: Perceptual criteria for image quality evaluation. In Handbook of Image and Video Processing, A. Bovik ed., Academic Press (2000)
4. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Attacks on copyright marking systems. In Proc. of Second International Workshop on Information Hiding, Oregon, USA, (1998) 219-239
5. Taubman, D.S., Marcellin, M.W.: JPEG2000: standard for interactive imaging. Proceedings of the IEEE, Vol. 90, No. 8 (2002) 1336-1357

6. Wallace, G.: The JPEG Still Picture Compression Standard. Communications of the ACM, Vol. 34, No. 4 (1991) 30-44
7. Wang, Z., Bovik A.C., Lu, L.: Why is image quality assessment so difficult?. IEEE International Conference on Acoustics, Speech, Signal Processing, Vol: 4, Orlando, FL, (2002) 3313-3316
8. Hao, P., Shi Q.-Y., Chen, Y.: Co-Histogram and Its Application in Remote Sensing Image Compression Evaluation. In Proceedings of International Conference on Image Processing (ICIP), Barcelona, Spain, Vol. 3 (2003) 177-180
9. Hao, P., Chen, Y.: Co-Histogram and Its Application in Video Analysis. In Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Taiwan, (2004)

MAP Signal Reconstruction with Non Regular Grids

João M. Sanches* and Jorge S. Marques

IST/ISR, Torre Norte, Av. Rovisco Pais, 1049-001, Lisbon, Portugal

Abstract. The estimation of a scalar function f using a regular grid has been extensively used in image analysis. This amounts to approximate f by a linear combination of known basis functions. However, this approach is usually not efficient. This paper proposes a more efficient algorithm, based on the use of a non regular grid, which achieves better accuracy with less basis functions. Experimental results are provided to illustrate the performance of the proposed technique.

1 Introduction

The representation of continuous functions using linear interpolation of a given set of basis functions is a common procedure in several areas [1,2,3]. Non-uniform grids are used in a large number of applications, e.g., finite elements mathematics, image format conversion, surface representation in computer graphics and curve re-sampling. Usually these non uniform sampled problems are converted to uniform sampled problems by re-sampling the data. For instance [6] presents an algorithm to re-sample data between arbitrary grids using an intermediary rectangular regular grid.

This procedure can be performed by using some optimality criterion, e.g. least squares reconstruction (LSR) error [4,5] . The main problem in these conversion operations is the representation error introduced when a given continuous or discrete function is represented using a different set of basis functions.

In this paper we will adopt a different strategy: instead of converting a non regular grid problem into a regular one we will start with a regular grid and change its geometry in order to minimize a given energy function.

We aim to represent a continuous function, $f(x)$, estimated from noisy observations at non uniform positions x_i. This function is represented as a linear combination of basis functions which vary during the estimation process. The proposed method uses the MAP criterion to estimate the coefficients associated to basis functions and to estimate its positions. This approach allows to estimate the basis functions as well as its coefficients.

* corresponding author: João Sanches, IST/ISR, Torre Notre, Av. Rovisco Pais, 1049-001 Lisboa, Portugal, Email:jmrs@alfa.ist.utl.pt, Phone:+351 21 8418195

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 204–211, 2004.
© Springer-Verlag Berlin Heidelberg 2004

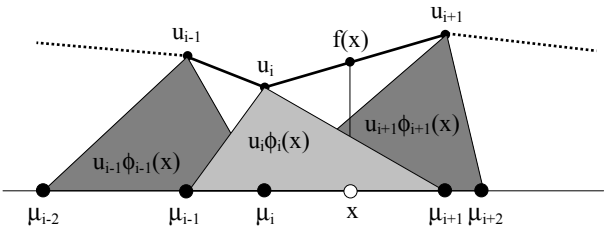


Fig. 1. Function representation with a non regular grid.

2 Problem Formulation

Let $f(x)$ be a scalar function and $V = \{(y_i, x_i)\}$ a set of M observations, where y_i are noisy observations of $f(x)$ taken at locations x_i. The goal is the estimation of $f(x)$ from the data V, by minimizing an energy function.

2.1 $f(x)$ Definition

It is assumed that the $f(x)$ belongs to a class of admissible functions defined in an interval $\Omega \subset R$, i.e., $f : \Omega \longrightarrow R$. Furthermore, it is assumed that the set of admissible functions is a finite dimension vector space F with unknown basis function, $\phi_i : \Omega \longrightarrow R$, belonging to a given class of functions. Each function $f \in F$ can be expressed as a linear combination of the basis functions,

$$f(x) = \Phi(x)^T U \tag{1}$$

where $\Phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_N(x)]$ is a vector of basis functions and $U = [u_1, u_2, ..., u_N]$ is a $N \times 1$ vector of coefficients.

Piecewise linear basis functions are used as shown in Fig.1. The shape and regions of support of the basis function are different depending on the distances between the μ_i's, where the μ_i parameter is the position of the maximum of the basis function ϕ_i (see Fig. 1). Therefore, the value of $f(x)$ at an arbitrary location x in the interval $[\mu_i, \mu_{i+1}]$ is obtained by interpolating the coefficients u_i and u_{i+1}, as follows

$$f(x) = u_i \frac{x - \mu_{i-1}}{\mu_i - \mu_{i-1}} + u_{i+1} \frac{\mu_{i+1} - x}{\mu_{i+1} - \mu_i} \tag{2}$$

All the basis functions are initialized with the same shape and at equally spaced positions, that is, forming a uniform regular grid, where $\mu_i = \min(x) + i * (\max(x) - min(x))/(N-1)$ with $0 \leq i \leq N-1$.

2.2 Observation Model

In this paper we assume that the observations are corrupted by Gaussian additive noise, $\eta_i \sim N(0, \sigma^2)$, with zero mean and standard deviation σ, i.e.,

$$y_i = f(x_i) + \eta_i \tag{3}$$

Assuming that the observations are independent this leads to the log likelihood function

$$l_y = C - \frac{1}{2\sigma^2} \sum_{j=1}^{M} (y_j - f(x_j))^2 \tag{4}$$

where $C = -\frac{1}{2} \log (2\pi\sigma^2)$.

2.3 Prior

The assumption that $f(x)$ is band limited, makes it possible to model the vector $U = \{u_i\}$ as a Markov Random Field, described by a Gibbs distribution. This approach introduces correlation between the values of neighboring nodes [7], which is expected for band limited signals.

Furthermore, in this paper, changes on the shape of the basis functions are performed. By changing the positions μ_i's of the basis functions we will maximize a given objective function. Constraints should be introduced to avoid the collapsing of all nodes in a one single point. Therefore, the vector $\Xi = \{\mu_i\}$ is also modeled as a Markov Random Field described by Gibbs distribution.

In this paper it will be used first order Markov Random Fields described by Gibbs distributions [7] with quadratic potential functions with parameters α and β for the vectors U and Ξ respectively, i.e.,

$$P(U) = \frac{1}{Z_u} e^{-\alpha \sum_{i=2}^{N} (u_i - u_{i-1})^2} \tag{5}$$

$$P(\Xi) = \frac{1}{Z_\mu} e^{-\beta \sum_{i=2}^{N} (\mu_i - \mu_{i-1})^2} \tag{6}$$

where Z_u and Z_μ are the partition functions associated to $P(U)$ and $P(\Xi)$ respectively.

The estimation of U and Ξ is performed by using the MAP criterion,

$$\begin{aligned} \hat{U} &= arg \max_{U} \log (P(Y|U, \Xi) P(U) P(\Xi)) \\ \hat{\Xi} &= arg \max_{\Xi} \log (P(Y|U, \Xi) P(U) P(\Xi)) \end{aligned} \tag{7}$$

2.4 Energy

The optimization problem defined by the equations (7) can be solved by minimizing the following energy function

$$E(U, \Xi) = -l_y - l_u - l_\mu \tag{8}$$

with respect to U and with respect to Ξ, where $l_u = \log(P(U))$ and $l_\mu = \log(P(\Xi))$. Replacing l_y, l_u and l_μ we obtain

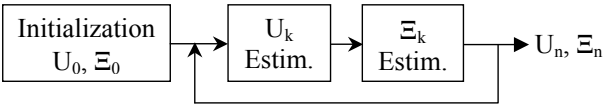


Fig. 2. Optimization process.

$$E(U,\Xi) = C + \frac{1}{2\sigma^2}\sum_{j=1}^{M}(y_j - f(x_j))^2 - \sum_{i=1}^{N}(\alpha(u_i - u_{i-1})^2 + \beta(\mu_i - \mu_{i-1})^2) \tag{9}$$

3 Optimization

The minimization of (9) is performed based on two steps: 1)optimization with respect to U and 2)optimization with respect to Ξ, as displayed in Fig.2. These two steps alternate during the iterative process until the convergence is achieved. In each step, the optimization is performed by using the ICM algorithm [8], where the energy function is minimized with respect to each unknown, keeping all the other constant. Ξ is initialized with a regular grid in Ω where $\mu_1 = x_{min}$ and $\mu_N = x_{max}$. Concerning the coefficients U, they are first initialized with a constant value, the mean of $\{y_i\}$.

The minimization of $E(U, \Xi 1)$ with respect to the parameters is performed by finding its stationary point with respect to u_i and μ_i, i.e., by solving the following equations

$$\begin{aligned}\frac{\partial E(U,\Xi)}{\partial u_i} &= 0\\ \frac{\partial E(U,\Xi)}{\partial \mu_i} &= 0\end{aligned} \tag{10}$$

Using the fixed point algorithm we obtain

$$\begin{aligned}u_i^{n+1} &= \frac{1}{2\alpha\sigma^2}\sum_{j=1}^{M}(y_j - f(x_j))\phi_i(x_j) + \bar{u}_i\\ \mu_i^{n+1} &= \frac{1}{2\beta\sigma^2}\sum_{j=1}^{M}(f(x_j) - y_j)\frac{df(x_j)}{d\mu_i} + \bar{\mu}_i\end{aligned} \tag{11}$$

where $u_i = (u_{i-1} + u_{i+1})/2$, $\mu_i = (\mu_{i-1} + \mu_{i+1})/2$ and

$$\frac{df(x)}{d\mu_i} = \begin{cases}\frac{(u_{i-1}-u_i)(x-\mu_{i-1})}{(\mu_i-\mu_{i-1})^2} & \mu_{i-1} < x \le \mu_i\\ \frac{(u_i-u_{i+1})(\mu_{i+1}-x)}{(\mu_{i+1}-\mu_i)^2} & \mu_i < x \le \mu_{i+1}\\ 0 & otherwise\end{cases}$$

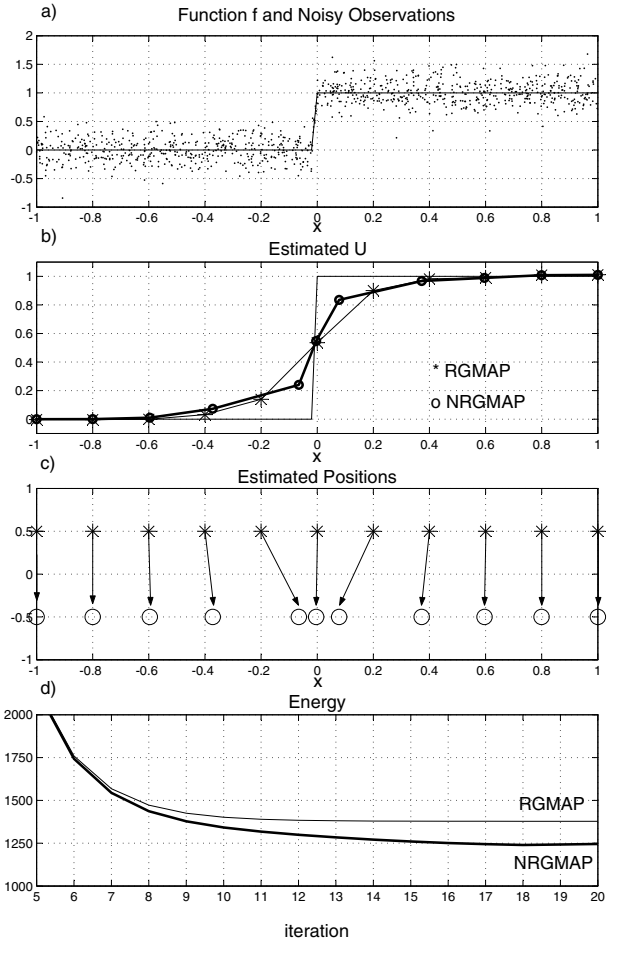


Fig. 3. Experiment with synthetic data. a)Original function and the observations. b)Estimated functions with the RGMAP (*) and NRGMAP (o). c)Node locations. d)Energy function.

During the optimization process μ_1 and μ_N, are not updated, preventing the collapse of the nodes into a single position. In fact, it is important to note that the prior associated to Ξ favors the approximation of the nodes due to the use of quadratic distances between consecutive node positions.

4 Experimental Results

In this sections we will present experiments using synthetic data and real data.

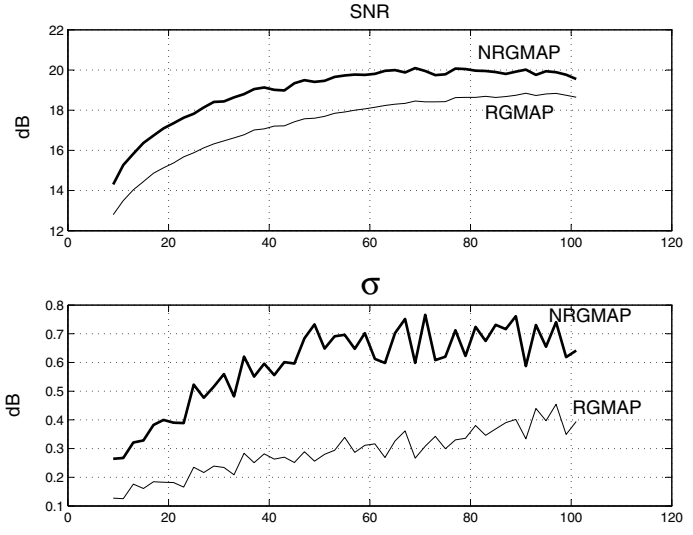


Fig. 4. Monte Carlo tests. Mean and standard deviation of the SNR obtained with 40 experiments for each number of nodes between N=11 and N=111.

4.1 Synthetic Data

The first experiment considers synthetic data. We have generated a set of 1000 points uniformly distributed in the interval $[-1, 1]$. The function $f(x)$, to be estimated, is a a step function with transition at $x = 0$. The observations, y_i, obtained at locations x_i are corrupted with additive Gaussian noise, with zero mean and standard deviation $\sigma = 0.2$ (see Fig.3.a)). The function $f(x)$ is defined using a linear combination of 11 basis functions.

The locations of the nodes associated to the basis functions are initialized as a regular grid, i.e., they are equally separated in the interval $[-1, 1]$. In a first experiment we have estimated only the coefficients u_i of the basis functions. In a second step we have jointly estimated the coefficients and the corresponding locations. Figs.3.b-d) show the results. Fig.3.b) shows the original function $f(x)$, the estimated function using the regular grid based MAP method (RGMAP), marked with asterisks, and the estimated function using the non regular grid based MAP method (NRGMAP), marked with circles. Fig.3.c) shows the final positions of the nodes for both methods, and Fig.3.d) shows the evolution of the energy function along the iterative process of estimation.

The NRGMAP method achieves lower values for the energy function ($E_{NRGMAP} = 1190$, $E_{RGMAP} = 1314$) and the nodes tend to concentrate at the transitions, as expected. The final result is clearly better.

4.2 Monte Carlo Tests

We have performed monte carlo tests, with the parameters used in the previous experiment, but with different number of nodes, starting with $N = 11$ and ending

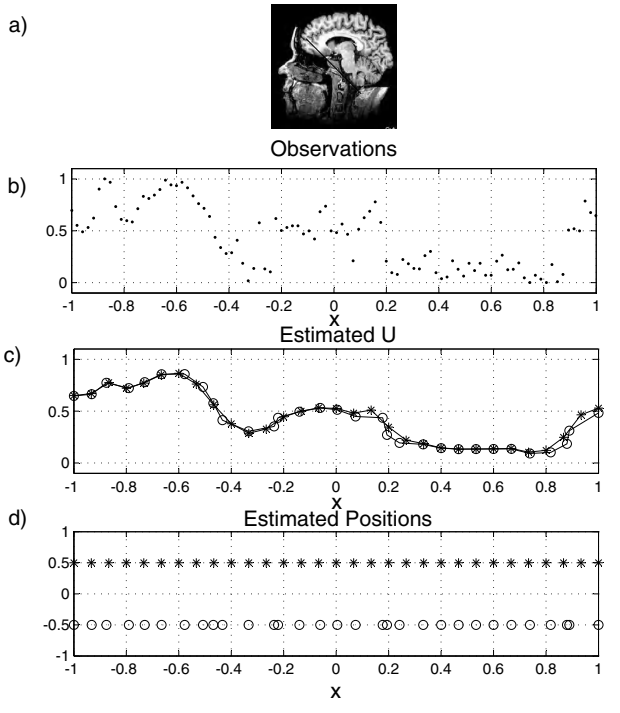


Fig. 5. Estimation of MRI profile: a) MRI image. b)Observations. c)Estimated functions. d)Node positions.

with $N = 111$. For each number of nodes we have performed 40 experiments and for each one we have computed the signal to noise ratio of the estimated function using both methods. The mean and standard deviation of the SNRs obtained in the 40 experiments were computed and the results are displayed in the Fig.4. These results show that the NRGMAP method always works better leading to a gain of approximated 2dB. This gain, however, decreases as the number of nodes increases, because when the number of nodes is large, the function is already well represented with a regular grid. The adjustments on the node positions, in this case, do not leads to relevant improvements in the representation of $f(x)$.

4.3 Real Data

In a second experiment we show a 1D profile obtained from a MRI image of the brain (see Fig.5.a)). We have used this profile to estimated a scalar function describing the anatomy along the profile, using 31 basis functions. The results are displayed in Fig.5. Once again, lower values of the energy function are achieved using the NRGMAP method, meaning a function that better represents the observations, and the nodes tend to approximate in the regions where the function to be estimated presents faster variations.

5 Conclusions

In this paper we have presented an algorithm to estimate a function from noisy observations sampled at arbitrary positions in a given interval. The function is described as a linear combination of a finite number of basis functions where the corresponding locations are optimized using the MAP criterion. We have shown that this strategy of jointly estimating the coefficients and corresponding positions leads to better results than estimating only the coefficients of the linear combination. Our goal is to apply this strategy to higher order problem by deforming 2D or 3D grids in order to obtain better representation of the estimated function with the small number of nodes as possible.

Acknowledgment. This work was partially supported by FCT under project HEART 3D (SAPIENS).

References

1. T. Lehmann, C. Spitzer, Survey:Interpolation Methods in Medical Image Processing, IEEE Trans. on Medical Imaging, vol.18, no.11, November 1999.
2. P. Thévenaz, T. Blu, M. Unser, Interpolation Revisited, IEEE TMI, vol.19, no.7, July 2000.
3. A. Jonas, N. Kiryati, Digital Representation Schemes for 3D Curves, Pattern Recognition, 1803-1816, 1997.
4. H. Sederat, D.G. Nishimura, On the Optimality of the Gridding Recosntuction Algorithm, IEEE TMI, vol.19, no.4, April 2000.
5. M. Barrutia, T. Blu, M. Unser, Non-Uniform to Uniform Grid Conversion Using Least-Squares Splines, Proceedings of the Tenth European Signal Processing Conference (EUSIPCO'00), Tampere, Finland, September 4-8, 2000, vol. IV, pp. 1997-2000
6. V. Rasche et al., Resampling of Data Between Arbitrary Grids Using Convolution Interpolation, IEEE TMI, vol.18, no.5, May 1999.
7. S. Geman, D. Geman, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE PAMI, 721-741, November 1984.
8. J. Besag, On the Statistical Analysis of Dirty Pictures, J. R. Statist. Soc. B, vol.48, no. 3, pp. 259-302, 1986.

Comparative Frameworks for Directional Primitive Extraction

M. Penas[1], M.J. Carreira[2], M.G. Penedo[1], M. Mirmehdi[3], and B.T. Thomas[3]

[1] Computer Science Department. Universidade da Coruña. Spain.
{infmpc00,cipenedo}@dc.fi.udc.es
[2] Dpto. Electrónica e Computación. Universidade de Santiago de Compostela. Spain.
mjose@dec.usc.es
[3] Department of Computer Science. University of Bristol, United Kingdom.
{majid,barry}@cs.bris.ac.uk

Abstract. This paper introduces two alternative computational frameworks for the extraction of the directional primitives present in an image. Both frameworks are divided into three stages: low level primitive extraction, organisation of low level primitives by means of dynamical neural networks (growing cell structures) and segment extraction through a pseudo-colour Hough transform. The alternative frameworks are compared and their relative advantages and disadvantages are outlined.

Keywords: Directional primitives, Gabor wavelets, Sobel filter, auto-organised structures, growing cell structures, chromaticity diagram.

1 Introduction

The boundaries of objects in an image often lead to oriented and localised changes in intensity called edges. Edge detection is the first step in many image analysis applications and it is of great importance as it constitutes the basis for the higher levels in the system. It has always been a fundamental problem in computer vision that the higher level processing stages suffer due to either too little or too much data from the lower levels of the processing. Thus, the quality of data available for further analysis is very critical.

In this paper, we investigate and compare directional properties of the popularly used, classical Sobel edge filter against those generated using Gabor filters. Both techniques are implemented in a novel computational framework using single and multi-layer *growing cell structures* followed by a line segment extraction stage using a pseudo-colour fuzzy Hough transform. The directional primitives extracted will be the basic features of a framework designed for the detection of objects in scenes through perceptual organisation [5].

The Gabor Wavelet decomposition framework presented here is computationally a more expensive process, but provides precise information about the orientation of image pixels and is independent of image type. Moreover, we have

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 212–219, 2004.
© Springer-Verlag Berlin Heidelberg 2004

implemented an approximation to Gabor Wavelets that reduces the computational time and memory requirements through the use of a pyramidal multi-scale Gabor Wavelet transform in the spatial domain, which is faster than conventional frequency domain implementations.

This paper is organised as follows. The next section describes the computational framework for the Gabor Wavelet decomposition, while section 3 outlines a similar framework for Sobel filtering. Finally, section 4 compares both frameworks, in qualitative and computational terms, and concludes the paper.

2 Directional Primitive Extraction Through Gabor Decomposition

The directional primitives in the image are obtained through a three-stage framework: (a) extraction of low level primitives implemented through Gabor Wavelet decomposition, (b) organisation of these primitives through a hierarchy of growing cell structures, (c) line segment detection through a pseudo-colour fuzzy Hough transform.

Gabor Wavelets [4] are complex exponential signals modulated by Gaussians with two important properties that make them good edge detectors: the optimisation of edge localisation [2] and the absence of image-dependent parameter tuning. Their most important drawback is their greedy demand both in memory and computational time. In a previous paper [7], we have developed a more efficient, multi-resolution spatial domain implementation of Gabor Wavelet decomposition, which we employ here, based on the convolution of 11 1D-component masks obtained through the decomposition of the 2D masks that define the wavelets. The implementation here utilises the good edge localisation property of Gabor Wavelets, with the exact position of an edge determined as a conjunction between a maximum in the modulus and a zero crossing in the even or the odd part of Gabor results.

In our Gabor decomposition, the input image is filtered with a bank of 16 filters centred at 2 frequencies ($\frac{1}{4}, \frac{1}{8}$) and 8 orientations ($\frac{k\pi}{8}, k = 0..7$) leading to 16 resulting images. A reduction of this output space dimensionality is necessary in the interest of efficiency. Auto-organised structures are a suitable instrument to achieve this dimensionality reduction as they allow simultaneously the reduction of the input space and the projection of the topological order in the input space to the output structure. The auto-organised structures are deployed in a two-layer hierarchy:

- The first layer is composed of two structures to analyse the results from individual frequency channels,
- The second layer is composed of a new structure that integrates the information from individual frequency channels and provides an unique result for the next processing step.

In [6], self-organised maps, growing cell structures and growing neural gas structures were investigated and compared for their power of dimensionality

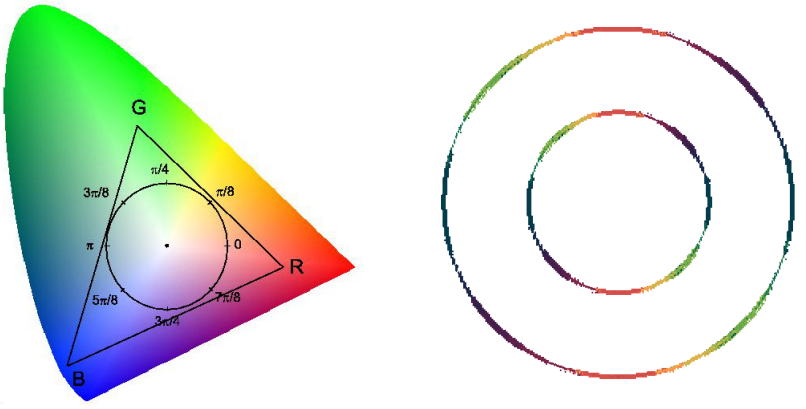


Fig. 1. Left: Colourmap inside the RGB triangle, Right: All orientations after second layer GCS.

reduction of Gabor decomposition results. Growing cell structures (GCS) [3] provided significantly better results. They are artificial neural networks based on self-organised maps that eliminate the restrictions of the *a priori* network size definition, incorporating a mechanism to add new processing elements when needed, while maintaining the network topology.

The input to the auto-organised structures in the first layer is a nine component vector assigned to each pixel (x, y) composed of: an eight component vector for each frequency channel, where each component corresponds to the result of Gabor decomposition centred at each main orientation, and a ninth component, named *response*, added to each vector in order to determine the presence of a directional feature:

$$response^i(x,y) = 0.5 + \frac{\arctan\left(\frac{(|H^i(x,y)| - |\bar{H^i}|)}{|\tilde{H^i}|}\right)}{\pi}, i = 0, 1$$

where $H^i, i = 0, 1$ are the Gabor features at the two frequencies, $|\bar{H^i}|$ is the mean and $|\tilde{H^i}|$ is the standard deviation. The first eight components were scaled such that their modulus was equal to the *response*, reducing inter-image and inter-frequency variability (as low-level frequency filters tend to produce a greater range of values).

To represent the different directionalities provided by the auto-organised structures, each processing element was assigned a colour from a colourmap to indicate its orientation. The colourmap was obtained from 8 equidistant points on the perimeter of the maximum circle inside the RGB triangle in the chromaticity diagram [8], centred at white (see Fig. 1-left). The right image in Fig. 1 shows the second layer GCS output from a ring demonstrating the colours of the entire direction space, i.e. $0 - 2\pi$.

The system was tested with many different kinds of images, three of which are shown in Fig. 2. The second and third rows of this figure show the results

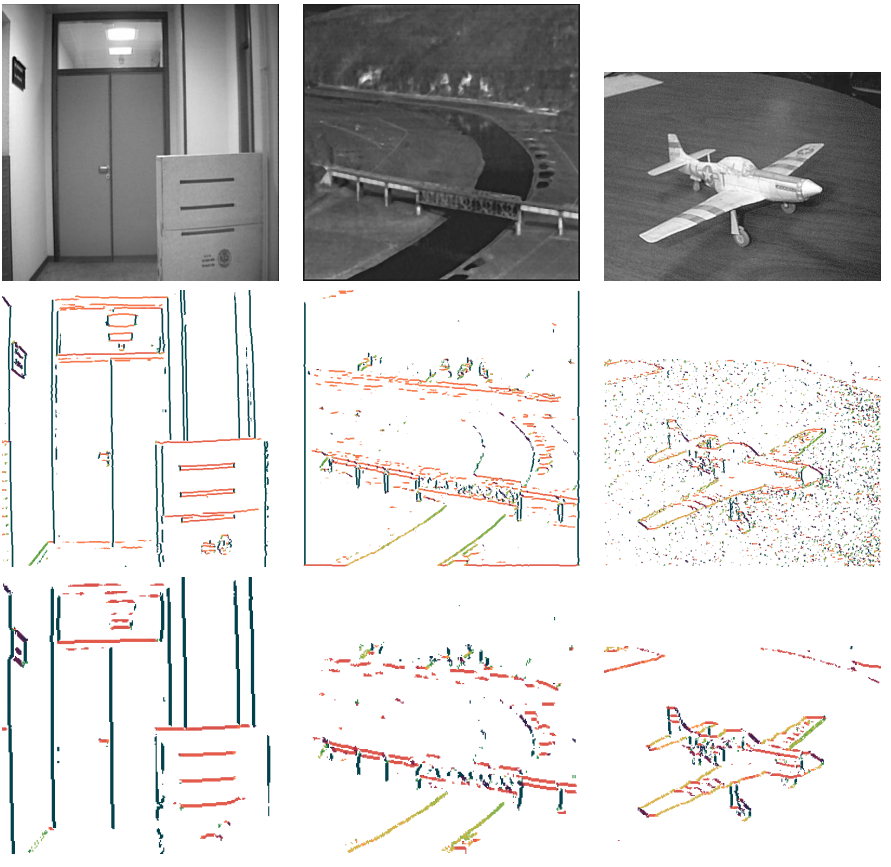


Fig. 2. First row: images 'room', 'bridge' and 'plane'. Second row: first frequency channel analysis. Third row: second frequency channel analysis.

from the first and second frequency channels after the first layer of the GCS. Although the first frequency channel seems to do enough to detect the directional features present in the image, results from images like the 'plane' image show that the second frequency channel provides very useful information in images where there may be detailed multi-directional textural background. The second higher frequency channel discards this information and reinforces the stronger image features.

In the second hierarchical layer of the auto-organised structures, the results from both frequency channels are integrated using a new GCS. The input to the structure is a six component vector for each pixel, i.e. the two RGB outputs from the two frequency channels both normalised to lie in the range $[0, 1]$. Fig. 3 shows the results of this integration, again using a colourmap to assign colours to the output of each processing element. This time, the colour assigned to a processing element is the arithmetic mean of the RGB colours that its weights represent.

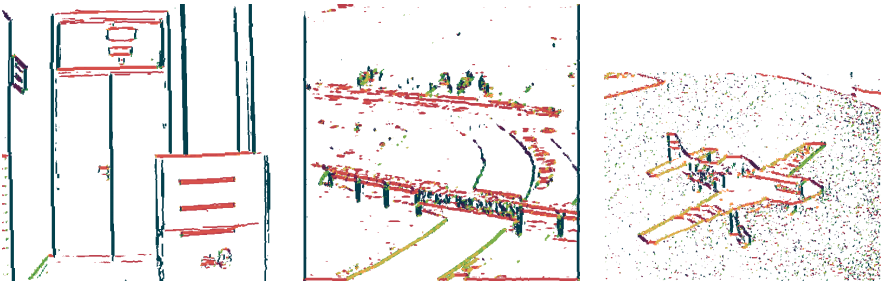


Fig. 3. Integration of both frequency channels from images in first row of Fig. 2.

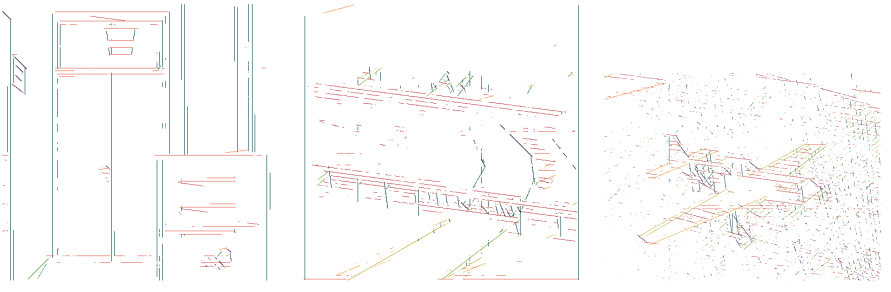


Fig. 4. Line segments from images in first row of Fig. 2 using our pseudo-colour fuzzy HT.

In the third and final stage, line segments are extracted using a pseudo-colour fuzzy Hough transform [1] (HT). Traditional HT implementations are based on the results of classical edge detectors like Sobel or Canny. We have designed and implemented a Hough transform based on pseudo-colour images where each colour represents a specific orientation. The final result of this process is an array of line segments, shown in Fig. 4. Each segment is defined by the polar coordinates of the line it belongs to, and its endpoints.

The advantages of the framework described are the precision of the results due to the richer orientational information provided by Gabor decomposition and the lack of need to tune parameters regardless of the nature of the image.

3 Directional Primitive Extraction Through Sobel Filtering

Directional primitive extraction through Sobel filtering is also carried out in three stages: (a) extract low-level primitives through Sobel filtering in two 'frequency' channels, (b) organise the results via a growing cell structure, and (c) extract lines using the pseudo-colour HT. This framework differs from the previous one in the first and second stages of the process, with the third remaining common to both.

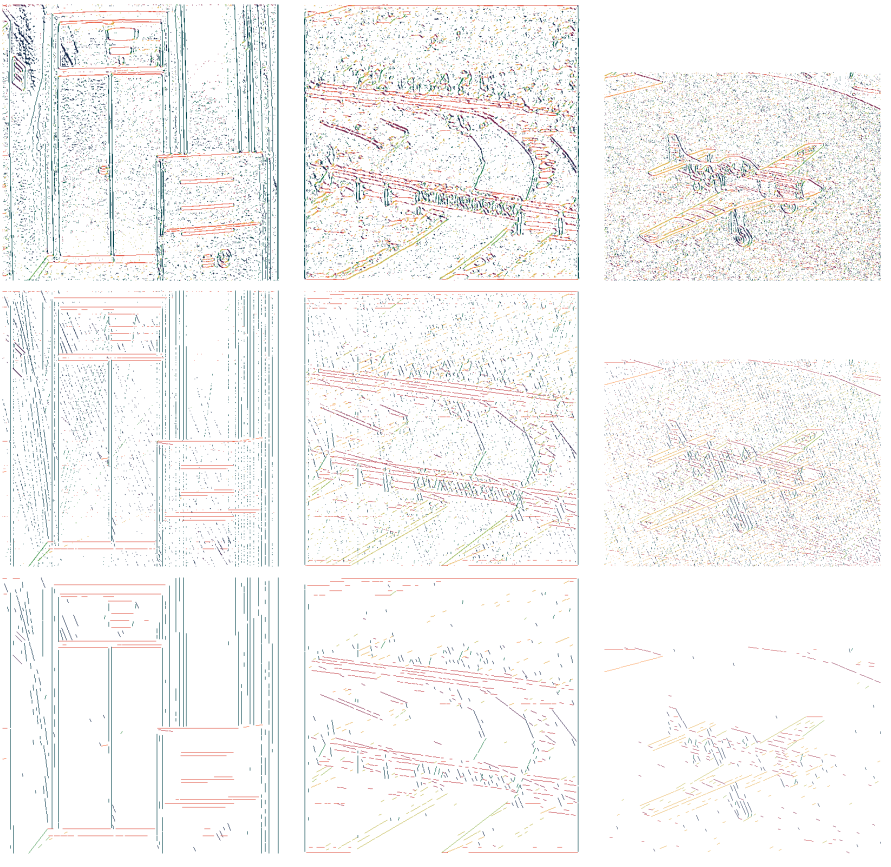


Fig. 5. First row: results of Sobel analysis after the GCS layer. Second row: results of segment detection from images in first row. Third row: results after the removal of segments with less than 5 pixels.

Initially, Gaussian smoothing is applied to the input image, once with $\sigma = 1$ for a first frequency channel and once with $\sigma = 2$ for a second frequency channel. This helps in removing part of the noise in the image and also to some extent replicates the multi-scale analysis of the Gabor decomposition method. Then, Sobel filtering is applied using 3×3 and 5×5 filters, for the first and second frequency channels respectively, resulting in two pairs of images $(Mag_i, Grad_i), i = 1, 2$ containing the magnitude and gradient of the edges found in the image. The analysis is completed with the implementation of a non-maximum suppression algorithm that operates in the direction indicated by $Grad_i$. The result of this processing is a new image Max_i that, with $Grad_i$, will be employed in the following stages of the process.

Note, only one auto-organised structure is necessary in this framework, as the amount of information provided by Sobel filtering is less than that provided

by Gabor decomposition, i.e. 4 instead of 16 sets of filtered data. The input to this structure is composed of Sobel filter outputs $(Max_1, Grad_1, Max_2, Grad_2)$, with Max_1 and Max_2 normalised as in Section 2 to reduce inter-image and inter-frequency variability. A colourmap has been used again for visualising the orientations, with the colour assigned to a processing element being the arithmetic mean of the colours corresponding to the gradients in its weights.

Results from the GCS analysis are shown in the first row of Fig. 5. As illustrated, the Sobel is quite sensitive to noise and even after the smoothing and the growing cell neural network analysis, the results are not encouraging. The second row of Fig. 5 shows the segments extracted by the pseudo-colour HT which results in a number of spurious line segments due to the noisy orientations. A post-processing stage to improve the HT results is shown in the last row of Fig. 5 where lines shorter than a predefined length threshold have been removed.

The results reiterate the dependency of the image type for a traditional edge filter. While the 'room' image results in some lines of reasonable quality (but not as good in comparison to the Gabor decomposition technique) that could be of use to higher-level processing, the 'plane' image encapsulates a much less useful set of lines. Even after post-processing, both results are still extremely noisy.

4 Conclusions

In this paper we considered two frameworks for the extraction of image primitives (i.e. directional pixel information) using growing cell structures and a pseudo-colour fuzzy HT with application to features extracted in a traditional manner and using multi-scale Gabor Wavelets. Two frequency channels were chosen in both implementations, centred at high and intermediate frequencies, as this is the frequency range where we can expect the most important image features to be located. Also, eight main orientations were chosen in the implementation based on the Gabor decomposition, as they adequately cover the space of all possible angles. We performed experiments over a wide set of test images, considering various numbers of frequency channels and orientations, before the current set was selected as the most suitable.

Table 1. Timing results in seconds for each and all stages of both implementations, for three example images.

	Gabor/GCS			**Sobel/GCS**		
	Room	**Bridge**	**Plane**	**Room**	**Bridge**	**Plane**
Filtering	40	37	47	7	7	9
GCS analysis	1126	945	1297	333	265	390
pseudo HT	6	5	12	7	5	15
Totals	1172	987	1356	347	277	414

Directional primitive extraction based on Sobel filtering is simple and computationally cheaper. However, by trading for computational complexity, we can use the Gabor Wavelet decomposition with growing cell structures to obtain not only more (subjectively judged) accurate results, but also without having to consider any parameters for the type of image being investigated. The main parts of both frameworks were timed on a 1.3GHz PC running Linux Debian and are shown in Table 1 for the 'room', 'bridge' and 'plane' images which were 512×512, 512×512 and 640×480 respectively. These images can be accessed in the URL: `http://www.lfcia.org/~infmpc00/ICIAR2004`. Note that the processing time of each stage of both frameworks is image-dependent: the amount of noise, the image size, and the number of features in the scene can all affect each stage. As expected, the GCS part is the main element of the computational complexity.

Acknowledgements. This work has been supported by Xunta de Galicia and MCyT through grants PGIDIT03TIC10503, TIC2003-04649-C02-0 and the Dirección Xeral de I+D.

References

[1] M.J. Carreira, M. Mirmehdi, B.T. Thomas, and M. Penas. Perceptual primitives from an extended 4D Hough transform. *Image and Vision Computing*, 20(13-14):969–980, 2002.

[2] J. Van Deemter and J. Du Buf. Simultaneous detection of lines and edges using compound Gabor filters. *Journal of Pattern Recognition and Artificial Intelligence*, 14(4):757–777, 2000.

[3] B. Fritzke. Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.

[4] B. Gabor. Theory of communication. *Journal of the Institute of Electronic Engineers*, 36(93):429–457, 1946.

[5] D. G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, 1985.

[6] M. Penas, M. J. Carreira, and M. G. Penedo. Auto-organised structures for extraction of perceptual primitives. *LNCS*, 2085:628–636, 2001.

[7] M. Penas, M. J. Carreira, and M. G. Penedo. Perceptual organization of directional primitives using a pseudo-color Hough transform. *LNCS*, 2749:893–898, 2003.

[8] G. Wyszecki and W. S. Stiles. *Color science, concept and methods, quantitative data and formulae.* John Wiley & sons, 1982.

Dynamic Content Adaptive Super-Resolution

Mei Chen

Hewlett-Packard Laboratories
1501 Page Mill Road, MS 1203, Palo Alto, CA 94304, U.S.A.
meichen@hp.com

Abstract. We propose an automatic adaptive approach to enhance the spatial resolution of an image sequence that allows different regions of the scene to be treated differently based on the content. Experimental results have shown its promise to avoid artifacts that otherwise might result from treating all regions of the scene in the same way during the resolution enhancement process. Moreover, it is able to dynamically tailor the image resolution enhancement process in an intelligent way. In particular, it can deploy processing resources to different regions of the scene at varying computational intensity levels to achieve high quality resolution enhancement in an efficient way.

1 Background and Motivation

As anyone who has ever tried to "blow up" a digital photograph knows, the resolution of the sensor is a fundamental limit. This limitation is even more pronounced for digital videos, for which a typical VGA video only has 0.3MByte pixels per frame. Interpolation can be used to show an image on a display with higher resolution, but this adds no additional information, and often results in visual degradation.

With still photographs interpolation is the best we can do, as information not captured by the sensor is permanently lost. However, this is not necessarily true for an image sequence. Imagine the scene as an array of N x N subpixels, with each of the subpixel much smaller than the M x M pixels captured by the camera sensor, and the camera is moving in a continuous manner relative to the scene. Each video frame will include a different M x M subset of the N x N subpixels. On the other hand, there will be macroscopic features of each sequential image that are related. As a result, each frame includes a mixture of new and previous information. Therefore, it should be possible to extract from each successive video frame the new information that could be used to enhance the resolution of a single frame, i.e. super-resolution. Applying super-resolution to image sequences allows a temporal integration of image information leading to imagery with higher resolution, improved sharpness and a better signal to noise ratio.

Enhancing image quality by synthesizing information from multiple frames is a difficult problem that has been extensively researched [1-11]. However, existing approaches attempt to combine image information in multiple frames without discriminating different content in the scene, thus tend to have sub-optimal performance for dynamic scenes. For example, for a scene with a slow moving pedestrian and a fast moving car, there are more consecutive frames that share

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 220–227, 2004.
© Springer-Verlag Berlin Heidelberg 2004

complementing information for the pedestrian than that for the car. Therefore, if all images sharing information of the pedestrian are used to enhance the resolution of the full frame, it's very likely that the result will be noisy around the car, due to higher possibility of errors in image correspondence. On the other hand, if only images sharing sufficient information of the car are used for resolution enhancement of the full frame, the result for the pedestrian will be compromised since much available information is wasted without being taken advantage of.

In this paper, we propose a dynamic approach that uses image cues such as motion to segment scene content, so as to adaptively synthesize information for different image content from relevant frames to render a resolution enhanced image. An additional advantage is that this allows selective enhancement of the image scene, which is efficient when there is limited computing resource (e.g. CPU time, memory, etc.)

2 System Overview

Our proposed system is consisted of an *inter-frame motion estimation module*, a *motion-based scene segmentation module*, and a *dynamic content adaptive super-resolution synthesis module*, as illustrated in Figure 1. To enhance the spatial resolution of an image in a sequence, the inter-frame motion estimation module first computes dense correspondence maps between the frame of interest, i.e. the reference frame, and a number of neighboring auxiliary frames. The motion-based scene segmentation module then classifies regions in the reference frame according to the estimated motion vectors. Specifically, regions with small motion vectors between the frame of interest and **all** the auxiliary frames are classified as having *low motion*; regions with large locally coherent motion vectors between the frame of interest and **any** of the auxiliary frames are classified as having *high motion;* whereas other regions are grouped as having *intermediate motion*. This motion segmentation process is done in a hierarchical manner to ensure region connectivity and continuity. Lastly, the dynamic content adaptive super-resolution synthesis module reconstructs a higher resolution image integrating complementing information from relevant neighboring auxiliary frames. In particular, for regions of *high motion*, fewer auxiliary frames are used because only closely adjacent video frames contain sufficient complementing information that can be reliably registered and synthesized; whereas for regions of *low motion*, a larger number of auxiliary images are employed as more frames share macroscopic feature with the reference frame can be correctly corresponded and integrated; regions with intermediate motion are synthesized using a medium number of auxiliary frames. A final refinement process is done on the synthesized image to ensure consistency between regions.

3 Algorithm Details

In this section, we discuss the aforementioned processes in more details.

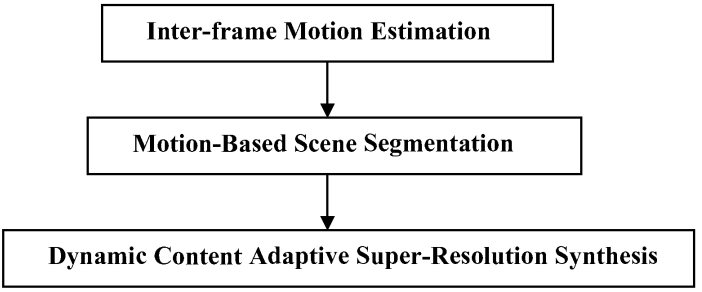


Fig. 1. Overall system diagram of the proposed approach

3.1 Inter-frame Motion Estimation

The motion estimation module uses a pyramid-based hierarchical image alignment technique to compute the motion between a pair of images (i.e., an auxiliary image and the reference image) [12]. Laplacian or Gaussian pyramids are constructed for the input images, and motion parameters are estimated in a coarse-to-fine manner. In particular, at each pyramid resolution, the sum of squared differences (SSD) measure integrated over regions of interest (which is initially the entire image) is used as the objective function:

$$E(U(P)) = \sum_{P} \left(I_r(P) - I_t(P - U(P)) \right)^2 \qquad (1)$$

where I is the Laplacian or Gaussian filtered image intensity (or the luminance channel of a color image). The sum is computed over all the points P within the region of interest, and is used to denote the SSD error of the entire motion field within that region. The motion field $U(p)$ is modeled by a set of global parameters (e.g., plane parameters) and local parameters (e.g., optical flow). Besides improving efficiency, another advantage of using image pyramids is the reduction of false matches caused mainly by the mismatches at higher resolutions for image regions under large motion. This is because larger displacements are first computed using images of lower spatial resolution, where they become small displacements due to sub-sampling. The result at a coarser resolution is used to initialize the estimation at a finer resolution. Images at higher spatial resolution are used to improve the accuracy of displacement estimation by incrementally calculating finer displacements.

Numerical methods such as Gauss-Newton minimization can be applied to the objective function (1) to compute the unknown motion parameters. Starting with some initial values (typically set to zero), the hierarchical estimation algorithm iteratively refines intermediate results in order to minimize the SSD error from coarse to fine spatial resolutions. After each iteration of motion estimation, the current set of parameters is used to warp the auxiliary image to the coordinate frame of the reference image in order to reduce the residual displacement between the image pair.

The motion vector at each pixel is assumed to be locally constant (i.e. *smooth*) within a small support window around that pixel. The motion for a pixel is estimated by using all the pixels in its support window. This process is repeated for each pixel and thus results in a smoothly varying motion vector field. Furthermore, motion vectors are computed using windows on and off-centered around each pixel under examination, and the estimate that produces the smallest error measure is kept as the best estimate for the pixel under consideration. This helps achieving crisper correspondence maps that are sub-pixel accurate at occluding image boundaries.

As a validation step, an alignment accuracy map for each pairing of the reference image and a respective auxiliary image is generated based on the computed correspondence maps. The current implementation computes normalized correlations between the reference image and each motion-compensated auxiliary image using respective estimated correspondence maps.

3.2 Motion-Based Scene Segmentation

The scene segmentation module assigns regions in the reference image to respective motion classes based on the computed image correspondence maps, as illustrated in Figure 2. It first computes motion magnitude maps from each estimated correspondence map; secondly, it builds multi-resolution motion pyramids for each of the computed motion magnitude maps. It then segments each down-sampled motion magnitude map into a respective set of motion classes (e.g., a high motion class, an intermediate motion class, and a low motion class) using a k-means clustering method. The segmentation is refined iteratively by assigning pixels to each class and re-computing the centroid of each cluster. The segmentation method iterates between the following steps:

1. Compute cluster centroids and use them as new cluster seeds; and
2. Assign each object to the nearest seed.

The final partition corresponds to a respective set of motion clusters for each coarse motion magnitude map in which the total distance between pixels' motion vectors and the centers of their respective clusters is minimized, while the distances between clusters are maximized.

The motion-based segmentation module iteratively groups pixels of similar motion from coarser resolution levels up to the original base image resolution level. The segmentation results from each respective coarser resolution are used as the starting point for the same segmentation process applied to the next higher resolution level.

The motion-based segmentation maps computed for each pairing of the reference image and respective auxiliary images are merged into a unified motion class segmentation map for the reference image. Specifically, it assigns a given reference image pixel to the low motion class in the unified motion class segmentation map when the given pixel is assigned to the low motion class in all of the separate motion-based segmentation maps; it assigns a given reference image pixel to the high motion class in the unified motion class segmentation map if the given pixel is assigned to the high motion class in any of the separate motion-based segmentation maps; and it

assigns a given reference image pixel to the intermediate motion class in the unified motion class segmentation map when the given pixel is unassigned to the high motion class in any of the separate motion-based segmentation maps and is unassigned to the low motion class in all of the separate motion-based segmentation maps.

3.3 Dynamic Content Adaptive Super-Resolution Synthesis

The dynamic content adaptive synthesis module first up-projects the reference image, the unified motion-based segmentation map, the image correspondence maps, and the image alignment accuracy maps from the base image resolution level to the target image resolution level. Secondly, it warps each of the auxiliary images to the coordinate frame of the reference image at the target resolution using the respective up-projected image correspondence maps. Next, the adaptive synthesis module generates synthesis maps for each of the motion-compensated auxiliary images. This is done by applying an empirically determined threshold to the normalized correlation values in the alignment accuracy maps. The synthesis maps are used to classify computed motion vectors in each up-projected motion map into valid and invalid motion vector classes. The purpose of this step is to ensure that only the image pixels that are sufficiently aligned with their corresponding counterparts in the reference image contribute to the reconstruction of the target image. Finally, the adaptive synthesis module computes pixel values for the target image based on their classifications in the unified motion-based segmentation map. In particular, it combines pixel value contributions from the reference image and a selected number of auxiliary images in accordance with the unified motion-based segmentation map and the synthesis maps. The number of auxiliary images selected and the degree of inter-frame motion are related by inverse variation. In our current implementation, only the two nearest neighbors of the reference image contribute pixel values to target image pixels in the high motion class, only the four nearest neighbors (i.e., two nearest neighbors on either side) of the reference image contribute pixel values to target image pixels in the intermediate motion class, and only the six nearest neighbors (i.e., three nearest neighbors on either side) of the reference image contribute pixel values to target image pixels in the low motion class.

For pixels in the auxiliary images that are identified as having invalid motion vectors according to the synthesis maps, their contribution to the target image is zero. For target image pixels whose corresponding auxiliary image pixels are all identified as having invalid motion vectors, the adaptive synthesis module computes the respective target image pixel values by intra-frame interpolation. For target image pixels with at least one corresponding auxiliary image pixel with valid motion vectors, the adaptive synthesis module computes the target pixel values by merging contributions from the reference image pixel and the *valid* pixel(s) in the respective motion-compensated auxiliary images (determined by the reference pixels' motion classes) using a weighted combination. The pixel-wise weights are proportional to the pixel's alignment quality measure in the corresponding synthesis map, and inversely proportional to the temporal distance between the reference image and the respective auxiliary image. Note that the weights may vary from pixel to pixel, and from image

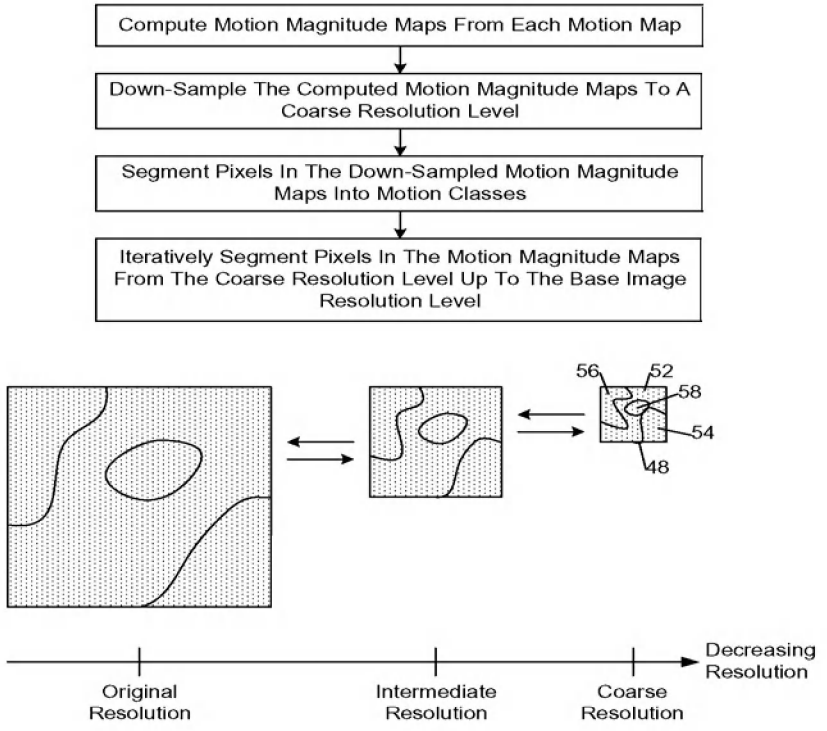


Fig. 2. An illustration of the motion-based scene segmentation module

to image. This alignment-quality-related weighting helps guarantee that only relevant and valid information from well-aligned images is used during the reconstruction process and that unreliable information is ignored. This super-resolution enhancement procedure may be applied to one or all of the luminance and chrominance components of the base images. Post-processing is done to ensure smoothness at motion segmentation boundaries.

4 Experimental Results

We have tested the proposed system on a number of videos with dynamic content, and achieved encouraging results. Figure 3 shows an example from a video with a girl riding a bike passing a parked van (with the text). The system was able to segment the fast moving object (the girl) from the relatively stable object (the van), and adaptively enhance their resolution by exploiting the appropriate information without compromising the performance. Note that the resolution for the van was significantly enhanced so that the previously illegible text is now clearly readable, and the quality of the girl's face is also considerably improved without introducing any motion artifact.

Low motion region: original video frame (left) and super-resolution enhanced to 2x (right)

High motion region: original video frame (left) and super-resolution enhanced to 2x (right)

Fig. 3. Experimental results

5 Discussions and Future Directions

We have proposed a system that adopts an adaptive approach to enhancing the spatial resolution of a video sequence. It allows different regions of a dynamic scene to be treated differently based on inter-frame motion, and is able to avoid artifacts that otherwise might result from treating all regions of the scene in the same manner. As an additional benefit, it is able to dynamically tailor the resolution enhance process in an intelligent way, by allocating computing resources to different regions in an image according to its content so as to achieve higher efficiency. Experimental results have shown to be promising.

In the future, we would like to further improve the ability to focus computing resources, e.g. time, memory, on image areas where information is most concentrated, such as areas of high textural content, or areas around human faces. One possibility is to add *a texture analysis module* to analyze textural statistics of the images, and marks the regions containing high textual information as regions of interest (ROIs). The original video stream and the information of the ROIs are given to the content adaptive super-resolution enhancement module, where only regions of interest in the

video frame(s) undergo the full enhancement process, with the rest of the image regions re-sampled to the higher resolution grid.

References

1. S. Baker and T. Kanade, "Hallucinating Faces," Technical Report CMU-RI-TR-99-32, The Robotics Institute, Carnegie Mellon University, 1999.
2. S. Baker and T. Kanade, "Limits on Super-resolution and How to Break Them," *Proceedings of CVPR'2000*, Hilton Head , South Carolina, 2000, pp. 372-279.
3. S. Borman and R.L. Stevenson, "Spatial Resolution Enhancement of Low-Resolution Image Sequences: A Comprehensive Review with Directions for Future Research," *Technical Report,* University of Notre Dame, 1998.
4. N.K. Bose, H.C. Kim, and H.M. Valenzuela, "Recursive Implementation of Total Least Squares Algorithm for Image Reconstruction from Noisy, Undersampled Multiframes," *Proceedings of ASSP'93*, Minneapolis, MN, vol. 5, pp. 269-272.
5. M. Elad and A. Feuer, "Restoration of Single Super-resolution Image from Several Blurred, Noisy and Down-sampled Measured Images," *IEEE Trans. on Image Processing*, Vol. 6, no.12, pp. 1646-58, 1997.
6. R.C. Hardie, K.J. Barnard and E.E. Amstrong, "Joint MAP Registration and High-resolution Image Estimation Using a Sequence of Undersampled Images," *IEEE Trans. on Image Processing*, Vol. 6, No. 12, pp. 1621-33, 1997.
7. N.J. Higham, "A Survey of Componentwise Perturbation Theory in Numerical Linear Algebra," *Mathematics of Computation 1943–1993: A Half Century of Computational Mathematics, volume 48 of Proceedings of Symposia in Applied Mathematics (W. Gautschi eds.)*, pp. 49-77, American Mathematical Society, Providence, RI, USA, 1994.
8. T.S. Huang and R. Tsai, "Multi-frame Image Restoration and Registration," *Advances in Computer Vision and Image Processing*, Vol. 1, pp. 317-339, 1984.
9. M. Irani and S. Peleg, "Improving Resolution by Image Restoration," *Computer Vision, Graphics, and Image Processing*, Vol. 53, pp. 231-239, 1991.
10. R. Schultz and R. Stevenson, "Extraction of High-resolution Frames from Video Sequences," *IEEE Trans. on Image Processing*, Vol. 5, No. 6, pp. 996-1011, 1996.
11. S. Xu, *The Theory and Methods of Matrix Computation (in Chinese)*, Peking University Press, 1995.

Efficient Classification Method for Autonomous Driving Application

Pangyu Jeong and Sergiu Nedevschi

Computer Science Department, Technical University of Cluj-Napoca,
Constantin Daicoviciu 15, 400020 Cluj-Napoca, Romania
{Pangyu.Jeong,Sergiu.Nedevschi}@cs.utcluj.ro

Abstract. This paper intends to propose a real-time and robust classification method against noise facts for extracting the road region in complex environments. A new approach based on the probability is presented aiming the reduction of the classification area and time. The process starts from initial seed inside sampled road region and stops when the seeds identify the road region borders. In order to increase accuracy of classification, a more powerful discrimination function is proposed based on the local difference probability. This method behaves like a supervised classification. However, it extracts a priori information from each processed image providing better tuning of the discrimination threshold to the image features.

1 Introduction

Many classification methods appeared so far related to road region detection. Computationally high algorithms, from mathematical point of view, are used in order to obtain more polished classification results by the time expensive applications: wavelet-based [1], [2], [3], [4] and filter-bank-based classification [5], [6], [7]. These methods use the frequency domain in order to extract the texture features from the image.

In the time-critical applications, the K-Mean family applying on the gray scale image is used with time domain features: K-Mean, Fuzzy-K-Mean, and Contiguity-K-Mean [8]. These methods can considerably reduce classification time as opposed to frequency-based classification. However it still doesn't satisfy the real-time condition.

In our previous works, we used the K-Mean and the local threshold method [9] in the combined feature space [10], color/gray and texture, basically, in order to classify the pixels in road or non-road. In the generic methods and in our previous methods, after the classification is finished, we have to decide which class is the road. Especially, in the highway case, it is a challenge to obtain well-detected road region, i.e. pavement region, because of the limited classifying ability of the discrimination function in the case when we use two classes classification. However, if we increase the class number to obtain abundant road classes, the classification time is increased, and more, the procedure of class merging is required.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 228–235, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper we focus our attention on reducing classification time and on obtaining accurate classification results in real-time conditions. The process starts from initial seed inside sampled road region and stops when the seeds identify the road region borders. In order to increase accuracy of classification, a more powerful discrimination function is proposed based on the local difference probability (LDP). The proposed method is divided into two processes mainly.

In the learning process, carried out for each image, the LDP is estimated on well-established road sample area, corresponding to the initial seeds, and discrimination threshold are established.

In the classification process, LDPs is computed in each candidate pixels, and compared against discrimination threshold. The detailed procedure of the classification method based on the LDP is described in Section 2.

The advantages of this method compared to generic methods, e.g. the K-Mean family, Gabor filters, and GMRF follows:

1) It doesn't use recursive operation; each pixel is used only one time in the classification procedure. The algorithm iteration cost is $\{O(n_p) \mid 1 \le n_p \le (n_r \times n_c)\}$. Where n_p is the total pixel count used in the image space, n_r and n_c are row and column size of image. Running the above-mentioned generic algorithm takes $O(n_c \times n_r \times c)$, where c is the classifier classes' number. The quantity of the saved classification time is $O(n_c \times n_r \times c) - O(n_p)$.

2) It is not necessary to extend feature vector in order to overcome the noise factors: shadow and high illumination. It has strong discrimination power that is based on the LDP.

2 The Probability-Based Classification

We change the gray scale feature vector space to the Gaussian feature vector space in order to apply the probability-based classification. There are two advantages. The first one is that it can reduce strong noise facts roughly. The dedicated noise reduction is presented in Section 2.1. The second one is that it emphasizes the pixel positions where the neighbors have great variation. The entire procedure of the LDP-based classification is presented in Figure 1.

2.1 Learning Process

The proposed LDP-based classification is a sort of supervised classification on the Gaussian feature vector space. The difference between the LDP-based classification and the most used supervised classification, namely Bayesian classification, is that the LDP-based classification uses current state information for a priori knowledge, and the pixels aren't classified by the similarity of the pixel values, but by the distance range of the similarity among the pixels based on the LDP. To perform of this, we have to solve two problems.

1) We need to decide on a well-established road sample region in order to extract a priori information. We assume that this area is placed in front of the ego vehicle.

2) We have to determine the size of well-established road sample region. We make an assumption. The image has 50% road pixels and 50% non-road pixels. We give the same probability for the road and the non-road pixels. It means that the results of the probability calculation are not affected from previous results. Then, we locate well-established road sample region inside 50% road pixels. If the randomly sampled area is placed on the road region, the size of this is not much influenced to the classification. To make it surely including in the road pixels, the size of this is determined within 25% of height and 25% of width of the image, it is placed on the in front of the ego-vehicle.

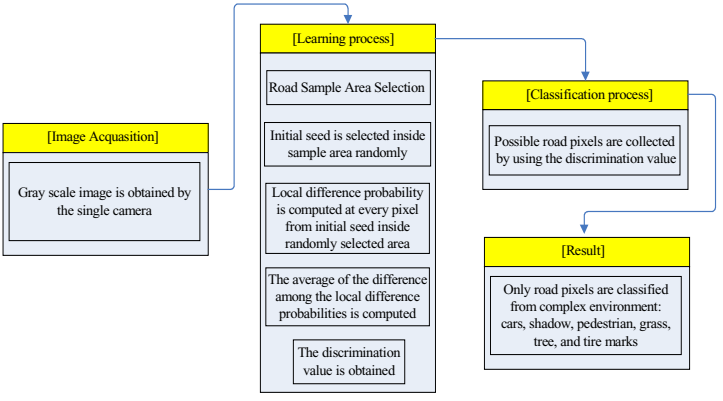


Fig. 1. The procedure of the local-difference-probability-based classification

If the position and the size of the sample road area are determined, we have to compute the discrimination threshold inside well-established road sample area. The procedure of obtaining discrimination threshold starts from the seed extension based on the probability. The theoretic procedure follows:

Let's take the randomly selected point $x_{i(r,c)}$, where r, c are row and column of the input image respectively. The four neighbors of the current pixel are $N_i = \{x_{i(r,c-1)}, x_{i(r-1,c)}, x_{i(r,c+1)}, x_{i(r+1,c)}\}$, where i is the neighborhood identifier. Each pixel of N_i is processed by applying the simple neighborhood-averaging filter processes.

$$x_{i(r,c)} = \frac{\sum_{k=(r-1)}^{(r+1)} \sum_{j=(c-1)}^{(c+1)} x_{i(k,j)}}{9}, \quad x_{i(r-1,c)} = \frac{\sum_{k=(r-2)}^{r} \sum_{j=(c-1)}^{(c+1)} x_{i(k,j)}}{9} x_{i(r,c+1)} = \frac{\sum_{k=(r-1)}^{(r+1)} \sum_{j=c}^{(c+2)} x_{i(k,j)}}{9}, \quad x_{i(r+1,c)} = \frac{\sum_{k=(r)}^{(r+2)} \sum_{j=(c-1)}^{(c+1)} x_{i(k,j)}}{9}$$

This procedure is performed at each pixel inside a 3x3 sub-window that is used for calculating the LDP. It means that the N9 neighborhood averaging isn't applied on the entire image, but it is applied on where the LDP is computed. It solves the below described over/under extension problem. The over/under extension is critical problem in the LDP-based classification. The application results, comparison results between Level set [11] and LDP method, are presented in Figure 3.

In general cases, the pixel values are changed smoothly in some directions. However, if the noise pixel is placed on those directions, this rule isn't satisfied.

In our case, we classify road pixels in four directions based on the discrimination threshold. Therefore if the noise pixels are placed on those directions, the discrimination-ability for the classification is degraded. The explanation of how the classification is defected by the noise pixel follows below.

1) The noise pixel is located on the classification direction $(r, c+1)$. The difference between the LDP of pixel (r, c) and the LDP of pixel $(r, c+2)$ is **greater** than the discrimination value. In this case, the difference between the LDP of pixel (r, c) and the LDP of pixel $(r, c+1)$ has to be greater than the discrimination threshold. However, that difference is smaller than the discrimination threshold due to the noise pixel. It brings about over classification.

2) The noise pixel is located on the classification direction $(r, c+1)$. The difference between the LDP of pixel (r, c) and the LDP of pixel $(r, c+2)$ is **smaller** than the discrimination threshold. In this case, the difference between the LDP of pixel (r, c) and the LDP of pixel $(r, c+1)$ has to be smaller than the discrimination threshold. However, that difference is greater than the discrimination threshold due to the noise pixel. It brings about under classification.

The detailed theoretic explanation of how the discrimination threshold is obtained, and of how the classification is performed is presented below.

The set of LDF in the N4 neighborhood around N_i is:

$$\begin{gathered} P_i(x) = \{ p_j(x) \mid j \in \{1, \cdots, 5\} \} \text{ or:} \\ P_i(x) = \{ p_1(x \mid x_{i(r,c)}),\ p_2(x \mid x_{i(r,c-1)}),\ p_3(x \mid x_{i(r-1,c)}),\ p_4(x \mid x_{i(r,c+1)}),\ p_5(x \mid x_{i(r+1,c)}) \} \end{gathered} \tag{1}$$

where the probability p_j must satisfy the following two conditions:

$$p_j \geq 0 \text{ and } \sum_{k=1}^{5} p_{jk} = 1 \tag{2}$$

where p_{jk} is the probability of neighbors around p_j.

The components of equation (1) are described below.

$p_1(x \mid x_{i(r,c)}) = e^{\frac{-(x_{i(r,c)} - \mu_{(r,c)})^2}{\sigma^2_{(r,c)}}}$, where $\mu_{(r,c)}$ is the mean of the sub-window around (r, c), $\sigma_{(r,c)}$ is the standard deviation of the sub-window around (r, c).

The first order derivative of the probability inside the sub-window around (r, c) is

$$dp_{(r,c)} = p_1'(x \mid x_{i(r,c)}) = (x_{i(r,c)} - \mu_{(r,c)})^2 \, p_1(x \mid x_{i(r,c)}) \tag{3}$$

$$\vdots$$

$$dp_{(r+1,c)} = p_5'(x \mid x_{i(r+1,c)}) = (x_{i(r+1,c)} - \mu_{(r+1,c)})^2 \, p_5(x \mid x_{i(r+1,c)}) \tag{4}$$

We rectify the distance between the central points and the mean around the central points using their power of 2 because we are interested in the probability distance between two pixels.

The four distances are obtained around (r,c) from equations (3) and (4).

$$d_i = \{d_{(r,c)->(r,c-1)}, d_{(r,c)->(r-1,c)}, d_{(r,c)->(r,c+1)}, d_{(r,c)->(r+1,c)}\} \tag{5}$$

Where $d_{(r,c)->(r,c-1)} = \left|dp_{(r,c)} - dp_{(r,c-1)}\right|$, $d_{(r,c)->(r-1,c)} = \left|dp_{(r,c)} - dp_{(r-1,c)}\right|$,

$d_{(r,c)->(r,c+1)} = \left|dp_{(r,c)} - dp_{(r,c+1)}\right|$, $d_{(r,c)->(r+1,c)} = \left|dp_{(r,c)} - dp_{(r+1,c)}\right|$.

We discard the smallest and largest distance values in the sets of distances corresponding to the road sample area. The average of distances is:

$$d_{th} = \frac{\sum_{i=1}^{M}\sum_{k=1}^{4} d_i(k)}{4*M-r} \tag{6}$$

where M is the number of sets. r is the number of discarded distances, and we consider that discarded distances are generated by the noise facts.

"d_{th}" will be used as the discrimination threshold.

2.2 Road Pixels Classification Process

Sometimes the road pixels aren't classified to the entire road region because the discrimination threshold is computed by randomly sampling the road area (well-established sample road region).

It means that the discrimination threshold doesn't satisfy all variance of the distance between two local pixel probabilities in the selected sample area. Therefore we need initial seed acceptance/rejection procedures. It is achieved by the following constraint condition.

We assume the extended contour is obtained from the most extended outer pixels, and has to meet the following condition. Actually, the LDP doesn't generate the contour because it is extended randomly.

The number of expanded points (x_i) in the contour has to be greater than the number of pixels (X) of the selected sample area. The seed position that satisfies equation (7) becomes the starting position of the seed.

$$\oint_{i \in M} x_i > X \tag{7}$$

If the initial seed satisfies Equation (7), the extension procedure is performed. If it doesn't satisfy Equation (7), a new randomly selected road sample area is required. Extending some directions mean that the classification is performed in those directions because the LDP-based classification is the extension-based classification. In order to decide the evolution from this position, the distance obtained from Equation (5) is compared to the discrimination threshold.

The evolution of the seed position has to satisfy the following condition.

$$E(x_i)=\begin{cases}1 & if \quad (\overleftarrow{d}_{(r,c)\to(r,c-1)} \le d_{th}) \| (\overuparrow{d}_{(r,c)\to(r-1,c)} \le d_{th}) \| (\overrightarrow{d}_{(r,c)\to(r,c+1)} \le d_{th}) \| (\overdownarrow{d}_{(r,c)\to(r+1,c)} \le d_{th}) \\ 0 & if \quad (\overleftarrow{d}_{(r,c)\to(r,c-1)} > d_{th}) \| (\overuparrow{d}_{(r,c)\to(r-1,c)} > d_{th}) \| (\overrightarrow{d}_{(r,c)\to(r,c+1)} > d_{th}) \| (\overdownarrow{d}_{(r,c)\to(r+1,c)} > d_{th}) \end{cases} \quad (8)$$

The result "1" represents that the seed can extend toward that direction. The pixel that is used for extension belongs to the road cluster. Otherwise, the seed stops the evolution. The pixel at which the extension is stopped represents the border of between the road and the non-road. Each extended seed has another four neighbors. This iteration is continuous until all extended seeds get "0". Therefore the LDP-Based classification is performed by extending the seed within the discrimination value. Finally, road area is constructed by the contour of the most extended seeds.

3 Experiments

In this section, we present the results obtained from the classification comparison among the K-Mean family, and the proposed LDP-based classification. The K-Mean family consists of K-Mean, Fuzzy K-Mean, and Contiguous K-Mean. It is applied on the gray intensity vector directly. The main focus of the comparison is:

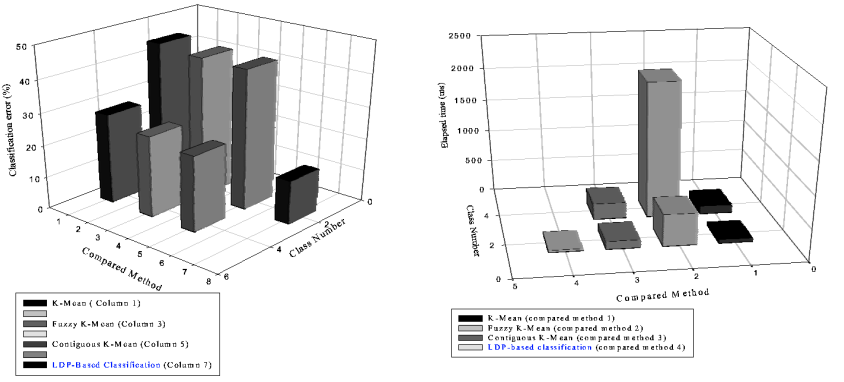
i) Classification ability and classification robustness: we will present the quantitative analysis of the classification. The proposed method uses two classes classifier because we are interested in finding only road and non-road region. The results of statistical analysis are presented in Figure 2. The quantitative analysis is presented in Table 1. The seed extension ability that is directly connected to classification ability is also compared with the Level-set-based seed extension and the LDP-based seed extension. It is presented in Figure 3. The LDP-based seed extension adopts in the changing environment by using the learning process, however, Level-set-based seed extension doesn't adopt in the changing environment because it uses only gradient values of image for extension. It causes the under/over extension related to image resolution.

ii) Classification time cost: we will display the time elapsed during classification. This elapsed time is obtained from relative time. It means that we obtained the time under the same testing environment. We used a Pentium-IV 2.1Ghz CPU, 256 Mbyte memory, and 4 Mbyte graphic memory. The results are presented in Table 1.

The results displayed in Figure 2 and in Table 1 come from a highway image sequence of about 3000 images. The chosen images consist of both high-resolution and low-resolution images. In each case, the image includes the high illumination, the shadow, and many clusters.

Table 1. The quantitative analysis of the classification ability and the relative time cost of the classification. (256x256 size, 3000 image).

Comparison Item	Used classes	K-Mean	Fuzzy K-mean	Configuous K-Mean	Proposed Method
Error (%) (average)	2 classes	44.82	43.51	43.3	**13.6**
	4 classes	27.96	25.21	23.5	No necessity
Time cost (average)	2 classes	55.7 ms	500 ms	125 ms	**31 ms**
	4 classes	125 ms	2104 ms	250 ms	No necessity



(a) The classification accuracy comparison error rate (%)

(b) The elapsed time comparison for classification (ms)

Fig. 2. The statistical comparison in accordance with Table 1.

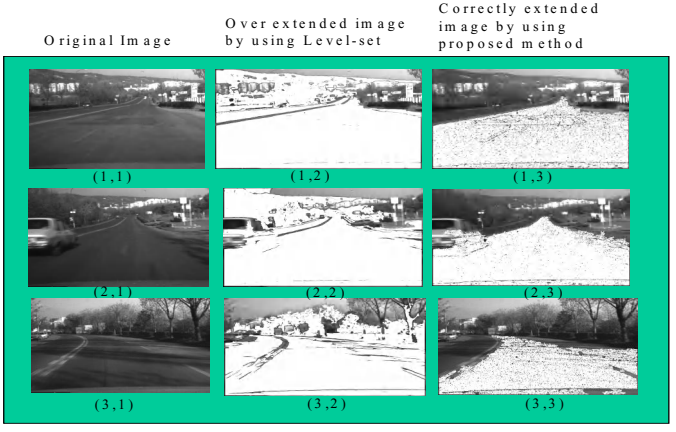


Fig. 3. The seed extension ability comparison in the case of the changing environment between the Level set and the proposed method (the white region is extended region)

Following the results from Figure 2 and Table 1, the proposed LDP-based classification is more efficient than the generic classification method in both time-cost and classification robustness point of view. The error rate includes the under classified classification and the over classified classification. In four-class classification of the generic method, we select the classes belonging to the road, manually, in order to compute the error rate. However, in the real application, the automatic selection of the road class is very difficult for the four-class classification. Furthermore, the classification of the proposed method doesn't depend on the image resolution. However the generic classification method produces higher error rates in the low-resolution image than in the high-resolution image. The only controversial thing of the proposed method is the selection of the well-established road sample area. However, the suggested method in this paper is reasonable in the case of the autonomous driving appli-

cation. We only present the classification result in the highway images. But it can be applied in another kinds of possible moving areas: rural roads and desert environment. GMRF and the Gabor filter are mentioned in the introduction. However we can't consider a time comparison because they aren't real-time methods, obviously. But they are useful if we want to extend the features in order to obtain the dedicated classification results. The contiguity K-Mean, sometimes, fails to classify the features in high-illumination condition and in the low-resolution image.

4 Conclusion

We proposed an efficient classification method for the autonomous driving application compared to generic classification methods in the aspects of the classification ability and of the relative classification time cost. We obtained much improved classification results and time-cost then in the generic classification method.

References

1. T.R.Reed, H. Wechsler: Segmentation of textured images and Gestalt organization using spatial/spatial-frequency representations. IEEE Trans. Pattern analysis and Machine Intelligent, vol. 12. pp. 1-12. Jan (1990)
2. C. Nikias: High Order Spectral Analysis. Advances in Spectrum Analysis and Array Processing. S.Haykin, Ed., pp 326-365. Prentice Hall, Englewood Cliffs, NJ (1991)
3. O. Rioul and M. Vetterli: Wavelet and signal processing. IEEE SP mag., pp 14-38, oct (1991)
4. G. Strang: Wavelet and dilation equation: a brief introduction, SIAMRev, vol. 31, pp. 614-627, Dec (1989)
5. L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg: Face recognition by elastic graph matching. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, July (1997)
6. R. O. Duda, Peter E. Hart, David G. Stork: Pattern classification, (2001)
7. S. Krishnamachari, and R. Chellappa: Multiresolution Gauss-Markov Random Field Models for Texture Segmentation. IEEE Trans. Image Processing, vol. 6, no. 2, Feb (1997)
8. J. Theiler and G. Gisler: A contiguity-enhanced K-Means clustering algorithm for unsupervised multispectral image segmentaion. Processing SPIE, vol. 3159, pp 108-118, (1997)
9. P Jeong, S. Nedevschi: Intelligent Road Detection Based on Local Averaging Classifier in Real-Time Environments. 12th IEEE International Conference on Image Analysis and Processing, pp. 245 – 249, Mantova, 17-19 September, (2003)
10. P. Jeong, S. Nedevschi: Unsupervised Muliti-classification for Lane detection using the combination of Color-Texture and Gray-Texture: vol 1, pp 216-221, CCCT 2003, August (2003)
11. Nikolaos PARAGIOS and Rachid DERICHE, "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects," *IEEE Transactions on pattern analysis and machine intelligent*, vol. 22, no 3. March 2000.

Parameterized Hierarchical Annealing for Scientific Models

Simon K. Alexander[1]*, Paul Fieguth[2], and Edward R. Vrscay[1]

[1] Department of Applied Mathematics
University of Waterloo, Waterloo, Ontario, Canada, N2L-3G1
{sk2alexa,ervrscay}@uwaterloo.ca
[2] Department of Systems Design Engineering
University of Waterloo, Waterloo, Ontario, Canada, N2L-3G1
pfieguth@uwaterloo.ca

Abstract. The accurate synthesis of binary porous media is a difficult problem. Initial applications of simulated annealing in this context with small data sets and simple energy functions have met with limited success. Simulated annealing has been applied to a wide variety of problems in image processing. Particularly in scientific applications such as discussed here, the computational complexity of this approach may constrain its effectiveness; complex, non-local models on large 2D and 3D domains may be desired, but do not lend themselves to traditional simulated annealing due to computational cost. These considerations naturally lead to a wish for hierarchical/multiscale methods. However, existing methods are few and limited. In this paper a method of hierarchical simulated annealing is discussed, and a simple parameterization proposed to address the problem of moving through the hierarchy. This approach shows significant gains in convergence and computational complexity when compared to the simulated annealing algorithm.

1 Introduction

We are interested in the problem of computational practicality of simulated annealing in large phase spaces. As is often the case, a constrained problem domain allows concentration on particular issues of interest. Hence, we choose as a motivational application the synthesis of binary porous media images. Figure 1 gives two examples, binary images representing density (white) and pore structures in a physical media (hence the name). Images such as these are important in the study of porous media [11]. The above examples, however, are physically imaged; researchers in the area are interested in ways to accurately synthesize such data. Resulting data sets can be used to perform many useful calculations [10].

The simulated annealing algorithm has been used to perform this sort of synthesis with some success [11]. Ultimately the computational complexity has limited the viability of this approach. Simulated annealing [6] has been used successfully in many

* This research was supported in part by the Natural Sciences and Engineering Council of Canada (NSERC), in the form of grants (P.F. and E.R.V.) and a Postgraduate Scholarship (S.K.A.). We also wish to thank M. Ioannidis, Dept. of Chemical Engineering University of Waterloo for providing sample porous media images and interesting discussions.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 236–243, 2004.
© Springer-Verlag Berlin Heidelberg 2004

(a) (b)

Fig. 1. Examples of binary porous media a) sintered glass spheres, b) sandstone

imaging applications, particularly in estimation problems. However, the computational cost of slow annealing schedules has limited its applicability, especially where large phase spaces are involved. In this work, we are interested in extending the discussion of *hierarchical* methods, primarily as a way to reduce the computational complexity.

Simulated annealing is controlled by a *cooling schedule* — a decreasing (although perhaps not strictly so) temperature parameter which affects the likelihood of non-energetically-favourable events. Theoretical results only exist for impractically slow cooling, or special cases [5]. Our work is motivated by a wish to leverage multiscale characteristics of a model to reduce the amount of computation needed.

The general approach of hierarchical annealing [3, 1, 2] is inspired in part, by renormalization group approaches in Markov random fields (e.g., [7]). A key realization is that while the scaling behavior of local-interaction models may be very difficult to analyze (hence renormalization difficult), other models may be proposed based on non-local quantities which are inherently renormalizable.

After briefly describing annealing approaches and how this example problem fits in, hierarchical annealing will be outlined. In this paper, we concentrate on the issue of parameterizing the hierarchical approach, and discuss the application to a particular model with comparison to 'flat' annealing methods. Empirical results are presented, along with a discussion of generalizations and future work.

2 Simulated Annealing and Hierarchical Approaches

Since its introduction to image processing by Geman & Geman [6], simulated annealing has been used for a large range of applications to image estimation and synthesis. The flexibility of this model is its main strength. If a problem can be stated in terms of minimum energy states of a Gibbs type density (1), or equivalently a Markov random field, simulated annealing will give a correct solution *given a long enough cooling schedule*. The usual way of describing such a density is

$$\pi_\beta(x) = \frac{e^{-\beta \mathcal{E}(x)}}{Z_\beta} \quad , \tag{1}$$

where $\beta = 1/T$ is the inverse temperature parameter. The partition function Z_β is not evaluated in simulated annealing.

However, the primary drawback is that while in theory this approach will work, this result relies on an intractably slow *logarithmic cooling schedule* in temperature T : $T_{n+1} \sim 1/\log T_n$.

Even computationally tractable cooling schedules (which are not proven to converge, in general) are expensive — especially for complex models and large domains. The majority of annealing implementations use a single ('flat') lattice at a particular resolution, annealed by brute force. For many interesting applications, this approach is simply too computationally expensive to be practical.

The fundamental reason for the slow convergence of such models is that local MRF models are inherently subject to a phenomena known as *critical slowing down*. Essentially what happens is the following: since all structure in the model is created by local interactions, all longer scale (i.e., non-local) structure *must* be created by many local interactions. Recall that we are obeying a *cooling schedule* which reduces the 'temperature' of the simulation over time. At *phase-transition* points (i.e. *critical temperatures*), non-local structure will appear. Once this temperature is reached, significant changes in the energy become more expensive, since the non-local structure must change through the cumulative effect of many local interactions. In order to address the cost of sampling such a model, attacking this critical slowing down is vital.

The method we are describing here, *hierarchical annealing* is an approach to do just this. Hierarchical approaches have been proposed in the literature. Primarily, they are of two types:

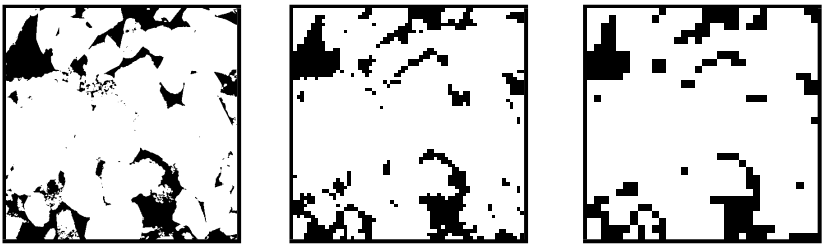
1. Hierarchical estimation (assuming a dense first scale), and
2. Region-based sampling methods.

Methods for region-based sampling have been applied successfully to the computation of local models in very particular cases – for example clustering methods for Ising or Ising-like models at or near the critical temperature. (Here Ising exhibits the interesting phase-change behaviour; studying this and the structures created near this temperature has driven much of the interest in these sorts of methods). Our models are in general quite different than this class, especially in that they exhibit complex structure at very low temperatures or, equivalently, for long temperature scales as we anneal.

There are several authors who have discussed hierarchical approaches to the related problem of accelerating image estimation. In particular, a 'label' pyramid may be built above the (full resolution, finest scale) image [9, 4, 8]. This is, however, quite different than our situation. Given a dense image at finest scale, this estimation problem tends to be well-conditioned, and converges well with very little information from coarser scales. In fact, often the coarse scale state is not annealed at all.

2.1 Hierarchical Annealing

The question then is: At any given level in this multiscale hierarchy, what image features are represented? In particular, we wish to work 'down' a hierarchy, from coarsest to finest resolution. How may we anneal in such a way that features are represented at the current level, and can be meaningfully projected to the the next finest level? Figure 2 illustrates this point.



(a) 512×512 (b) 128×128 (c) 64×64

Fig. 2. A porous media image viewed at several resolutions: How do local and non-local features scale?

Our approach to hierarchical annealing [1] is outlined as follows. Consider a hierarchy $\{X_n\}_{n=0}^{M}$ of coarse-grainings of the configuration space where each increase in level represents decimation by a factor of two (X_0 is finest resolution). At each higher level in the hierarchy, the energy function for that level is $\mathcal{E}^s$. We denote projection (coarse to fine) from level s to level $s-1$ as P_{s-1}. Annealing is performed as shown in Algorithm 1.

Algorithm 1 Hierarchical Annealing

```
k ⇐ 0
for scale s from coarsest to finest do
  while E^s(X_s) not converged do
    β ⇐ 1/T_k
    X_s ⇐ apply Gibbs sampler to X_s
    k ⇐ k + 1
  end while
  X_{s-1} ⇐ P_{s-1}(X_s)        {map to next finer resolution}
end for
```

There are two sources of computational benefit in this approach. First, the size of the coarse domains is small, allowing rapid iterations of the sampler. Second, as suggested in the previous section, at an intermediate scale the algorithm needs to iterate only long enough to allow relatively local structure to converge, since the larger structures converged at coarser scales. So we work in a decimated configuration space (at less computational cost) until some appropriate condition is reached, and then project onto the next larger space and continue annealing.

Annealing with too low an initial temperature (not enough energy) approaches a greedy algorithm, that is, it will be prone to getting stuck in local minima. On the other hand, too high an energy may destroy larger structures (all such structure may be taken apart if you anneal for long enough). Since this holds true at any level in the hierarchy, clearly there is a delicate balance in achieving computational gains while retaining good optimization performance (i.e., low temperatures).

Analysis of the interactions between the annealing (cooling) schedule, the scale-to-scale projection, and the convergence of the stochastic sampler is not straightforward. In this work, we present a discussion of a simple parameterization of the process, which is in no way presumed to be optimal. Empirical results will show this approach to be beneficial when applied to our example. Further work will refine these methods; from the preceding it should be clear that there are several related avenues for improvement.

2.2 Parameterization for H-A

Due to the aforementioned difficulties, choosing a 'best' way to parameterize the hierarchical annealing is not straightforward. In this work, we have explored a simple parameterization and present the benefits when compared to 'flat' annealing.

Parameterization is made more difficult by the fact that it is difficult to distinguish the effect of varying the cooling schedule and the number of iterations of the system. To simplify things, in the following we have taken a fixed, *geometric* cooling schedule:

$$T_{n+1} = \alpha^n T_0, \qquad 0 < \alpha < 1 \ . \tag{2}$$

In this approach, the process is broken into three steps. There is an initialization step, which is flat annealing, at the coarsest scale, from an initial temperature t_{init} to some t_0 according to a particular cooling schedule. At this point, the second step, begins: The hierarchical stepping of the system is begun. This process is controlled by two parameters θ_0, the number of iterations per step, and θ_1, the 'iteration offset' upon projection. For each scale, until the final, high-resolution scale is reached, the system is iterated under the given cooling schedule for θ_0 steps. At this point, the system is projected to the next highest scale, but the temperature is adjusted by θ_1 *steps* (i.e. as if n has changed in 2). The process is repeated until the final scale is reached. Finally, the system may be iterated at the finest scale until it has converged. Algorithm 2 describes this process:

Algorithm 2 Parametrized Hierarchical Annealing

```
{Initialize system at coarsest scale}
{Anneal until temperature t_0 is reached}
for scale s from coarsest to finest do
  for θ_0 iterations do
    β ⇐ 1/t
    X_s ⇐ apply Gibbs sampler to X_s
    t ⇐ α^n T_0
    n ⇐ n + 1
  end for
  X_{s-1} ⇐ P_{s-1}(X_s)        {map to next scale}
  n ⇐ n + θ_1
end for
{Final iterations done at highest resolution scale}
```

2.3 Models/Energy Functions

The statistical sampling methods discussed here are characterized in terms of an energy function $\mathcal{E}$ in (1) . When considering a hierarchical approach, the key problem is how to define it at coarser scales. We are considering binary dense/pore structures [12] in an image. We can denote $I^s(x)$ for the index function for our binary image at scale s, (yielding 0 for pore, and 1 for density). If we let $< \cdot >$ denote a spatial average over the image, then the average density ϕ (or 'one-point' correlation can be denoted as

$$S^s(r) = \ < I^s(x+r) > \ = \phi^s \ . \tag{3}$$

Similarly, two-point correlation:

$$S^s(r_1, r_2) = \ < I^s(x+r_1) I^s(x+r_2) > \ . \tag{4}$$

Quantities of particular interest in the study of porous media are the above mentioned correlation functions, and the distribution of chordlengths [12]. Notationally, let us take $\hat{\cdot}$ to denote trained/target values. Furthermore, denote the lattice size of the image at a particular scale s as $\mathcal{O}(s)$ (then for initial lattice of $N \times N$ $\mathcal{O}(s) = N/2^s$). When considering chordlength distributions, if we restrict ourselves to the horizontal and vertical directions again, this is essentially the distribution of length of contiguous 'runs' of density pixels in these directions. Denoting these probability mass functions as p_C^h and p_C^v for the horizontal and vertical directions, respectively, we may construct an energy function at scale s:

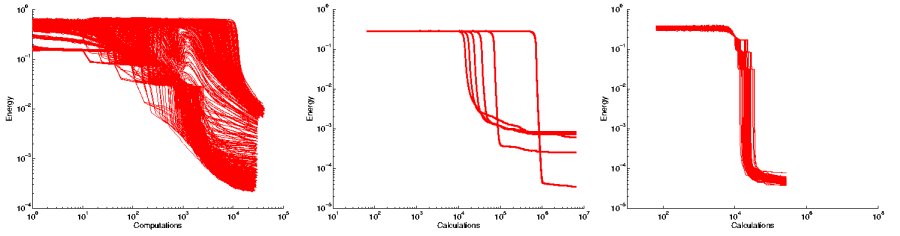
$$\mathcal{E}^s = \sum_{n=1}^{\mathcal{O}(s)} \|\hat{p}_C^h(n) - p_C^h(n)\| + \|\hat{p}_C^v(n) - p_C^v(n)\| \ . \tag{5}$$

Here the sample pmf's are estimated by histograms from the image data. Taken alone, this puts no constraint on the *amount* of pore or density in the image. Hence a constraint based on (3) must be added; either explicitly in the energy function, or implicitly by using a sampling algorithm that conserves density by exchanging pixels rather than flipping single sites. Other energy functions are of interest for this application [3], in particular two-point correlation (4) [12]. However, in the results following this section, we concentrate on chordlength distribution to construct energy functions like (5).

3 Results and Conclusions

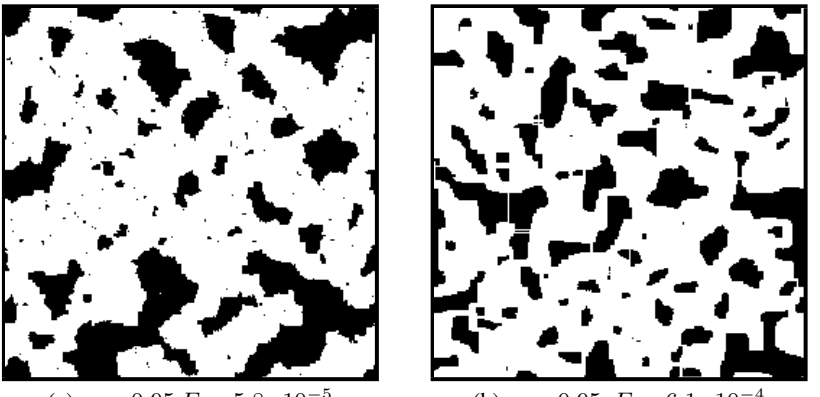
In order to demonstrate the efficacy of the ideas introduced in this work, we present some empirical results for a particular model, comparing the behaviour of 'flat' and (parameterized) 'hierarchical' annealing methods. The energy function used here is based on (5) with l_2 distance between histograms for both horizontal and vertical chordlength distributions (i.e., mean squared error). The chordlength distribution is important in the study of porous media, and has been used in annealing methods in the porous media literature [12, 11].

This model demonstrates how critical slowing down affects the flat annealing process. Figure 3 shows three plots: 3(a) gives energy vs. calculations results for a large number



(a) (b) (c)

Fig. 3. Flat and hierarchical sampling. Energy vs. Computations for (a) many hierarchical parameters, (b) flat annealing for several values of α, and (c) several 'good' hierarchical parameters for comparison to flat annealing results. Note similar low-temperatures are reached in (c) with much less computation than (b).



(a) $\alpha = 0.95, E = 5.8 \cdot 10^{-5}$ (b) $\alpha = 0.95$, $E = 6.1 \cdot 10^{-4}$

Fig. 4. Synthetic images from hierarchical (a), and flat (b) annealing. Energy E, and coefficient of geometric cooling (2), α, is given. Note the clear morphological differences between the right (flat) and left (hierarchical) images. Blockiness and line segments on the right are due to annealing too quickly.

of parameterizations of a chordlength model (without periodic boundaries) demonstrating a range of results; unsurprisingly a poor choice of parameterization leads to poor convergence. Figures 3(b) and 3(c) show for a similar model (with periodic boundaries) collection of good parameterizations (low final energy) to a number of flat annealing sessions with different cooling schedule parameters α,respectively.

Note that all energies are calculated at the finest scale, and results are shown on a log-log plot. Hierarchical annealing was done with the cooling parameter $\alpha = 0.95$. For flat annealing, α ranged from 0.9 to 0.999. Computations are normalized to the cost of a full 'sweep' of the Gibbs sampler on a 32×32 image. The longest flat annealing run takes approximately 2 days on a 3 Ghz. Pentium 4 class machine.

These plots show that hierarchical annealing may reach much lower temperatures for the same computational cost. Conversely, we may reach the same temperatures with far fewer computations. It should be noted that none of the above results are particularly good, morphologically speaking, when compared to the class of images they are trying

to reconstruct. Training images were taken from a set of glass sphere images (see Fig 1-a). However, Figure 4 shows that the lower energy images have significantly different characteristics from those resulting from 'flat' annealing. The problem of critical slowing down is illustrated in the right images (flat annealing). In particular, note that breaking a 'bad' long chord into two smaller chords need not be energetically favourable, hence such chord can be difficult to get rid of. Flat annealing images (right hand side) exhibit thin vertical and horizontal chords which are not seen in the (lower energy) hierarchical results.

In conclusion, we have discussed the need for hierarchical or multiscale methods in annealing, and related our proposed method to the literature. We have demonstrated the efficacy of a simple parameterized approach to hierarchical annealing when compared to standard, 'flat' annealing. Our method is shown to exhibit significant improvements to a practical application. It is, however, not in any way considered optimal and further analysis of the process is needed.

References

1. S. K. Alexander, P. Fieguth, and E. R. Vrscay, *Hierarchical annealing for random image synthesis*, EMMCVPR'03, LNCS, no. 2683, Springer, 2003.
2. _____, *Image sampling by hierarchical annealing*, ICIP'03, IEEE, 2003.
3. _____, *Hierarchical annealing for scientific models*, ICASSP'04, IEEE, 2004.
4. C. Bouman and M. Shapiro, *A multiscale random field model for Bayesian image segmentation*, IEEE Image Processing **3** (1994), no. 2, 162–177.
5. Pierre Brémaud, *Markov chains: Gibbs fields, monte carlo simulation, and queues*, Springer, 1998.
6. S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **6** (1984), 721–741.
7. Basilis Gidas, *A renormalization group approach to image processing problems*, IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989), no. 2, 164–180.
8. T. Hofmann, J. Puzicha, and J.M. Buhmann, *Unsupervised texture segmentation in a deterministic annealing framework*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **20** (1998), no. 8, 803–818.
9. Zoltan Kato, Marc Berthod, and Josiane Zerubia, *A hierarchical Markov random field model and multitemperature annealing for parallel image classification*, Graphical Models and Image Processing **58** (1996), no. 1, 18–37.
10. Z. Liang, M.A. Ioannidis, and I Chatzis, *Geometric and topological analysis of three-dimensional porous media: Pore space partitioning based on morphological skeletonization*, Journal of Colloid and Interface Science **221** (2000), 13–24.
11. _____, *Reconstruction of 3d porous media using simulated annealing*, Computational Methods in Water Resources XIII (Balkema, Rotterdam) (Bentley et al., ed.), 2000.
12. M. S. Talukdar, O. Torsaeter, and M. A. Ioannidis, *Stochastic recontruction of particulate media from two-dimensional images*, Journal of Colloid and Interface Science **248** (2002), 419–428.

Significance Test for Feature Subset Selection on Image Recognition

Qianren Xu[1], M. Kamel[1], and M.M.A. Salama[2]

[1] Dept. of System Design Engineering, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada
q3xu@engmail.uwaterloo.ca, mkamel@pami.uwaterloo.ca
[2] Dept. of Electrical and Computer Engineering, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada
m.salama@ece.uwaterloo.ca

Abstract. This paper proposes a novel feature selection method based on significance test (ST). Statistical significant difference between (or among) classes, such as t statistic in Student test and F statistic in ANOVA, is utilized to measure pattern recognition ability of individual features. The feature significance level during a feature selecting procedure is used as feature selection criterion, which is determined by the product of the significant difference level and the independent coefficient of the candidate feature. An algorithm of maximum significant difference and independence (MSDI) and strategies of monotonically increasing curve (MIC) are proposed to sequentially rank the feature significance and determine the feature subset with minimum feature number and maximum recognition rate. Very good performances have been obtained when applying this method on handwritten digital recognition data.

Keywords. Feature selection, maximum significant difference and independence (MSDI), significance test, t-test, ANOVA, correlation

1 Introduction

Feature selection is the process of determining relevant features from the original feature set for a pattern recognition problem. One way to select the optimal features from given initial feature subset is to exhaustively evaluate the recognition rate for all possible combinations of features. But, in the case of large number of features, such as some image recognition that may have hundreds of features, this exhaustive search strategy is impractical because of its time consuming. In order to avoid the exhaustive search, intermediate results or feature selection criteria have been introduced for this purpose. For instance, features can be selected based on the mutual information criterion [1], or by using fuzzy region analysis [2]. However, it has been showed that no non-exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset [3]. This paper proposes a new feature selection method of the non-exhaustive sequential selection type in order to obtain better pattern recognition performance.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 244–252, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In the following section, the measurements related with the criteria of feature selection are presented in Section 2.1 and 2.2, and selection procedures based on the significance level of features are proposed in Section 2.3 and 2.4. Experimental results and a comparison with the mutual information feature selection method are discussed in Section 3.

2 Methodology

The significance test (ST) method for feature selection involves a sequential selection procedure of maximum significant difference and independence (MSDI). MSDI first selects the feature with the most significant difference between (or among) classes as the first feature, and then every new feature is selected by maximizing the product of individual pattern separability and independence coefficient between the candidate and already selected feature. After the procedure of MSDI, monotonically increasing curve strategy (MIC) can be used to delete useless features in the feature subset selected by MSDI.

2.1 Measurement of Pattern Separability on Individual Features

The significance of difference between (or among) classes is used to estimate the ability of pattern recognition of a feature. Various statistical significance tests can be used to determine the level of significant difference.

Significant Difference Between Two Classes

It is straightforward that the significance of the difference (*sd*) between classes is in proportion to the difference of their means and in inverse proportion to the variance of their distribution,

$$sd = \frac{difference\ \ between\ \ two\ \ classes}{distribution\ \ variance}. \tag{1}$$

Assuming data set has two classes $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ with the number *m* of features, and $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ have $n_{(1)}$ and $n_{(2)}$ instances respectively,

$$\mathbf{x}_{(1)} = \begin{bmatrix} x_{(1)11} & x_{(1)12} & \cdots & x_{(1)1n_{(1)}} \\ x_{(1)21} & x_{(1)22} & \cdots & x_{(1)2n_{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(1)m1} & x_{(1)m2} & \cdots & x_{(1)mn_{(1)}} \end{bmatrix}, \mathbf{x}_{(2)} = \begin{bmatrix} x_{(2)11} & x_{(2)12} & \cdots & x_{(2)1n_{(2)}} \\ x_{(2)21} & x_{(2)22} & \cdots & x_{(2)2n_{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(2)m1} & x_{(2)m2} & \cdots & x_{(2)mn_{(2)}} \end{bmatrix}. \tag{2}$$

If the two classes are distributed normally, the significance of the difference between the two classes on feature *i* can be statistically estimated by the t statistic of the Student test [4]

$$t_i = |\bar{x}_{(1)i} - \bar{x}_{(2)i}| \Big/ \sqrt{s^2_{(1)i}/n_{(1)} + s^2_{(2)i}/n_{(2)}}, \quad i = 1, 2, \ldots, m, \tag{3}$$

where s is the standard deviation used to measure the variance of the distribution

$$s_i = \sqrt{\left[\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2\right] \Big/ (n-1)}, \quad i = 1, 2, \ldots, m. \tag{4}$$

Significant Difference Among More than Two Classes

When there are more than two classes, the overall significant difference can be estimated by the variance among these classes and variance within classes,

$$sd = \frac{variance\ \ among\ \ classes}{variance\ \ within\ \ classes}. \tag{5}$$

We assume that there are c classes $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(c)}$ with the number m of features. F statistic in ANOVA (analysis of variance) [4] can be used to measure the significant difference (5) for the all classes on feature i,

$$F_i = ms^2_{ai} / s^2_{wi}, \quad i = 1, 2, \ldots, m, \tag{6}$$

where s_{ai} and s_{wi} are the standard deviations used to measure the variance of the distribution among classes and within classes respectively

$$s_{ai} = \sqrt{\left[\sum_{k=1}^{c} (\bar{x}_{(k)i} - \bar{x}_i)^2\right] \Big/ (c-1)}, \quad s_{wi} = \sqrt{\left[\sum_{k=1}^{c}\sum_{j=1}^{m} (x_{(k)ij} - \bar{x}_{(k)i})^2\right] \Big/ c(m-1)}. \tag{7}$$

2.2 Measurement of Independence Between (or Among) Features

We define independence between features $\mathbf{x}_u$ and $\mathbf{x}_v$ as

$$ind_{uv} = \sqrt{1 - r^2_{uv}}, \tag{8}$$

where r_{uv} is the Pearson correlation coefficient [4],

$$r_{uv} = \left[\sum_{j=1}^{n} (x_{uj} - \bar{x}_u)(x_{vj} - \bar{x}_v)\right] \Big/ \sqrt{\sum_{j=1}^{n} (x_{uj} - \bar{x}_v)^2 \sum_{j=1}^{n} (x_{uj} - \bar{x}_v)^2}. \tag{9}$$

The independence between multiple features $\mathbf{x}_{u1}, \mathbf{x}_{u2}, \ldots \mathbf{x}_{up}$ and $\mathbf{x}_v$ can be estimated by the independence between multiple regression $\mathbf{x}_{\hat{v}}$ and $\mathbf{x}_v$

$$ind_{u1u2\ldots up \cdot v} = \sqrt{1 - \left[\sum_{j=1}^{n} (x_{\hat{v}j} - \bar{x}_{\hat{v}})(x_{vj} - \bar{x}_v)\right]^2 \Big/ \left[\sum_{j=1}^{n} (x_{\hat{v}j} - \bar{x}_{\hat{v}})^2 \sum_{j=1}^{n} (x_{vj} - \bar{x}_v)^2\right]}, \tag{10}$$

where $u1, u2, ..., up$ represent the already-selected features $x_{u1}, x_{u2}, ..., x_{up}$; $x_{\hat{v}}$ is the regression predicted feature ($x_{\hat{v}} = a + b_1 x_{u1} + b_2 x_{u2} + ... + b_1 x_{up}$); x_v is the candidate feature.

The independence above is estimated within class, and the overall independence on the all classes can be estimated by the mean of these independence levels.

$$ind_{uv} = \frac{1}{c}\sum_{k=1}^{c} ind_{(k)uv} \cdot \tag{11}$$

2.3 Measurement of Feature Significance

We defined the feature significance as the contribution of a candidate on improving pattern recognition when the candidate feature is added to feature subset. The measurement of feature significance can be reasonably estimated by the product of significant difference between (or among) classes and the independence level between the candidate feature and the already-selected feature(s),

$$sf = sd \times ind \ . \tag{12}$$

The significant difference (*sd*) represents the pattern separability of individual features, which can be estimated by t statistic (t-test) or F statistic (ANOVA). The independence level *ind* ($0 \le ind \le 1$) can be viewed as the weight of feature significance in a specific procedure. When $ind = 1$ (the feature is totally not correlative with the already-selected features), the feature significance will be equal to the significance difference value ($sf = t$ or $sf = F$). When $ind = 0$ (the feature can be obtained from the linear combination of the already-selected features), the feature significance will be set to 0 ($sf = 0$), as this kind of feature is useless for improving pattern recognition even when its calculated significance of difference might be high.

2.4 Maximum Significant Difference and Independence Algorithm

Maximum significant difference and independence (MSDI) is a forward stepwise procedure; it starts from the first feature with the most significant difference in the original feature set, and then adds the most significant feature at a time. The selection procedure of the MSDI algorithm is described as follows:

1) Compute the significant difference (*sd*) between (or among) classes on all original features;
2) Choose the first feature with the maximum *sd*

$$1st_feature \Leftarrow \max(sd); \tag{13}$$

3) Compute the independence (*ind*) between every candidate feature and the already-selected feature;

4) Compute the level of feature significance ($sf = sd \times ind$) on every candidate feature, then select the feature with the maximum *sf* value as the second feature;

$$2nd_feature \Leftarrow \max(sf); \tag{14}$$

5) If the already-selected feature subset reaches the maximum number of features, terminate the procedure, otherwise go to step 3.

2.5 Monotonically Increasing Curve Strategy

After the original features are ranked by MSDI, a performance curve can be drawn by the recognition rate vs the size of the feature subset. Generally, the curve tends to increase (in overall trend) at first, after reaching the maximum, and then it tends to decrease. The curve can be divided into two parts by the maximum point. Ideally, the curve monotonically increases in the first part and decreases in the second part. However, the selected feature subset (the first part of curve) usually still has a few "bad" or "not good" features, and the curve is not perfectly monotonic. In order to reach maximum pattern recognition rate with the minimum size of feature subset, we can delete the features that have "no good" contribution to the recognition performance, and make the curve monotonic.

One strategy to make the curve monotonically increasing is to sequentially delete the "not good" features as follows:

1) Delete the leftmost feature that has "no good" contribution to the increasing of recognition rate, the index of the deleting feature (i_{dl}) is determined by

$$i_{dl} = \min(i_d),\ \ i_d \in \{i \mid f(i) \le f(i-1)\}, \tag{15}$$

where *i* is the rank of feature significance in already-selected feature subset, i_d is the index set of "not good" feature, *f(i)* is the rate of recognition;

2) Plot the performance curve again with the new feature subset;
3) Repeat the procedure until the part of the curve before the maximum point is strictly monotonically increasing.

Another more computationally effective is as follows:

1) Delete all of the features that has "no good" contribution to the increasing of recognition rate, the index set (i_d) of deleting features is determined by

$$i_d \in \{i \mid f(i) \le f(i-1)\}, \tag{16}$$

where *i* is the rank of feature significance in already-selected feature subset, *f(i)* is the rate of recognition;

2) Plot the performance curve again with the new feature subset;
3) Repeat the procedure until the part of curve before the maximum point is strictly monotonically increasing.

3 Experimental Results

We have used the dataset of handwritten numeral recognition from UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html) It consists of 649 features on handwritten numerals (`0'--`9'). These 649 features distribute over the following feature sets: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages in 2 x 3 windows, 47 Zernike moments, 6 morphological features. There are 200 patterns per class (for a total of 2,000 patterns) in the data set, and we randomly use half as training set and remaining data as testing set on every class. We used a probabilistic neural network (PNN) [5] as a classifier.

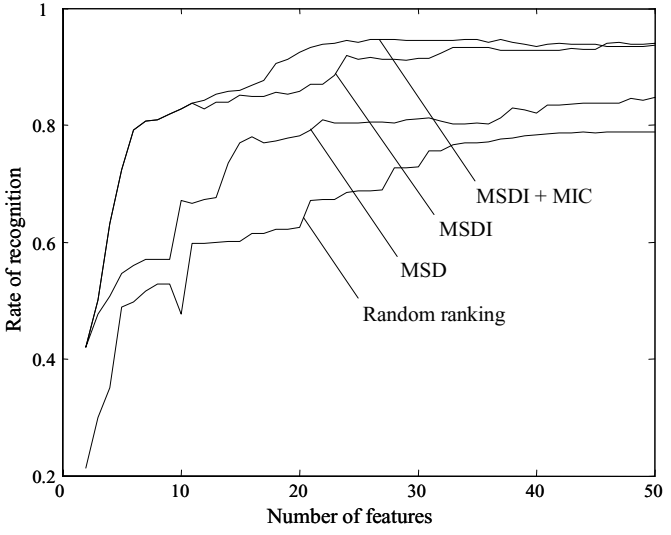


Fig. 1. The performance curve of pattern recognizing the handwritten numerals by maximizing significant difference (*MSD*), maximizing significant difference and independence (*MSDI*), and monotonically increasing curve (*MIC*) strategy

Fig. 1. shows the 50 relation between the rate of recognition and the number of features used. Only first features are plotted in the figure, and the performance of PNN for every number of features is on the optimal condition. We can see that the recognition performance of maximizing significant difference (MSD) is much better than that of random ranking, because MSD selected the feature subset with maximum individual pattern seperability. The performance of maximizing significant difference and independence (MSDI) is further improved comparing that of maximizing significant difference (MSD). MSDI considers both of the pattern separability of individual features and the correlation between the candidate feature and the already-selected features, which discards redundant features and thus reaches better rate of recognition with less number of features than MSD does. Fig. 1 also shows the results after the

monotonically increasing curve (MIC) strategy is applied on the feature subset selected by MSDI, which makes the performance curve monotonic and obtains higher rate of recognition with less number of features than MSDI only.

The mutual information (MI) feature selection method [1] is applied in the same data set for comparison,

$$I_v = I_v(y_c; x_v) - \beta \times \sum_{u=1}^{p} I_v(x_u; x_v), \tag{17}$$

where β regulates the relative importance of the MI between the candidate feature and the already-selected features with respect to the MI between the feature and the output class. β is recommended between 0.5 and 1.0 [1]. However, by searching β between all range of 0 to 1, we find out that the optimal value is on β = 0.2 and the performance on β = 0.5 (within the range 0.5 to l.0) is only suboptimal. Fig. 2. shows that our MSDI is much better than MI in this handwritten numeral recognition problem on two aspects: a) MSDI (before the MIC procedure) obtains better recognition than MI does in the range of large feature number (feature number >5); b) MSDI is more computationally effective than MI method, one reason is that the MI need to search , which is time consuming since it involves classifier learning.

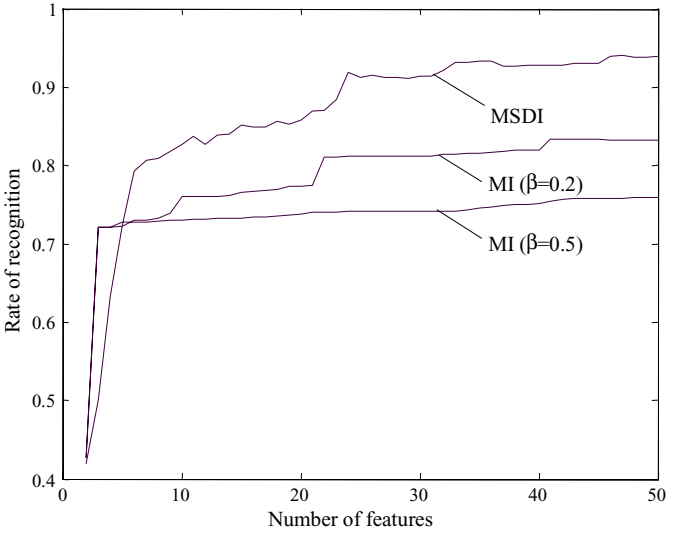


Fig. 2. The performance curve of pattern recognizing the handwritten numeral by maximizing significant difference and independence (*MSDI*) and mutual information (*MI*)

We have also used another handwritten digital dataset also from UCI Machine Learning Repository. 32x32 bitmaps are divided into non-overlapping blocks of 4x4 size, the number of pixels is counted in each block in the range 0 to16, and it generates 8x8 matrix (that is 64 features). Training set in the database has 3823 patterns totally. We randomly use half of every class in the training set as actually training set

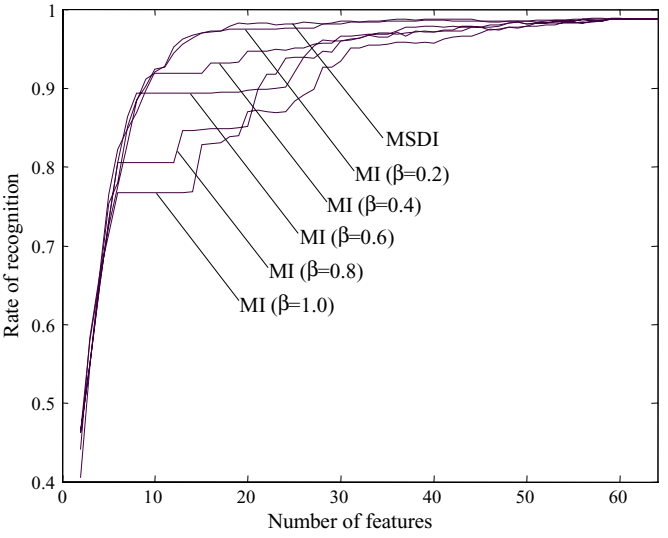


Fig. 3. The performance curve of pattern recognizing the optical handwritten digits by maximizing significant difference and independence (*MSDI*) and mutual information (*MI*)

and remaining data as testing set. Fig. 3. shows that the curves are almost monotonic increase in all range of number of features, which means that almost all the feature positively contribute to pattern recognition. Although all features are "good" features, they have different significance to contribute on pattern recognition procedure. Fig. 3. also shows the results of MI at different β values, and the optimal β is between 0.1 ~ 0.2. The performance of MI at the optimal condition is close to that of MSDI on this example.

4 Conclusion

The paper shows that the significance test (ST) feature selection has very good performance in the image field (the handwritten digital recognition), and MSDI (even before the MIC procedure) can obtain better performance with more effective computation than MI. In fact, the ST can be applied to any supervised classification, and the significant difference can utilize different statistical model in accordance with the properties of the original data set.

References

1. Battiti, R.: "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Network. vol 5, (1994) 537-550

2. Thawonmas, R., Abe, S.: "A novel approach to feature selection based on analysis of fuzzy regions," IEEE Trans. Systems, Man, and Cybernetics Part B, vol 27, (1997) 196-207
3. Cover, T.M., Van Campenhout, J.M.: "On the Possible Orderings in the Measurement Selection Problem," IEEE Trans. Systems, Man, and Cybernetics, vol. 7, (1977) 657-661
4. Wonnacott, R.J., Wonnacott, T.H.: Introductory Statistics. 4th ed. John Wiley & Sons (1996)
5. Wasserman, P.D.: Advanced Methods in Neural Computing, Van Nostrand Reinhold, New York (1993) 35-55

Image Recognition Applied to Robot Control Using Fuzzy Modeling

Paulo J. Sequeira Gonçalves[1], L.F. Mendonça[2], J.M.C. Sousa[2], and J.R. Caldas Pinto[2]

[1] Instituto Politécnico de Castelo Branco
Escola Superior de Tecnologia
Dept. of Industrial Engineering
6000-767 Castelo Branco, Portugal
pgoncalves@est.ipcb.pt

[2] Technical University of Lisbon,
Instituto Superior Técnico
Dept. of Mechanical Engineering, GCAR/IDMEC
1049-001 Lisboa, Portugal
mendonca,j.sousa,jcpinto@dem.ist.utl.pt

Abstract. A new approach to eye-in-hand image-based visual servoing based on fuzzy modeling and control is proposed in this paper. Fuzzy modeling is applied to obtain an inverse model of the mapping between image features errors and joints velocities, avoiding the necessity of inverting the Jacobian. An inverse model is identified for each trajectory using measurements data of a robotic manipulator, and it is directly used as a controller. The control scheme contains an inverse fuzzy model, which is applied to a robotic manipulator performing visual servoing, for a given profile of image features errors. The obtained experimental results show the effectiveness of the proposed control scheme: the fuzzy controller can follow a point-to-point pre-defined trajectory faster (or smoother) than the classic approach.

1 Introduction

In image-based visual servoing [4], the controlled variables are the image features, recognized from the image containing the object. In eye-in-hand image-based visual servoing, the Jacobian plays a decisive role in the convergence of the control, due to its analytical model dependency on the selected image features. Moreover, the Jacobian must be inverted on-line, at each iteration of the control scheme. Nowadays, the research community tries to find the right image features to obtain a diagonal Jacobian [8]. The obtained results only guarantee the decoupling from the position and the orientation of the velocity screw. This is still a hot research topic, as stated very recently in [8].

In this paper, the previous related problems in the Jacobian are addressed using fuzzy techniques, to obtain a controller capable to control the system. A fuzzy model to derive the inverse model of the robot is used to compute the joints and end-effector velocities in a straightforward manner. A two degrees of freedom planar robotic manipulator is controlled, based on eye-in-hand image-based visual servoing using fuzzy control systems.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 253–260, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The paper is organized as follows. Section 2 describes briefly the concept of image-based visual servoing. Section 3 presents fuzzy modeling and identification. The obtained results are presented in Section 4. Finally, Section 5 presents the conclusions and the future research.

2 Image-Based Visual Servoing

In image-based visual servoing, the choice of different image features induces different control laws, and its number depends also on the number of degrees of freedom (DOF) of the robotic manipulator under control. The robotic manipulator used as test-bed in this paper is depicted in Fig. 1, and it has 2 DOF. Thus, the image features s consist of the coordinates x and y of one image point, which are needed to perform the control.

Fig. 1. Planar robotic manipulator with eye-in-hand, camera looking up.

2.1 Modeling the Image-Based Visual Servoing System

Image-based visual servoing is used in an eye-in-hand system [4], where the camera is fixed at the robotic manipulator end-effector. The kinematic modeling of the transformation between the image features velocities and the joints velocities must be found.

The kinematic modeling of the transformation between the image features velocities, $\dot{s}$, and the joints velocities $\dot{q}$ is defined as follows [2]:

$$\dot{s} = J(x, y, Z, q) \cdot \dot{q}\,, \tag{1}$$

where J is the total Jacobian, defined as:

$$J(x, y, Z, q) = J_i(x, y, Z) \cdot {}^{c}W_e \cdot {}^{e}J_R\,(q) \tag{2}$$

where Z is the depth between the camera and object frames; $J_i(x, y, Z)$ is the image Jacobian; ${}^{c}W_e$ is defined as the transformation between the camera and end-effector frames velocities and ${}^{e}J_R$ is the robot Jacobian for the planar robotic manipulator, [2].

2.2 Controlling the Image-Based Visual Servoing System

One of the classic control scheme of robotic manipulators using information from the vision system, is presented in [1]. The global control architecture is shown in Fig. 2, where the block *Robot inner loop* law is a PD control law, with sample time of 1 ms.

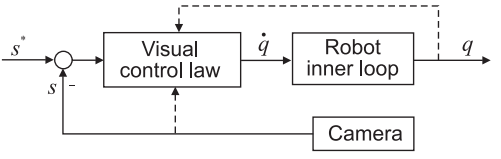


Fig. 2. Control loop of image-based visual servoing.

The robot joint velocities $\dot{q}$ to move the robot to a predefined point in the image, s^* are derived using the *Visual control law*, [2], where an exponential decayment of the image features error is specified:

$$\dot{q} = -K_p \cdot \hat{J}^{-1}(x, y, Z, q) \cdot (s - s^*) \,. \tag{3}$$

K_p is a positive gain, that is used to increase or decrease the decayment of the error velocity.

2.3 Problem Statement

To derive an accurate global Jacobian, J, a perfect modeling of the camera, the image features, the position of the camera related to the end-effector, and the depth of the target related to the camera frame must be accurately determined.

Even when a perfect model of the Jacobian is available, it can contain singularities, which hampers the application of a control law. To overcome these difficulties, a new type of differential relationship between the features and camera velocities was proposed in [7]. This approach estimates the variation of the image features, when an increment in the camera position ${}^cP$ is given, by using a relation G. This relation is divided into G_1 which relates the position of the camera to the image features, and F_1 which relates their respective variation:

$$s + \delta s = G({}^cP + \delta^c P) = G_1({}^cP) + F_1({}^cP, \delta^c P). \tag{4}$$

Considering only the variations in (4):

$$\delta s = F_1({}^cP, \delta^c P), \tag{5}$$

let the relation between the camera position variation $\delta^c P$, the joint position variation, δq and the previous position of the robot q be given by:

$$\delta^c P = F_2(\delta q, q). \tag{6}$$

The two previous equations can be composed because the camera is rigidly attached to the robot end-effector, i.e., knowing q, ${}^cP$ can easily be obtained from the robot direct kinematics. Thus, an inverse function F^{-1} is given by:

$$\delta q = F^{-1}(\delta s, q)\,, \tag{7}$$

and it states that the joint velocities depends on the image features errors and the previous position of the robot manipulator. Equation (7) can be discretized as

$$\delta q(k) = F_k^{-1}(\delta s(k+1), q(k)). \tag{8}$$

In image-based visual servoing, the goal is to obtain a joint velocity, $\delta q(k)$, capable of driving the robot according to a desired image feature position, $s(k+1)$, with an also desired image feature error, $\delta s(k+1)$, from any position in the joint spaces. This goal can be accomplished by modeling the inverse function F_k^{-1}, using inverse fuzzy modeling as presented in Section 3. This new approach to image-based visual servoing allows to overcome the problems stated previously regarding the Jacobian inverse, the Jacobian singularities and the depth estimation, Z.

3 Inverse Fuzzy Modeling

3.1 Fuzzy Modeling

Fuzzy modeling often follows the approach of encoding expert knowledge expressed in a verbal form in a collection of if–then rules. Parameters in this structure can be adapted using input-output data. When no prior knowledge about the system is available, a fuzzy model can be constructed entirely on the basis of system measurements. In the following, we consider data-driven modeling based on fuzzy clustering [6].

We consider rule-based models of the Takagi-Sugeno (TS) type. TS models consist of fuzzy rules describing a local input-output relation, typically in an affine form:

$$R_i : \textbf{If } x_1 \text{ is } A_{i1} \textbf{ and } \ldots \textbf{ and } x_n \text{ is } A_{in} \textbf{ then } y_i = \mathbf{a}_i\mathbf{x} + b_i\,, \quad i = 1, 2, \ldots, K. \tag{9}$$

Here R_i is the ith rule, $\mathbf{x} = [x_1, \ldots, x_n]^T$ are the antecedent variables, $A_{i1}, \ldots, A_{in}$ are fuzzy sets defined in the antecedent space, and y_i is the rule output variable. K denotes the number of rules in the rule base, and the aggregated output of the model, $\hat{y}$, is calculated by taking the weighted average of the rule consequents:

$$\hat{y} = \frac{\sum_{i=1}^{K} \beta_i y_i}{\sum_{i=1}^{K} \beta_i}\,, \tag{10}$$

where β_i is the degree of activation of the ith rule: $\beta_i = \Pi_{j=1}^{n} \mu_{A_{ij}}(x_j)$, $i = 1, \ldots, K$, and $\mu_{A_{ij}}(x_j) : \mathbb{R} \rightarrow [0, 1]$ is the membership function of the fuzzy set A_{ij} in the antecedent of R_i.

To identify the model in (9), the regression matrix X and an output vector $\mathbf{y}$ are constructed from the available data: $\mathrm{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, $\mathbf{y}^T = [y_1, \ldots, y_N]$, where $N \gg n$ is the number of samples used for identification. The number of rules, K,

the antecedent fuzzy sets, A_{ij}, and the consequent parameters, $\mathbf{a}_i, b_i$ are determined by means of fuzzy clustering in the product space of the inputs and the outputs [6]. Hence, the data set Z to be clustered is composed from X and $\mathbf{y}$: $\mathrm{Z}^T = [\mathrm{X}, \mathbf{y}]$. Given Z and an estimated number of clusters K, the Gustafson-Kessel fuzzy clustering algorithm [3] is applied to compute the fuzzy partition matrix U.

The fuzzy sets in the antecedent of the rules are obtained from the partition matrix U, whose ikth element $\mu_{ik} \in [0, 1]$ is the membership degree of the data object $\mathbf{z}_k$ in cluster i. One-dimensional fuzzy sets A_{ij} are obtained from the multidimensional fuzzy sets defined point-wise in the ith row of the partition matrix by projections onto the space of the input variables x_j. The point-wise defined fuzzy sets A_{ij} are approximated by suitable parametric functions in order to compute $\mu_{A_{ij}}(x_j)$ for any value of x_j.

The consequent parameters for each rule are obtained as a weighted ordinary least-square estimate. Let $\theta_i^T = \left[\mathbf{a}_i^T; b_i\right]$, let X_e denote the matrix $[\mathrm{X}; \mathbf{1}]$ and let W_i denote a diagonal matrix in $\mathbb{R}^{N \times N}$ having the degree of activation, $\beta_i(\mathbf{x}_k)$, as its kth diagonal element. Assuming that the columns of X_e are linearly independent and $\beta_i(\mathbf{x}_k) > 0$ for $1 \leq k \leq N$, the weighted least-squares solution of $\mathbf{y} = \mathrm{X}_e\theta + \epsilon$ becomes

$$\theta_i = \left[\mathrm{X}_e^T \mathrm{W}_i \mathrm{X}_e\right]^{-1} \mathrm{X}_e^T \mathrm{W}_i \mathbf{y}\,. \tag{11}$$

3.2 Inverse Modeling

For the robotic application in this paper, the inverse model is identified using input-output data from the inputs $\dot{q}(k)$, outputs $\delta s(k+1)$ and the state of the system $q(k)$, following the procedure described in [5]. A commonly used procedure in robotics is to learn the trajectory that must be followed by the robot. From an initial position, defined by the joint positions, the robotic manipulator moves to the predefined end position, following an also predefined trajectory, by means of a PID joint position controller. This specialized procedure has the drawback of requiring the identification of a new model for each new trajectory. However, this procedure revealed to be quite simple and fast. Moreover, this specialized identification procedure is able to alleviate in a large scale the problems derived from the close-loop identification procedure. The identification data is obtained using the robot-camera configuration shown in Fig. 3.

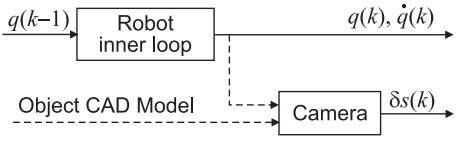


Fig. 3. Robot-camera configuration for model identification.

Note that we are interested in the identification of the inverse model in (8). Fuzzy modeling is used to identify an inverse model. In this technique, only one of the states of the original model, $\dot{q}(k)$, becomes an output of the inverted model and the other state, $q(k)$, together with the original output, $\delta s(k+1)$, are the inputs of the inverted model.

This model is then used as the main controller in the visual servoing control scheme. Therefore, the inverse model must be able to find a joint velocity, $\dot{q}(k)$, capable to drive the robot following a desired image feature error in the image space, $\delta s(k+1)$, departing from previous joint positions, $q(k)$.

4 Results

This section presents the experimental results obtained for the robotic manipulator. First, the identification of the inverse fuzzy model of the robot is described. Then, the control results using the fuzzy model based controller introduced in this paper, i.e. the inverse fuzzy control, are presented.

4.1 Inverse Fuzzy Modeling

In order to apply the controller described in this paper, first an inverse fuzzy controller must be identified. Note that a model must be identified for each trajectory. The profile chosen for the image features error moves the robot from the initial joints position $q_i = [-\frac{\pi}{2}; \frac{\pi}{2}]$ to the final position $q_f = [-\frac{\pi}{6}; \frac{\pi}{6}]$, in eight seconds, starting and ending with zero velocity. An inverse fuzzy model (8) for this trajectory is identified using the fuzzy modeling procedure described in Section 3.1. The measurements data is obtained from the real planar robotic manipulator eye-in-hand system.

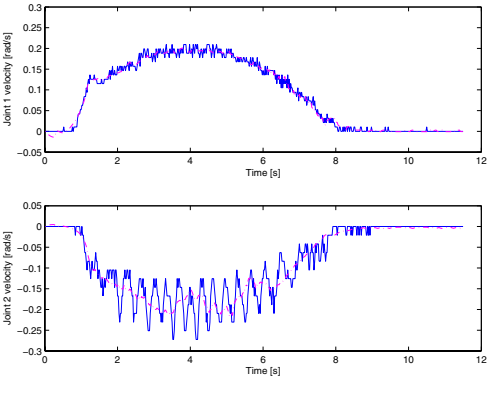
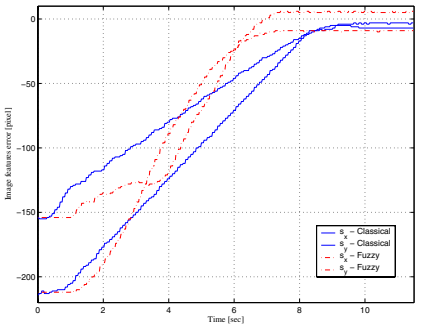
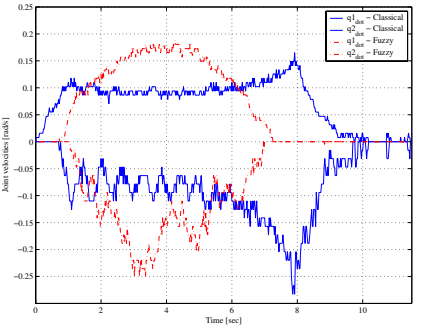


Fig. 4. Validation of the inverse fuzzy model (joint velocities δq). Solid – real output data, and dash-dotted – output of the inverse fuzzy model.

The validation of the inverse fuzzy model is shown in Fig. 4, where the joint velocities δq are depicted. Note that two fuzzy models are identified, one for each joint velocity. It is clear that the model is quite good. Considering, e.g. the performance criteria *variance accounted for* (VAF), the models have the VAFs of 99.01% and 87.95%. When a perfect match occur, this measure has the value of 100%. Then, the inverse model for the joint velocity δq_1 is very accurate, but the inverse model for δq_2 is not so good. This was



(a) Image features error, s. Solid – fuzzy visual servo; dashed – classical visual servo.



(b) Joint velocities, δq. Solid – fuzzy visual servo; dashed – classical visual servo.

Fig. 5. Comparison between fuzzy and classical image features errors and joint velocities.

expectable as the joint velocity δq_2 varies a bit more than δq_1. However, this model is sufficient to obtain an accurate controller, as is shown in Section 4.2.

In terms of parameters, four rules (clusters) revealed to be sufficient for each output, and thus the inverse fuzzy model has 8 rules, 4 for each output, δq_1 and δq_2. The clusters are projected into the product-space of the space variables and the fuzzy sets A_{ij} are determined.

4.2 Control Results

This section presents the obtained control results, using the classical image-based visual servoing presented in Section 2, and the fuzzy model-based control scheme using inverse model control presented in Section 3. The implementation was performed in the planar robotic manipulator eye-in-hand system.

The chosen profile for the image features error moves the robot from the initial image features position $s = [-154; -213]$ to the final position $s^* = [0; 0]$ in eight seconds. The comparison of the image features error for both the classic and the fuzzy visual servoing controllers is presented in Fig. 5a. In this figure, it is shown that both controllers can achieve the vicinity of the desired position. However, the fuzzy controller is slightly faster, and reaches the vicinity of the desired position before the classical controller. This vicinity is defined with 10 pixels error around the desired final position, which corresponds to the joint actuators dead-zone. To better compare the two approaches, it is necessary to check the joint velocities in order to check their smoothness. Thus the joint velocities are depicted in Fig. 5b, where it is clear that the classical controller presents more oscillations in the joint velocities. Moreover, at the end of the planned trajectory, eight seconds, the classical controller has not yet achieved the steady-state and the desired position is achieved using two more seconds. Speeding up the classical controller, i.e. increasing the proportional gain, the desired position can be achieved decreasing the two seconds needed at the cost of increasing the joint velocities oscillations. This case is clearly undesirable under a control point of view.

5 Conclusions

This paper introduces an eye-in-hand image-based visual servoing scheme based on fuzzy modeling. The fuzzy modeling approach was applied to obtain an inverse model of the mapping between image features error trajectory and joints velocities. This inverse model is directly used as the controller of a robotic manipulator performing visual servoing for a given image features trajectory profile. The obtained experimental results showed that both the classical and the fuzzy controllers can achieve the desired image features position. However, the joint velocities behavior during the control and the time needed for the classical approach to achieve the steady-state, lead us to the conclusion that the inverse fuzzy control proposed in this paper performs better.

As future work, the proposed fuzzy model based control scheme will be extended to a 6 DOF robot manipulator and implemented in a PUMA robot. Note that an off-line identification of the inverse fuzzy model must first be performed. The complete automation of this identification step is also under study.

Acknowledgements. This work is supported by the "Programa de Financiamento Plurianual de Unidades de I&D (POCTI) do Quadro Comunitário de Apoio III", by program FEDER, by the FCT project POCTI/EME/39946/2001, and by the "Programa do FSE-UE, PRODEP III, acção 5.3, no âmbito do III Quadro Comunitário de apoio".

References

1. B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3):313–326, 1992.
2. P. S. Gonçalves and J. C. Pinto. Camera configurations of a visual servoing setup, for a 2 dof planar robot. In *Proceedings of the 7th International IFAC Symposium on Robot Control, Wroclaw, Poland.*, pages 181–187, Wroclaw, Poland, Sept. 2003.
3. D. E. Gustafson and W. C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings IEEE CDC*, pages 761–766, San Diego, USA, 1979.
4. S. Hutchinson, G. Hager, and P. Corke. A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation*, 12(5):651–670, 1996.
5. J. M. Sousa, C. Silva, and J. Sá da Costa. Fuzzy active noise modeling and control. *International Journal of Approximate Reasoning*, 33:51–70, April 2003.
6. J. M. C. Sousa and U. Kaymak. *Fuzzy Decision Making in Modeling and Control*. World Scientific Pub. Co., Singapore, 2002.
7. I. Suh and T. Kim. Fuzzy membership function based neural networks with applications to the visual servoing of robot manipulators. *IEEE Trans. on Fuzzy Systems*, 2(3):203–220, 1994.
8. O. Tahri and F. Chaumette. Application of moment invariants to visual servoing. In *Proc.of the IEEE Int. Conf. on Robotics and Automation*, pages 4276–4281, Taipeh, Taiwan, 2003.

Large Display Interaction Using Video Avatar and Hand Gesture Recognition

Sang Chul Ahn, Tae-Seong Lee, Ig-Jae Kim, Yong-Moo Kwon, and Hyoung-Gon Kim

Imaging Media Research Center, KIST
39-1 Hawolgok-dong, Sungbuk-gu, Seoul KOREA 136-791
{asc, kij, lts, ymk, hgk}@imrc.kist.re.kr

Abstract. This paper presents a new system for interacting with a large display using live video avatar of a user and hand gesture recognition. The system enables a user to appear on a screen as a video avatar and to interact with items therein. The user can interact with the large display remotely by walking and touching icons through his video avatar. In order to build the system, we developed live video composition, active IR vision-based hand gesture recognition, and 3D human body tracking system, and incorporated a voice recognition system, too. Using this system, a user can interact efficiently with an embedded computer in a space that equipped with a large display.

1 Introduction

HCI(Human Computer Interaction) has been one of the most important research issues since the computer became one of the necessities of our daily lives. Many researchers are working on developing new interaction paradigms for efficient computer usage over mice and keyboards. As related works, there have been some researches to control computers by vision-based interaction. For instance, Kjeldsen tried to use gestures for computer interface[1]. The Perceptual Window offered a head motion based interaction technique[2]. Further, HCI techniques have been applied to virtual environment control[3], and building a smart interactive space[4,5].

Nowadays, as computers are getting smaller and smaller, and are embedded inside a lot of things around us. This embedded computing environment doesn't allow us to use mouse and keyboard type interface any more. We need a new type of interaction mechanism with things that have computing capability. In this environment, we don't care where the computers are located, but only confirm the result of our commands through some output devices like a large screen display. In this case, interaction in a space is required. The smart interactive spaces[4,5] are the examples of this kind of interaction.

This paper proposes a new mechanism of interaction with a large display using live video avatar of user and computer vision based hand gesture recognition. The proposed system enables a user to appear on the screen as a video avatar, and to walk and navigate across the screen. The system also allows the user to interact with items or icons in the screen by touching them with a hand. The proposed system replaces the

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 261–268, 2004.
© Springer-Verlag Berlin Heidelberg 2004

mouse cursor with live video avatar of a user. The user can control the computer by walking, navigating, and touching icons while seeing himself as a video avatar on the screen. In order to build the system, we have developed live video composition, active IR vision-based hand gesture recognition systems, and incorporated voice recognition system, too. Later we added 3D human body tracking to the system for 3D interaction. The proposed system can be applied to any tasks that can be done in a computer, but it is more efficient in some cases such as games and Powerpoint presentations.

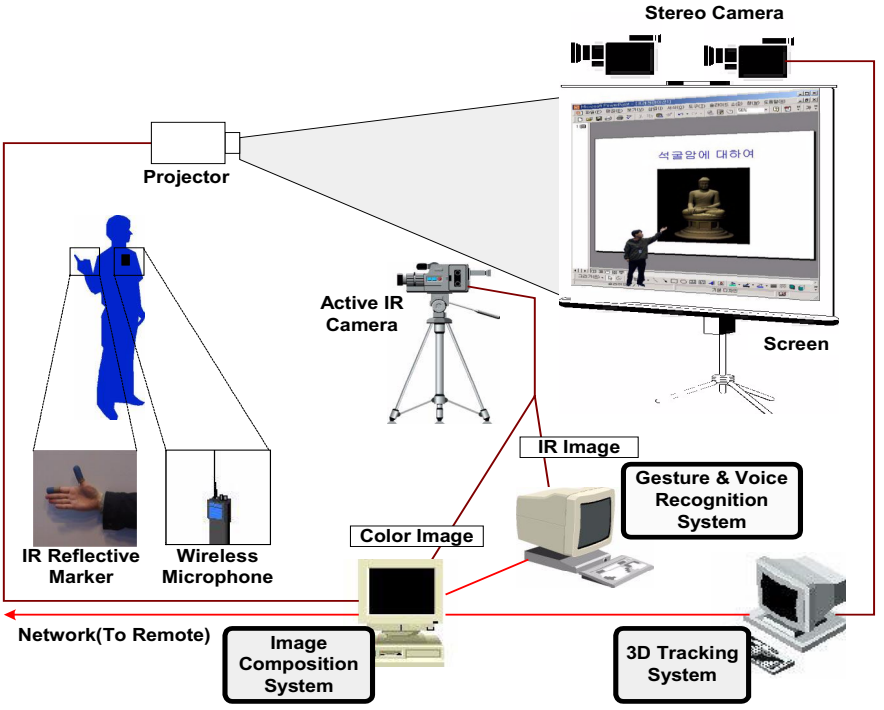


Fig. 1. Configuration of the proposed large screen interaction system

2 System Overview

The proposed system consists of three subsystems: *Image composition subsystem, Gesture & voice recognition subsystem,* and *3D Tracking subsystem*. Figure 1 shows the configuration of the system, and three computers represent the subsystems. The user interacts with the large screen in front of a camera while looking at the screen. The image composition subsystem extracts the user image and overlays it on the screen. So, the user sees himself as a video avatar on the screen during interaction. The user can interact with icons and items within the screen. He can use hand gesture and voice to control them. For robust hand gesture recognition, we used an IR (Infra Red) camera and IR reflective markers. The 3D tracking subsystem uses a stereo

camera and tracks the user's position. The user can use this subsystem for 3D interaction such as walking around a 3D object. As a whole, this proposed system gives us a more intuitive interaction method with a large screen.

3 Video Avatar

The proposed system uses a live video avatar of a user to give more intuitive sense of interaction. Since the image of a user is overlaid on a screen and moves across the screen, the user can feel as if he is in the computer screen. The live video avatar also has the role of mouse cursor in controlling and interacting with the screen.

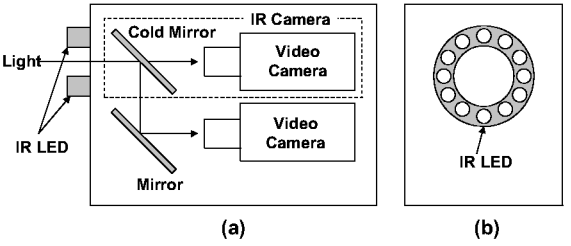


Fig. 2. Active IR camera, (a) side cut view, (b) front view

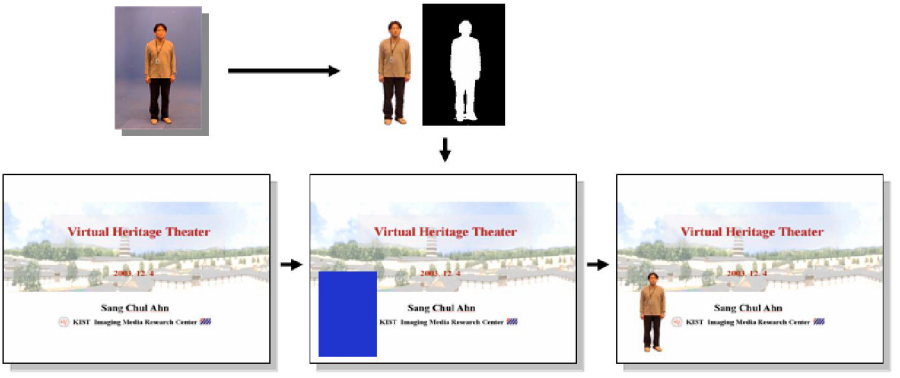


Fig. 3. Video avatar composition

In order to make a live video avatar, the image composition subsystem captures the images of a user with video camera. In Figure 1, we can see that an active IR camera is capturing the images of the user. This active IR camera is a combination of an IR camera, an IR light source, and color video camera. The structure of the active IR camera is shown in Figure 2. The IR camera and IR light source are used in the gesture & voice recognition subsystem, which will be explained in the next section. The image composition subsystem uses the output of the color video camera. The image composition subsystem extracts the image of a user from the background, and uses it as a live video avatar. Background subtraction or chroma-keying method can be used

in this process. The video avatar is overlaid on the screen. The composition of video avatar can be done with a video processing hardware or a chroma-keying hardware. However, we found that we could implement it in realtime using a chroma-keying function of the Windows XP. The user can control the position of the video avatar by walking around the stage as long as he remains within the field of view of the video camera. Figure 3 shows the image composition mechanism of video avatar.

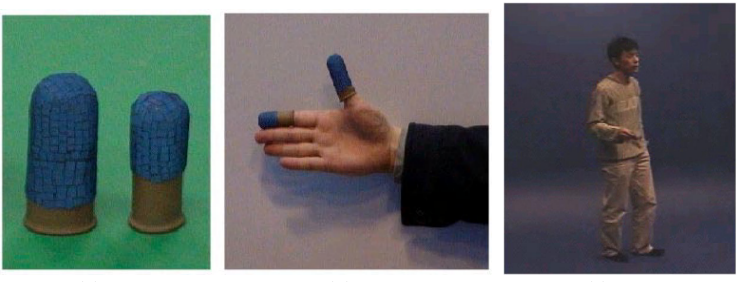
4 Gesture and Voice Interaction

The proposed system uses hand gesture and voice as the main interaction tools. The gesture & voice recognition subsystem enables the user interaction by vision-based hand gesture recognition and voice recognition. For instance, a user can issue a command to play music by saying "Music Start" to a microphone. By hand gesture "Up / Down / Left / Right", he can also move his video avatar across the screen to select or play an item. This is useful when it is too far from him to reach by walking within the field of view of camera. The implemented user commands are shown in Table 1. As can be seen, there are many voice commands, and hand gesture commands are included in the voice commands. However, hand gesture commands are more useful for 2D/3D motion than voice commands. Voice commands are more useful for discrete actions such as start or end of some action. Some of the commands are implemented for interactive Powerpoint presentation.

Table 1. The implemented commands for interaction

Type	Commands	Comments
Hand gesture	Left / Right Up / Down	2D avatar move or 3D navigation
	Forward / Backward	3D navigation
	Select	Left mouse button Double click
Voice	Left / Right Up / Down	2D avatar move or 3D navigation
	Slideshow Next / Previous Select / Close	Powerpoint full screen Proceed or go back Selection/Window close
	Pen Mode Pen Mode Exit	Powerpoint pen mode
	Screen Keyboard	
	Navigation Forward / Backward Navigation Stop	3D navigation start 3D navigation 3D navigation stop
	Music Start Equalizer	Media Player start Media Player equalizer
	Help	Call a helping MS Agent

Although computer vision-based hand gesture recognition has been widely studied, it also inherits the shortcoming of most computer vision algorithms: sensitiveness to lighting condition. Thus, we adopted active IR based recognition method for robust recognition. As mentioned before, an active IR(Infra Red) camera is used for capturing a user's hand gesture. The IR camera can be made with a normal video camera and an IR filter. In Figure 2, the upper part composes the IR camera. The cold mirror is an IR filter that absorbs IR rays while reflecting visible rays. We used a cold mirror that absorbs the rays of above 800nm in wave length. Additionally we made two IR reflective thimbles for user's hand. The IR reflective thimbles were made with retro-reflective material so that they can be viewed best from the camera with IR light source. Figure 4 shows the IR reflective thimbles and a user's hand wearing them. Since thumb and index finger are mostly used for selecting and pointing, we decided to use the thimbles for thumb and index finger.



(a) **(b)** **(c)**

Fig. 4. IR reflective thimbles, (a) thimbles, (b) a hand wearing thimbles, (c) user interacting by hand gesture.

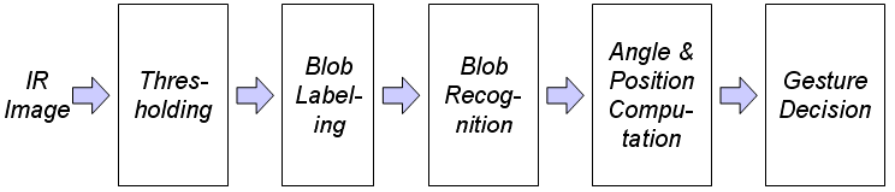


Fig. 5. The flow chart of hand gesture recognition

Figure 5 shows the flowchart of hand gesture recognition algorithm. Since the IR reflective thimbles look white in an IR image, we can extract the regions by simple thresholding. Then, we apply the labeling operation to the regions, and find thimble regions of thumb and index finger by size. The hand gesture is recognized by relative position and direction of thimbles. The center position (x_c, y_c) of a thimble is determined using the 0^{th} and 1^{st} moments as follows,

$$(x_c, y_c) = \left(\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right), \qquad M_{ij} = \sum_x \sum_y x^i y^j I(x, y),$$

where $I(x,y)$ is the intensity value of IR image at (x,y) position. M_{ij} is x and y directional moment. We can also compute directional angle using 2nd moment as follows,

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{b}{a-c}\right), \quad a = \frac{M_{20}}{M_{00}} - x_c^2, \quad b = 2\left(\frac{M_{11}}{M_{00}} - x_c y_c\right), \quad c = \frac{M_{02}}{M_{00}} - y_c^2 .$$

Then, we check the following 5 conditions to determine relative position.

Condition 1: $y_Y - y_G \geq \frac{H_Y + H_G}{2} + \alpha_1$; Yellow blob is below green blob

Condition 2: $y_G - y_Y \geq \frac{H_Y + H_G}{2} + \alpha_1$; Yellow blob is above green blob

Condition 3: $x_Y - x_G \geq \frac{W_Y + W_G}{2} + \alpha_2$; Yellow blob is right of green blob

Condition 4: $x_G - x_Y \geq \frac{W_Y + W_G}{2} + \alpha_2$; Yellow blob is left of green blob

Condition 5: $|x_G - x_Y| \geq \frac{W_Y + W_G}{2}$; Two blobs are close each other vertically

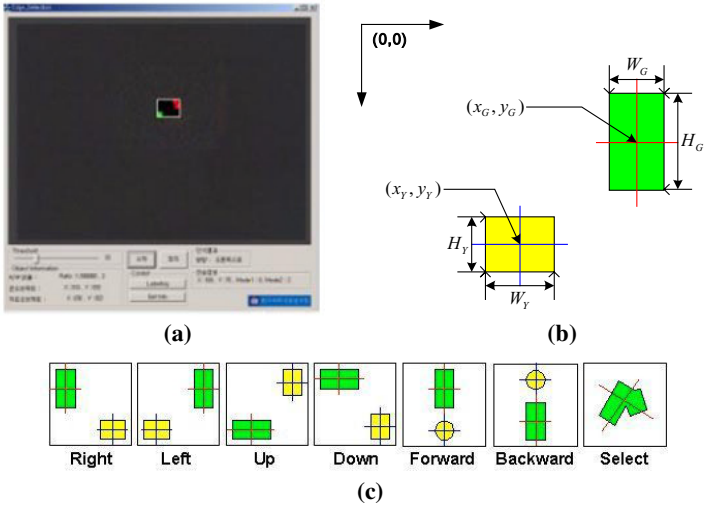
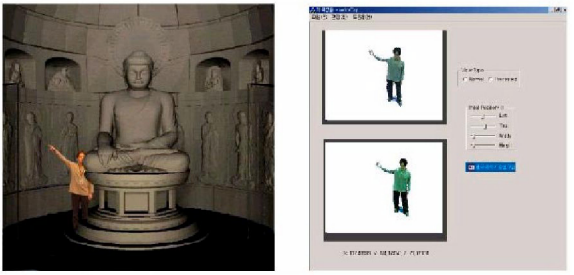


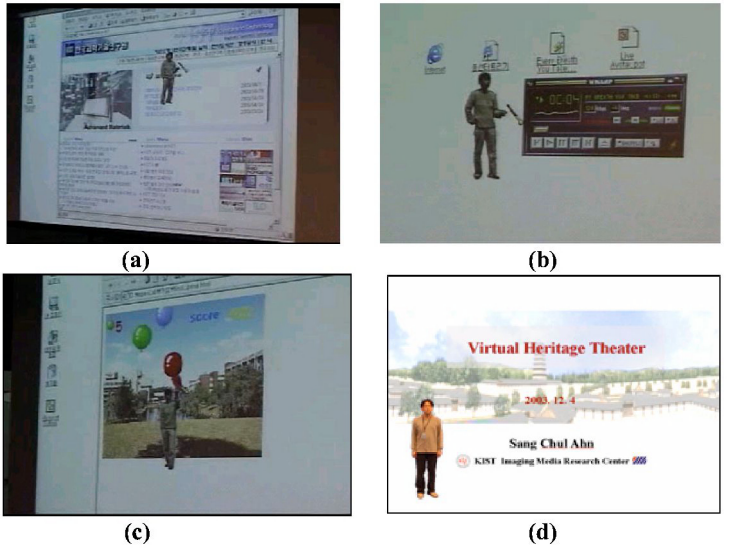
Fig. 6. (a) Screen shot of the hand gesture recognition system, (b) Metrics of thimbles, (c) typical configuration of hand gesture image.

Finally we interpret hand gestures as shown in Figure 6. Note that the green(longer) and yellow blobs represent thimbles for thumb and index finger, respectively. α_1, α_2 are predefined small values to make up the distance between thumb and index finger. For instance, if all the directional angles of yellow and green blobs are 90 degree, and the condition 1 and 5 are satisfied, the hand gesture is recognized as the "forward" command.



(a) (b)

Fig. 7. 3D tracking result, (a) a user walking around a statue of Buddha, (b) screen shot of the3D tracking system.



(a) (b)

(c) (d)

Fig. 8. Screen shots of various applications, (a) internet surfing (b) music control (c) game (d) Powerpoint presentation.

Once the hand gesture is recognized, it is converted to a command and sent to image composition subsystem by network. Then, image composition subsystem applies an appropriate action to the screen. Here we have to make the IR image have the same coordinates to that of normal video image. That's why we use the active IR camera structure as in Figure 2.

5 3D Interaction

In the proposed system, a user can use 3D interaction. Figure 7 shows one of the examples, where a user shows a 3D model of Buddha and his video avatar is walking around it. It can be seen that part of the avatar is occluded by the statue. In this way a video avatar can walk and navigate the 3D space. We implemented it by invoking our

3D model viewer program. The 3D tracking subsystem extracts a user out of the images from stereo camera, and gets his 3D position. The information is sent to image composition subsystem so that it can overlay a video avatar at appropriate 3D position. Figure 7(b) shows a screen shot of 3D tracking system.

6 Application

We applied the proposed system to build an intelligent room called "Smart Studio". In the Smart Studio we could interact with a large screen and use an embedded computer in the environment. We could enjoy music and navigate internet as we want. We could enjoy some games with whole body action, too. From our experiments, we noticed that the proposed system provides efficient interaction mechanism in space and that it is quite effective in games and Powerpoint presentation applications. Figure 8 shows some of the screen shots of various applications.

7 Conclusion

In this paper, we have presented a new system for interacting with a large screen, which enables a user to appear on the screen and to interact with icons and items therein. The proposed system incorporates live video composition, voice recognition, active IR vision-based hand gesture recognition, and 3D human body tracking. The screen output of the system looks similar to TV news, but is different in that a user can control every item on the screen in realtime. The proposed system has the following advantages. First, a user can feel somewhat immersive sense since he appears as a video avatar on the screen and can touch icons by hands. Secondly, a user can feel naturalness with controlling by walking or touching and picking icons and items. We applied the system to building an intelligent space and found that it worked efficiently. As embedded computing environment is coming to our lives, this type of interaction is expected to be used. Currently, we are extending this system to a teleconferencing system that shows video avatars of every participant and shares the same workspace.

References

1. R. Kjeldsen, J. Kender: Toward the use of gesture in traditional user interfaces. Proceeding of Automatic FGR'96. (1996) 151-156
2. J.L. Crowley, J. Coutaz, and F. Berard: Things that see. Communications of the ACM. vol. 43. no. 3. Mar. (2000) 60–61
3. I.J. Kim, S. Lee, S.C. Ahn, Y.M. Kwon, H.G. Kim: 3D tracking of multi-objects using color and stereo for HCI. Proceeding of ICIP2002. (2002)
4. A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, C.S. Pinhanez, L.W. Campbell, Y.A. Ivanov, A. Schutte, and A. Wilson: The Kids Room. Communications of the ACM. vol. 43. no. 3. Mar. (2000) 60–61
5. Brumitt, B., Meyers, B., Krumm, J., Kern, A., and Shafer, S.: EasyLiving: Technologies for Intelligent Environments. Handheld and Ubiquitous Computing, September (2000)

Optimal Transform in Perceptually Uniform Color Space and Its Application in Image Coding*

Ying Chen[1,2], Pengwei Hao[1,2], and Anrong Dang[3]

[1] Center for Information Science, Peking University, Beijing, 100871, China
phao@cis.pku.edu.cn

[2] Department of Computer Science, Queen Mary, University of London, E1 4NS, UK
{ying, phao}@dcs.qmul.ac.uk

[3] Center for Science of Human Settlements, Tsinghua University, Beijing, 100084, China
danrong@tsinghua.edu.cn

Abstract. To find an appropriate color transform is necessary and helpful for the applications of color images. In this paper, we proposed a new scheme to find color transforms close to the optimal transform and agree with human vision system for comparison. We first apply the perceptually uniform color space transform to convert RGB components into uniform CIE LAB components, and then use principal components analysis (PCA) in the uniform space to find the image-dependent optimal color transforms (KLT) for each test image group and for all the images in all the groups. With the KLTs, an approximate but image-independent transform in CIE LAB space is presented, namely LAR, which is just the LAB space rotated and has an elegant and simple form. Finally, we apply it to image compression, and our experiment shows that LAR performs better (PSNR about 4dB higher) than the default color transform in JPEG 2000.

1 Introduction

Color images are everywhere in science, technology, medicine and industry. Color images are acquired and reproduced based on tristimulus values whose spectral composition is carefully chosen according to the principles of color science. Color space transform is critical for color feature extraction and data redundancy reduction. To find an appropriate color transform is necessary and very helpful in many color image applications, such as image display, processing, retrieval, recognition, and compression.

In 1931, Commission Internationale de l'Eclairage (CIE) defined three standard primaries, called X, Y and Z, to replace red, green and blue (RGB) and with positive weights to match all the colors we see. It took CIE a decade or more to find a transformation of CIE XYZ into a reasonable perceptually uniform space. So far, CIE standardized two perceptually uniform systems, CIE LUV and CIE LAB. Besides, many color standards enable users to have the freedom to choose the color space in

* This work was supported by the Foundation for the Authors of National Excellent Doctoral Dissertation of China, under Grant 200038.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 269–276, 2004.
© Springer-Verlag Berlin Heidelberg 2004

which to represent their data. RGB, CMYK, YIQ, HSV, CIE 1931 XYZ, CIE LUV, CIE LAB, YES, CCIR 601-2 YCbCr, and SMPTE-C RGB are proposed for diverse requirements [1]. However, in many applications, we need some appropriate color transforms and we also wish the transformed or the inverse-transformed components are inter-comparable and the comparison done by computers agrees with that we do by our human visual system. Therefore, we need compare the results in a perceptually uniform color space after applying inverse of our specific color space transforms. Our idea is to find some optimal color transforms in the uniform space.

In this paper, a new scheme to find an optimal color transform is proposed. We transform color images into three components in the uniform space CIE LAB, and then use principal components analysis (PCA) to find image-dependent optimal color transforms, Karhunen-Lòeve Transform (K-L transform, or KLT). Finally, we take the optimal transform obtained from all the analyzed images as an image-independent color transform and apply it to image compression of some other test images with JPEG 2000.

2 Principal Components Analysis in CIE LAB Color Space

CIE LAB color space is used in this paper as a perceptually uniform space and then the K-L transforms are found in this space. The work is implemented with five different groups of color images and by three main steps:

Step1. Transform all the images from the original image color space into CIE LAB color space.

Step2. Compute the covariance matrix in CIE LAB space with all the pixels of all the images in a group.

Step3. Find the three eigenvalues and their corresponding eigenvectors of the covariance matrix and then make the K-L transform with the three eigenvectors.

Among the steps, Step2 and Step3 are actually principal components analysis.

2.1 RGB to CIE LAB

The conversion from RGB color values into CIE LAB components is implemented by converting from nonlinear RGB space into linear RGB space first, then into CIE XYZ space, and finally into CIE LAB space.

The original color space of the images we generally come across is nonlinear RGB space. Nonlinear RGB, also called gamma-corrected RGB, may be stored in a file as three 8-bit integers ranging from 0 to 255. A simple scaling conversion is required to transform the three integer values from 0 to 255 into floating-point numbers in [0, 1].

There are several possible choices for the conversion between nonlinear RGB and linear RGB, the transform and its inverse used in this paper are:

From the gamma-corrected RGB values to a linear RGB values:

$$\begin{aligned} &\text{For } R_N, G_N, B_N \geq 0.0,\ R_L = R_N^{2.2},\ G_L = G_N^{2.2},\ B_L = B_N^{2.2} \\ &\text{For } R_N, G_N, B_N < 0.0,\ R_L = -|R_N|^{2.2},\ G_L = -|G_N|^{2.2},\ B_L = -|B_N|^{2.2} \end{aligned} \tag{1}$$

From the linear RGB values to gamma-corrected nonlinear values:

For $R_L, G_L, B_L \geq 0.0$, $R_N = R_L^{1/2.2}$, $G_N = G_L^{1/2.2}$, $B_N = B_L^{1/2.2}$ (3)

For $R_L, G_L, B_L < 0.0$, $R_N = -|R_L|^{2.2}$, $G_N = -|G_L|^{2.2}$, $B_N = -|B_L|^{1/2.2}$ (4)

The following is a linear transform from linear RGB to CIE XYZ (D50).

$$\begin{bmatrix} X_{D50} \\ Y_{D50} \\ Z_{D50} \end{bmatrix} = \boldsymbol{A} \cdot \begin{bmatrix} R_L \\ G_L \\ B_L \end{bmatrix} \quad \text{where} \quad \boldsymbol{A} = \begin{bmatrix} 0.4298 & 0.3967 & 0.1377 \\ 0.2142 & 0.7141 & 0.0718 \\ 0.0146 & 0.0903 & 0.7199 \end{bmatrix} \tag{5}$$

where the transform matrix **A** is derived by the matrix multiplication of two matrices:

$$\boldsymbol{B} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} 1.0161 & 0.0553 & -0.0522 \\ 0.0061 & 0.9956 & -0.0012 \\ 0 & 0 & 0.7576 \end{bmatrix} \tag{6}$$

where $\boldsymbol{A}=\boldsymbol{DB}$, $\boldsymbol{B}$ is the matrix to transform from linear RGB into XYZ (D65) [2], $\boldsymbol{D}$ is the matrix to transform from XYZ (D65) into XYZ (D50) [1].

The conversion from CIE XYZ (D50) into CIE LAB is computed as follows, which can be found in [1]:

$$\begin{aligned} L^* &= 116(Y/Yn)^{1/3} - 16, (Y/Yn) > 0.008856 \\ L^* &= 903.3\,(Y/Yn), (Y/Yn) < 0.008856) \\ a^* &= 500[f(X/Xn) - f(Y/Yn)] \\ b^* &= 200[f(Z/Zn) - f(Y/Yn)] \end{aligned} \tag{7}$$

where $f(t) = 7.787\mathrm{t} + 16/116$ if $\mathrm{t} < 0.008856$, and $f(t) = t^{1/3}$ if $\mathrm{t} > 0.008856$.

In this paper, we use CIE XYZ (D50) as our XYZ space, and at the white point, Xn, Yn and Zn have the values of 0.9642, 1.0, and 0.8249, respectively.

There are ways to speed up the transform, such as to approximate the cube root [4].

There are also some methods using the nonlinear functions in conversion between CIE XYZ and CIE LAB [5] and in RGB gamma-correction, simpler and less time-consuming. A system using linear interpolation to transform RGB to CIE LAB can be found in [6].

2.2 Principal Components Analysis

Having been converted into the CIE LAB color space, the images are analyzed with principal components analysis (PCA) method [3].

Consider $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is a collection of pixel samples in the CIE LAB color space. Each sample, $\boldsymbol{x}_i=[x_{i1}, x_{i2}, x_{i3}]^T$, has 3 elements, which are 3 components L^*, a^* and b^* respectively. We use these samples to calculate the covariance matrix of $\boldsymbol{X}$, denoted by $\boldsymbol{S}$.

$$\boldsymbol{S} = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^T \tag{8}$$

where $\boldsymbol{m}$ is the mean vector of the n sample pixels, $\boldsymbol{m} = (1/n)\sum_{k=1}^{n} \boldsymbol{x}_k$.

$\boldsymbol{S}$ is a 3×3 symmetric matrix, and the solutions of the principal components analysis satisfy the equation: $\boldsymbol{Sv}=\lambda\boldsymbol{v}$, where λ is one of the 3 eigenvalues of matrix $\boldsymbol{S}$ and $\boldsymbol{v}$ is the corresponding eigenvector.

For each covariance matrix, we can obtain three eigenvalues λ_1, λ_2 and λ_3, and their eigenvectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$.

Then, the K-L transform is defined as $\boldsymbol{y}=\boldsymbol{V}\boldsymbol{x}$, where $\boldsymbol{x}$ is a pixel in the CIE LAB space, $\boldsymbol{y}$ is the resultant pixel after transformation, and $\boldsymbol{V}=[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T$.

The samples to make matrix $\boldsymbol{X}$ may be all the pixels of an image or all the pixels of all the images in a group. The computation of the covariance matrix is the most complex work and it has a temporal complexity of $O(n^2)$. A covariance matrix from an image gives a K-L transform that is optimal for the single image. The matrix generated from all the images in a group provides a K-L transform that is optimal for the group images as a whole.

The three eigenvalues are useful if we want to know whether a component is more significant than another. The larger an eigenvalue is, the more significant its corresponding component is. In this paper, the eigenvector that corresponds to the maximum eigenvalue is called *principal eigenvector*. By using the principal eigenvector in the K-L transform, the most important component can be found. This is helpful for many color image applications, such as color image coding and color image retrieval.

3 PCA Experiments

In this paper, a database of 5 image groups is used for our experiments to obtain the optimal transforms.

1. Architectures: 56 images in JPEG format of Asian architectures in Japan, Iran and Indonesia.
2. Sports: 53 JPEG images of football and soccer games.
3. Animals: 50 JPEG images of feral horses.
4. Plants: 125 JPEG images of many diverse plants.
5. Landscapes: 99 images in bitmap format. Landscape pictures of Greenland.

The images in Group 1, 2 and 5 are from the image database of the Department of Computer Science & Engineering, University of Washington (www.cs.washington.edu /research/imagedatabase). Those in Group 3 and 4 are from the collections of the digital library of the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley (http://elib.cs.berkeley.edu).

For the JPEG files [7], each pixel is converted from YCbCr color space back into RGB color space first during decoding process.

The principal components analysis takes place in the CIE LAB color space, so the eigenvectors and the K-L transform are also in the CIE LAB space. The K-L transform matrix $\boldsymbol{V}$ and three eigenvectors have a relation of $\boldsymbol{V}=[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T$. The eigenvalues are sorted in the magnitude-descending order, and $\boldsymbol{v}_1$ is the principal vector.

In order to better understand the proportions between the eigenvalues, we use a rough *eigPercent* to stand for the percentages of eigenvalues relative to the sum of all three.

3.1 The Optimal Transforms of the 5 Individual Groups

The K-L transform matrix (transpose of eigenvectors matrix) and eigenvalues of image group 1 (architectures) are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9919 & 0.0211 & -0.1251 \\ 0.0614 & 0.7834 & 0.6184 \\ 0.1111 & -0.6211 & 0.7758 \end{bmatrix} \tag{9}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{10} \times [1.5191, 0.3070, 0.2217] \quad eigPercent = [74,15,11] \tag{10}$$

The K-L transform matrix and eigenvalues of image group 2 (sports) are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9831 & -0.1832 & -0.0036 \\ 0.1118 & 0.5842 & 0.8039 \\ 0.1452 & 0.7907 & -0.5948 \end{bmatrix} \tag{11}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{10} \times [1.0332, 0.2603, 0.1647] \quad eigPercent = [71,18,11] \tag{12}$$

The K-L transform matrix and eigenvalues of image group 3 (animals) are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9890 & 0.1406 & -0.0468 \\ -0.0619 & 0.6792 & 0.7314 \\ 0.1346 & -0.7204 & 0.6804 \end{bmatrix} \tag{13}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{10} \times [1.1709, 0.5858, 0.4026] \quad eigPercent = [54,27,19] \tag{14}$$

The K-L transform matrix and eigenvalues of image group 4 (plants) are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9178 & -0.1138 & 0.3803 \\ 0.3032 & 0.8193 & -0.4867 \\ -0.2562 & 0.5620 & 0.7865 \end{bmatrix} \tag{15}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{10} \times [2.1028, 1.2671, 0.9036] \quad eigPercent = [49,30,21] \tag{16}$$

The K-L transform matrix and eigenvalues of image group 5 (landscapes) are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9814 & -0.1130 & 0.1553 \\ 0.1888 & 0.7154 & 0.6727 \\ 0.0352 & -0.6895 & 0.7234 \end{bmatrix} \tag{17}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{10} \times [4.2005, 1.3912, 0.7246] \quad eigPercent = [67,22,11] \tag{18}$$

It is very interesting that the principal eigenvectors of the 5 image groups are all close to $(1,0,0)^T$. It implies that, in the CIE LAB space, the principal component of the images is in the direction of L^*.

3.2 The Optimal Transform of All the Pictures in All the Groups

We use all the test images to find the general K-L transform for all the groups. The K-L transform matrix (transpose of eigenvectors matrix), eigenvalues and their proportional percentages from all the images in all the groups are:

$$\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]^T = \begin{bmatrix} 0.9992 & -0.0301 & 0.0275 \\ -0.0018 & -0.7065 & -0.7077 \\ 0.0408 & 0.7070 & -0.7060 \end{bmatrix} \tag{19}$$

$$[\lambda_1, \lambda_2, \lambda_3] = 10^{11} \times [1.1990, 0.5694, 0.5060] \quad eigPercent = [53,25,22] \tag{20}$$

The K-L transform for all the images in our database also implies that the principal vector in CIE LAB space is very close to the vector $(1,0,0)^T$, and the other components are primarily in the AOB plane of the CIE LAB coordinate system, where O is the origin in CIE LAB space. Both vectors have a rotation angle of about $\pi/4$ to CIE LAB axes except a sign.

Therefore, through our experiments, an approximate optimal transform can be given as:

$$\boldsymbol{V} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi/4) & \sin(\pi/4) \\ 0 & -\sin(\pi/4) & \cos(\pi/4) \end{bmatrix} \tag{21}$$

It looks so elegant and simple. More importantly, the color space after this transform is still perceptually uniform, and we believe that it is a better choice for perceptually uniform color space standard. Since the new space is just LAB space rotated, we name the transformed space as LAR.

From the proportional percentages between the eignevalues, we know that the significance between the components in CIE LAB space is roughly 2:1:1. This also looks very nice and will be very useful in image applications.

The percentages of the three eigenvalues give the information of significance of components, which can be used in image retrieval or image compression during quantization.

4 Application in Image Coding

The optimal transform in CIE LAB has many applications. In this paper, we test lossy compression of color images in the image coding framework of the new international standard JPEG 2000. Our transform gives better performance than the standardized color transform.

4.1 Intercomponent Transforms

We use our optimal transform in CIE LAB space presented above as an intercomponent transform in the new image coding standard framework, JPEG 2000 [8].

We assume RGB components are in a cube of $256 \times 256 \times 256$, and each component can be any integer from 0 to 255. For all possible RGB colors, the ranges of our transformed components LAR are: $L \in [0, 100.004]$, $A \in [-49.317, 106.296]$, and $R \in [-114.222, 111.811]$. Therefore, we use the new bounding cube to normalize

LAR components into [0, 255] by a simple linear mapping: $f(x)=255*(x-x_{\min})/(x_{\max}-x_{\min})$.

In the standard JPEG 2000, two color transforms are employed. The one for lossy image coding is in YCbCr space. The forward and the inverse of the transform are defined as:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & -0.000 & 1.402 \\ 1 & -0.344 & -0.714 \\ 1 & 1.772 & 0.000 \end{bmatrix} \begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} \tag{22}$$

The other color transform is integer reversible for lossless compression, and is a rough approximation of YCbCr color transform.

4.2 Compression Evaluation

The quality of a compressed image is generally evaluated by peak signal-to-noise ratio (PSNR) objectively and human visual system subjectively.

In order to make our objective comparison agree with the subjective evaluation, PSNR is calculated with the difference between the original image and the reconstructed image, and the difference is found in CIE LAB:

$$PSNR = 10\log 10 \frac{(L_{\max} - L_{\min})^2 + (A_{\max} - A_{\min})^2 + (B_{\max} - B_{\min})^2}{MSE} \tag{23}$$

$$MSE = \frac{1}{MN} \sum_{x,y} \left(L_1(x,y) - L_2(x,y)\right)^2 + \left(A_1(x,y) - A_2(x,y)\right)^2 + \left(B_1(x,y) - B_2(x,y)\right)^2 \tag{24}$$

Since our optimal color transform is just a 45-degree rotation in AOB plane, and L is the same as L^* in LAB space, we can estimate PSNR in our LAR space directly.

4.3 Evaluation Experiments

In our evaluation experiments, we use 6 popular RGB color test images, `baboon`, `barbara`, `goldhill`, `lena`, `peppers` and `airplane`. The bit rates we tested are 0.5, 0.25, 0.125, 0.0625, and 0.03125 bits per pixel (bpp). All the PSNRs of our experiments are listed in Table 1.

Table 1. Performance of lossy image compression in JPEG 2000 framework (PSNR, dB)

Name	0.5 bpp		0. 25 bpp		0.125 bpp		0. 0625 bpp		0. 03125 bpp	
	YcbCr	LAR	YcbCr	LAR	YcbCr	LAR	YcbCr	LAR	YcbCr	LAR
`lena`	28.11	**31.09**	25.80	**29.29**	23.44	**27.43**	21.30	**25.50**	18.88	**23.40**
`baboon`	18.30	**22.06**	16.66	**21.08**	15.54	**20.38**	14.70	**19.77**	13.99	**18.96**
`barbara`	25.61	**29.57**	22.46	**27.18**	19.75	**25.40**	17.94	**23.82**	16.37	**22.15**
`goldhill`	26.66	**30.50**	24.66	**28.99**	22.98	**27.53**	21.43	**26.26**	19.89	**24.96**
`peppers`	29.96	**33.94**	26.57	**30.52**	23.11	**26.64**	19.69	**23.00**	16.60	**19.99**
`airplane`	28.62	**32.16**	25.44	**29.26**	22.50	**26.54**	19.83	**24.10**	17.81	**22.22**
Average	26.21	**29.89**	23.60	**27.72**	21.22	**25.65**	19.15	**23.74**	17.26	**21.95**

Table 1 shows that our color space transform LAR performs better for all the images at all the bit rates, average 4.3dB higher, minimum 3dB and maximum 6dB better than the default color transform of JPEG 2000.

By subjective comparison, the compression with our color transform also looks better than that with the default YCbCr color transform.

5 Conclusions

Principal components analysis is applied to obtain the optimal color transforms for the images of 5 groups in the perceptually uniform color space, CIE LAB. The eigenvectors of the 5 image groups are close. The principal eigenvectors corresponding to the maximum eigenvalues are all approximately in the direction of the L^*. The other two eigenvectors are almost perpendicular to the L^* direction.

The color space K-L transform for an image group is useful for image retrieval and image coding as pre-processing. When an image is given in practice, an appropriate fixed optimal color transform can be applied after we simply recognize what class the image belongs to.

With hundreds of test images of diverse content, a general optimal K-L transform is obtained and can be used as a fixed optimal color transform for all the images. Our approximately optimal K-L transform LAR is actually elegant and simple. It is just 45-degree rotated LAB, and can be used as an image-independent color transform. As in our experiments with JPEG 2000, LAR performs better for lossy color image compression, and the average PSNR is about 4 dB higher than the default YCbCr.

With our LAR and a faster color space transform between RGB and CIE LAB, an approximate linear transform from RGB to LAR can be found for real-time applications and is more efficient and applicable, so it is our next investigation.

References

1. Kasson, J.M., Plouffe, W.: An Analysis of Selected Computer Interchange Color Spaces. ACM Transactions on Graphics, Vol. 11, No. 4 (1992) 373-405
2. Poynton, C.: Frequently Asked Questions about Color. www.inforamp.net/~poynton, (1999)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley InterScience (1999)
4. Pratt, W.K.: Spatial transform coding of color images. IEEE Transactions on Communication Technology, Vol. 19 (1971) 980-912
5. Connolly, C., Fliess, T.: A Study of Efficiency and Accuracy in the Transformation from RGB to CIELAB Color Space. IEEE Transactions on Image Processing, Vol. 6, No. 7 (1997) 1046-1048
6. Asakawa, K., Sugiura, H.: High-precision color transformation system. IEEE Transactions on Consumer Electronics, Vo.41, No.2 (1995) 304-312
7. Wallace, G.K.: The JPEG Still Picture Compression Standard. Communications of the ACM, Vol. 34, No. 4 (1991) 30-44
8. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 still image coding system: An overview. IEEE Trans. Consumer Electronics, Vol. 46, No. 4 (2000) 1103-1127

Lossless Compression of Color-Quantized Images Using Block-Based Palette Reordering

António J.R. Neves and Armando J. Pinho*

Dept. de Electrónica e Telecomunicações / IEETA
Universidade de Aveiro, 3810–193 Aveiro, Portugal
an@ieeta.pt, ap@det.ua.pt

Abstract. It is well-known that the lossless compression of color-indexed images can be improved if a suitable reordering of the palette is performed before encoding the images. In this paper, we show that, if this reordering is made in a block basis, then further compression gains can be attained. Moreover, we show that the use of block-based palette reordering can outperform a previously proposed block packing procedure. Experimental results using a JPEG-LS encoder are presented, showing how different reordering methods behave.

Keywords: Image compression, reordering techniques, lossless image coding

1 Introduction

Color-indexed images are represented by a matrix of indexes (the index image) and by a color-map or palette. For a particular image, the mapping between index values and colors is not unique — it can be arbitrarily permuted, as long as the corresponding index image is changed accordingly. Although equivalent in terms of representation, for most continuous-tone image coding techniques, such as JPEG-LS [1, 2] or lossless JPEG 2000 [3, 4], different mappings may imply dramatic variations in the compression performance.

With the aim of minimizing this drawback, several techniques have been proposed. Basically, they rely on finding an appropriate reordering of the color-map, such that the corresponding image of indexes becomes more amenable to compression. Regrettably, the problem of finding the optimal mapping seems to be computationally intractable [5]. This motivated several sub-optimal, lower complexity, proposals, such as those reported in [6, 5, 7, 8, 9, 10] (for a survey, that includes complexity comparison, see [11]).

In this paper, we show that the performance of palette reordering methods can be improved if used in a block basis. Moreover, we show that, in some cases, the use of block-based palette reordering can outperform the block packing procedure described in [12, 13]. To perform this evaluation, we used the JPEG-LS standard.

The remainder of this paper is organized as follows. In Section 2, we provide a brief presentation of the reordering methods that will be used. In Section 3 we describe the proposed method. In Section 4 we provide experimental results showing how the proposed method performs. Finally, in Section 5 we draw some conclusions.

* This work was supported in part by the Fundação para a Ciência e a Tecnologia (FCT).

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 277–284, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Palette Reordering

2.1 Luminance

The use of the luminance information conducts to a simple palette reordering method that is based only on the color palette. This method was proposed by Zaccarin *et al.* [6] in the context of lossy compression. It relies on the assumption that, generally, a given pixel has neighbors of similar luminance and, therefore, colors with similar luminance should have similar indexes. Reordering is performed by sorting the colors according to its luminance, computed according to

$$Y = 0.299R + 0.587G + 0.114B,$$

where R, G and B denote the intensities of the red, green and blue components, respectively.

2.2 The Modified Zeng's Method (mZeng)

Zeng *et al.* proposed an one-step look-ahead greedy approach for palette reordering, aiming at increasing the lossless compression efficiency of color-indexed images [7]. This algorithm starts by finding the index that is most frequently located adjacent to other indexes, and the index that is most frequently found adjacent to it. This pair of indexes is the starting base for an index list P_n that is constructed, one index at a time, during the operation of the reordering algorithm. If we denote by l_j the indexes already assigned to the index list and by u_i those still unassigned, then, just before starting the iterations, $P_2 = (l_1, l_2)$, where l_1 and l_2 form the pair of indexes mentioned above. New indexes can only be attached to the left or to the right extremity of the list.

We denote by u_L the index that satisfies

$$u_L = \arg\max_{u_i} D_L(u_i, N), \quad \text{with} \quad D_L(u_i, N) = \sum_{j=1}^{N} w_j \, C(u_i, l_j) \tag{1}$$

and by u_R the index satisfying

$$u_R = \arg\max_{u_i} D_R(u_i, N), \quad \text{with} \quad D_R(u_i, N) = \sum_{j=1}^{N} w_{N-j} \, C(u_i, l_j). \tag{2}$$

The function $C(i,j) = C(j,i)$ denotes the number of occurrences corresponding to pixels with index i that are spatially adjacent to pixels with index j. The weights w_j control the impact of the $C(i,j)$ on $D_L(u_i, N)$ and $D_R(u_i, N)$, and the summations are performed over all the N indexes already located in the index list $P_N = (l_1, l_2, \ldots, l_N)$.

The new index list will be given by $(u_L, l_1, l_2, \ldots, l_N)$ if $D_L(u_L, N) > D_R(u_R, N)$ or by $(l_1, l_2, \ldots, l_N, u_R)$ otherwise. Finally, the indexes in the index list are relabeled, creating $P_{N+1} = (l_1, l_2, \ldots, l_{N+1})$. This iterative process continues until all indexes are assigned to the index list. Then, the reordered image is constructed by applying the mapping $l_j \mapsto j - 1$ to all image pixels, and changing the color-map accordingly.

Zeng *et al.* [7] suggested that a reasonable choice for the weights w_j's is given by $w_j = \log_2(1+1/j)$, where j corresponds to the distance between the current left position of the index list and the position of index l_j in the index list.

A theoretical analysis of Zeng's method for the case of Laplacian distributed differences of neighboring pixels was presented in [10], leading to a set of parameters differing from that suggested in [7]. It was found that, under that Laplacian model, the process of building P_{N+1} from P_N has to be conducted in two steps. First, the index u satisfying (1) (or, equivalently, (2)) is determined, where the weights w_j (or w_{N-j}) are all equal (they can be set to one, for example). Then, the correct side of P_N to which the new index u should be attached is determined based on the sign of

$$\Delta = \sum_{j=1}^{N} (N+1-2j)C(u,l_j),$$

where if $\Delta > 0$ then the left side of P_N should be chosen, otherwise choose the right side.

2.3 Memon's Method

Memon *et al.* formulated the problem of palette reordering within the framework of linear predictive coding [5], with the objective of minimizing the zero-order entropy of the prediction residuals. They noticed that, for image data, the prediction residuals are often well modeled by a Laplacian distribution and that, in this case, minimizing the absolute sum of the prediction residuals leads to the minimization of the zero-order entropy of those residuals. For the case of a first-order prediction scheme, the absolute sum of the prediction residuals reduces to

$$\sum_{i=1}^{M}\sum_{j=1}^{M} C(i,j)|i-j|, \tag{3}$$

where, in this case, $C(i,j)$ denotes the number of times index i is used as the predicted value for a pixel whose color is indexed by j.

The problem of finding a palette reordering that minimizes (3) can be formulated as the optimization version of the linear ordering problem (also known as the minimum linear arrangement), whose decision version is known to be NP-complete [5]. In fact, if we consider a complete non-directed weighted graph $G(V,E,w)$, where each vertex in $V = \{v_1, v_2, \ldots, v_{|V|}\}$ ($|V| = M$) corresponds to a palette color, and $w(v_i,v_j) = C(i,j) + C(j,i), (v_i,v_j) \in E$ corresponds to the weight associated to the edge defined between vertices v_i and v_j, then the goal is to find a permutation $\sigma : V \to \{1,2,\ldots,|V|\}$ satisfying

$$\sigma = \arg\min_{\sigma_n} \sum_{(v_i,v_j)\in E} w(v_i,v_j)|\sigma_n(v_i) - \sigma_n(v_j)|. \tag{4}$$

For finding approximate solutions to this problem, Memon *et al.* proposed a heuristic, called pairwise merge, which is based on repeatedly merging ordered sets of colors until

obtaining a single reordered set. Initially, each color is assigned to a different set. Then, each iteration consists of two steps. First, the two sets A and B maximizing

$$\sum_{v_i \in A} \sum_{v_j \in B} w(v_i, v_j)$$

among all possible pairs of ordered sets are chosen. Then, a number of merging combinations of the sets A and B are tested (ideally, all possible combinations), and is chosen the one minimizing

$$\sum_{i=1}^{k} \sum_{j=1}^{k} w(u_i, u_j)|i-j|,$$

where $u_1, u_2, \ldots, u_k$ is the combined ordered set under evaluation. To alleviate the computational burden involved in selecting the best way of merging the two ordered sets, Memon *et al.* proposed to use a reduced number of configurations [5]. If $a_1, a_2, \ldots, a_r$ and $b_1, b_2, \ldots, b_s$ are the two ordered sets under evaluation, and if $r, s > 1$, then the following configurations are considered:

$$\begin{array}{c} a_1, a_2, \ldots, a_r, b_1, b_2, \ldots, b_s \\ a_r, \ldots, a_2, a_1, b_1, b_2, \ldots, b_s \\ b_1, b_2, \ldots, b_s, a_1, a_2, \ldots, a_r \\ b_1, b_2, \ldots, b_s, a_r, \ldots, a_2, a_1 \end{array}$$

Alternatively, if one of the sets has size one, then the following configurations are tested (without loss of generality, we consider $r = 1$):

$$\begin{array}{c} a_1, b_1, b_2, \ldots, b_s \\ b_1, a_1, b_2, \ldots, b_s \\ b_1, b_2, a_1, \ldots, b_s \\ \vdots \\ b_1, b_2, \ldots, b_s, a_1 \end{array}$$

3 Block-Based Approach

Previous work [12, 13] has shown that the performance of global palette reordering techniques can be improved if, in addition to reordering, we consider techniques that operate locally, in particular, block-based histogram packing methods.

Histogram packing can be seen as a preprocessing method that constructs a particular mapping of the image intensity values, where the order is preserved [14]. On the other hand, palette reordering also tries to find a mapping of the image intensity values but, in this case, more generic, and where the order is not necessarily preserved. This difference has a great impact on the side information necessary to reconstruct the original images. For histogram packing, only an indication of the existence of the intensities is needed (one bit per intensity/color suffices), whereas for the case of palette reordering the whole permutation has to be stored ($\log_2 M$ bits per intensity/color, for a M-color image).

Block-based palette reordering consists on applying reordering to each block of a given image partition. The basic idea is to address the palette reordering problem from a local point of view, i.e., on image regions, instead of addressing it globally.

Table 1. Compression results, in bits per pixel (bpp), of the color-indexed "Kodak" images. The lossless compression of the index images was performed using JPEG-LS, after reordering the index images by luminance. "Global" indicates global palette reordering, whereas "Block Packing" indicates histogram block packing after global palette reordering by luminance.

		Block Packing	
Image	Global	bpp	%
01	5.973	5.286	11.5
02	5.762	5.384	6.5
03	3.633	2.434	33.0
04	4.771	3.723	21.9
05	5.667	4.612	18.6
06	4.956	4.268	13.8
07	4.202	3.251	22.6
08	5.566	4.869	12.5
09	4.510	3.650	19.0
10	4.652	3.805	18.2
11	5.119	4.116	19.5
12	4.327	3.295	23.8
13	6.452	5.766	10.6
14	5.347	4.219	21.0
15	4.058	3.255	19.7
16	4.669	3.910	16.2
17	4.550	3.951	13.1
18	5.899	4.917	16.6
19	5.080	4.101	19.2
20	3.371	3.192	5.3
21	5.257	4.330	17.6
22	5.223	4.357	16.5
23	3.599	2.631	26.8
Average	**4.898**	**4.058**	**17.1**

4 Experimental Results

The experimental results that we present in this paper are based on a set of 23 true color images (768 columns $\times$ 512 rows) that we refer to as the "Kodak" images[1].

Using version 1.2.5 of the "`gimp`" program,[2] each image was color-quantized based on an image-dependent palette of 256 colors (generated by "`gimp`"). After color quantization, the index images have been reordered using the methods described is this paper, in a block basis. Results using global reordering and block-based histogram packing are presented for comparison.

The compression results that are presented include, besides the size of the encoded index image, the (uncompressed) size of the color table ($6\,144$ bits) and, for the block-based palette reordering approach, the overhead needed for storing the (uncompressed)

[1] These images can be obtained from `http://www.cipr.rpi.edu/resource/stills/kodak.html`.

[2] `http://www.gimp.org`.

Table 2. Compression results, in bits per pixel (bpp), of the color-indexed "Kodak" images. The lossless compression of the index images was performed using JPEG-LS after palette reordering using mZeng's technique. "Global" indicates global palette reordering. "Block Packing" refers to block histogram packing after global palette reordering using mZeng's method, whereas "Block Reordering" and "Block Reordering (Best)" indicate, respectively, the results of applying the proposed approach on square blocks of 128×128 pixels and of using square blocks with size indicated in column "bSize".

		Block Packing		Block Reordering		Block Reordering (Best)		
Image	Global	bpp	%	bpp	%	bpp	%	bSize
01	5.506	5.500	0.1	5.388	2.1	5.353	2.8	224
02	4.883	4.892	-0.1	4.786	1.9	4.786	2.0	128
03	2.618	2.586	1.2	2.283	12.7	2.283	12.8	128
04	4.175	3.879	7.0	3.641	12.8	3.632	13.0	96
05	4.956	4.841	2.3	4.699	5.1	4.694	5.3	112
06	4.824	4.676	3.0	4.341	9.9	4.296	10.9	160
07	3.493	3.475	0.4	3.323	4.8	3.270	6.4	192
08	5.659	5.407	4.4	5.163	8.7	5.101	9.9	176
09	3.915	3.894	0.5	3.719	5.0	3.645	6.9	176
10	4.703	4.262	9.3	3.969	15.6	3.954	15.9	112
11	4.479	4.358	2.6	4.110	8.2	4.083	8.8	144
12	3.847	3.639	5.4	3.269	15.0	3.262	15.2	112
13	6.318	6.079	3.7	5.648	10.5	5.648	10.6	128
14	4.586	4.492	2.0	4.206	8.2	4.206	8.3	128
15	3.538	3.420	3.3	3.128	11.5	3.107	12.2	144
16	4.567	4.385	3.9	4.021	11.9	4.021	11.9	128
17	4.668	4.535	2.8	4.261	8.7	4.250	9.0	112
18	5.228	5.044	3.5	4.764	8.8	4.764	8.7	128
19	4.819	4.620	4.1	4.164	13.5	4.164	13.6	128
20	3.436	3.426	0.2	3.178	7.4	3.167	7.8	192
21	4.684	4.610	1.5	4.289	8.4	4.269	8.9	112
22	4.791	4.524	5.5	4.393	8.3	4.365	8.9	112
23	2.991	2.767	7.5	2.565	14.2	2.560	14.4	144
Average	**4.465**	**4.318**	**3.2**	**4.057**	**9.1**	**4.038**	**9.5**	**–**

permutation of the colors for each block ($2\,048$ bits per block, a total of $49\,152$ bits for 128×128 blocks). The results referring to the block-based histogram packing procedure include the overhead needed for storing the (uncompressed) list of intensities that occur inside each image block (256 bits for each block).

Results using reordering by luminance are provided in Table 1. We note that, for this reordering method, if the palette is globally reordered, then it is also locally reordered, independently of the size and shape of the region. Therefore, it makes no sense to use block-based palette reordering by luminance, since only overhead is added without gaining in compression performance. However, the same does not apply to reordering methods that are based on statistics of the index images, such as Memon's and mZeng's methods, as the results in Tables 2 and 3 demonstrate. In fact, as can be observed, an increase in compression efficiency was obtained for all images, in comparison with

Table 3. Compression results, in bits per pixel (bpp), of the color-indexed "Kodak" images. The lossless compression of the index images was performed using JPEG-LS after palette reordering using Memon's technique. "Global" indicates global palette reordering. "Block Packing" refers to block histogram packing after global palette reordering using Memon's method, whereas "Block Reordering" and "Block Reordering (Best)" indicate, respectively, the results of applying the proposed approach on square blocks of 128×128 pixels and of using square blocks with size indicated in column "bSize".

		Block Packing		Block Reordering		Block Reordering (Best)		
Image	Global	bpp	%	bpp	%	bpp	%	bSize
01	5.368	5.412	-0.8	5.170	3.6	5.153	4.0	144
02	4.737	4.781	-0.9	4.666	1.5	4.634	2.2	160
03	2.433	2.445	-0.4	2.166	10.9	2.149	11.7	224
04	3.812	3.632	4.7	3.419	10.2	3.410	10.5	112
05	4.861	4.705	3.2	4.530	6.8	4.530	6.8	128
06	4.659	4.395	5.6	4.220	9.4	4.161	10.7	160
07	3.266	3.261	0.1	3.103	4.9	3.072	5.9	176
08	5.278	4.981	5.6	4.979	5.6	4.955	6.1	176
09	3.747	3.810	-1.6	3.606	3.7	3.571	4.7	320
10	4.387	4.076	7.0	3.764	14.1	3.764	14.2	128
11	4.206	4.148	1.3	3.960	5.8	3.960	5.8	128
12	3.629	3.444	5.0	3.144	13.3	3.144	13.4	128
13	6.013	5.782	3.8	5.569	7.3	5.569	7.4	128
14	4.177	4.136	1.0	4.032	3.4	4.013	3.9	176
15	3.265	3.177	2.7	2.983	8.6	2.960	9.4	176
16	4.148	4.022	3.0	3.847	7.2	3.847	7.3	128
17	4.128	4.088	0.9	3.989	3.3	3.976	3.7	240
18	5.046	4.893	3.0	4.522	10.3	4.522	10.4	128
19	4.271	4.245	0.6	3.974	6.9	3.973	7.0	144
20	3.131	3.219	-2.8	3.063	2.1	3.011	3.8	192
21	4.399	4.373	0.5	4.092	6.9	4.092	7.0	128
22	4.726	4.417	6.5	4.254	9.9	4.254	10.0	128
23	2.747	2.591	5.6	2.442	11.1	2.424	11.8	144
Average	**4.193**	**4.088**	**2.5**	**3.891**	**7.2**	**3.876**	**7.5**	**–**

both the globally palette reordered approach and the block-based histogram packing approach. The columns under the label "Block Reordered (Best)" in Tables 2 and 3 contain the compression values corresponding to the size of the blocks that provide the highest compression for each image.

5 Conclusions

Color palette reordering is a very effective approach for improving the compression of color-indexed images. In this paper, we provided experimental results showing that the lossless compression of color-indexed images can be further improved if palette reordering is applied in a block basis. With this new approach, additional lossless compression improvements of around 9% for the modified Zeng's technique and of 7% using Memon's

technique have been attained, in relation to global palette reordering. In absolute terms, we also conclude that Memon's technique provides the best results. Nevertheless, we should have into account that the computational time necessary to perform the reordering task is proportional to the number of image blocks, which might put the other approaches into an overall competitive position.

References

1. ISO/IEC 14495–1 and ITU Recommendation T.87: Information technology - Lossless and near-lossless compression of continuous-tone still images. (1999)
2. Weinberger, M.J., Seroussi, G., Sapiro, G.: The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS. IEEE Trans. on Image Processing **9** (2000) 1309–1324
3. ISO/IEC International Standard 15444–1, ITU-T Recommendation T.800: Information technology - JPEG 2000 image coding system. (2000)
4. Skodras, A., Christopoulos, C., Ebrahimi, T.: The JPEG 2000 still image compression standard. IEEE Signal Processing Magazine **18** (2001) 36–58
5. Memon, N.D., Venkateswaran, A.: On ordering color maps for lossless predictive coding. IEEE Trans. on Image Processing **5** (1996) 1522–1527
6. Zaccarin, A., Liu, B.: A novel approach for coding color quantized images. IEEE Trans. on Image Processing **2** (1993) 442–453
7. Zeng, W., Li, J., Lei, S.: An efficient color re-indexing scheme for palette-based compression. In: Proc. of the 7th IEEE Int. Conf. on Image Processing, ICIP-2000. Volume III., Vancouver, Canada (2000) 476–479
8. Battiato, S., Gallo, G., Impoco, G., Stanco, F.: A color reindexing algorithm for lossless compression of digital images. In: Proc. of the IEEE Spring Conf. on Computer Graphics, Budmerice, Slovakia (2001) 104–108
9. Spira, A., Malah, D.: Improved lossless compression of color-mapped images by an approximate solution of the traveling salesman problem. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-2001. Volume III., Salt Lake City, UT (2001) 1797–1800
10. Pinho, A.J., Neves, A.J.R.: A note on Zeng's technique for color reindexing of palette-based images. IEEE Signal Processing Letters **11** (2004) 232–234
11. Pinho, A.J., Neves, A.J.R.: A survey on palette reordering methods for improving the compression of color-indexed images. IEEE Trans. on Image Processing (2004 (in press))
12. Pinho, A.J., Neves, A.J.R.: Block-based histogram packing of color-quantized images. In: Proc. of the IEEE Int. Conf. on Multimedia and Expo, ICME-2003. Volume 1., Baltimore, MD (2003) 341–344
13. Pinho, A.J., Neves, A.J.R.: JPEG 2000 coding of color-quantized images. In: Proc. of the IEEE Int. Conf. on Image Processing, ICIP-2003. Volume 2., Barcelona, Spain (2003) 181–184
14. Pinho, A.J.: An online preprocessing technique for improving the lossless compression of images with sparse histograms. IEEE Signal Processing Letters **9** (2002) 5–7

Fovea Based Coding for Video Streaming

Çağatay Dikici[1], H. Işıl Bozma[1], and Reha Civanlar[2]

[1] Intelligent Systems Laboratory,
Electrical and Electronics Engineering Department
Boğaziçi University, 34342, Bebek, Istanbul, Turkey
{dikicica, bozma}@boun.edu.tr
http://www.isl.ee.boun.edu.tr
[2] Computer Engineering Department
Koç University, 34450, Istanbul, Turkey
rcivanlar@ku.edu.tr

Abstract. Attentive robots, inspired by human-like vision – are required to have visual systems with fovea-periphery distinction and saccadic motion capability. Thus, each frame in the incoming image sequence has nonuniform sampling and consecutive saccadic images have temporal redundancy. In this paper, we propose a novel video coding and streaming algorithm for low bandwidth networks that exploits these two features simultaneously. Our experimental results reveal improved video streaming in applications like robotic teleoperation. Furthermore, since the algorithm employs the Gaussian-like resolution of human visual system and is extremely simple to integrate with the standard coding schemes, it can also be used in applications such as cellular phones with video.

1 Introduction

Motivated by biological vision systems, there has been a growing trend to have robots explore their environment and look around in an attentive manner – thereby minimizing the amount of collected information and thus reducing the required computation considerably [1,2,3]. In order to remotely communicate with such a robot and see its visual surroundings, one needs real-time video transmission – preferably using standard coding schemes. Interestingly, video broadcasted from a robot with such attention capabilities has two intrinsic properties which are not exploited in general by the standard compression algorithms: i.) Fovea-periphery distinction, which leads to varying spatial resolution within each frame; and ii.) Saccadic motion between foveae, as a result of which there is much overlapping information between consecutive frames. For example, in H.263 [4] all blocks in a frame and between frames are coded with the same priority. In this paper, we present a real-time video algorithm that exploits both the spatial and temporal redundancy that exist in the video sequences and thus can be used for attentive robot-video streaming over internet. The novelty of this work is twofold: i)Simultaneous foveal and temporal compression; and ii) Easy integration with any of the standard compression technology – without

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 285–294, 2004.
© Springer-Verlag Berlin Heidelberg 2004

requiring any alterations in these standards. Furthermore, this approach can be used in general in any low-bandwidth video transmission since it matches the fall-off resolution of the human visual system.

1.1 Related Literature

Attentive robots explore their surroundings in a loop of pre-attention and attention [2]. The aim of the pre-attention stage is to determine the next attention regions. This is achieved through the fovea-periphery mechanism [5]. Unlike traditional cameras, the distribution of receptor cells on the retina is Gaussian-like with a small variance, resulting in a loss of resolution as we move away from the optical axis of the eye [6]. The fovea is the small region of highest acuity around the optical axis and the rest of the retina is called periphery. Saccades - very rapid jumps of optical axis - are used to bring images of chosen objects to fovea where resolution of fine visual detail is at its best. Non-orthogonal spatial filter responses can be used for determination of saccadic movements of foveae in still images [7]. The remote access (in particular internet based) and teleoperation of such a robot requires real-time transmission of thus generated video – which has varying resolution and redundancy. Consequently, video streaming methods that exploit these properties become crucial.

The application of varying resolution to video coding and streaming is relatively new [8]. As outlined therein, this approach presents several distinct advantages such as guaranteed compression ratios and speed. Foveal and peripheral regions are coded differently in spatial domain, and the priority assignment of the ATM cells are used for transmitting the regions of video frames with varying priorities [8]. However, in the case of a network congestion, peripheral information which attracts relatively lower attention are first lost. However, since the approach depends solely on the Quality of Service (QoS) parameters, it will potentially have problems on best effort systems like internet. Furthermore, the redundancy along the temporal dimension is not utilized at all. An approach that applies both spatial & temporal processing on MPEG2 streams has been presented in [9]. DCT coefficients of the periphery are quantized in order to reduce the length of the bitstream. However, since DCT coefficients are calculated on 8×8 blocks, the foveal region definition is limited with rectangular shape. In order to minimize the blocking artifacts between the foveal and the peripheral regions, image pyramids and raised-cosine blending are used in [10]. However, such an approach requires the generated pyramids also to be transmitted through the channel then implied increased bandwidth. Space-variant approaches such as log-polar mapping of original frame can also be applied for foveation [11], but the increased computation for achieving such transformations impede real-time applications. Finally, if methods that take particular coding algorithms like H.263, MPEG-4 and and JVT into account do not offer compression independent solutions.

1.2 Problem Statement

Suppose that the visual task involves a robot looking at a scene in an attentive manner and a video is generated meanwhile. The objective can be defined as real-time transmission of this data over the Internet so that users can see precisely what the robot is seeing. Moreover, the system should also: i.) allow real-time streaming and ii.) be usable with any particular video compression algorithm.

2 Approach

Consider an incoming image sequence. Let I_v^t denote the visual field image at time t. The function $c^t : I_v^t \rightarrow C$ maps each pixel in this region to a value from the color space C. The fovea is represented by $I_f^t \subset I_v^t$. The next fovea I_f^{t+1} at time $t+1$ is chosen from the the set of candidate foveae $C(I_v^t)$ – as determined from the visual field. For each candidate fovea $I_c \in C(I_v^t)$, an attention criteria $a : I_c \rightarrow R^+$ is computed. The attention criteria is a scalar valued function of interest based on the presence of simple features with low computational requirements. This function is determined by the measure of interest on that frame and its definition will vary depending on the measure of interest such as color, intensity or human facial features. The candidate fovea that maximizes this measure is then designated to be the next fovea as:

$$I_f^{t+1} = arg \max_{\forall I_c \in C(I_f^t)} a(I_c) \tag{1}$$

The robot then moves its camera as to fixate on this newly determined fovea. Such camera movements correspond to saccadic eye movements in humans. As a result, a raw image sequence $\{I_v^0, ..., I_v^t, I_v^{t+1}, ...\}$ is generated.

In general, most video coding algorithms achieve compression by applying transforms on the original sequence that exploit spatial redundancies such as discrete cosine transform (DCT). In attentive vision, foveal and peripheral regions of the incoming visual field each incoming visual field I_v^t is processed differently. The visual data as defined by the color map c in the incoming fovea is preserved in the fovea while that of the periphery $I_p^t = I_v^t - I_f^t$ is transmitted with lower resolution. Consequently, a new color map $\tilde{c}_S^t : I_v^t \rightarrow C$ is defined as:

$$\tilde{c}_S^t(x) = \begin{cases} c^t(x) & \text{if } x \in I_f^t \\ f_S * c^t(x) & \text{if } x \in I_p^t \end{cases} \tag{2}$$

where $f_S : I_p^t \rightarrow \tilde{I}_p^t$ is a spatial filter function. The main idea in choosing this filter is that since the peripheral pixels do not attract our attention, the high frequency contained therein is not important and thus can be removed from the data. Consequently, the corresponding image areas can be coded with fewer DCT coefficients. For example, low pass filters such as Gaussian and blurring filters can be utilized.

However, with such a definition, spatial edge artifacts will appear in the reconstructed image after transmission. In order to minimize this, the color map $\tilde{c}_S$ is modified to include α-blending:

$$c_S^t(x) = \alpha^t(x)c^t(x) + (1 - \alpha^t(x))f_S * c^t(x) \tag{3}$$

The blending function $\alpha^t : I_v^t \to [0, 1]$ is time dependent function whose value at time t varies between 0 and 1. The values of α^t are 1 or close to 1 on the fovea and converge towards 0 for peripheral pixels as a function of their proximity to the fovea. Consequently, a gradual color transition is introduced in the fovea periphery neighborhood. Furthermore, introducing such a blending enables us to define non-rectangular shaped foveal regions.

As the camera saccades from the current fovea to the next, an image sequence $\{I_v^0, ..., I_v^t, I_v^{t+1}, ...\}$ is generated. If temporal sampling is fast enough, there is much temporal redundancy between saccadic image frames. In general, most video coding algorithms also exploit temporal redundancies in their compression schemes. In motion compensation based coding, each frame I_v^{t+1} can be represented as the difference of the current frame and the previous one I_v^t, and thus be coded using Motion Vectors (MV). As expected, the efficiency of this type of coding goes up with increased temporal redundancy. In attentive vision, we can increase the temporal redundancy between frames depending on whether foveal or peripheral regions are under consideration. Since the fovea is subject to close scrutiny, all minute detail changes between frames should be preserved and no new temporal redundancy can be introduced. However, this is not the case for periphery. Since changes between frames in the peripheral regions are ignorable to some extent, temporal redundancy can be increased by applying filters across temporal dimension. For example, color maps in the periphery can be updated with every K saccades.In doing so, since the current peripheral region can be estimated from the previous one, the length of the bitstream using MVs is reduced considerably. In this case, the temporal color map $c_T^t : I_v^t \to C$ is defined as:

$$c_T^t(x) = \begin{cases} c^t(x) & \text{if } x \in I_f^t \\ c_S^{t-tmodK}(x) & \text{if } x \in I_p^t \end{cases} \tag{4}$$

If both spatial and temporal redundancies are taken into account, the resulting color map becomes a composition of the two functions c_T^t and c_S^t as $c_T^t \circ c_S^t : I_v^t \to C$.

3 Experiments

Our approach is implemented on video streaming from APES - an attentive robot developed in our laboratory [12,13]. APES can be remotely controlled and teleoperated and streams its acquired image sequence over the internet. Hence, any registered user can connect and watch the visual field of the APES robot as it explores its current surroundings. One authorized user can also control the APES remotely while watching its captured video in real time. In this setup, Helix^{TM}

Project [14] is used as video coding and streaming framework. RTP/UDP/IP is used for transmission of real-time data, and Real Time Session Protocol(RTSP) is used for session initiation and control of the video stream. Furthermore, the remote control of the APES (pan & tilt controls) is sent via TCP/IP for lossless data transmission.

In order to quantify the visual quality of the foveal system, two metrics are used: The first metric is Foveal Mean Square Error(FMSE) which is similar to Mean Square Error (MSE), but is defined only on the fovea[1]

$$FMSE = \frac{1}{\mid I_f^t \mid} \sum_{n=1}^{x \in I_f^t} [c_T^t \circ c_S^t - c^t]^2 \tag{5}$$

However, it is well known that MSE value has a clear physical meaning in statistical sense, but it may not always reflect perceived visual quality [15].

As an alternative, Structural Similarity Measure (SSIM) - a metric capturing the amount of structural degradation between two images has been proposed in [15]. In this metric, luminance, contrast, and structural components of the two images are weighted and a quality index is generated. Physical meaning of SSIM can be explained such that it is a metric based on the comparison of mean, standard deviation and correlation coefficient of the normalized variance of two images. However, since the focus is on the fovea, we use a modified version Foveal Structural Similarity Measure(FSSIM) which is defined only on the fovea:

$$FSSIM = \frac{(2\mu_{c_T^t \circ c_S^t}\mu_{c^t} + C_1)(2\sigma_{(c_T^t \circ c_S^t)c^t} + C_2)}{(\mu^2_{c_T^t \circ c_S^t} + \mu^2_{c^t} + C_1)(\sigma^2_{c_T^t \circ c_S^t} + \sigma^2_{c^t} + C_2)} \tag{6}$$

μ, and σ are the mean and variance of the respective images within the fovea. C_1 and C_2 are used in order to prevent unstable conditions if the values $\mu^2_{c_T^t \circ c_S^t} + \mu^2_{c^t}$ or $\sigma^2_{c_T^t \circ c_S^t} + \sigma^2_{c^t}$ are very close to 0. $C_1 = 0.01$ and $C_2 = 0.03$. Note that $FSSIM \leq 1$ with equality holding if and only if the source and the target images are identical.

Finally, note that FMSE value indicates the mean square error of the fovea region. If the FMSE value increases, so does the quantity of error increase. On the other hand, FSSIM value indicates the similarity measure of the fovea region by using HVS properties. The more similar the input and the reference frames are, its value gets closer to '1'; and similarly this value being '0' indicates the contrary case.

In order to evaluate the performance of the algorithm, an extensive statistical comparative study was conducted. First, the APES Video Database was created by making the robot look at scenes consisting primarily of a person – talking and mimicking at 3 different distances (long, intermediate, short distances) and 3 different poses (left, right, frontal views)[16]. Each incoming video

[1] FMSE is a newly defined quality metric. We leave it to the vision science researchers to check its validity but for our applications, based on visual observations, it seems to be meaningful.

was recorded in 384 × 288 resolution, 20 second long, $25 frames/sec$ RGB video in uncompressed AVI format without any preprocessing. Next, the videos in this database were subjected to the following preprocessing:

1. Twelve video sequences are selected randomly from the APES Database - with 4 of long, intermediate and short distance category respectively.
2. Next, for each video sequence, the foveal area is determined after visual examination of the video sequence and ensuring that the fovea overlaps with the image area containing the person's face. For each category, 2 different fovea sizes are considered:
   - For *long-distance* ($l > 4m$) sequences, these are taken to be 100 × 100 and 130 × 130 pixels;
   - For *intermediate distance* ($l = 4m$) sequences, they are 130 × 130 and 160 × 160 pixels;
   - For *short distance sequences* ($l = 2m$), they are taken to be 160 × 160 190 × 190 pixels respectively.
3. Each sequence is first only spatially processed - using the 2 different fovea sizes. In spatial processing, $5x5$ box blur filter is selected as spatial filter function f_S. For α-blending the transition width is chosen as 5.
4. Each sequence is next spatio-temporally processed for the two different fovea sizes. K is selected as three in this process.
5. The original raw video and the two preprocessed videos are then are then encoded with Real-Media [14] codec with bit-rates $25k/sec$ and $35k/sec$. A sample frame is shown in Fig. 1

Fig. 1. Sample frames - Left to right: No preprocessing ,spatial only and spatio-temporal preprocessing.

In analysis part, all the encoded video frames are compared with the original input raw frames. The comparison is performed as follows: First, for each encoded frame, MSE,FMSE,SSIM,FSSIM values are calculated. Since the number of encoded frames and the number of input raw frames may vary because of the encoding process, minimum FMSE values within 5 frame neighborhood of input raw frames are selected as reference frames. For each sequence, the first 450 frame statistics are stored. Figure 2 presents FMSE and FSSIM values of a video sequence with a 130 × 130 sized fovea with classical coding, with spatial and spatio-temporal coding. As expected, compared to classical coding, spatial

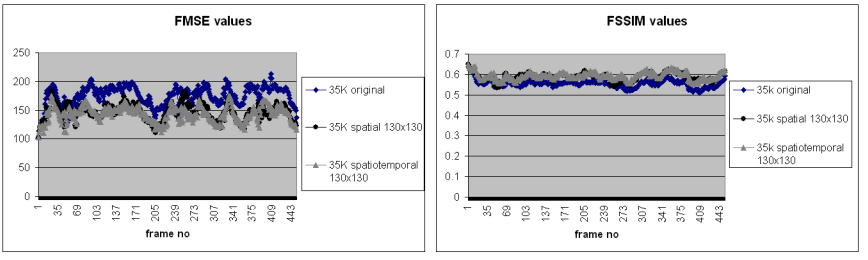


Fig. 2. FMSE &FSSIM Values.

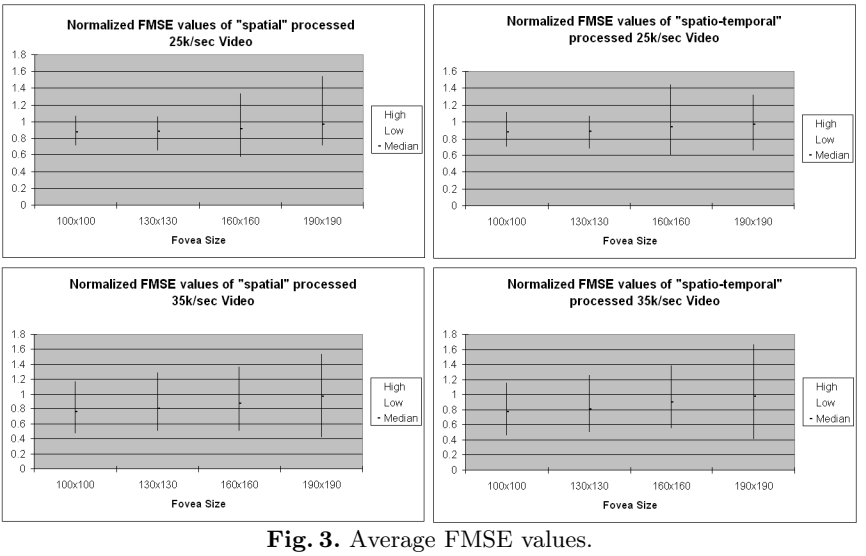


Fig. 3. Average FMSE values.

and spatio-temporal coding improve the FMSE and FSSIM values considerably. However, there is not much added performance between just spatial and spatio-temporal coding schemes – possibly due to the temporal filtering selected for our particular application. The performance should improve with a more appropriately selected filter.

Next, statistical metrics are gathered for each processed sequence. FSSIM and FMSE values are first normalized by dividing each frame's FSSIM and FMSE values by their originally coded frames FSSIM and FMSE values respectively and then averaged over the video. Finally, average normalized FSSIM and normalized FMSE values for each fovea size and encoding bit-rate is computed. For each group, the mean, minimum and maximum values are calculated – using the 2700 frames in each group. Fig. 3 presents the FMSE results. The FMSE values for both the spatial and spatial-temporal processing are all less than one. As expected, the quality goes down as the fovea size is increased. Furthermore, these values are improved for higher bit rates as can be seen by comparing the results for the 25k/sec and 35k/sec video streaming respectively. Similarly, the

FSSIM values are shown in Fig. 4. Averaged FSSIM values are greater than one, which is consistent with the FMSE outputs.

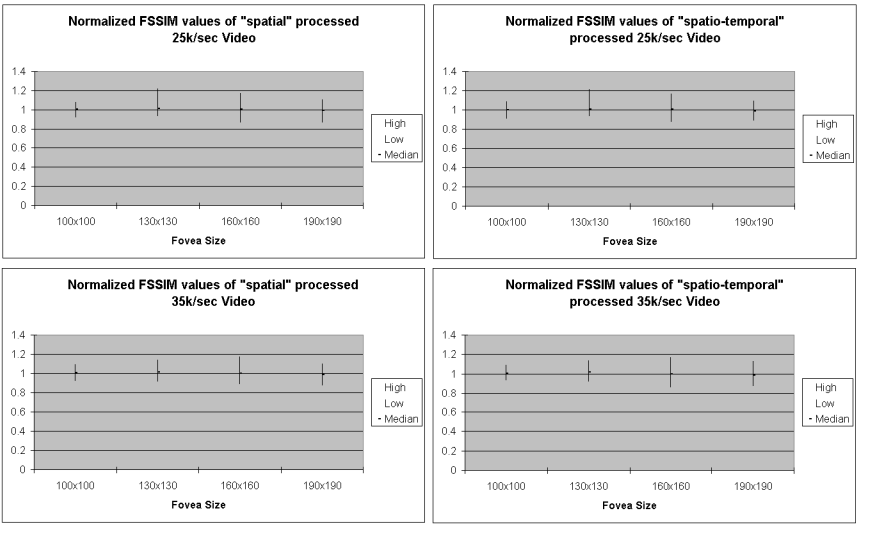


Fig. 4. Average FSSIM values.

We also computed the required processing overhead for the spatial and spatial-temporal coding schemes in order to check its suitability for real-time applications (on a Pentium II/1000 Mhz with 224 MB RAM). Each fovea size was considered seperately using randomly selected 200 frames. The results are as shown in Table 1. First of all, it is observed that the worst is about 10msec – which is quite acceptable with the frame rate of 25 frames/sec. As expected, the overhead is reduced considerably with the spatial-temporal coding.

Table 1. Computational overhead statistics.

Fovea size	Overhead (msec) for spatial coding	Overhead (msec) for spatial-temporal coding (K=3)
100 × 100	9.39	5.3
130 × 130	9.64	5.29
160 × 160	9.47	5.48
190 × 190	9.48	5.55

4 Conclusion

In this paper, we present a fovea based coding scheme for video streaming through low bandwidth networks that exploits two important aspects of human vision: fovea-periphery distinction and saccadic motion. Thus, each frame in the acquired image sequence has nonuniform sampling and consecutive saccadic images have temporal redundancy. Such a coding scheme is suitable for applications such as video broadcasting from attentive robot or cellular phones with video where the perceiver fixates on objects in a continual manner. Our experimental results indicate that compared to classical coding, spatial and spatio-temporal coding improve the transmission quality. For our future work, we will work on methods for increasing temporal redundancy through careful generation of the saccadic movements.

Acknowledgment. This work has been supported by Bogazici University Scientific Research Projects Grant #BAP02S106. We gratefully acknowledge Doğan Can for his meticulous collection of the APES video database and helping with the experiments. We also gratefully acknowledge Oksijen Teknoloji A.S. for supporting Çağatay Dikici's project and completing his masters thesis.

References

1. Akins, K. (ed.): Perception, pages 290-316. Oxford University Press, (1996)
2. Ballard, D.H., Brown, C.M.: "Principles of Animate Vision", CVIP: Image Understanding, (1992), 56
3. Koch, C, Itti, L.: "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention", Vision Research, (2000)
4. ITU-T Recommendation H.263: "Video coding for low bit rate communication", (1996)
5. Kowler, E. (ed.): "Eye Movements and Their Role in Visual and Cognitive Processes", Elsevier, (1990)
6. Gouras, P., Bailey, C.H.: "The retina and phototransduction". In J. H. Schwartz and E. R. Kandel, (eds.), Principles of Neural Science. Elsevier, (1986)
7. Takacs, D., Wechsler, H.: "A Dynamic and Multiresolution Model of Visual Attention and Its Application to Facial Landmark Detection", Computer Vision and Image Understanding, Vol. 70. No. 1 ,(1998), 63-73
8. Wiebe, K., Basu, A.: "Improving image and video transmission quality over ATM with foveal priorization and priority dithering" Pattern Recognition Letters, (2001), 22
9. Reeves, T.H., Robinson, J.A.: "Rate Control of Foveated MPEG Video", CCECE (1997)
10. Geisler, W.S., Perry, J.S.: "A Real-time Foveated Multiresolution System for Low-bandwidth Video Communication", SPIE Proceedings: Human Vision and Electronic Imaging, Vol. 3299, (1998), 294-305
11. Grosso, E, Manzotti, R., Tiso, R., Sandini, G.: "A Space-Variant Approach to Oculomotor Control", Proceedings of International Symposium on Computer Vision, (1995), 509-514

12. Soyer, C., Bozma, H.I., Istefanopulos, Y.: "Apes: Actively Perceiving Robot", Proceedings of IEEE/RSJ International Conference on Robots and Systems, Lausanne, Switzerland, (2002)
13. Soyer, C., Bozma, H.I., Istefanopulos,Y.: "Attentional Sequence Based Recognition: Markovian and Evidential Reasoning", IEEE Transactions on Systems, Man and Cybernatics, Vol. 33, No. 6, (2003), 937-950
14. Helix Encoder, http://www.helixcommunity.org
15. Whang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: "Image Quality Assessment: From Error Measurement to Structural Similarity", IEEE Transactions on Image Processing, Vol. 13, (2004), No. 1
16. "APES Video Database", http://www.isl.ee.boun.edu.tr/Apes/VideoDatabase.html

Influence of Task and Scene Content on Subjective Video Quality

Ying Zhong[1], Iain Richardson[1], Arash Sahraie[2], and Peter McGeorge[2]

[1] Image Communication Technology Group, The Robert Gordon University, UK
{y.zhong; i.g.richardson} @rgu.ac.uk
http://www.rgu.ac.uk/eng/ict/
[2] Vision Research Lab, University of Aberdeen, UK
{a.sahraie; mcgeorge}@abdn.ac.uk

Abstract. The influences of task and scene content on a viewer's subjective opinion of compressed video quality are investigated. A group of test subjects are presented with a number of compressed video clips, with or without an initial "task" instruction. Subjective quality ratings and selected eye movement tracking results are recorded. The results indicate that subjective quality is strongly influenced by the presence or absence of distortion in foreground human figures in a video scene and is also influenced by the presence or absence of an initial task. The implications of these results for subjective quality testing and for the design of video compression systems are discussed.

1 Introduction

Video compression is a key technology for multimedia communication applications such as videoconferencing and video streaming. Existing and emerging video coding standards such as MPEG-4 Visual [1] and H.264 [2] offer efficient compression performance but designers are still faced with the challenge of optimising the trade-off between video quality and network bitrate [3]. A common goal of a video coding system is to provide optimal visual quality within constraints of coded bitrate and computational complexity. In order to optimize the performance of a video CODEC or to compare alternative coding algorithms, it is necessary to accurately measure or model visual quality. Fundamentally, this is the quality of a visual scene as perceived by a human observer and so a complete "picture" of visual quality requires human visual perception mechanisms to be taken into account.

The Human Visual System (HVS) is complex, multi-layered and not fully understood. At a low level of operation, the HVS responds to movement, colour, texture, orientation and shape in ways that can be measured and defined [4]. At a higher level, visual perception is strongly influenced by factors such as the content of a visual scene, the task carried out by the observer (e.g. passive observation or an active search task) and previous experience or learning. For example, perceiving a visual scene typically involves a series of fixations (gaze points), each bringing part of the

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 295–301, 2004.
© Springer-Verlag Berlin Heidelberg 2004

scene into the relatively small foveal region of vision (in which the scene is perceived at a high spatial resolution). It has been demonstrated that the pattern of visual search (i.e. the fixation pattern) depends on the user's task and the scene content [5]. It is therefore reasonable to assume that a user's perception of visual quality is also likely to be related to these factors.

The established practice in current video compression research is to evaluate and optimize performance in terms of objective measures such as Peak Signal to Noise Ratio (PSNR), a measure based on the mean squared error (MSE) between a source and decoded video frame. For example, [6] describes a method of controlling a video encoder to optimize rate and distortion, where distortion is measured as the sum of squared differences (SSD) between original and decoded samples. Metrics based on MSE or SSD incorporate an implicit assumption that all samples of a video frame (and all frames of a video sequence) are equally "important" to visual quality, an assumption that is far removed from what is currently known about the operation of the Human Visual System. A video CODEC that is optimized with respect to MSE or SSD (i.e. minimizing squared distortion across a video sequence) may therefore be sub-optimal in terms of visual quality perceived by a human observer [7].

More sophisticated objective quality measures have been proposed [8], but to date there is no clear alternative to subjective assessment of visual quality. Subjective video quality measurement methods have been standardized by the ITU [9,10]. A typical test scenario involves a human observer watching a series of video clips and rating each clip in terms of its visual quality (with or without comparison to a "reference" version of the clip). These standardized test methods do not take account of the task that a viewer may have to perform (e.g. searching for an object, reading text, following a conversation, etc).

The aim of this research is to investigate the influence of task instructions and scene content on a human observer's opinion of visual quality. There is some evidence that very specific tasks influence attention and therefore the perception of quality in different parts of a visual scene, for example in [11] it was demonstrated that deaf people observing sign language attend almost exclusively to the signer's face and therefore perceive the face in high visual acuity and other parts of the scene in peripheral vision. At the same time, various techniques have been proposed to facilitate "prioritized" video coding in which regions or objects in a video scene may be coded with varying levels of compression and hence varying decoded quality, including object-based coding and foveated coding [12]. A better understanding of the relationships between task, scene content and perceived quality may make it possible to apply this type of coding technique in order to improve the subjective quality of coded video.

2 Method

In this experiment, participants are asked to rate the perceived quality of a set of video clips. Participants are given instructions prior to watching some (but not all) of the clips. 9 test video clips were used in the experiment (fig.1), each with one or two

human figures on the left or right of the picture and a static or moving background. In each case, the main figure is talking "to" the camera. Each clip is 10 seconds long (captured at 25 frames per second) with a luma resolution of 352x288.

Fig. 1. The 9 test sequences (A, B, C, D, E, F, G, H, I)

Three versions of each of the 9 clips were produced: "low compression", "highly compressed figure" and "highly compressed background". The "low compression" version was encoded and decoded using an H.263 CODEC with a quantization parameter of 4. The section of the scene with the main figure (left or right hand side) was encoded and decoded with a quantization parameter of 24 (high compression) and the remainder of the scene (the background) was encoded with a quantization parameter of 4 to produce the "highly compressed figure" version. For the "highly compressed background" version, the background was encoded with quantization parameter 24 and the figure with quantization parameter 4.

Fig.2 shows a sample frame from each of the 3 versions of sequence E. The original audio track for each clip was retained without any compression or modification.

Fig. 2. Sequence E (highly compressed figure; highly compressed background; low compression)

During the experiment, each video clip was displayed full-screen on a 15" LCD display with a viewing distance of approximately 3H (where H is the height of the displayed picture) [10]. A total of 81 observers took part in the experiment. Each observer was presented with the 9 video clips (each participant saw only one of the three coded versions of each clip) and given one of the following task instructions prior to watching each clip:

(a) Concentrate on the main figure in the video clip.
(b) Concentrate on the background of the video clip.
(c) No instruction (neutral case).

After viewing the clip, the observer was asked to rate the subjective quality of the clip using a 7-point scale from 1 (very bad quality) to 7 (excellent quality) using the Single Stimulus test method described in [10]. In addition, the observer was asked to describe what they could remember of the scene.

In order to verify whether observers were carrying out the given instructions, the eye movements of 4 observers were captured whilst watching 2 of the test sequences. An Arrington Research ViewPoint eye tracking system was used to record eye movements at a sample rate of 60Hz.

3 Experimental Results

3.1 Subjective Test Results

The subjective rating results for all participants and all sequences were combined to produce the mean subjective ratings shown in fig.3.

As expected, the "low compression" sequences (QP=4) received the highest mean rating, with a high mean score when no initial instruction was given and a lower score when the subject was instructed to concentrate on the background or on the foreground.

The sequences with a highly-compressed background (QP=24 for the background, QP=4 for the main figure) received a lower mean quality rating than the "low compression" sequences. For these sequences, a similar mean score was obtained for no instruction or for the instruction "concentrate on the main figure" whilst a lower mean score was obtained for the instruction "concentrate on the background".

The lowest mean score was given to the sequences with a highly-compressed main figure (QP=4 for the background, QP=24 for the main figure). Within this category, the instruction "concentrate on the background" produced the highest score and the instruction "concentrate on the main figure" produced the lowest mean score.

3.2 Eye Movement Results

Table 1 summarizes the eye movement data collected from 4 subjects observing two of the test video sequences. The Table records the mean time spent observing the main figure or the background of each clip. These results demonstrate that the observers tended to follow the initial instructions, spending most of the time looking at the required area (figure or background).

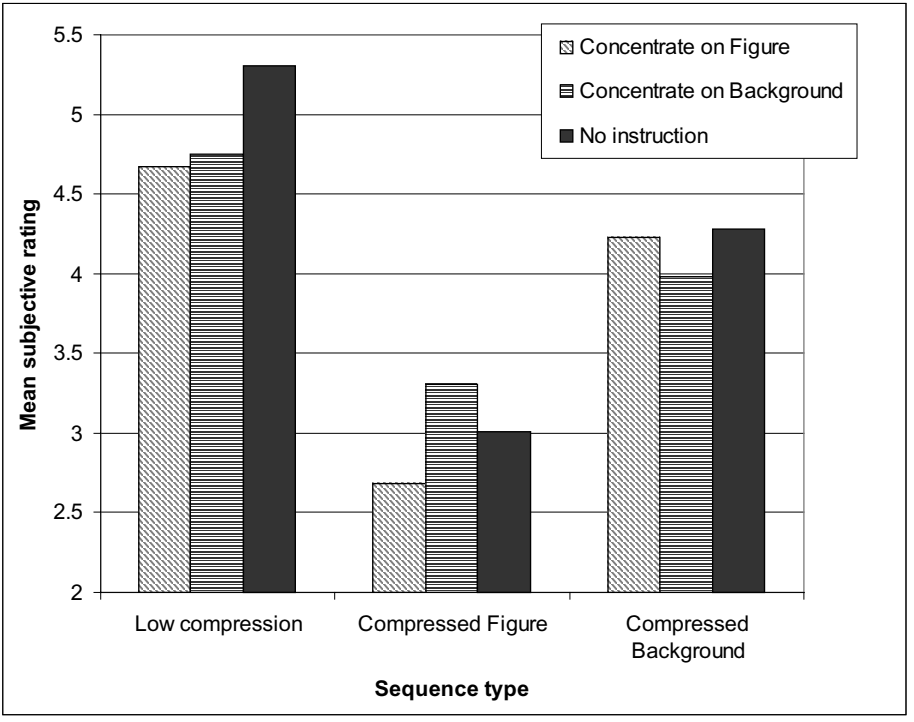


Fig. 3. Mean subjective rating (81 participants and 9 sequences)

It is interesting to note that the subjects spent 15% of the time looking at the figure when instructed to concentrate on the background, but spent only 6% of the time looking at the background when instructed to look at the figure. In addition, when no instruction was given, the subjects spent more time looking at the figure than at the background. This implies that a human figure in the foreground of a scene is likely to attract the attention of an observer.

Table 1. Results of eye movement tracking

Instruction	Time spent observing main figure	Time spent observing background
Concentrate on main figure	94.2%	5.8%
Concentrate on background	15.2%	84.8%
No instruction	61.9%	38.1%

4 Discussion

The results presented in Section 3 show two interesting outcomes. First, the observers' subjective opinion of visual quality appears to be strongly influenced by the

quality of the main foreground figure in each sequence. The lowest quality rating was awarded to the version of the clips with a highly-compressed (i.e. distorted) main figure, regardless of the task instruction given to the viewer. This is perhaps to be expected if no instruction is given, or if the subject is instructed to attend to the main figure, but is somewhat surprising for the case where the subject is instructed to concentrate on the background. Eye movement tracking results confirm that the observers do tend to follow the initial instructions but typically spend a small amount of time looking at the figure when instructed to look at the background. The presence of distortion in a foreground figure appears to have a significant influence on subjective quality, even when the figure is not the main focus of attention.

Second, the results indicate that subjective quality is affected by the task carried out by the observer (in this case, to attend to certain parts of the scene). The mean quality rating given to the sequence with a compressed foreground figure (fig.3) is lowest when the observer is instructed to concentrate on the figure. Similarly, the mean rating given to the sequence with a compressed background is lowest when the observer is instructed to concentrate on the background. This implies that distortion in an area of the scene has a stronger influence on subjective quality if the viewer is specifically concentrating on that area of the scene.

5 Conclusions

The results described in this paper demonstrate that (a) foreground human figures in a video clip have a significant (perhaps dominant) influence on subjective quality rating, regardless of whether the viewer is concentrating on the human figure and (b) an observer tends to rate subjective quality lower when their task involves concentrating on a distorted area of a scene. The first outcome implies that it may be possible to improve subjective quality in a decoded video scene by minimizing the distortion of human figures. This could be carried out by prioritizing regions of a scene containing human figures during encoding, using methods such as reducing quantization step size, object-based coding or pre-filtering. The second outcome implies that subjective testing of coded video sequences may give more realistic results if the expected task of the end-user is taken into account. For example, if the user is specifically attending to a foreground figure (e.g. during a videoconference session) or to the background of a scene (e.g. observing security camera footage), the perception of subjective quality is especially influenced by the presence of distortion in that region of the scene.

Further work is in progress to investigate the subjective mechanisms implied by these results in more detail. This will involve carrying out subjective tests using a wider range of video material (for example, without the presence of human figures; with human figures but without a main “foreground” figure; etc). Based on the outcomes of this work, a prioritized coding scheme will be developed with the aim of optimizing subjective (rather than objective) visual quality.

References

1. ISO/IEC 14496-2: Coding of Audio-Visual Objects: Part 2 (MPEG-2 Visual). (July 2000)
2. ISO/IEC 14496-10 / ITU-T Rec. H.264, :Advanced Video Coding for Generic Audiovisual Services .(May 2003)
3. Richardson, I.: Video Codec Design. John Wiley & Sons (April 2002)
4. Gazzaniga, M., Ivery, R. and Mangun, G.: Cognitive neuroscience, the biology of the mind. the 2nd edition, Norton (2002)
5. Findlay, J.M., and Gilchrist, I.: Active Vision: the Psychology of Looking and Seeing. Oxford University Press (Augest 2003)
6. Wiegand, T., Schwarz, H., Joch, A., Kossentini, F. and Sullivan, G.: Rate-constrained coder control and comparison of video coding standards. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13 no. 7 (July 2003)
7. Zhong, Y., Richardson, I., Sahraie, A. and McGeorge,, P.: Qualitative and quantitative assessment in video compression. The 12th European Conference on Eye Movements, (August 2003) Dundee, Scotland
8. Richardson, I.: H.264 and MPEG-4 Video Compression. John Wiley & Sons (August 2003)
9. Recommendation BT.500-10: Methodology for the subjective assessment of the quality of television pictures, ITU-R (2000)
10. Recommendation P.910: Subjective video quality assessment methods for multimedia applications. ITU-T (1999)
11. Muir, L., Richardson, I., Leaper, S.: Gaze tracking and its application to video coding for sign language. Picture Coding Symposium 2003, Saint Malo – France (April 2003)
12. Schaar, M., Lin, Y.: Content-based selective enhancement for streaming video. Proc. of IEEE International Conference on Image Processing (ICIP) (October 2001)

Evaluation of Some Reordering Techniques for Image VQ Index Compression

António R.C. Paiva and Armando J. Pinho*

Dept. de Electrónica e Telecomunicações / IEETA
Universidade de Aveiro, 3810–193 Aveiro, Portugal
antrafael@ieeta.pt, ap@det.ua.pt

Abstract. Frequently, it is observed that the sequence of indexes generated by a vector quantizer (VQ) contains a high degree of correlation, and, therefore, can be further compressed using lossless data compression techniques. In this paper, we address the problem of codebook reordering regarding the compression of the image of VQ indexes by general purpose lossless image coding methods, such as JPEG-LS or CALIC. We present experimental results showing that techniques available for palette reordering of color-indexed images can also be used successfully for improving the lossless compression of images of VQ indexes.

Keywords: Image compression, vector quantization, reordering techniques, lossless image coding

1 Introduction

Vector quantization (VQ) aims at reducing the redundancy of an image by encoding a block of pixels (or vector) with an index pointing to a similar block (vector) stored in a codebook. Although this process has some loss of information, the analysis of the pixels as a whole eliminates most of the redundancy while keeping the mutual information which is essential to properly reconstruct the image.

By increasing the block size, the quantization process becomes more efficient. However, for blocks greater than 4×4 pixels the computational complexity becomes high. Thus, still much redundancy exists between VQ indexes. Considering this problem of memoryless VQ, many methods have been proposed to further reduce the bitrate by suppressing this interblock redundancy. Among them, we find predictive VQ [1], finite-state VQ [2], address VQ [3] and conditional entropy coding of VQ indexes (CECOVI) [4]. These techniques, although more effective than memoryless VQ, are generally also much more complex, and some of them requiring large probability tables both at the encoder and decoder, that have to be obtained by training.

In this paper, we address the problem of codebook reordering regarding the compression of the image of VQ indexes by general purpose lossless image coding methods, such as JPEG-LS [5, 6]or CALIC [7]. Therefore, our objective is to create an image of VQ indexes that is more appropriate for compression by these techniques, through a suitable reordering of the codebook.

* This work was supported in part by the Fundação para a Ciência e a Tecnologia (FCT).

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 302–309, 2004.
© Springer-Verlag Berlin Heidelberg 2004

(a) (b) (c) (d)

Fig. 1. Images of VQ indexes of Lena (512×512, 8 bpp, 29.25 dB) obtained with full-search memoryless VQ, and (a) unsorted codebook, (b) codebook energy reordered, (c) codebook reordered with modified Zeng's and (d) Memon's algorithm.

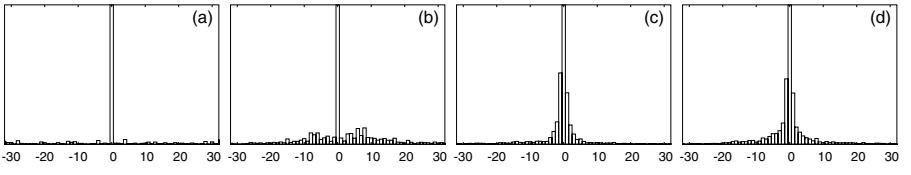


Fig. 2. Histograms of the first order differences for the images presented in Fig. 1, respectively. The histograms have been truncated to the $[-32, 32]$ interval to enhance the central region.

The problem is that, due to the limited block size in VQ still much unnecessary interblock correlation remains to be suppressed. The interblock correlation can be made visible by placing the VQ indexes in an image structure. Looking at the resulting image it is possible to discover some of the original image outline (see Fig. 1(a)). However, the image of VQ indexes usually does not have a Laplacian distribution of the first order prediction residuals as the image coders expect (as can be seen in Fig. 2(a)), which causes the general purpose lossless images coders to perform poorly. Hence, the purpose of image reordering in this context is to reorder the codebook, resulting in an image having a more adequate distribution to lossless compression by the general purpose image coders, therefore, improving the compression results. The effect of reordering can be observed in Figs. 1(b–d) and Figs. 2(b–d).

A similar problem to the one addressed in this paper can be found in coding color-indexed images (a survey can be found in [8], with references to other work on the subject). There, the aim is to reorder the palette of colors such that the reordered image can be better compressed. Given the similarity, we will address some of the methods that have been proposed in that framework, and evaluate how they perform in the VQ context.

The remainder of this paper is organized as follows. In Section 2 we briefly introduce the reordering techniques that we will be evaluating. Section 3 presents the results of simulations on several images. Finally, some conclusions are drawn in Section 4.

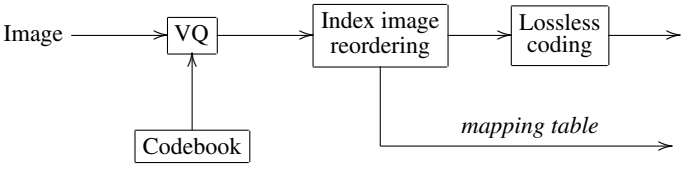


Fig. 3. Block diagram of the encoding process

2 Reordering Techniques

For reordering the images of VQ indexes we will consider one vector-based and two index-based techniques. Vector-based methods try to approximate the distribution to that of the original image which is known to be smooth (for natural images) and with the intended Laplacian distributed differences, and depend only on the codebook that is used. Therefore, needs only to be applied when the codebook is changed. On the other hand, index-based methods analyze the image statistics, and construct a mapping of reduced variance based on these statistics. But, since the index-based techniques greatly depend on the image, a mapping table needs to be transmitted for each image with the information to reverse the reordering.

Figure 3 shows the general case block diagram of the encoding process. In this process the codebook reordering is a preprocessing method aiming to improve the efficiency of the lossless image compression. At the decoder, the process is reversed, by applying the reverse mapping to the image of VQ indexes, before the VQ reconstruction.

2.1 Codebook Reordering by Energy

The concept behind energy reordering is very similar to that of luminance reordering in color-indexed images [9]. The energy of a vector gives us an idea of the vector intensity and, therefore, essentially allows to reconstruct the image contrast structure. Figure 1(b) shows precisely this behaviour, with the image vector quantized with the energy ordered codebook looking very much as a subsampled version of the original image by the block size dimensions (this is, in this case, reduced by a factor of 4 in both directions, since the VQ block size used was 4×4). With only this very simple step the image histogram of the differences have a different appearance, more similar to a Laplacian distribution, as can be seen in Fig. 2(b). This distribution reflects the new statistics of the image of VQ indexes which translates in a much more efficiently lossless coded image.

The codebook sorting is made by calculating the energy of each vector in the codebook,

$$E = \sum_{i=1}^{M} \sum_{j=1}^{N} v_{ij}^2,$$

where v_{ij} is the pixel at position (i, j) of the vector, considering vectors of $M \times N$ pixels. Then a sorting algorithm is applied to the codebook where the sorting criterion compares the vector energy.

2.2 Codebook Reordering with the Modified Zeng's Technique

Zeng *et al.* [10] proposed a re-indexing technique with the the aim of improving the compression of color-indexed images. This technique performs a re-indexing of the indexes independently of their physical meaning, relying only on the statistical information of the image of indexes. Later, a theoretical analysis of Zeng's method by Pinho *et al.* [11] for the case of Laplacian distributed differences conducted to a new set of parameters, different from those originally suggested by Zeng. In the same paper the authors also showed that these proposed parameters provide better results. Since the modified Zeng method depends only on the image statistics, independently of their meaning, it can be used without modification to the reordering of images of VQ indexes. Figure 1(c) shows an image reordered with the modified Zeng technique.

The modified Zeng algorithm performs as follows [11]. The algorithm starts by finding the index that is most frequently found contiguous to other different indexes, and the index most frequently found adjacent to it. These indexes are the initial values of an index list P_N, to be constructed iteratively by the algorithm, where N is the number of elements in the list. In each iteration an index is selected from the set of indexes still not in the list and appended to the list either to the right or to the left. We denote by l_j the indexes already in the list and by u_i those still to be assigned. Thus, the initial list is $P_2 = (l_1, l_2)$. In each iteration, the algorithm calculates the index σ which satisfies

$$\sigma = \arg\max_{u_i} \sum_{j=1}^{N} C(u_i, l_j),$$

where N is the list size, and $C(i,j)$ is the number of occurrences (on the initial index image) of the index i found adjacent of index j. Then, to choose the side of the list where to place the index, we calculate

$$\Delta = \sum_{j=1}^{N} (N + 1 - 2j) C(u_i, l_j).$$

If this expression is positive ($\Delta > 0$) then the index should be placed on the left side of the list, otherwise it should be placed on the right side.

2.3 Codebook Reordering with Memon's Technique

Like the modified Zeng's method described previously, the method by Memon *et al.* [12] is also an index-based reordering algorithm developed in the framework of color-indexed images. Visually (see Fig. 1(d)), the results are similar to modified Zeng, but the image histogram of the differences (Fig. 2(d)) shows some improvements.

As it was referred previously, it is a known fact that the prediction residuals are well modeled by a Laplacian distribution, and that most lossless coders depend on this property to compress the image. Knowing this, the Memon algorithm tries to reorder the indexes so that the prediction residuals are minimized. This corresponds to reducing the

zero-order entropy of the residuals. Globally, this is made by the minimization of the sum of the prediction residuals given by

$$\sum_{i=1}^{M}\sum_{j=1}^{M} C(i,j)|i-j|, \tag{1}$$

where M represents the number of indexes. In this case, $C(i,j)$ is the number of times index i is used as the prediction value by a pixel with index j. For example, if only first-order causal prediction is used, then $C(i,j)$ is just the number or times index i if found directly after j.

The solution to the problem from the perspective presented in (1) is very difficult, if not impossible. Hence, Memon *et al.* have proposed two heuristics: one based on simulated annealing, and one called "pairwise merging". The "pairwise merging" technique, which we will use, is much faster to compute. Nevertheless, it is computationally more demanding (order $\mathcal{O}(M^4)$) than the modified Zeng's algorithm (order $\mathcal{O}(M^3)$).

In the pairwise merge heuristic, ordered sets of indexes are merged until only one set exists. Through the resulting ordered set one can determine a mapping table to perform the reordering of the image pixels, and therefore, to determine the re-indexed images. Initially, each index is assigned to a different set. Next, the algorithm performs two steps on each iteration. First, the two sets A and B maximizing

$$\sum_{v_i\in A}\sum_{v_j\in B} w(i,j)$$

are selected from all possible pairs of ordered sets, where $w(i,j) = C(i,j) + C(j,i)$. Then, several combinations of the sets A and B are evaluated, and is chosen the one minimizing

$$\sum_{i=1}^{k}\sum_{j=1}^{k} w(u_i,u_j)|i-j|,$$

where $u_1, u_2, \ldots, u_k$ is the ordered set of the vectors from A and B under evaluation. Since the evaluation of all the possible combinations is impractical, Memon *et al.* [12] suggested some combinations that should perform reasonably well in most situations. If $a_1, a_2, \ldots, a_r$ and $b_1, b_2, \ldots, b_s$ are the two ordered sets under evaluation, with $r, s > 1$, then the following combinations are considered:

$$\begin{gathered} a_1, a_2, \ldots, a_r, b_1, b_2, \ldots, b_s \\ a_r, \ldots, a_2, a_1, b_1, b_2, \ldots, b_s \\ b_1, b_2, \ldots, b_s, a_1, a_2, \ldots, a_r \\ b_1, b_2, \ldots, b_s, a_r, \ldots, a_2, a_1 \end{gathered}$$

Otherwise, if any of the sets has size one, then the following configurations are tested:

$$\begin{gathered} a_1, b_1, b_2, \ldots, b_s \\ b_1, a_1, b_2, \ldots, b_s \\ b_1, b_2, a_1, \ldots, b_s \\ \vdots \\ b_1, b_2, \ldots, b_s, a_1 \end{gathered}$$

3 Experimental Results

In this Section, we present some compression results of applying general purpose lossless image coders to the VQ index images. The lossless coders JPEG-LS and CALIC were applied over 23 images of VQ indexes, and for each reordering method: energy, modified Zeng (referred in the tables as mZeng) and Memon. For comparison, we have also included the results of the lossless coders applied directly to the images vector quantized before any reordering.

The VQ index images where obtained with a basic full-search memoryless vector quantizer, applied over the 23 images of the kodak images test set[1], with block size 4×4 (vector dimension 16) and codebook size 256. Each image of the kodak set has dimensions 768×512, and was not in the codebook training set. To the generation of the codebook we used the generalized Lloyd algorithm [13], also known as LBG algorithm from the article authors, with a training set made up of 13 natural images. The image quality obtained with this VQ method is far from being the best. However, since we are focusing our attention on the increase of efficiency of the lossless compression performed after the reordering, and other VQ methods can be used to enhance the image quality, we believe that this is not really important on the evaluation of the reordering techniques with general purpose lossless coding methods.

First, the whole kodak set was vector quantized using the unsorted codebook that resulted directly from the generalized Lloyd algorithm. Then, the codebook was sorted by energy and a new set of index images was created by vector quantizing again the kodak set. Finally, the two other sets of images were created by applying the modified Zeng and the Memon reordering methods, respectively, to the images of VQ indexes obtained using the unsorted codebook. The implementations of modified Zeng and Memon are the same used in [8] and were provided by the authors of that article. All the other tools where implemented by us.

Table 1 shows the results of image compression, in bits per pixel (bpp), using JPEG-LS[2] and CALIC[3]. The values are for the entire process, i.e., they represent the total number of bits resulting from the lossless compression relative to the number of pixels in the original (before VQ) image. For the index-based methods, which perform the re-indexing for each image, the 256 bytes of mapping table (which depends on the size of the codebook) is also considered. The row labeled "Average" provides overall results for each reordering method. Because of space limitations, we do not present were the tests made with JPEG 2000[4] in lossless mode. The average results for energy, modified Zeng's and Memon's, are 0.346 bpp, 0.299 bpp and 0.294 bpp, respectively.

[1] These images are available at `http://www.cipr.rpi.edu/resource/stills/kodak.html` in raster format. The images where then converted to PGM with the Linux utility `rasttopnm`.

[2] This coded can be obtained at `http://www.ece.ubc.ca/spmg/research/jpeg/jpeg_ls/jpegls.html` (version 2.2).

[3] The codec used is available at `http://compression.graphicon.ru/download/i_glless.html`.

[4] This coded is available at `http://www.ece.uvic.ca/~mdadams/jasper/` (version 1.700.5).

Table 1. Compression resuls, using a JPEG-LS and CALIC codec, on the kodak image test set. The results for modified Zeng and Memon reordering methods include the mapping table of 256 bytes to be transmitted.

	JPEG-LS				CALIC			
Image	Unsorted (bpp)	Energy (bpp)	mZeng (bpp)	Memon (bpp)	Unsorted (bpp)	Energy (bpp)	mZeng (bpp)	Memon (bpp)
01	0.495	0.374	0.346	0.342	0.447	0.360	0.336	0.333
02	0.385	0.253	0.186	0.185	0.304	0.222	0.177	0.177
03	0.366	0.250	0.203	0.196	0.294	0.221	0.192	0.184
04	0.458	0.312	0.262	0.246	0.385	0.286	0.246	0.234
05	0.530	0.421	0.398	0.394	0.485	0.405	0.387	0.384
06	0.434	0.314	0.277	0.271	0.384	0.301	0.264	0.261
07	0.400	0.304	0.271	0.262	0.351	0.284	0.259	0.253
08	0.497	0.389	0.403	0.402	0.453	0.372	0.383	0.382
09	0.336	0.243	0.212	0.205	0.262	0.207	0.189	0.184
10	0.403	0.282	0.245	0.236	0.344	0.259	0.231	0.225
11	0.449	0.314	0.278	0.269	0.397	0.301	0.268	0.262
12	0.362	0.233	0.177	0.170	0.284	0.206	0.168	0.163
13	0.528	0.412	0.385	0.381	0.481	0.401	0.372	0.370
14	0.514	0.375	0.348	0.333	0.463	0.364	0.334	0.324
15	0.366	0.238	0.211	0.202	0.305	0.213	0.191	0.184
16	0.432	0.288	0.245	0.231	0.375	0.273	0.236	0.224
17	0.464	0.317	0.281	0.271	0.403	0.300	0.266	0.259
18	0.487	0.354	0.321	0.319	0.431	0.337	0.309	0.306
19	0.469	0.327	0.280	0.262	0.404	0.303	0.265	0.249
20	0.250	0.181	0.169	0.162	0.227	0.174	0.165	0.158
21	0.367	0.282	0.259	0.243	0.320	0.262	0.248	0.236
22	0.462	0.327	0.278	0.271	0.403	0.312	0.267	0.261
23	0.399	0.264	0.198	0.189	0.326	0.235	0.184	0.178
Average	0.428	0.306	0.271	0.262	0.370	0.286	0.258	0.251

Not surprisingly, Memon is the reordering method which performs better, for all lossless coders, followed by the modified Zeng. The coding of the index images resulting from a codebook sorted by energy produced the worst results. Nevertheless, the energy reordering allows an average increase of the coding efficiency of 29%, 30% and 23%, comparatively to the unsorted coding, with JPEG-LS, JPEG 2000 and CALIC, respectively. Moreover, all of this is made without any complexity increase at the coding / decoding process since the reordering is made offline. Comparatively to Memon reordering, it has an efficiency decrease of 17%, 18% and 14%, for the same coders. These results of energy codebook reordering are quite interesting since they show how a simple reordering can dramatically improve lossless compression. Comparing modified Zeng with Memon, we see that this method performs only 2–3% worse. Considering these results and the fact that modified Zeng is much faster than Memon, the choice between the two should consider the tradeoff between computational complexity and compression efficiency.

4 Conclusions

Reordering techniques provide a simple yet very effective way to improve the compression of images of VQ indexes. In this paper, we evaluated the performance of some reordering techniques when applied to images of VQ indexes that are subsequently losslessly coded, and presented results for three lossless image coders.

Comparing the average results for CALIC and JPEG-LS with Memon reordering with other state-of-the-art VQ coding methods, we see that these are considerably good results. Moreover, this is achieved using only, correctly combined together, *off-the-shelf* algorithms for reordering and lossless image compression.

From the experimental results obtained, we also conclude that Memon's method is the best reordering algorithm for the three image coders. Nevertheless, the modified Zeng's algorithm should also be taken into account since it also performs very good (only a few percent worse than Memon's), but much faster. So, if computational complexity (or time) is a stringent factor, then the modified Zeng's method can provide an efficiency / computation tradeoff very well balanced.

References

1. Hang, H.M., Woods, J.W.: Predictive vector quantization of images. IEEE Trans. on Communications **33** (1985) 1208–1219
2. Foster, J., Gray, R.M., Dunham, M.O.: Finite-state vector quantization for waveform coding. IEEE Trans. on Information Theory **31** (1985) 348–359
3. Nasrabadi, N.M., Feng, Y.: Image compression using address-vector quantization. IEEE Trans. on Communications **38** (1990) 2166–2173
4. Wu, X., Wen, J., Wong, W.H.: Conditional entropy coding of VQ indexes for image compression. IEEE Trans. on Image Processing **8** (1999) 1005–1013
5. ISO/IEC 14495–1 and ITU Recommendation T.87: Information technology - Lossless and near-lossless compression of continuous-tone still images. (1999)
6. Weinberger, M.J., Seroussi, G., Sapiro, G.: The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS. IEEE Trans. on Image Processing **9** (2000) 1309–1324
7. Wu, X., Memon, N.: Context-based, adaptive, lossless image coding. IEEE Trans. on Communications **45** (1997) 437–444
8. Pinho, A.J., Neves, A.J.R.: A survey on palette reordering methods for improving the compression of color-indexed images. IEEE Trans. on Image Processing (2004 (in press))
9. Zaccarin, A., Liu, B.: A novel approach for coding color quantized images. IEEE Trans. on Image Processing **2** (1993) 442–453
10. Zeng, W., Li, J., Lei, S.: An efficient color re-indexing scheme for palette-based compression. In: Proc. of the 7th IEEE Int. Conf. on Image Processing, ICIP-2000. Volume III., Vancouver, Canada (2000) 476–479
11. Pinho, A.J., Neves, A.J.R.: A note on Zeng's technique for color reindexing of palette-based images. IEEE Signal Processing Letters **11** (2004) 232–234
12. Memon, N.D., Venkateswaran, A.: On ordering color maps for lossless predictive coding. IEEE Trans. on Image Processing **5** (1996) 1522–1527
13. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. on Communications **28** (1980) 84–95

Adaptive Methods for Motion Characterization and Segmentation of MPEG Compressed Frame Sequences

C. Doulaverakis, S. Vagionitis, M. Zervakis, and E. Petrakis*

Department of Electronic and Computer Engineering
Technical University of Crete
Chania, Crete, Greece
{doulaver,bagsta,michalis}@systems.tuc.gr, petrakis@ced.tuc.gr

Abstract. A fast and accurate method for scene change detection and classification of camera motion effects in MPEG compressed video is proposed. The method relies on adaptive threshold estimation and on the analysis and combination of various types of video features derived from motion and intensity information. This analysis is also applied for cleaning-up false shot boundaries due to camera motion effects. Two techniques for adaptive threshold estimation are also proposed and evaluated.

1 Introduction

Temporal video segmentation is intended to partition a video into consecutive shots [1, 2]. The transitions between consecutive shots can be abrupt or gradual. Video effects caused by camera panning, titling or zooming result in interframe signal or histogram transitions of the same order of magnitude as gradual transitions. Moreover, transitions caused by sharp changes in camera motion direction are of the same order of magnitude as camera breaks. Both types of effects introduce false shot boundaries (false positives). Detecting the exact type of video effects and cleaning-up the false positives is a difficult task [3]. This is exactly the focus of this work.

This work also focuses on gray-scale video processing and analysis directly on the compressed domain. MPEG-2 compressed information is computed for 8x8 pixel regions called *blocks,* whereas motion vectors are computed for 16x16 pixel regions called *macroblocks*. In this work, intensity and motion information is encoded in the same coherent way regardless of frame type (i.e., DC intensity and forward predicted motion vector for each macroblock for any *I*, *P* or *B* frame) [4].

The following summarizes the contributions of this work:

- An approach is proposed for the segmentation of MPEG compressed video. It allows for more reliable video segmentation than intensity histogram thresholding by correctly identifying camera motion effects and by utilizing adaptive threshold estimation in detecting such effects.

* Corresponding author.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 310–317, 2004.
© Springer-Verlag Berlin Heidelberg 2004

- The novelty of the method relies on the analysis and combination of various types of video features derived from motion information directly in the compressed domain. This analysis is also applied for cleaning-up false shot boundaries due to camera motion effects.
- Following the example of [5, 6] two adaptive threshold estimation methods are proposed and evaluated. Compared to existing adaptive methods they are more theoretically principled and are easily integrated within a proposed simple, intuitive and fast video segmentation algorithm.

In the rest of this paper, threshold selection methods are discussed in Sect. 3; the proposed video analysis method is presented in Sect. 4; experimental results are presented and discussed in Sect. 5 followed by conclusions in Sect. 6.

2 Threshold Selection

The state-of-the-art approach for automatic threshold selection is referred to as "Twin-Comparison" (TC) approach [7]. TC requires that the whole video must be scanned once prior to segmentation and for detection of gradual transitions computes two thresholds. As globally optimum, these thresholds do not adapt to local properties of the histogram differences. The proposed threshold estimation techniques do not require preprocessing, compute only one threshold and adapt the threshold to local properties of the input signal.

2.1 Sliding Window (SW) Threshold

A threshold T_a is computed over a small range of W frames (W=15 in this work). The SW method starts by taking the first W histogram differences D and for video partitioning works as follows:

Repeat until end of video stream:

1. Compute the mean $\mu(i)$ and variance $\sigma(i)$ within a range of W frames (window W), where i is the rightmost difference $D(i)$ value of W.
2. Compute threshold as $T_a(i) = \mu(i) + \alpha\sigma(i)$, where α is a user defined parameter as in TC.
3. Compare $T_a(i)$ with the next histogram difference $D(i+1)$ outside the window.
4. If $T_a(i) > D(i+1)$ advance W to the right (increase i) to include $D(i+1)$ and go to *step 1*.
5. if $T_a(i) < D(i+1)$ a transition is found. Increase i (move W to the right) until $D(i) < T_a(i)$. Count the number n of skipped positions. If $n < w$ ($w = 5$ in this work) a *camera break* is declared, if $n > w$ a *gradual transition* is declared.

Algorithm 1: Sliding window approach for video partitioning

Parameter a, is user defined and depends on video properties. In Fig. 1 a camera break is declared at frame 150 and a gradual transition between frames 300 and 325.

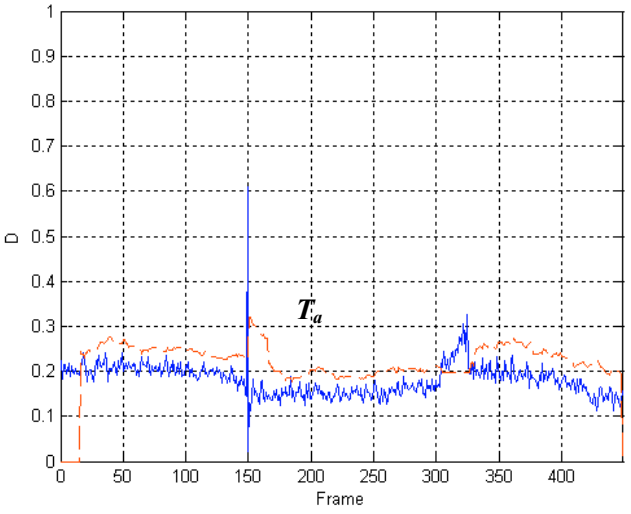


Fig. 1. Sliding window approach. The dashed line represents local values of threshold T_a

2.2 Adaptive Window (AW) Threshold

This method works the same way as Algorithm 1, but T_a is defined as in an autoregressive form

$$T_a(i) = \mu(i) + \alpha\sigma(i) \quad \text{with}$$
$$\mu(i) = \mu(i-1) - c\,(\mu(i-1) - D(i)), \quad \sigma(i) = |\mu(i)^2 - \lambda(i)|^{1/2} \text{ and}$$

$\lambda(i) = \lambda(i-1) - c\,(\lambda(i-1) - D(i))^2$. Moreover, $\mu(1) = D(1)$, $\lambda(1) = D(1)^2$, $\sigma(1) = 0$. Parameter c controls the sensitivity of the threshold to signal changes ($c = 0.05$).

3 Video Shot Characterization by Motion Information

A forward predicted vector is computed to each macroblock [4] except for *intracoded* macroblocks for which no matching macroblock can be detected in the next frame. In the present consideration, macroblocks with magnitude of motion vector less than 1 are called *static* (they include *skipped* macroblocks with 0 motion). In the following, comparisons involving motion information are normalized with respect to the number of macroblocks with identified motion (no intracoded or static).

3.1 Direction Histogram

A motion vector is described by a pair (u,v) representing horizontal and vertical displacement of the macroblock respectively. From this pair, the angle (direction) of motion is computed as $\theta = \tan^{-1}(v/u)$, taking values in $[0,2\pi]$. All angles are quantized into 8 directions, multiples of $\pi/4$. The notion or *direction histogram* is introduced as

a tool for global motion analysis. Each bin in this histogram counts number of macroblocks in each angle range. An additional bin (bin 1) is also added in the direction histogram representing number of static macroblocks (with $|(u,v)| < 1$) in the frame. Fig. 2 illustrates the direction histogram of the scene on its left.

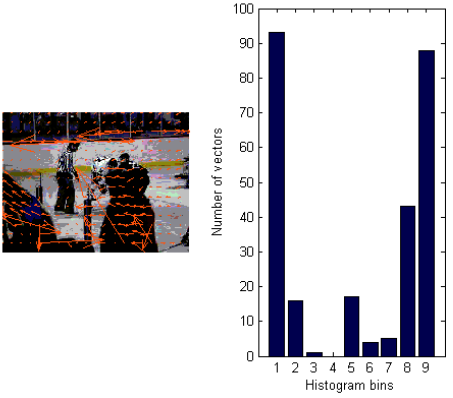


Fig. 2. Scene and its corresponding direction histogram

3.2 Camera Zooming, Panning, and Tilting

The analysis is based on the following observations:

- **Zooming**: The motion vectors are equally spread on the direction histogram.
- **Panning or Tilting**: The motion vectors tend to concentrate at a single histogram bin. The position of this bin denotes the direction of camera motion.
- **Static Camera**: Most motion vectors are concentrated at bin 0; the distribution of motion vectors in the remaining bins is irrelevant.

The analysis of various video types indicates that for reliable prediction of camera motion effects, at least 40% of the total number of macroblocks in a frame must be motion predicted (no intracoded or static). The variance of the histogram provides the means for analyzing the structure of the direction histogram:

$$\sigma_h^2 = \sum_{i=1}^{9} p_i (i - \bar{h})^2 \text{ with } p_i = \frac{h_i}{\sum_{i=1}^{9} h_i} \text{ and } \bar{h} = \sum_{i=1}^{9} p_i i \text{ where}$$

h_i involves the histogram values at bin i and p_i represents the probability of occurrence of motion vectors with angle i (bin 1 indicates static regions). The maximum variance is encountered when the motion vectors are equally spread along the histogram while, the minimum variance is encountered when all motion vectors are concentrated at a single histogram bin. The variance is further normalized with respect to the actual number of motion vectors in each frame:

$$\sigma_{h,norm}^2 = \gamma \sigma_h^2 \text{ where } \gamma = 1 - \frac{IntracodedVectors + StaticVectors}{TotalMacroblocks\ in\ Frame}.$$

Detection of camera zooming, panning and tilting relies on the application of appropriate threshold values on the plot of normalized variance (computed from bin 2 through 9 that is, over the part of the histogram representing angles of motion). The SW method or the AW method can be applied.

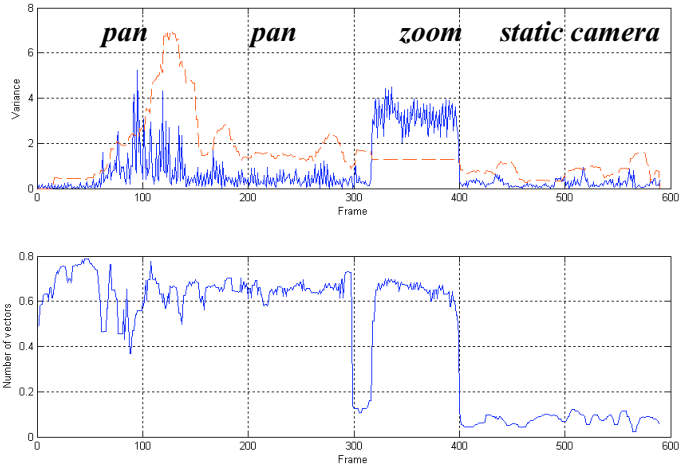


Fig. 3. Normalized variance of angles of motion vectors with two panning and one zoom areas. The diagram below illustrates number of actual motion vectors per frame

Fig. 3 illustrates the behavior of this measure on pan sequences with small variance, zoom sequences with large variance and static camera regions with small variance but with small number of motion vectors. The lower figure illustrates the number of actual motion vectors (no static and intracoded) in each frame. The dashed line plots the SW threshold T_a. It correctly detects the zoom section, while it falsely detects isolated camera changes before frame 100, which are due to camera instability.

Algorithm 2 summarizes the above scheme (γ is the percentage of motion predicted vectors and T_v is set to 40% of the frame size measured in macroblocks).

- If $\gamma > T_v$ and $\sigma_{h,norm} < T_a$, the frame belongs to a panning or tilting sequence.
- If $\gamma > T_v$ and $\sigma_{h,norm} > T_a$, the frame belongs to a zoom sequence.
- If $\gamma < T_v$ none from the above holds (no camera motion is detected).

Algorithm 2. Detection of zooming, panning and tilting frame sequences

3.3 Camera Direction Changes

This kind of camera effects is characterized by changes in the distribution of motion vectors. They are detected by direct comparison of direction histograms which for two histograms f_{n-1} and f_n is defined as

$$D_H(f_n, f_{n-1}) = \sum_{i=1}^{9} \left| p_i^{f_n} - p_i^{f_{n-1}} \right| \text{ where } p_i = \frac{h_i}{\sum_{i=1}^{9} h_i}.$$

The static vectors are included in the difference, since they reflect a motion property that needs to be measured (e.g., in the transition between static camera and pan). Camera motion changes are detected by thresholding on the plot D_H differences.

3.4 Video Segmentation Using Motion Information

The above approach can also be applied to enhance existing video segmentation methods by cleaning-up false shot boundaries due to camera motion effects.

3.4.1 Camera Breaks

Traditionally, camera breaks are detected when the number of intracoded vectors exceeds a threshold [8]. The SW or the AW approach is applied on the plot of the number of intracoded macroblocks of a frame sequence. The method of Sect. 3.3 is applied to clean-up the false shot boundaries. Algorithm 3 summarizes this approach

1. If the condition for camera motion changes is not satisfied (Sect. 3.3) and
2. If the number of intracoded vectors exceeds the threshold then
3. A camera break is detected.

Algorithm 3. Detection of camera breaks by combining motion information

3.4.2 Gradual Transitions

When the transition between frames is extended over time (gradual transition), the majority of motion vectors behave randomly. Such gradual transitions are detected based on the variance of the magnitude of motion vectors. The variance σ_l of the magnitude l of motion vectors in each frame is computed as

$$\sigma_l^2 = \frac{1}{n_v} \sum_{i=1}^{n_v} (l_i - \bar{l})^2 \text{ where } \bar{l} = \frac{1}{n_v} \sum_{i=1}^{n_v} l_i \text{ and}$$

l_i is the magnitude of the i-th vector and n_v is the number of motion vectors with $|(u,v)| > 1$. The variance is normalized with respect to the number of actual motion vectors.

$$\sigma_{l,norm}^2 = \gamma \sigma_l^2 \text{ where } \gamma = 1 - \frac{IntracodedVectors + StaticVectors}{TotalMacroblocks\ in\ Frame}$$

Adaptive thresholding (by SW or AW) is applied on normalized variance to detect peaks (potential gradual transitions). Notice that $\sigma_{l,norm}$ exhibits the same behavior in gradual transitions as in zoom sequences (the vectors have small magnitude at the center of the frame and high magnitude near the edges). Zooming can be distinguished from gradual transitions by Algorithm 2. For cleaning-up false positives due to camera motion, it is required that both the intensity histogram difference and the motion magnitude variance are above their thresholds. Algorithm 4 summarizes the above approach.

1. If the condition for zoom sequence (Algorithm 2) is not satisfied and
2. If the intensity condition (Algorithm 1) along with the condition on the normalized variance of the magnitude of motion vectors: $\sigma_{l,norm} > T_a$ are both satisfied then
3. A gradual transition is detected.

Algorithm 4. Detection of gradual transitions by combining motion and intensity information

4 Experimental Results

The effectiveness of each method is measured by the average (over 17 videos) precision and *recall*. Each method is represented by its precision and recall as a function of the threshold parameter *a*.

The plot on the left of Fig. 4 demonstrates that the AW method is particularly effective for the detection of panning or tilting achieving precision close to 1 (*a=2*) or recall close to 1 (*a=4.5*). The plot on the right demonstrates that AW is also effective for zoom detection. Higher precision is achieved for higher α and higher recall is achieved for lower α. The opposite was observed in panning and tilting since it is detected as $\sigma_{h,norm} > T_\alpha$ while zooming is detected as $\sigma_{h,norm} > T_\alpha$ (Sect. 3.2).

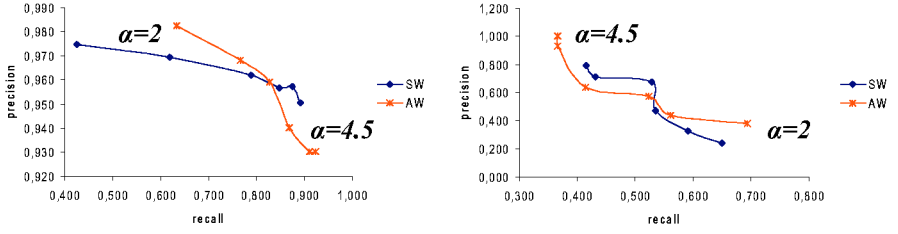


Fig. 4. Average precision and recall for panning or tilting (left) and zooming detection (right)

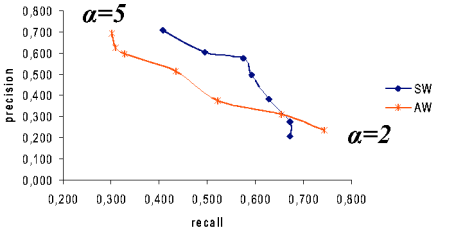


Fig. 5. Average precision and recall for camera direction change detection.

The plot of Fig. 5 demonstrates that the SW method achieves up to 10% better recall and almost always better precision in detecting camera motion changes.

The left plot of Fig. 6 illustrates that video segmentation by the SW method on motion information (Sect. 4.3) is at least as accurate as the TC method (the test data did not contain enough sequences with camera direction changes). The plot on the right demonstrates the superiority of the proposed approach combining motion and intensity information for detecting gradual transitions.

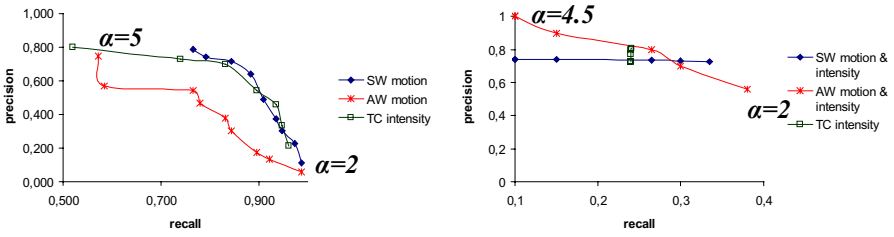


Fig. 6. Average precision and recall for detection of camera breaks (left) and gradual transitions (right).

The above experiments revealed the independence of the performance of all methods on the parameter *a*. Maximum precision and recall are achieved for the same values of *a*.

5 Conclusions

Two methods for automatic and adaptive threshold estimation on video information have been tested and proven particularly effective for video segmentation and for eliminating false positives caused by various camera effects. The novelty of the proposed approach relies on the analysis and combination of motion information in the compressed domain and adaptive threshold estimation on various signal patterns. The experimental results provide good support to the claims of efficiency.

References

1. Lefevre, S., J. Holler, and N. Vincent, *A Review of Real Time Segmentation of Uncompressed Video Sequences for Content-Based Search and Retrieval.* Real Time Imaging, 2003(9): p. 73-98.
2. Lienhart, R. *Comparison of Automatic Shot Boundary Detection Algorithms*. in *Image and Video Processing*. 1999.
3. Truong, B.T., C. Dorai, and S. Venkatesh. *New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation*. in *ACM Multimedia Conference*. 2000. California, USA.
4. Kobla, V., *et al. Compressed Domain Video Indexing Techniques Using DCT amd Motion Vector Information in MPEG Video*. in *Storage and Retrieval for Image and Video Databases*. 1997.
5. Yusoff, Y., W. Christmas, and J. Kittler. *Video Shot Cut Detection Using Adaptive Thresholding*. in *British Machine Vision Conference*. 2000. Bristol, U.K.
6. Yeo, B.L. and B. Liu, *Rapid Scene Analysis on Compressed Video.* IEEE Trans. on Circuits and Systems for Video Technology, 1995(6): p. 533-544.
7. Zhang, H.J., A. Kankanhali, and S.W. Smoliar, *Automatic Partitioning of Full Motion Video.* Multimedia Systems, 1993. **1**(1): p. 10-28.
8. Zhang, H.-J., L.C. Yong, and S.W. Smoliar, *Video Partitioning and Browsing Using Compressed Data.* Multimedia Tools and Applications, 1995(1): p. 91-113.

On the Automatic Creation of Customized Video Content

José San Pedro[1], Nicolas Denis[2], and Sergio Domínguez[1]

[1] UPM - DISAM, Universidad Politécnica de Madrid
(http://www.disam.upm.es/vision)
[2] Omnividea Multimedia - Madrid, http://www.omnividea.com

Abstract. Creating new multimedia content is normally a task out of the reach of automatic tools. The exception is the creation of video summaries, using an existing video as source. In this paper, a tool for creating new video content to fit a determinate purpose is introduced. It uses a Mpeg-7 database of video descriptors and annotations to select the segments that best fit that purpose. The required descriptors can be automatically extracted, but additional manual annotations can be included to pre-filter the database using application domain knowledge. The tool is able to use segments from several videos and automates all the creation process, following the parameters defined by a supervisor in a program specification document. This document provides enough flexibility to customize the new program to whatever purpose is intended, with parameters to set low level features (color, motion, ...) as well as production related parameters.

1 Introduction

Creating new multimedia content is becoming an important task in a wide range of fields, as the creation of video summaries or personalized programs based on profiles. Nowadays, new content generation must be manually supervised. Existing automatic tools allow only to perform production tasks: transitions between shots, on screen text, ... The supervisor is still in charge of the most time consuming tasks, as choosing input video content, select its right order within the program, define transition types between shots, ... Not only is this generation process expensive, but very time consuming due to the heavy load of work left to be performed manually. On the other hand, the quality of the final result is normally very high, due to the supervisor knowledge about the purpose of the new program being created (semantic knowledge).

In this paper, a fully automatic system to create new video content is presented. The system uses an annotated database of video clips based on the Mpeg7 standard. The Mpeg7 standard defines a common way to annotate multimedia features as a series of descriptors in a xml document. This standard is indeed much more complex than what we need for our purposes, allowing the annotation of other kind of information in a comfortable centralized model. Therefore, each video is described from the lowest level of abstraction (color, motion, ...)

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 318–326, 2004.
© Springer-Verlag Berlin Heidelberg 2004

to the highest (manual annotation about genre, ...). The system creates new content using a description of the desired features of the final result provided by the supervisor. Taking advantage of his knowledge about the project and the content of the source videos, the operator can provide the system with a specific subset obtain by means of a query to the Mpeg-7 database and a simple description that will guide it towards the goal it was intended.

Section 2 contains an overview of previous related work. In section 3, the proposed system is analyzed. A summary of results of the generation engine is provided in section 4 and section 5 contains some conclusions.

2 Related Work

The creation of new video content is not a very common issue in previous bibliography. However, it is becoming quite popular nowadays. The main goal of such texts is the creation of video summaries (also known as trailers). In [5], the concept of "Video Abstract" is introduced in analogy of "Abstracts" in papers.

The author proposes the creation of small video clips that summarize the essence of original the videos; an audiovisual summary will be always much richer than a brief text annotation.

Several summary generation systems have been created. The "Video Skimming" system summarizes documentary and news programs. It uses natural language recognition over the audio channel to look for important words and selects the sequence where they appear by detecting shot boundaries. Another simple, yet interesting, system is described in [3], where the goal is to produce a video summary with minimal visual redundancy applying a clustering process to the shots of the video. The most complete system in the bibliography is called "MoCA" (Movie content Analysis) [7]. The goal of the project is to "understand" the content of the video and experiment with applications that use this information: detection of film genre, detection of advertisements, video summary, ... One of the main features of the MoCA project is that it performs automatic face and text recognition in frames in order to include them in the final summary. Other interesting issues introduced by this project are:

- Summary type: Different types of summaries for different purposes.
- Input content: Manual information helps to create better summaries.
- Restriction of content: Certain content should be included (main characters) or not included(ending of a film).

3 Program Generation System

Analysis and processing performed upon video content is barely exploited. This paper focuses in the exploitation and use of the information gathered in a previous analysis stage, and annotated in the Mpeg-7 format. The goal is to create new video content in a supervised fashion. Often, new content must not only include a certain content (depending on its purpose), but it also must meet some

restrictions: duration, resolution, structure, ... The user creates a program pattern that will guide the module in the creation of the program. With this input, the module is able to perform its task automatically.

3.1 Automatic Video Analysis

This section reviews some low level features that can be used in the generation process. Note that although our implementation uses only these low level descriptors, the architecture allows to include easily any other descriptor (low or high level).

Shot Boundaries. In this stage the video is split in its simplest elements, shots, so video analysis can be simplified assuming certain intra-shot features: homogeneous semantic content, low inter-frame differences, ... About 90% of the shots are hard-cuts, fades, or dissolves[6]. Due to the different nature of each of these transitions, algorithms are normally specialized in the detection of just one kind. The sudden change of a hard-cut can be easily detected analyzing color histogram differences. Fades and dissolves are more difficult to detect because the transition is very smooth, so frame differences are meaningless by themselves. A deeper study of the evolution is, therefore, necessary for them([5]).

Motion. Computing motion at its finest level is a very complex problem. However, in some contexts, it is enough to use a gross representation: a measurement of the amount of motion which is present in the frame sequence. Such representation is used in indexing, retrieval and non-linear video browsing tasks. Mpeg7 includes a very appropriate descriptor: motion activity. Motion activity measures the amount of motion in a video sequence. It can be computed very fast, both in the compressed and the uncompressed domain ([8]).

Color. Color information is commonly used by image processing algorithms. A great number of different representations are available to work with color (histograms, . . .). A very convenient one is included in the Mpeg-7 standard: Dominant color. It describes colors in a region of arbitrary shape. Its computation is easy and fast and can be complemented with the addition of a spatial coherence descriptor, a measure of the concentration/dispersion of the color in the region

3.2 Program Pattern

This XML document defines the rules (requirements and restrictions) that the generation engine must meet when creating content, so the supervisor can take decisions about the approximate output that will be generated.

Filtering Parameters. Filtering parameters guide the selection of video sequences from the input video database. Sequences of every input are matched to defined Mpeg-7 based filters and are scored or even discarded based on the result the supervisor expects by providing such filters.

High level filters. These filters are used to prune the subset of the database that the module will use as a direct input. High level filters normally work with manual annotation provided by the database administrator. The goal is to allow the selection of videos by its content or semantical information to lead the final program towards it. If the new program is meant to be used as a sport event advertisement, the filter could state to use only videos of other events of the same sport. These filters discard videos that does not satisfy them. This way, the database subset used in the following stages is smaller to improve performance.

Low level filters. These filters are used to score segments of the videos in the database subset after the high level filtering stage. Any Mpeg-7 visual descriptor can be added to the program pattern, in a multiple filter scheme. In this implementation the filters included are dominant color and motion activity. Other filters such as texture or face detection (supported by the Mpeg-7 standard) could also be used. Note that these filters does not discard videos out of the process but compute a similarity score for them.

Program structure parameters. This set of parameters defines the rate of shot boundaries to be included in the new program. Three presets are provided: slow, medium or fast program. Each of these presets scores input video segments differently: while slow programs score long segments with higher values, fast programs score them with very low values. Program, structure parameters are also used to decide the kind of transitions (hard-cuts for fast programs, fades for slow programs) and their duration.

Production parameters. Production parameters guide the generation of the program (program duration, program size, shot duration, shot boundaries type, ...).

Temporal restriction parameters They define parts of input content that must not be included in the new program. Forbidden ranges can be defined (forbidding the last 20.0% of the video ensures that the ending of a movie is not included).

Program requirements. Definition of spatial and temporal characteristics:

- Duration Defines the temporal duration of the new program. Three parameters can be defined: preferred, minimum and maximum duration. The generation engine will try to get a final duration as close as possible to the preferred value, while meeting minimum and maximum parameters. Notice that other parameters in the program pattern can affect the duration of the program.
- Spatial resolution Defines the width and height of the new program. Three choices are provided: resolution of the largest input video, resolution of the smallest video or custom resolution.

Input content parameters. Define the way that input content is used.

- Entire Sources Flag When this flag is enabled, the generation engine uses complete input videos. This means that when a segment of a video appears in the final program, the rest of this video segments also appear.

- Sequential Order Flag When this flag is enabled, segments belonging to the same video must be sorted chronologically.
- Mixed Sources Flag When this flag is enabled, the generator can mix segments of different videos. If it is disabled, all the selected segments of each video must appear together in the new program.
- Use All Sources Flag When this flag is enabled, at least one segment of each input video appears in the new program.

Transition effects parameters. This set of parameters defines the rate of appearance of each shot boundary type (hard cuts, dissolves, fades). Together with the program structure parameters and the final number of selected shots, the number of hard cuts, dissolves and fades will be computed.

Speed parameters. This set of parameters defines the playing speed of input content in the new program. The speed value must be in the $[0.0, 10.0]$ range. A value of 1.0 means no speed change. Values above 1.0 causes the input content to be played faster while values below 1.0 slower.

3.3 Program Generation Process

Initialization. Some internal values are initialized:

- Minimum/Preferred/Maximum duration Scaled versions of these values are computed taking into account the speed parameters, which allows other modules to ignore them. If speed is $\alpha \in (0, 10]$, then provided durations, T_{min} T_{pref} and T_{max}, are rescaled as $T'_{min} = \alpha T_{min}$, $T'_{pref} = \alpha T_{pref}$ and $T'_{max} = \alpha T_{max}$. These values are stored in the *EstimatedTime* structure.
- Fade/Dissolve duration A similar argument can be used for these two temporal values. In this case, *ProgramStructure* parameters must be taken into account as transition durations are affected by them.

After that, every video is assigned a VideoCandidate structure, which stores information useful for this and the rest of stages. That information includes a set of *VideoSegmentCandidate* structures, which store relevant information for every segment of the current video as shown in the following table.

state	*Potential* could be included in the new program. Initial state. *Rejected* cannot be included in the new program *Selected* will be included in the new program
mediaTime	Beggining time point for the current segment
mediaDuration	Duration of the current segment
lowLevelSimilarity	Score value of low level filtering
structureSimilarity	Score value of program structure preferences
keySequenceImportance	Score value of keyframe content

High level filtering. High level filtering relies in manual annotation provided by the supervisor of the system and stored together with automatic annotations in the Mpeg-7 description. These are, by nature, application dependant and often based on natural language. The authors are firmly convinced that

high level filtering must be separated from the main process, using a native XML DataBase Management System, to get better extensibility and re-usability behavior. Besides, these applications are optimized to find textual information in XML documents, getting a great performance even with large databases. Using this idea, the generation process can be divided in the following stages:

- Creation of annotation for input content automatic/manual analysis tools.
- Selection of input content for the microprogram achieved using native XML DBMS search engines. Using the knowledge about the program, the user selects the most appropriate content from the database.
- Program creation the generation engine receives the pruned database and continues with the process.

Low level filtering. In this stage, a low level features based filtering is performed to the input content. The goal is to score each segment of incoming content, so the most relevant are chosen to be included in the program. Similarity fields of VideoSegmentCandidates store scores for the different matching processes every segment must come across.

The LowLevelSimilarity value is computed as $LLS = \omega_1 C + \omega_2 M$, where $0 \leq LLS \leq 1$, ω_1 is the Color similarity (C) relevance weight and ω_2 is the Motion similarity relevance weight(M). C is computed by comparing the Dominant Color descriptor provided in the program pattern and the one in the Mpeg-7 descriptor of this segment. A similar process is done to compute M using, this time, Motion Activity descriptors. Both values are in the $[0.0, 1.0]$ range.

To compute the StructureSimilarity value, SS, score functions are defined that assigns, for each type of program structure, a score value based on the length of the segment. For fast programs, the highest score values will be assigned to short segments while for slow programs, it will be assigned to long segments.

A combined score is then computed as $Score = \omega_{lls} LLS + \omega_{ss} SS$ where ω_{lls} and ω_{ss} are relevance weights assigned to each similarity measure (i.e if a low level filter is not set in the program pattern, the value of ω_{lls} will be zero).

Temporal requirements management. In this stage, those segments which does not meet the temporal requirements defined in the program pattern are marked as rejected. MediaTime and MediaDuration fields of each VideoSegmentCandidate are compared to defined forbidden intervals, marking invalid as rejected.

Segment selection. The generation engine is now ready to select the segments that will be part of the new program. The selection process is performed in different ways depending on the value of the *EntireSources* flag (recall that when this flag is enabled, all segments of the input videos must be included in the program).

When the EntireSources flag is enabled, input videos are not considered as a set of segments but as a whole. Input videos longer than the maximum duration are rejected. If any video is rejected, the *UseAllSources* flag must be disabled or

an error will be generated. Then, the average value of the score of their segments is assign to each video. The generation engine tries to find the highest scored combination of VideoCandidates that meets EstimatedTime requirements. For that reason, every possible combination of input videos in the EstimatedTime defined range (the order is not relevant) will be considered. If UseAllSources flag is enabled, there is only one combination.

When the EntireSources flag is disabled, the generation engine selects segments to build the new program by their score. When no filters are defined (no color, motion or structure preferences) in the program pattern, a random value is generated and assigned to the missing score values. If UseAllSources is enabled, a pre-stage is performed where the best segment of each input video source is selected to make sure that at least one segment of each video is used. Then, the generation engine continues selecting segments (this time, based only on their score) until the additive duration reaches the preferred value. When UseAllSources flag is disabled, the selection pre-stage is omitted.

Segments and transition sorting. In this stage, selected segments and transitions are placed in the time line of the new program. Depending on the flags:

- AllowMixing=False, SequentialOrder=True All segments belonging to a video must be placed together and chronologically sorted. After placing all the segments of a VideoCandidate, it proceeds to the following.
- AllowMixing=False, SequentialOrder=False A ramdom index array, V, is generated for each VideoCandidate, which defines the order of its segments. For a VideoCandidate with n segments: $V = s_1, s_2, \ldots, s_n$, where $1 \leq s_i \leq n$ y $1 \leq i \leq n$. Going through this array sequentially provides a ramdom sort of segments. After placing all the segments of a VideoCandidate, it proceeds to the following.
- AllowMixing=True, SequentialOrder=True A sorted index array is generated for each VideoCandidate. The generation engine chooses a VideoCandidate randomly and places its next segment. Sequential order is ensured because in every VideoCandidate array, segments are sorted; segments from different videos are mixed because VideoCandidate are randomly selected.
- AllowMixing=true, SequentialOrder=False using a random index array for each VideoCandidate.

After that, the number of transitions is known, so using TransitionParams the module computes the necessary number of each kind. They are placed randomly.

Temporal settings management.

When the duration is above maximum time, the program can be cut by making the overlapping in each dissolve longer or making segments shorter. The former case is possible because during a dissolve frames of both shots are placed at the same time point. If necessary, a percentage some segments will be cut (those with the lowest score).

Video coding. This is the last stage of the program generation. The final list of segments and transitions is used to create the final video file. Taking into account speed options and knowing the list of segments, the generator chooses the appropriate frame and encondes it into the program. It will be spatially rescaled if necessary.

4 Results

A test application was created to evaluate the process performance and the quality of the result. Its GUI ease the creation of the XML program pattern.

One of the main concerns in every video processing application is performance. Using Mpeg-7 annotations helps by allowing to reuse analysis results. Table 1 shows the time consumed by each stage of the generation for a set of input videos. The most time consuming task is the creation of the output video file.

Another important result is the validity of fields in the program pattern to obtain customized results. Tests performed showed that the use of many restrictions (low level filters, ...) normally leads to non-desired results (few segments will meet all restrictions at once). The generator accurately follows Dominant Color filter, as shown in figure 1. Motion filtering highlights the activity nature of the program (slow, fast) when combined with the appropriate Program Structure filter (i.e. selecting high activity motion with fast program structure yields short segments full of motion). Setting a coherent and simple set of filters produces appropriate results.

About segment placement, enough configuration options are provided for most purposes (combining *InputContent* flags). The main drawback is the placement of transition, now randomly generated, which should follow some rules(i.e. fades are preferred between low activity motion shots, ...).

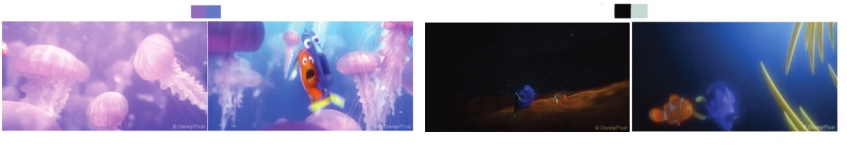


Fig. 1. Two dominant color filtering samples: at the left, magenta+blue; at the right, black+white (filters are shown at the top)

Table 1. Generation time for each stage (sec)

Length	LowLevel	Structure	Selection	Sorting	Timing	Creation
20	0,59	0,75	1,46	0,25	0,02	32,71
200	0,82	0,96	3,23	0,57	0,02	320,92

5 Conclusions

A system for creating new programs using automatically generated Mpeg-7 descriptors has been presented. Using the program pattern, the user can customize the result: segment selection (high/low level filtering), placement in the program, kind of transitions, ... Results obtained are quite promising. Additional manual configuration in video production tasks could help to improve the quality.

References

1. Tekalp AM Ferman AM. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia*, 5:244–256, 2003.
2. Y. Fu, A. Ekin, A. Tekalp, and R. Mehrotra. Temporal segmentation of video objects for hierarchical objectbased motion description, 2002.
3. Yihong Gong and Xin Liu. Video summarization with minimal visual content redundancies. In *Computer Vision and Pattern Recognition*. Proceedings IEEE, 2001.
4. MPEG Requirements Group. Introduction to mpeg-7: Context, objectives and technical roadmap. iso/mpeg m6602, 54th mpeg meeting, la baule, october 2000, 2000.
5. R. Lienhart and C. Kuhm nch abd W. Effelsberg. On the Detection and Recognition of Television Commercials. In *Proceedings of the International Conference on Multimedia Computing and Systems*, 1997.
6. Rainer Lienhart. Comparison of Automatic Shot Boundary Detection Algorithms. In *Image and Video Processing VII*. SPIE, 1999.
7. Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Communications of the ACM*, 40(12):54–62, 1997.
8. JungHwan Oh and Praveen Sankuratri. Computation of Motion Activity Descriptors in Video Segments. Technical report, Department of Computer Scince and Engineering, University of Texas at Arlington, 2001.

Graph Pattern Spaces from Laplacian Spectral Polynomials

Bin Luo, Richard C. Wilson, and Edwin R. Hancock

Department of Computer Science, University of York, York Y010 5DD, UK.

Abstract. Graph structures have proved computationally cumbersome for pattern analysis. The reason for this is that before graphs can be converted to pattern vectors, correspondences must be established between the nodes of structures which are potentially of different size. To overcome this problem, in this paper we turn to the spectral decomposition of the Laplacian matrix. We show how the elements of the spectral matrix for the Laplacian can be used to construct symmetric polynomials that are permutation invariants. The co-efficients of these polynomials can be used as graph-features which can be encoded in a vectorial manner. We explore whether the vectors of invariants can be embedded in a low dimensional space using a number of alternative strategies including principal components analysis (PCA), multidimensional scaling (MDS) and locality preserving projection (LPP).

1 Introduction

The analysis of relational patterns, or graphs, has proved to be considerably more elusive than the analysis of vectorial patterns. One of the challenges that arises in these domains is that of knowledge discovery from large graph datasets. The tools that are required in this endeavour are robust algorithms that can be used to organise, query and navigate large sets of graphs. In particular, the graphs need to be embedded in a pattern space so that similar structures are close together and dissimilar ones are far apart. Moreover, if the graphs can be embedded on a manifold in a pattern space, then the modes of shape variation can be explored by traversing the manifold in a systematic way. The process of constructing low dimensional spaces or manifolds is a routine procedure with pattern-vectors. A variety of well established techniques such as principal components analysis, multidimensional scaling and independent components analysis, together with more recently developed ones such as locally linear embedding [5], isomap [6] and locality preserving projection [7] exist for solving the problem. However, there are few analogous methods which can be used to construct low dimensional pattern spaces or manifolds for sets of graphs.

There are two reasons why pattern-vectors are more easily manipulated than graphs. First, there is no canonical ordering for the nodes in a graph, unlike the components of a vector. Hence, correspondences to a reference structure must be established as a prerequisite. The second problem is that the variation in the graphs of a particular class may manifest itself as subtle changes in structure, which may involve different numbers of nodes or different edge structure. Even if the nodes or the edges of a graph could be encoded in a vectorial manner, then the vectors would be of variable length. One

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 327–334, 2004.
© Springer-Verlag Berlin Heidelberg 2004

way of circumventing these problems is to develop graph-clustering methods in which an explicit class archetype and correspondences with it are maintained [9]. One of the problems that hinders graph clustering is the need to define a class archetype that can capture both the salient structure of the class and the modes of variation contained within it. For instance, the random graphs of Wong, Constant and You [11] capture this distribution using a discretely defined probability distribution, and Bagdanov and Worring [10] have overcome some of the computational difficulties associated with this method by using continuous Gaussian distributions.

However, one of the criticisms that can be aimed at these methods for learning the distribution of graphs is that they are in a sense brute force because of their need for correspondences either to establish an archetype or to compute graph similarity. For noisy graphs (those which are subject to structural differences) this problem is thought to be NP-hard. Although relatively robust approximate methods exist for computing correspondence [1], these can prove time consuming. In this paper, we are interested in the problem of constructing pattern spaces for sets of graphs. To do this we vectorise graphs and use standard manifold learning methods to project the vectors into a metric space. Our aim is to construct the vectors in a way which does not require the computation of node correspondences. We hence turn to a spectral characterisation of graph-structure [2]. Although existing graph-spectral methods have proved effective for graph-matching [3] and indexing [8], they have not made full use of the available spectral representation, and are restricted to the use of either the spectrum of eigenvalues or a single eigenvector. Moreover, in prior work we have demonstrated that simple hand-crafted spectral features that reflect the "shape" of graphs can be used for clustering [4]. The aims here are more ambitious since our characterisation is based on permutation invariants computed from the elements of the full spectral matrix. The representation is hence both more principled and richer.

2 Graph Spectra

In this paper we are concerned with the set of graphs $G_1, G_2, .., G_k, ..., G_N$. The kth graph is denoted by $G_k = (V_k, E_k)$, where V_k is the set of nodes and $E_k \subseteq V_k \times V_k$ is the edge-set. These graphs are undirected. Our approach in this paper is a graph-spectral one. For each graph G_k we compute the adjacency matrix A_k. This is a $|V_k| \times |V_k|$ matrix whose element with row index i and column index j is

$$A_k(i,j) = \begin{cases} 1 & \text{if } (i,j) \in E_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The aim in this paper is to perform spectral analysis. Unfortunately, the adjacency matrix can have negative eigenvalues. Instead we turn our attention to the Laplacian matrix, since it is positive semi-definite and therefore has positive or zero eigenvalues. For the graph indexed k the Laplacian matrix is $L_k = A_k - D_k$ where D_k is diagonal degree matrix with elements of $D_k(i,i) = \sum_{j \in V_k} A_k(i,j)$. The spectral decomposition of the Laplacian matrix is $L_k = \sum_{i=1}^{n} \lambda_i^k \boldsymbol{e}_i^k \boldsymbol{e}_i^{k\,T}$, where λ_i^k is the ith eigenvalue and $\boldsymbol{e}_i^k$ is the corresponding eigenvector of the Laplacian matrix L_k of the graph indexed k. With the eigenvalues and eigenvectors at hand the spectral matrix is given by

$\Phi_k = \left(\sqrt{\lambda_1^k}\boldsymbol{e}_1^k|\sqrt{\lambda_2^k}\boldsymbol{e}_2^k|\ldots|\sqrt{\lambda_n}\boldsymbol{e}_{|V_k|}^k\right)$. Since the eigevalues of the Laplacian matrix are real and positive or zero, the corresponding roots are real. The graph spectral matrix Φ_k is a complete representation of a graph G_k in the sense that the original Laplacian matrix which depicts the adjacency of graph nodes can be reconstructed using the relationship $L_k = \Phi_k\Phi_k^T$. Moreover, Φ_k is unique for a simple graph G (without repeated eigenvalues) up to a permutation of columns. Hence, we can not use the elements of Φ_k to construct useful graph features, since they are not permutation invariants. To overcome this problem we use symmetric polynomials to compute permutation invariants. The polynomials are defined as follows

$$\begin{aligned} S_1(x_1,\ldots x_n) &= \sum_{i=1}^{n} x_i \\ &\vdots \\ S_r(x_1,\ldots x_n) &= \sum_{i_1<i_2<\ldots<i_r} x_{i_1}x_{i_2}\ldots x_{i_r} \\ &\vdots \\ S_n(x_1,\ldots x_n) &= \prod_{i=1}^{n} x_i \end{aligned}$$

To overcome the problem that the resulting distributions may be extremely long-tailed, we take logarithm of the elements of S as the new features,

$$F_{i,j}^k = \mathrm{s}ignum(S_{i,j}^k)\ln(1+|S_{i,j}^k|) \tag{2}$$

We form a feature vector B_k for graph G_k using the standard vectorisation operation on the elements of the matrix F

$$\boldsymbol{B}_k = (F_{1,1}^k,\ldots,F_{1,n}^k,F_{2,1}^k,\ldots,F_{2,n}^k,\ldots,F_{n,1}^k,\ldots,F_{n,n}^k)^T. \tag{3}$$

In order to accommodate graphs of different sizes, we need to be able to compare representations of different sizes. This is achieved by expanding the representation. Consider two graphs of size m and n, $m<n$. If we add $n-m$ nodes with no connections to the first graph, we obtain two graphs of the same size. The edit cost in terms of edge insertions and deletions between these two graphs is identical to the original pair. The effect on the spectral representation is merely to add trailing zeros to each eigenvector and additional zero eigenmodes. As a consequence, the first m elementary symmetric polynomials are unchanged, and the subsequent $n-m$ are zero. The new representation in S can therefore be easily calculated from the original feature vector.

3 Graph Embedding Methods

We explore three different methods for embedding the graph feature vectors in a pattern space.

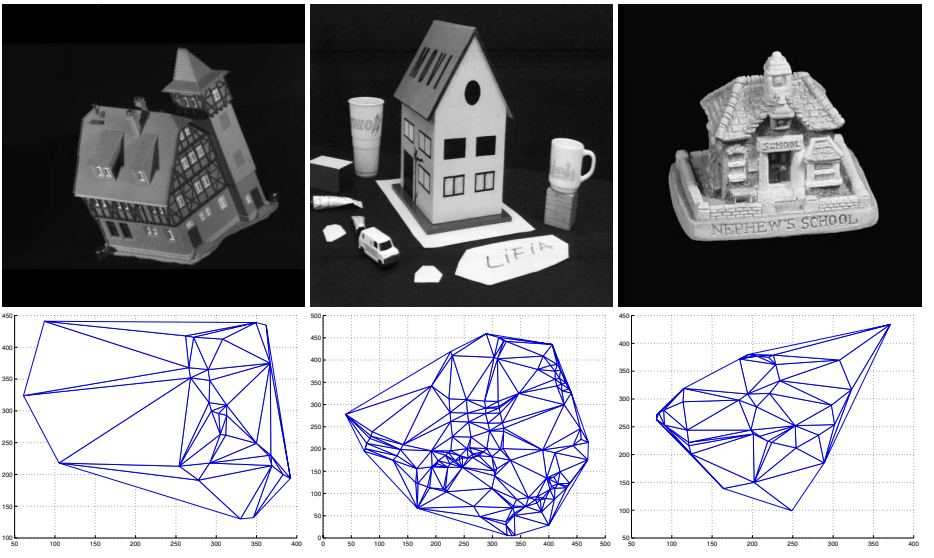


Fig. 1. Examples images and corresponding graphs.

3.1 Principal Component Analysis

We commence by constructing the matrix $S = [B_1|B_2|\ldots|B_k|\ldots|B_N]$. with the graph feature vectors as columns. Next, we compute the covariance matrix for the elements of the feature vectors by taking the matrix product $C = SS^T$.. We extract the principal components directions by performing the eigendecomposition $C = \sum_{i=1}^{N} l_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ on the covariance matrix C, where the l_i are the eigenvalues and the $\boldsymbol{u}_i$ are the eigenvectors. We use the first s leading eigenvectors (2 or 3 in practice for visualisation purposes) to represent the graphs extracted from the images. The co-ordinate system of the eigenspace is spanned by the three orthogonal vectors $Ue = (\boldsymbol{u}_1, \boldsymbol{u}_2, .., \boldsymbol{u}_s)$. The individual graphs represented by the long vectors $B_k, k = 1, 2, \ldots, N$ can be projected onto this eigenspace using the formula $\boldsymbol{x}_k = U^T B_k$.. Hence each graph G_k is represented by an s-component vector $\boldsymbol{x}_k$ in the eigenspace.

3.2 Multidimensional Scaling

Multidimensional scaling(MDS) is a procedure which allows data specified in terms of a matrix of pairwise distances to be embedded in a Euclidean space. Here we intend to use the method to embed the graphs extracted from different viewpoints in a low-dimensional space. To commence we require pairwise distances between graphs. We do this by computing the L2 norms between the spectral pattern vectors for the graphs. For the graphs indexed $i1$ and $i2$, the distance is $d_{i1,i2} = (B_{i1} - B_{i2})^T (B_{i1} - B_{i2})$. In this paper, we use the classical multidimensional scaling method to embed the view-graphs in a Euclidean space using the matrix of pairwise dissimilarities D. The first step of MDS is to calculate a matrix T whose element with row r and column c is given by $T_{rc} = -\frac{1}{2}[d_{rc}^2 - \hat{d}_{r.}^2 - \hat{d}_{.c}^2 + \hat{d}_{..}^2]$, where $\hat{d}_{r.} = \frac{1}{N}\sum_{c=1}^{N} d_{rc}$ is the average dissimilarity

value over the rth row, $\hat{d}_{.c}$ is the similarly defined average value over the cth column and $\hat{d}_{..} = \frac{1}{N^2}\sum_{r=1}^{N}\sum_{c=1}^{N} d_{r,c}$. is the average dissimilarity value over all rows and columns of the matrix T.

We subject the matrix T to an eigenvector analysis to obtain a matrix of embedding co-ordinates X. If the rank of T is $k, k \leq N$, then we will have k non-zero eigenvalues. We arrange these k non-zero eigenvalues in descending order, i.e. $l_1 \geq l_2 \geq \ldots \geq l_k > 0$. The corresponding ordered eigenvectors are denoted by $\boldsymbol{u}_i$ where l_i is the ith eigenvalue. The embedding co-ordinate system for the graphs obtained from different views is $X = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_s]$, where $\boldsymbol{f}_i = \sqrt{l_i}\boldsymbol{u}_i$ are the scaled eigenvectors. For the graph indexed i, the embedded vector of co-ordinates is $\boldsymbol{x}_i = (X_{i,1}, X_{i,2}, ..., X_{i,s})^T$..

3.3 Locality Preserving Projection

Our next pattern space embedding method is He and Niyogi's Locality Preserving Projections(LPP) [7]. LPP is a linear dimensionality reduction method which attempts to project high dimensional data to a low dimensional manifold, while preserving the neighbourhood structure of the data set. The method is relatively insensitive to outliers and noise. This is an important feature in our graph clustering task since the outliers are usually introduced by imperfect segmentation processes. The linearity of the method makes it computationally efficient.

The relational structure of the data is represented by a proximity weight matrix W with elements $W(i1, i2) = \exp[-kd(i2, i2)]$, where k is a constant. If Q is the diagonal degree matrix with the row weights $Q(k, k) = \sum_{j=1}^{N} W(k, j)$ as elements, then the relational structure of the data is represented using the Laplacian matrix $J = Q - W$. The idea behind LPP is to analyse the structure of the weighted covariance matrix SWS^T. The optimal projection of the data is found by solving the generalised eigenvector problem $SJS^T\boldsymbol{u} = lSQS^T\boldsymbol{u}$.. We project the data onto the space spanned by the eigenvectors corresponding to the s smallest eigenvalues.

4 Experiments

Our experimental vehicle is provided by 2D views of 3D objects. We have collected sequences of views for a number of objects. For the different objects the image sequences are obtained under slowly varying changes in viewer angle. From each image in each view sequence, we extract corner features. We use the extracted corner points to construct Delaunay graphs. In our experiments we use three different sequences. Each sequence contains images with equally spaced viewing directions. For each sequence we show a sample of the images and the resulting graph structure in Figure 1.

In Figure 2 we compare the results obtained with the different embedding strategies and different graph features. In the left-hand column, we show the results obtained when PCA is used, the middle column when MDS is used, and the right-hand column when LPP is used. The top row show the results obtained using a standard spectral feature, namely the spectrum ordered eigenvalues of the Laplacian i.e. $B_k = (\lambda_1^k, \lambda_2^k, ...)^T$. The second row shows the embedding obtained by computing the symmetric polynomials, using the spectral matrix for the graph *adjacency* matrices. The third row shows the results

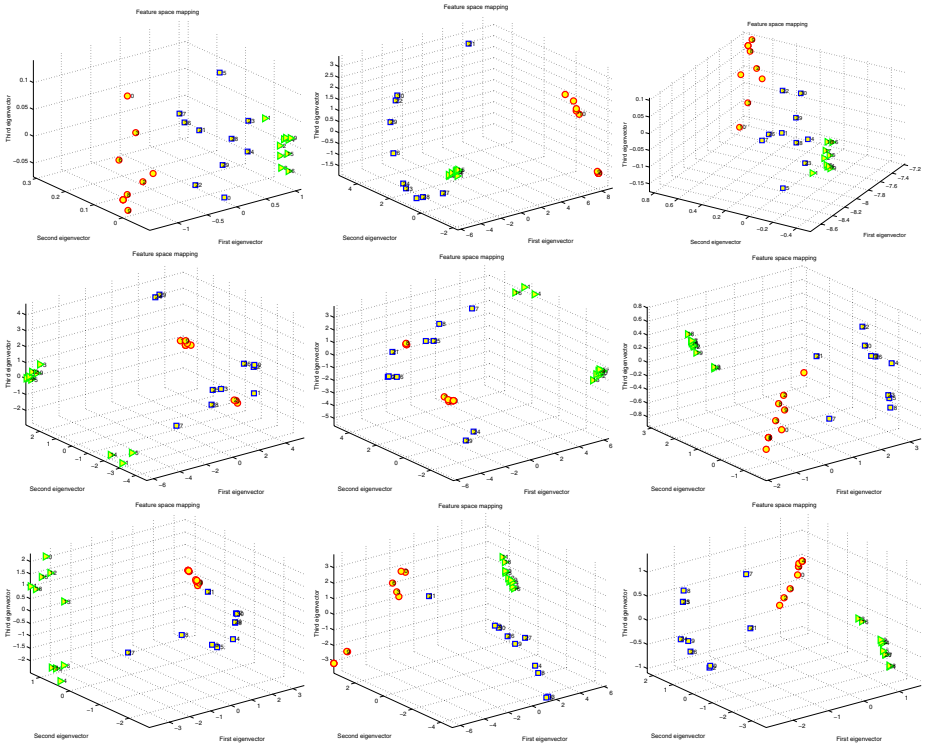


Fig. 2. 3D: Clustering CMU, MOVI and chalet sequences, Row 1: Eigenvalues, Row 2: Adjacency Matrix Polynomials, Row 3: Laplacian Matrix Polynomials; Col 1: PCA, Col 2: MDS, Col 3: LPP

obtained when the symmetric polynomials are computed using the spectral matrix for the *Laplacian*.

There are a number of conclusions that can be drawn from this plot. First, the most distinct clusters are produced when either MDS or LPP are used. Second, the best spectral features seem to be the symmetric polynomials computed the Laplacian spectral matrix. By contrast, when the spectral matrix of the adjacency matrix is used the cluster structure is less well defined. To display the cluster-structure obtained, in Figure 3 we visualise the results obtained using LPP and the Laplacian polynomials by placing thumbnails of the original images in the space spanned by the leading two eigenvectors.

To take this study further we investigate the effects of applying the embedding methods to two of the sequences separately. In Figure 4 for the chalet sequence, the top row shows the results obtained with the adjacency matrix polynomials while the bottom row shows the results obtained with the Laplacian polynomials. The left-hand column is for PCA, the middle column for MDS and the right-hand column for LPP. LPP results in a smooth trajectory as the different views in the sequence are traversed.

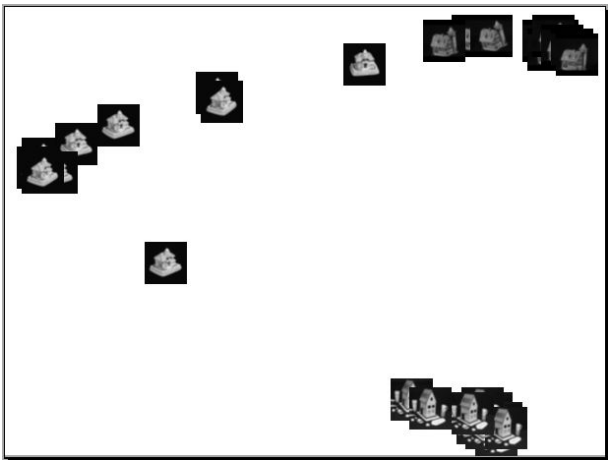


Fig. 3. LPP space clustering using the graph polynomials.

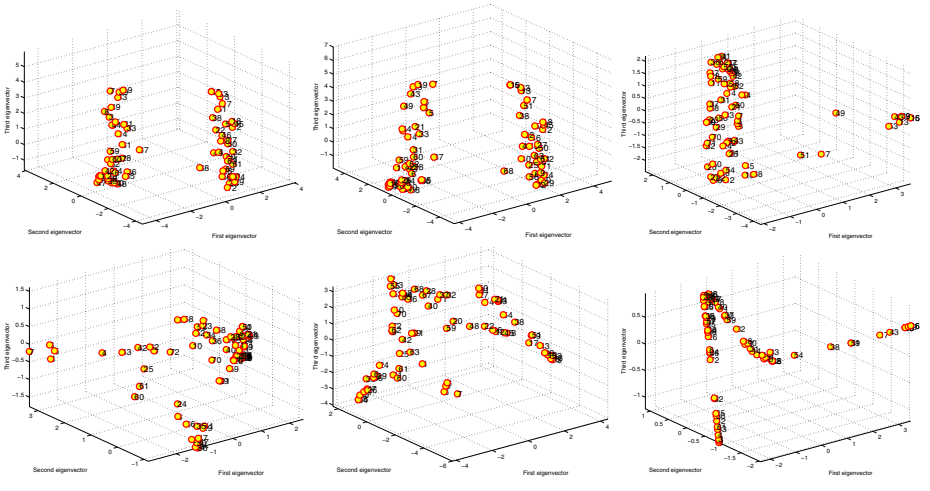


Fig. 4. Comparison of using graph adjacency matrices and Laplacian matrices for the chalet sequence.

5 Conclusions

In this paper we have shown how graphs can be converted into pattern vectors by utilising the spectral decomposition of the Laplacian matrix, and basis sets of symmetric polynomials. The Laplacian matrix appears to give better performance than the adjacency matrix. We then investigate how to embed the vectors in a pattern space, sutaible for clustering the graphs (i.e PCA, MDS and LPP). Results show that LPP gives the best clusters.

References

1. W. J. Christmas, J. Kittler and M. Petrou, "Structural Matching in Computer Vision using Probabilistic Relaxation", *IEEE PAMI*, **17**, pp. 749–764, 1995.
2. F.R.K. Chung, "Spectral Graph Theory", *American Mathmatical Society Ed.*, CBMS series 92, 1997.
3. S. Umeyama, "An eigen decomposition approach to weighted graph matching problems", *IEEE PAMI*, **10**, pp. 695–703, 1988.
4. B. Luo, R. C. Wilson and E. R. Hancock, "Spectral Embedding of Graphs", *Pattern Recognition*, **36**, pp. 2213–2230, 2003
5. S. Roweis and L.Saul, "Non-linear dimensionality reduction by locally linear embedding", *Science*, **299**, pp. 2323-2326, 2002.
6. J.B. Tenenbaum, V.D. Silva and J.C.Langford, "A global geometric framework for non-linear dimensionality reduction", *Science*, **290**, pp. 586–591, 2000.
7. X.He and P. Niyogi, "Locality preserving projections", *to appear in NIPS03*.
8. A. Shokoufandeh, S. Dickinson K. Siddiqi and S. Zucker, "Indexing using a Spectral Coding of Topological Structure", *CVPR*, pp. 491–497, 1999.
9. J. Segen, "Learning graph models of shape", in J. Laird, editor, *Proceedings of the Fifth International Conference on Machine Learning*, pp. 29–25, 1988.
10. A.D. Bagdanov and M. Worring, "First Order Gaussian Graphs for Efficient Structure Classification", *Pattern Recogntion*, **36**, pp. 1311-1324, 2003.
11. A.K.C Wong, J. Constant and M.L. You, "Random Graphs", *Syntactic and Structural Pattern Recognition*, World Scientific, 1990.

A Hierarchical Framework for Shape Recognition Using Articulated Shape Mixtures

Abdullah Al Shaher and Edwin R. Hancock

University of York,York YO1 5DD, UK.
{abdullah,erh}@minster.cs.york.ac.uk

Abstract. This paper describes a statistical framework for recognising 2D shapes with articulated components. The shapes are represented using both geometrical and a symbolic primitives, that are encapsulated in a three layer hierarchical architecture. Each primitive is modelled so as to allow a degree of articulated freedom using a polar point distribution model that captures how the primitive movement varies over a training set. Each segment is assigned a symbolic label to distinguish its identity, and the overall shape is represented by a configuration of labels. We demonstrate how both the point-distribution model and the symbolic labels can be combined to perform recognition using the hierarchical mixture of experts algorithm. This involves recovering the parameters of the point distribution model that minimise an alignment error, and recovering symbol configurations that minimise a structural error. We apply the recognition strategy on sets of Arabic characters.

1 Introduction

The task of recognising articulated shapes has attracted considerable interest in computer vision. The main problem is how to robustly recover correspondence when the object being tracked undergoes deformations and the detected feature points defining the object are subject to noise. One of the most effective ways of developing matching techniques is to draw on probabilistic and statistical methods. This approach has lead to the development of point distribution models [1], deformable templates [2] and condensation [3]. One of the most important modes of variation in a moving shape, especially biological forms, is that of articulation.

There are a number of ways in which object articulation can be modelled. Perhaps the simplest of these is to decompose the shape into a skeletal form, consisting of limbs or branches, and to model the articulation of the branches. The mechanical movement of the resulting shape can be captured by the rotation of the components. However, in order to constrain the change in shape to be physically realistic bounds, or distributions, must be imposed on the rotation angles [4]. Hence, the mechanical constraints on articulation must be combined with a statistical model of limb movement. In addition to movement of the limbs, the articulated shape also has a structural composition, since the limbs can be

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 335–343, 2004.
© Springer-Verlag Berlin Heidelberg 2004

assigned labels to distinguish them, and the arrangement of the labels used to provide further constraints for shape-recognition.

The aim in this paper is to develop a statistical framework that can be used to recognise articulated shapes using information concerning limb movement and symbolic constraints concerning the overall shape structure. To do this we turn to a simplified version of the hierarchical mixture of experts algorithm [5]. Each shape is represented as an arrangement of articulated limbs. The movement of the limbs is represented by a polar point distribution model. The structural component of the model is represented by a configuration of limb-labels. The recognition architecture has two intercommunicating layers. The first of these is concerned with limb alignment, and this aims to recover the lengths and polar angles of the limbs. The second aims to assign limb-labels so that the overall structure is recovered consistently. The fitting of this two-level model to data is effected using a variant of the expectation-maximisation algorithm.

2 Shape Representation

Our aim is to use point distribution models to account for shape deformations due to limb articulation. The model is a two component one. First, we have a limb-model. This accounts for the variations in shape of each of the individual limbs using a point distribution model to describe the modes of variation of the landmark points about a mean shape. Second, we have a limb arrangement model. This is an augmented point distribution that describes the arrangement of the centre points of the limbs, and their polar angles.

We are concerned with recognising 2D shapes by modelling segment movement around the centre of the shape. The shape under study is assumed to be segmented into a set of K jointed and non-overlapping limbs. The k^{th} limb is represented by a long-vector of consecutive landmark points

$$X_k = (x_1^k, y_1^k, x_2^k, y_2^k, ...x_{n_k}^k, y_{n_k}^k)^T \tag{1}$$

The centre-of-gravity of the limb indexed k is

$$\boldsymbol{c}_k = \frac{1}{n_k} \sum_{i=1}^{k_n} (x_i^k, y_i^k)^T \tag{2}$$

The overall shape is represented by a long-vector of consecutive limb centres $C = (\boldsymbol{c}_1^T, \boldsymbol{c}_2^T, .., \boldsymbol{c}_K^T)^T$. The centre of articulated shape is computed by averaging the centre of the limbs

$$\boldsymbol{U} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{c}_k \tag{3}$$

To model the articulated shape, we use a polar variant of the standard point distribution model [6]. This model allows the primitives to move about the centre of articulation. According to this model the shape is viewed as an arrangement of non-overlapping primitives. Each primitive is represented by mean point c_k. Limb articulation is represented by a set of limb-angles. For the k^{th} limb the angle is defined to be

$$\theta_k = \tan^{-1} \frac{U(y) - c_k(y)}{U(x) - c_k(x)} \tag{4}$$

and the angular arrangement of the limbs is represented by the vector $\Theta = (\theta_1, \theta_2, ..., \theta_K)^T$. The global configuration of the limbs within a shape is specified by the set of angles and the centres-of-articulation, i.e. by $S = (\Theta^T, C^T)^T$.

To augment the geometric information, we assign symbols to the articulated components. Each training pattern is assigned to a shape class and each component primitive is assigned to a primitive class. The set of shape-labels is Ω_c and the set of articulated component or limb labels is Ω_s. The symbolic structure of each shape is represented by a permissible arrangement of limb-labels. For shapes of class $\omega \in \Omega_c$ the permissible arrangement of limbs is denoted by $\Lambda_\omega =< \lambda_1^\omega, \lambda_2^\omega, ... >$.

The first component of our model is a set of point distribution models to represent the variations in the shape of the different limb classes. For the limb of class λ, we compute a mean long-vector $\hat{X}_\lambda$ and a long-vector covariance matrix Σ_λ. The eigenvector expansion of the covariance matrix is $\Sigma_\lambda = \Phi_\lambda \Xi_\lambda \Phi_\lambda$ where Φ_λ is the matrix with the ordered unit eigenvectors as columns and Ξ_λ is the diagonal matrix of ordered eigenvalues. Suppose that $\tilde{\Phi}_\lambda$ is the result of truncating the eigenvector matrix after the first m columns corresponding to the most significant eigenvalues. The point distribution involves deforming the positions of the landmark points in the directions of the eigenvectors of the covariance matrix. The long-vector of deformed point positions is $\tilde{X}_\lambda = \hat{X}_\lambda + \tilde{\Phi}_\lambda \gamma$ where γ is a parameter vector of length m. If the point distribution model is fitted to a set of landmark points $\boldsymbol{z}_k$ for the k^{th} limb, then the least-squares estimate of the parameter vector is $\gamma_\lambda^k = (\tilde{\Phi}_\lambda + \tilde{\Phi}_\lambda^T)(\boldsymbol{z}_k - \hat{X}_\lambda)$.

The second component of the model is a point distribution model for the augmented long-vector of limb-centres and limb angles. The vector of limb parameters for the shape class ω is denoted by Γ_ω. However, we need to constrain the parameters corresponding to the limb angles. Suppose that the mean-vector for the limb-angles is $\hat{\Theta}_w$ and the corresponding covariance matrix is Σ_w. The angular deformations are constrained to avoid flipping by limiting the deformation vector. We use the variance associated with the eigen-modes to constrain the deformation. The k^{th} component of the parameter vector is constrained to fall in the interval $-3\sqrt{l_k} \leq \Gamma_{(k)} \leq 3\sqrt{l_k}$. The articulation angles lie in the range $-180°$ to $180°$ to avoid discontinuities associated with the flip from $0°$ to $360°$.

3 Hierarchical Architecture

With the limb-articulation and limb-centre point distribution models to hand, our recognition method proceeds in a hierarchical manner. Our aim is to classify the set of landmark long-vectors $X = (\boldsymbol{z}_1, ..., \boldsymbol{z}_k, ..., \boldsymbol{z}_K)$ representing a test-shape. To commence, we make maximum likelihood estimates of the best-fit parameters of each limb-model to each set of limb-points. The best-fit parameters γ_λ^k of the limb-model with class-label λ to the set of points constituting the limb indexed k is

$$\gamma_\lambda^k = \arg\max_\gamma p(z_k|\Phi_\lambda, \gamma) \tag{5}$$

We use the best-fit parameters to assign a label to each limb. The label is that which has maximum a posteriori probability given the limb parameters. The

label assigned to the limb indexed k is

$$l_k = \arg\max_{\lambda \in \Omega_s} P(l|\boldsymbol{z}_k, \gamma_\lambda, \Phi_\lambda) \tag{6}$$

In practice, we assume that the fit error residuals follow a Gaussian distribution. As a result, the class label is that associated with the minimum squared error. This process is repeated for each limb in turn. The class identity of the set of limbs is summarised the string of assigned limb-labels $L =< l_1, l_2, ... >$. Hence, the input layer is initialised using maximum likelihood limb parameters and maximum a *posteriori* probability limb labels.

The shape-layer takes this information as input. The goal of computation in this second layer is to refine the configuration of limb labels using global constraints on the arrangement of limbs to form consistent shapes. The constraints come from both geometric and symbolic sources. The geometric constraints are provided by the fit of a limb-centre point distribution model. The symbolic constraints are provide by a dictionary of permissible limb-label strings for different shapes.

The parameters of the limb-centre point distribution model are found using the EM algorithm [7]. Here we borrow ideas from the hierarchical mixture of experts algorithm, and pose the recovery of parameters as that of maximising a gated expected log-likelihood function for the distribution of limb-centre alignment errors $p(X|\Phi_\omega, \Gamma_\omega)$. The likelihood function is gated by two sets of probabilities. The first of these are the a *posteriori* probabilities $P(\lambda_k^\omega|\boldsymbol{z}_k, \gamma_{\lambda_k^\omega}, \Phi_{\lambda_k^\omega})$ of the individual limbs. The second are the conditional probabilities $P(L|\Lambda_\omega)$ of the assigned limb-label string given the dictionary of permissible configurations for shapes of class ω. The expected log-likelihood function is given by

$$\mathcal{L} = \sum_{\omega \in \Omega_c} P(L|\Lambda_\omega) \left\{ \prod_k P(\lambda_k^\omega|\boldsymbol{z}_k, \gamma_{\lambda_k^\omega}, \tilde{\Phi}_{\lambda_k^\omega}) \right\} \ln p(X|\tilde{\Phi}_\omega, \Gamma_\omega) \tag{7}$$

The optimal set of limb-centre alignment parameters satisfies the condition

$$\Gamma_\omega^* = \arg\max_{\Gamma} P(L|\Lambda_\omega) \left\{ \prod_k P(\lambda_k^\omega|\boldsymbol{z}_k, \gamma_{\lambda_k^\omega}, \tilde{\Phi}_{\lambda_k^\omega}) \right\} \ln p(X|\tilde{\Phi}_\omega, \Gamma_\omega) \tag{8}$$

From the maximum likelihood alignment parameters we identify the shape-class of maximum *a posteriori* probability. The class is the one for which

$$\omega^* = \arg\max_{\omega \in \Omega_c} P(\omega|X, \tilde{\Phi}_\omega, \Gamma_\omega^*) \tag{9}$$

The class identity of the maximum *a posteriori* probability shape is passed back to the limb-layer of the architecture. The limb labels can then be refined in the light of the consistent assignments for the limb-label configuration associated with the shape-class ω.

$$l_k = \arg\max_{\lambda \in \Omega_s} P(\lambda|\boldsymbol{z}_k, \gamma_l^k, \tilde{\Phi}_\lambda) P(L(\lambda, k)|\Lambda_\omega) \tag{10}$$

Finally, the maximum likelihood parameters for the limbs are refined

$$\gamma_k = \arg\max_{\gamma} p(\boldsymbol{z}_k|\tilde{\Phi}_{l_k}, \gamma, \Gamma_\omega^*)) \tag{11}$$

These labels are passed to the shape-layer and the process is iterated to convergence.

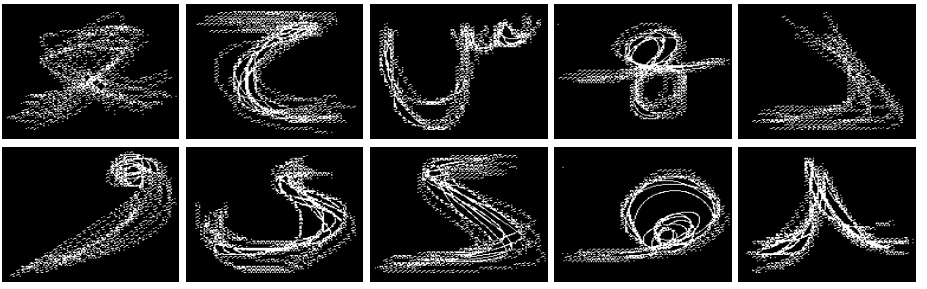


Fig. 1. Training Sets

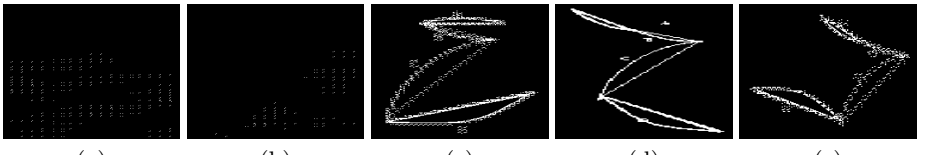


Fig. 2. Articulation example: (a) Sample shape; (b) Primitives movement; (c,d,e) Test shapes with articulated primitives

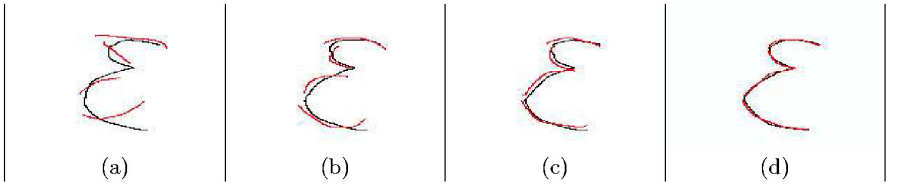


Fig. 3. Articulated Model Alignment: (a) iteration 1, (b) iteration 2, (c) iteration 4, (d) iteration 6

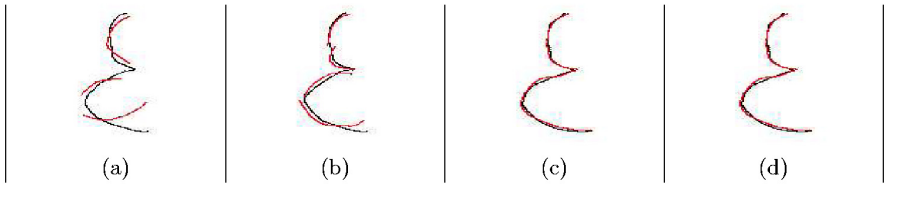


Fig. 4. Articulated Model Alignment: (a) iteration 1, (b) iteration 3, (c) iteration 6, (d) iteration 10

4 Models

To develop a useful alignment algorithm we require a model for the measurement process. Here we assume that the observed position vectors, i.e. $\boldsymbol{z}_k$ are

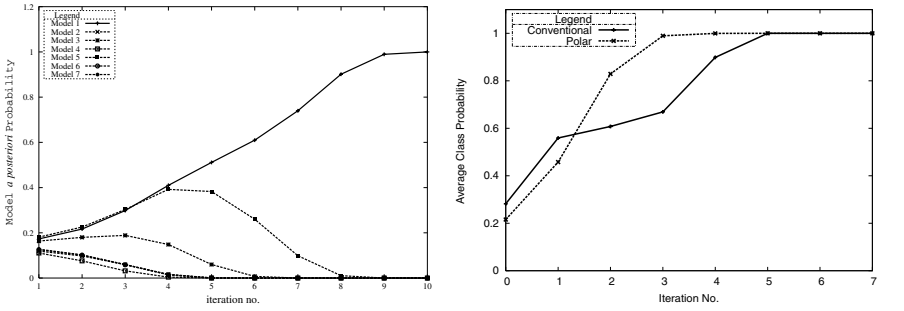


Fig. 5. Alignment: (a) Alignment as a function per iteration number; (b) Comparison between Cartesian and Polar Models

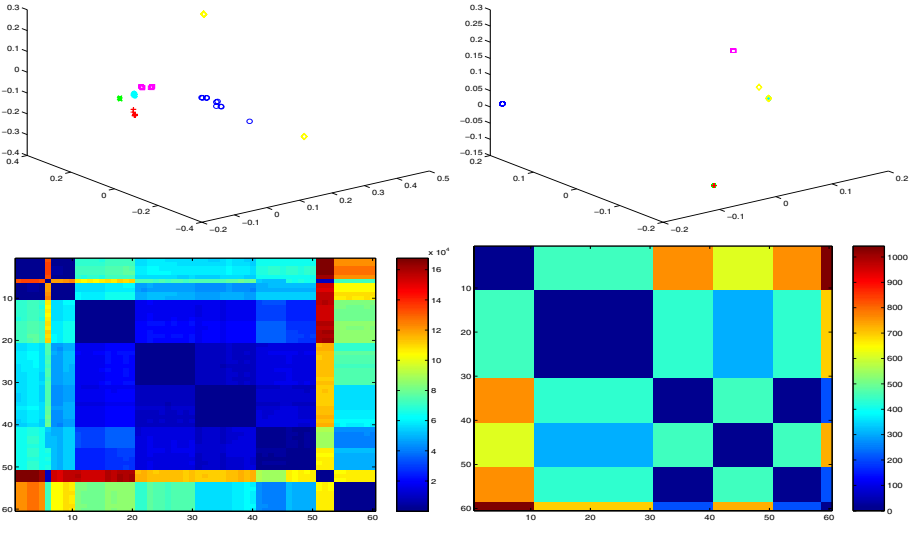


Fig. 6. Principal Component Analysis Classification: top raw is shows embedding of the shapes and bottom raw shows the distance matrix; (a) iteration 1, (b) iteration 4

derived from the model points through a Gaussian error process. According to our Gaussian model of the alignment errors,

$$p(\boldsymbol{z}_k|\tilde{\Phi}_\lambda,\gamma_\lambda) = \frac{1}{2\pi\sigma}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{z}_k - \hat{X}_\lambda - \tilde{\Phi}_\lambda\gamma_\lambda)^T(\boldsymbol{z}_k - \hat{X}_\lambda - \tilde{\Phi}_\lambda\gamma_\lambda)\right] \quad (12)$$

where σ^2 is the variance of the point-position errors which for simplicity are assumed to be isotropic. The maximum likelihood parameter vector is given by

$$\gamma_\lambda^k = \frac{1}{2}\left(\tilde{\Phi}_\lambda^T\tilde{\Phi}_\lambda\right)^{-1}\left(\tilde{\Phi}_\lambda + \tilde{\Phi}_\lambda^T\right)(\boldsymbol{z}_k - \hat{X}_\lambda) \quad (13)$$

A similar procedure may be applied to estimate the parameters of the limb centre point distribution model.

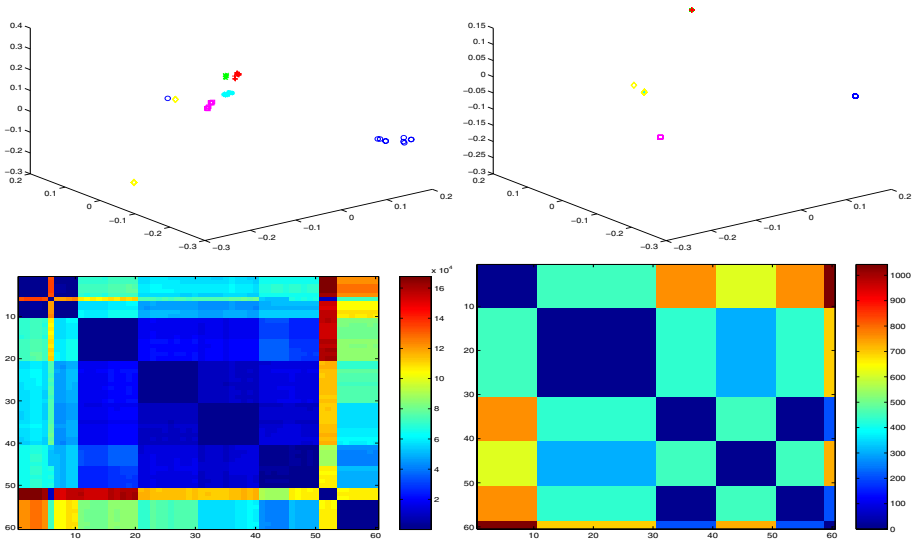


Fig. 7. Multidimensional Scaling Classification: top raw is shows embedding of the shapes and bottom raw shows the distance matrix; (a) iteration 1, (b) iteration 4

4.1 Label Assignment

The distribution of label errors is modelled using the method developed by Hancock and Kittler [8]. To measure the degree of error we measure the Hamming distance between the assigned string of labels L and the dictionary item Λ. The Hamming distance is given by

$$H(L, \Lambda_\omega) = \sum_{i=1}^{K} \delta_{l_i, \lambda_i^\omega} \tag{14}$$

where δ is the DiracDelta function. With the Hamming distance to hand, the probability of the assigned string of labels L given the dictionary item Λ is

$$P(L|\Lambda_\omega) = K_p \exp[-k_p H(L, \Lambda_\omega)] \tag{15}$$

where $K_p = (1-p)^K$ and $k_p = \ln \frac{1-p}{p}$ are constants determined by the label-error probability p.

5 Experiment

We have evaluated our approach on sets of Arabic characters. Figure 1 shows some of the data used for the purposes of learning and recognition. In total we use, 23 distinct classes of Arabic characters and for each class there are 200

samples and 48 classes of stroke articulated models. For a single character, Figure 2 illustrates the modes of articulation for a sample shape.

In Figures 3 and 4, we illustrate the fitting of the model learned from the training data to two shapes falling outside the training-set. The difference between the two examples is that the upper limb or stroke has different articulations with respect to the lower limb. In both cases the model fits the data well after 4 iterations.

In Figure 5(a) we show the a posteriori probabilities for the different shape classes used in training. It is clear that the best-fit shape dominates the alternatives after the fifth iteration. To investigate the advantages offered by the polar articulation model in Figure 5(b) we show the a posteriori probabilities of the best-fit shape as a function of iteration when both a conventional and an articulated model are used. In the case of the articulated model the fitting process converges more rapidly, than in the conventional case.

To explore the capacity of the method to identify clusters of shapes, we have applied Principal Component Analysis and Multidimensional Scaling to the vectors of fitted model parameters extracted using our recognition method. Figures 6 and 7 respectively show the result of applying the PCA and MDS to the alignment parameters for 60 sample shapes drawn at random from the set of shape classes. In the top row of the two figures, we show the result of projecting the data onto the leading three eigenvectors. In the bottom row of the two plots, we show the matrix of pairwise distances computed from the projection of the data into the three dimensional eigenspace. In the left-hand column of each plot, we show the results obtained with the initial parameter vectors, while the second column shows the results with the parameter vectors at convergence. The main effect of iterating the recognition method is to improve the clusters of shapes. This is reflected both by the distribution of data in the eigenspace, and the block structure in the pairwise distance matrices. Moreover, a better cluster structure emerges when MDS is used rather than PCA. However, a deeper analysis of the data reveals that PCA gives a better cluster structure when the shapes are subject to scaling. Table 1 shows the recognition rates per class for 2400 sample shapes. Table 1 compare recognition rates of shapes treated as a whole entity and shapes treated as continuous segment. We have achieved a recognition rate of 97.2% for primitives modelling and 92.3% for character modelling.

Table 1. Recognition Rate for shape-classes

Shape												
Character	190	195	192	190	193	177	184	183	178	186	187	182
Stroke	197	198	197	199	198	189	196	192	194	190	193	188

6 Conclusion

In this paper, we have described a method for fitting articulated shape-models to landmark point data. The shape deformation process adopted is based on point distribution models. The model representation is a hierarchical one. There is a Cartesian deformation model for the limbs and the limb-centres, together with a polar model which represents point articulation. We develop a probabilistic framework for fitting a mixture of articulated models to data. The method is demonstrated to deliver good results on Arabic character data.

References

1. Cootes T.; Taylor C. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–409, 1995.
2. Duta N.; Jain A.; Dubuisson P. Learning 2d shape models. *International Conference on Computer Vision and pattern Recognition*, 2:8–14, 1999.
3. Michael Isard; Andrew Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. European Conf. on Computer Vision*, pages 343–356, 1996.
4. J. Gonzales; J. Varona; F. Roca; and J. Villanueva. aspace: Action space for recognition and synthesis of human actions. *2nd IWAMDO, Spain*, pages 189–200, 02.
5. Jordan M.; Jacobs R. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.
6. Heap T.; Hogg D. Extending the point distribution model using polar coordinates. *Image and Vision Computing*, 14:589–599, 1996.
7. Dempster A.; Laird N.; Rubin D. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Soc. Ser.*, 39:1–38, 1977.
8. Edwin R. Hancock; Josef Kittler. Edge-labelling using dictionary-based relaxation. *IEEE Transaction on PAMI*, 12(2):165–181, 1990.

A New Affine Invariant Fitting Algorithm for Algebraic Curves

Sait Sener and Mustafa Unel

Department of Computer Engineering, Gebze Institute of Technology
Cayirova Campus
44100 Gebze/Kocaeli Turkey
{ssener, munel}@bilmuh.gyte.edu.tr

Abstract. In this paper, we present a new affine invariant curve fitting technique. Our method is based on the affine invariant Fourier descriptors and implicitization of them by matrix annihilation. Experimental results are presented to assess the stability and robustness of our fitting method under data perturbations.

1 Introduction

Automatic recognition of free-form objects is an important problem in pattern recognition and computer vision. Implicit algebraic models or so-called Implicit Polynomial (IP) curves and surfaces proved to be very useful for modeling 2D curves and 3D surfaces [1-7]. To build recognition and positioning systems based on implicit curves and surfaces, it is imperative to solve the fitting problem. Euclidean fitting was used for the estimation of general curves and surfaces to edge and range data. Using Euclidean distance is more stable and useful than algebraic distance commonly used in least-squares fittings. Curve fitting is usually done through non-linear optimization, which has high computational complexity and computational cost [1-3]. The 3L-fitting algorithm, which is repeatable and Euclidean invariant, is another new approach to the fitting problem. This algorithm also improves implicit curve fitting in terms of accuracy and stability [8]. Continuous improvements have been made on the stability of algebraic curve fitting for obtaining IP models [9].

On the other hand, an important link between the Fourier descriptor and quadratic B-spline function is established in [10]. A new matrix annihilation method for converting parametric representations to the implicit algebraic curves has recently been introduced in [11]. This method is numerical and computationally efficient.

In this paper, we present a new affine invariant fitting algorithm based on the affine invariant Fourier descriptors [13, 14] and the implicitization of them by matrix annihilation method [11]. The algorithm is applicable for all free-form shapes. The resulting implicit algebraic curves are quite robust with respect to data perturbations. Fitting is computationally efficient and allows high degree curve fitting. In Section 2, we review implicit polynomial curves and the fitting problem. In section 3, we de-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 344–351, 2004.
© Springer-Verlag Berlin Heidelberg 2004

scribe affine invariant Fourier descriptors and their implicitization using matrix annihilation method. In Section 4, we present experimental results to assess robustness of our fitting algorithm under noise, and missing data due to partial occlusion. Finally in Section 5, we conclude with some remarks and indicate possible future directions.

2 Fitting Algebraic Curves to Data

Fitting an algebraic curve to data is basically an error-minimization problem over the set of data points $D = \{z_i = (x_i, y_i)\}_{j=1}^{N}$, namely

$$\min_{\mathbf{a}} \sum_{i=1}^{N} dist^2(f_n(x,y), z_i) \tag{1}$$

where $\mathbf{a}$ is the coefficient vector of the curve, $f_n(x,y) = \sum_{0 \le p+q \le n} a_{pq} x^p y^q$ is the two variable (binary) implicit polynomial whose zero set is the algebraic curve, and *dist* is an appropriate distance function. In the classical least-squares fitting, $dist(f_n(x,y), z_i) = f_n(z_i) = f_n(x_i, y_i) = \sum_{0 \le p+q \le n} a_{pq} x_i^p y_i^q$. To avoid the trivial solution, the norm of the solution vector is constrained to be 1, i.e. $\|\mathbf{a}\| = 1$. Resulting fit is affine invariant, but usually contains extra pieces and does not respect the shape continuity. Stability is another issue because it does not provide any stable results. Since there is no closed-form expression for computing the geometric distance of a point to the curve, a first order distance approximation, namely

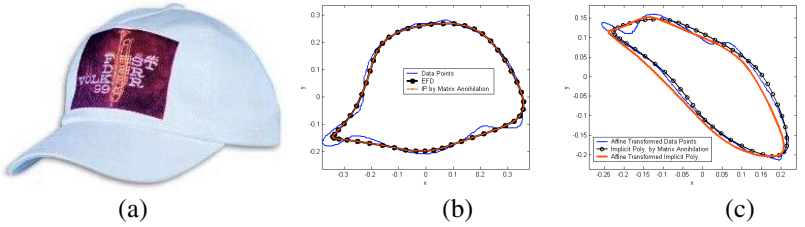
$$dist = \frac{\left| f_n(x_j, y_j) \right|}{\left\| \nabla f_n(x_j, y_j) \right\|} \tag{2}$$

is usually employed. This is a non-linear optimization problem and some important algorithms have been presented in the literature [1-3]. These algorithms have high computational complexity and in most cases do not provide stable and meaningful fits. 3L algorithm [8], which is a modified least squares fitting method, is more robust and gives better curves than the classical least squares algorithms. A solution has been offered for more stability and robustness of 3L algorithm to find the best implicit polynomials [9]. Although resulting fits are much more stable and superior than 3L fits, they are only rotation invariant and therefore cannot be used for images taken from different viewpoints.

3 Affine Invariant Fitting by Matrix Annihilation

Matrix annihilation method [11] established a link between the elliptic Fourier descriptors (EFDs) and implicit polynomials (IP). EFDs are arc-length parameteriza-

tions, which are made invariant to changes in location, orientation and scale, that is, they are similarity invariant [12]. However, Euclidean arc-length is not preserved under affine transformations. Fig. 1 depicts this fact. Both the original and the affine transformed data points are fitted by implicit curves obtained through the implicitization of EFDs by matrix annihilation method. Note in particular that the implicit curve obtained by the matrix annihilation method and the affine transformed version of the implicit curve in (b) are not coincided in (c).



(a) (b) (c)

Fig. 1. (a) Hat Object, (b) Data points (line), EFD (circles-line) and implicit curve (dot-line) obtained by matrix annihilation method are superimposed. (c) Affined data (line), implicit curve (circles-line) obtained by matrix annihilation method and affine transformed implicit curve (dot-line) in (b) are superimposed.

Instead of using similarity invariant EFDs, one could start with an affine invariant Fourier parameterization, which will then yield an affine invariant implicit polynomial curve as a result of matrix annihilation method. This is exactly what we are developing in this paper. By an affine invariant parameterization, we mean that the parameterization must be linear under an affine transformation and the parameterizing function must yield the same parameterization independent of the initial representation of the contour. Arbter et al. introduced affine invariant Fourier descriptors (AIFDs) by using the first derivatives of data curve as:

$$t = \frac{1}{2}\int_C \left|\det\left(\mathbf{x}(\xi), \mathbf{x}_\xi\right)\right| d\xi = \frac{1}{2}\int_C \left|x(\xi) y_\xi - y(\xi) x_\xi\right| d\xi \tag{3}$$

where x_ξ , y_ξ are the first derivatives of the components $x(\xi)$ and $y(\xi)$, and C is the path along the curve.

This parameterization will not be invariant under translation. Translation can be eliminated by initially moving the coordinate system to the area center, defined by

$$\mathbf{x}_s = \frac{2}{3}\frac{\oint_c \mathbf{x}(\xi)\det\left(\mathbf{x}(\xi), \mathbf{x}_\xi\right) d\xi}{\oint_C \det\left(\mathbf{x}(\xi), \mathbf{x}_\xi\right) d\xi} \tag{4}$$

The area center of an affine contour is the affine transform of the area center. This is just because the affine transformation transforms areas with a constant scale det(A) [15]. To calculate the coefficients, let $(u_i, v_i) : i = 0, \ldots, N-1$, be the coordinates of

the N vertices of an oriented polygon in $\Re^2$, and t_i their parameter values. With the periodicity, $(u_N, v_N) = (u_0, v_0)$ and $t_N = t_0 + T$. The Fourier transformation based on area parameterization is defined as [13,14]:

$$\begin{bmatrix} U_k \\ V_k \end{bmatrix} = \frac{T}{(2\pi k)^2} \sum_{i=0}^{N-1} \frac{1}{t_{i+1} - t_i} \begin{bmatrix} u_{i+1} - u_i \\ v_{i+1} - v_i \end{bmatrix} (\Phi_{k,i+1} - \Phi_{k,i})[1 - \delta(t_{i+1} - t_i)] + \frac{j}{2\pi k} \sum_{i=0}^{N-1} \begin{bmatrix} u_{i+1} - u_i \\ v_{i+1} - v_i \end{bmatrix} \Phi_{k,i} \delta(t_{i+1} - t_i) \tag{5}$$

where $\Phi_{k,i} = \exp\{-j2\pi k t_i / T\}$ and $\delta(t_{i+1} - t_i) = \begin{cases} 1 & if \ t_{i+1} = t_i \\ 0 & if \ t_{i+1} \neq t_i \end{cases}$.

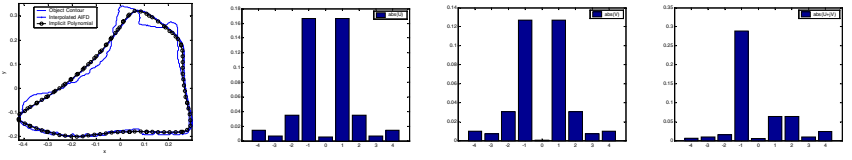
In contrast to the FFT, this transformation takes the indefinite set of all points on the polygon into account, does not need constant parameter intervals and allows the transformation of discontinuous functions. The area parameterization for polygons is given by

$$t_0 = 0 \ ; \ t_{i+1} = t_i + \frac{1}{2} \left| (u_i' v_{i+1}' - u_{i+1}' v_i') \right| , \qquad i = 0, 1, 2, \ldots, N-1 \ ; \ T = t_N , \tag{6}$$

with

$$\begin{bmatrix} u_i' \\ v_i' \end{bmatrix} = \begin{bmatrix} u_i - u_{AC} \\ v_i - v_{AC} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} u_{AC} \\ v_{AC} \end{bmatrix} = \sum_{i=0}^{N-1} (u_i v_{i+1} - u_{i+1} v_i) \begin{bmatrix} u_i + u_{i+1} \\ v_i + v_{i+1} \end{bmatrix} / (3 \sum_{i=0}^{N-1} (u_i v_{i+1} - u_{i+1} v_i)) \tag{7}$$

Since this parameterization is affine invariant, we can use the coefficients of this Fourier parameterization as an input to the implicitization method based on matrix annihilation ideas detailed in [11]. We note that the interpolated curve of affine-invariant Fourier descriptors might exhibit Gibbs phenomenon. Consider the boot depicted in Fig. 2. AIFD and IP curves are superimposed in Fig. 2 along with the amplitude of AIFDs.



(a) (b) (c) (d)

Fig. 2. (a) Object contour and AIFDs (dash-dot line) and IP curve (circles-line) obtained by MA method are superimposed. Amplitude spectrums of U_k, V_k and $U_k + jV_k$ are in (b,c,d).

Suppose that $U(k)$ and $V(k)$ are the Fourier coefficients based on the above parameterization. Matrix annihilation method determines the implicit polynomials by using the Fourier representations as inputs, namely $AIFD \overset{implicitization}{\Rightarrow} IP$. So, we start with the following EFDs:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_0 \\ c_0 \end{bmatrix} + \sum_{k=1}^{n} \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} \begin{bmatrix} \cos kt \\ \sin kt \end{bmatrix} \tag{8}$$

To transform the affine invariant Fourier descriptors into (8), we separate the Fourier coefficients into real and imaginary parts as

$$a_0 = U_0\,,\ c_0 = V_0\,,\ a_k = \mathrm{Re}\{U(k)\}\,,\ b_k = \mathrm{Im}\{U(k)\}\,,\ c_k = \mathrm{Re}\{V(k)\}\,,\ d_k = \mathrm{Im}\{V(k)\} \tag{9}$$

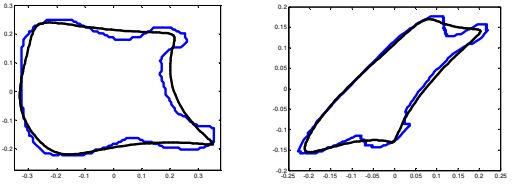
The separated coefficients are then taken as inputs to the matrix annihilation method. Using more harmonics increases the degree of the curve which in turn will improve the accuracy of the resulting fit. However, in such a case one should be careful about overfit problems. The obtained curve with k harmonics will be of degree $d=2k$.

4 Experimental Results

Shapes from a different viewpoint are generated according to

$$A = \begin{bmatrix} \cos\beta & \sin\alpha\sin\beta & tr_x \\ 0 & \cos\alpha & tr_y \\ 0 & 0 & 1 \end{bmatrix}$$

where α and β are the change of angles in x and y axes (pitch and yaw), tr_x and tr_y are the translations in x and y directions. Fig. 3 (a) depicts the data and the IP curve obtained from our method. Using above transform matrix A with $\alpha = 45°$ and $\beta = 70°$, one can verify that the IP fit on the data contour in (b) and the affine version of the original fit in (a) are exactly the same.

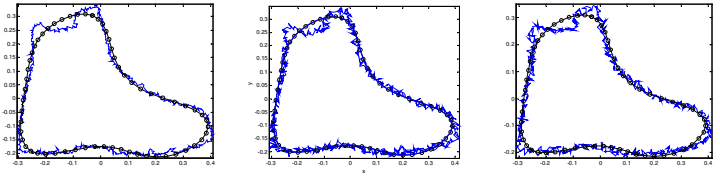


(a) (b)

Fig. 3. (a) Object contour (line) and the implicit curve (dashed-line) obtained by our fitting algorithm are superimposed. (b) Affine transformed contour and the IP fit.

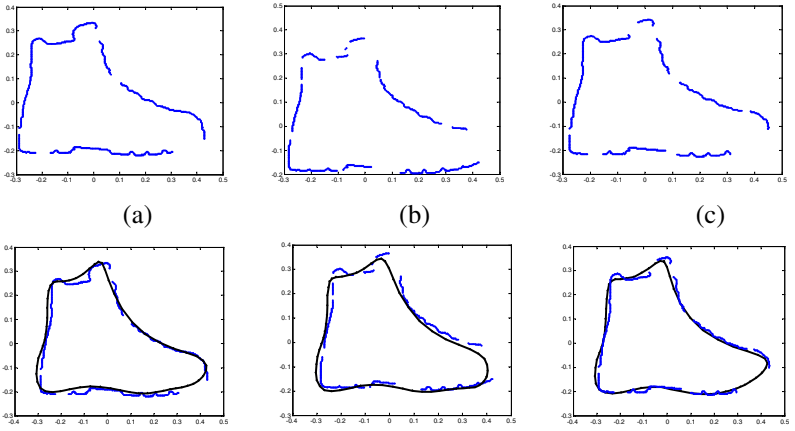
To realize distortions occurred during the segmentation process, we perturbed the objects before applying our fitting procedure. Experiments are based on the white noise distortion and missing data due to partial occlusion on the object boundaries. The random white noise is added to each point in the data. Fig. 4 (a), (b) and (c) show the perturbed shapes with white noise having the standard deviations 0.0001, 0.0005 and 0.005, respectively.

Robustness to missing data due to partial occlusion crucially depends on a good representation. Fig. 5 shows the 6th-degree IP fits on occluded data (a-d), (b-e) and (c-f) with %10, %15 and %20 missing data, respectively.



(a) (b) (c)

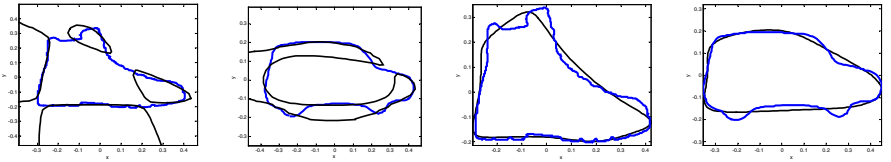
Fig. 4. Fits under white noise with standard deviations 0.0001, 0.0005 and 0.005, respectively.



(a) (b) (c)

(d) (e) (f)

Fig. 5. The 6th-degree IP fits on occluded data (a-d), (b-e) and (c-f) with %10, %15 and %20 missing data, respectively.

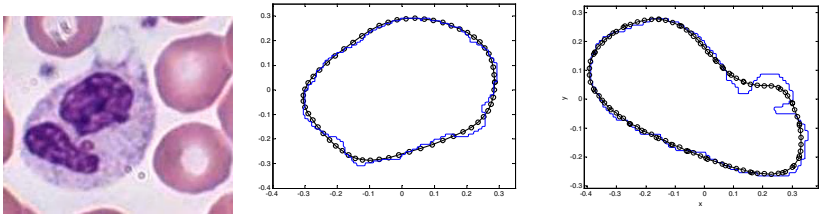


(a) (b) (c) (d)

Fig. 6. (a) and (b) are data points with superimposed IP curves by least-squares fit algorithm. (c) and (d) are the same data sets with superimposed IP curves by our fit algorithm.

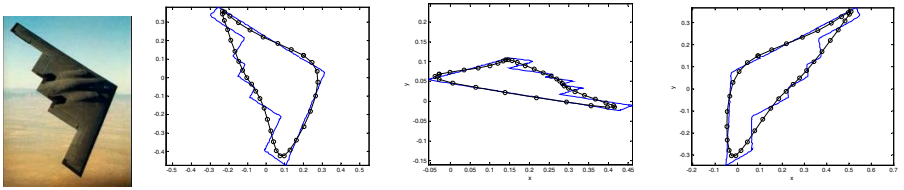
Although the classical least-squares fit algorithm is affine invariant, it does not provide meaningful fits. The resulting implicit curves do not respect data continuity and are not stable, and usually involve unbounded extra components. Fig. 6 shows this fact by comparing the least-squares fit results with ours.

Algebraic (or IP) curves obtained by our algorithm can be used in an Image-Guided Decision Support (IGDS) system designed to assist pathologists to discriminate abnormal blood cells. Fig. 7 shows an abnormal blood cell and the associated IP curves. They can also be very useful in motion estimation and aircraft recognition systems because of the affine invariant property of our fitting algorithm. See Fig. 8.



(a) (b) (c)

Fig. 7. (a) blood cell, (b) the first infected part and the 6th degree curve (circle-line), (c) the second infected part and the 6th degree curve (circle-line).



(a) (b) (c) (d)

Fig. 8. A 6th degree fit on a B2 aircraft contour and its affine versions.

5 Conclusion

We have now presented an affine invariant fitting algorithm for modeling 2D free-form curves based on affine invariant Fourier descriptors and matrix annihilation theory. Proposed algorithm is fast, repeatable, numerically stable, and robust to data perturbations.

Acknowledgment. This research is supported from GYTE research grant BAP #2003A23.

References

[1] Taubin, G., F. Cukierman, S. Sullivan, J. Ponce and D.J. Kriegman, Parameterized Families of Polynomials for Bounded Algebraic Curve and Surface Fitting. IEEE Trans. Pattern Analysis and Machine Intelligence, 16(3):287-303, March 1994.

[2] Keren, D., D. Cooper and J. Subrahmonia, Describing Complicated Objects by Implicit Polynomials, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 16, pp. 38-53, 1994.

[3] G. Taubin, Estimation of Planar Curves, Surfaces and Nonplanar Space Curves Defined by Implicit Equations, with Applications to Edge and Range Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, pp. 1115-1138, 1991.

[4] Wolovich, W. A. and Mustafa Unel, The Determination of Implicit Polynomial Canonical Curves, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20 (8), October 1998.

[5] Unel, Mustafa and W. A. Wolovich, On the Construction of Complete Sets of Geometric Invariants for Algebraic Curves. Advances in Applied Mathematics 24, 65-87, January 2000.

[6] Unel, Mustafa and W. A. Wolovich, A New Representation for Quartic Curves and Complete Sets of Geometric Invariants, Int. Jour. of Pattern Recognition and Artificial Intelligence, Vol. 13 (8), 1999.

[7] Subrahmonia, J., D. B. Cooper, and D. Keren, Practical Reliable Bayesian Recognition of 2D and 3D Objects using Implicit Polynomials and Algebraic Invariants, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(5):505-519, 1996.

[8] Lei, Z., M. M. Blane, and D. B. Cooper, 3L Fitting of Higher Degree Implicit Polynomials, In proceedings of Third IEEE Workshop on Applications of Computer Vision, pp. 148-153, Florida 1996.

[9] T. Tasdizen, T. Tarel and D. B. Cooper, Improving the Stability of Algebraic Curves for Applications, IEEE Transactions in Image Processing, 9(3):405-416, 2000.

[10] Wu-Chih Hu and Hsin-Teng Sheu, Quadratic B-spline for Curve Fitting, Proc. Natl. Sci. Counc. ROC(A), vol.24 No:5 pp. 373-381, 2000.

[11] H. Yalcin, M. Unel, W. A. Wolovich, Implicitization of Parameteric Curves by Matrix Annihilation, International Journal of Computer Vision, Vol. 54, pp. 105-115, 2003.

[12] Kuhl, F.P. and C.R. Giardina, Elliptic Fourier Features of a Closed Contour, Computer Graphics and Image Processing, vol.18, pp.236-258, 1982.

[13] K. Arbter, W. E. Synder, H. Burkhardt and G. Hirzinger, Application of Affine-Invariant Fourier Descriptors to Recognition of the 3D Objects, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.12 (7) pp. 640-647, 1990.

[14] K. Arbter, Affine-Invariant Fourier Descriptors, in From Pixels to Features, Amsterdam, The Netherlands: Elseiver Science, 1989.

[15] F. E. Pollick, G. Sapiro, Constant Affine Velocity Predicts the 1/3 Power Law of Planar Motion Perception and Generation, Elseiver Science, Vol. 37, No, 3, pp. 347–353, 1997.

Graph Matching Using Manifold Embedding

Bai Xiao, Hang Yu, and Edwin Hancock

Department of Computer Science, University of York, York Y010 5DD, UK.

Abstract. This paper describes how graph-spectral methods can be used to transform the node correspondence problem into one of point-set alignment. We commence by using a heat kernel analysis to compute geodesic distances between nodes in the graphs. With geodesic distances to hand, we use the ISOMAP algorithm to embed the nodes of a graph in a low-dimensional Euclidean space. With the nodes in the graph transformed to points in a metric space, we can recast the problem of graph-matching into that of aligning the points. Here we use a variant of the Scott and Longuet-Higgins algorithm to find point correspondences. We experiment with the resulting algorithm on a number of real-world problems.

1 Introduction

Graphs are general and powerful data structures that can be used to represent both physical objects and abstract concepts. Graph matching is a task of pivotal important in high-level vision since it provides a means by which abstract pictorial descriptions can be matched to one another. There is a considerable literature on the problem of graph matching, and many contrasting methods including search [12], relaxation [13] and optmisation [9] have been successfully used. However, the main challenge in graph matching is how to deal with differences in node and edge structure. One of the most elegant recent approaches to the graph matching problem has been to use graph spectral methods [2], and exploit information conveyed by the eigenvalues and eigenvectors of the adjacency matrix. There have been successful attempts to use spectral methods for both structural graph matching [11], and for point pattern matching [3]. For instance Umeyama [11] has developed a method for finding the permutation matrix which best matches pairs of weighted graphs of the same size, by using a singular value decomposition of the adjacency matrices. Scott and Longuet-Higgins [3], on the other hand, align point-sets by performing singular value decomposition on a point association weight matrix. Shapiro and Brady [7] have reported a correspondence method which relies on measuring the similarity the eigenvector of a Gaussian point-proximity matrix.

Although Umeyama's algorithm [11] is elegant in its formulation and can deal with both weighted or unweighted graphs, it can not be applied to graphs which contain different numbers of nodes and for weighted graphs the method is susceptible weight errors. One way to overcome these problems is to cast the

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 352–359, 2004.
© Springer-Verlag Berlin Heidelberg 2004

problem of recovering correspondences in a statistical setting using the EM algorithm [8]. However, the resulting algorithm is time consuming because of its iterative character. In this paper we take an alternative view of spectral matching algorithms to develop an efficient method that is robust to differences in structure. Spectral methods can be viewed as embedding the nodes of a graph in a space spanned by the eigenvectors of the adjacency matrix. In the case of the Umeyama algorithm [11], matching is effected by finding the transformation matrix that best aligns the embedded points. The Shapiro and Brady [7] algorithm finds correspondences by seeking the closest embedded points. Kosinov and Caelli [10] have improved this method by allowing for scaling in the eigenspace.

Our aim in this paper is to seek an embedding of the nodes of a graph which allows matching to be effected using simple point-pattern matching methods. In the mathematics literature, there is a considerable body of work aimed at understanding how graphs can be embedded on a manifold so as to minimize a measure of distortion. Broadly speaking there are three ways in which the problem can be addressed. First, the graph can be interpolated by a surface whose genus is determined by the number of nodes, edges and faces of the graph. Second, the graph can be interpolated by a hyperbolic surface which has the same pattern of geodesic (internode) distances as the graph [5]. Third, a manifold can be constructed whose triangulation is the simplicial complex of the graph [1]. A review of methods for efficiently computing distance via embedding is presented in the recent paper of Hjaltason and Samet [4]. Recently, there has been considerable interest in the pattern analysis community in how to embed complex relational data in a low dimensional manifold. Collectively [6], these methods are known as manifold learning theory. Their collective aim is to develop variants of the classical methods of PCA and MDS, that can be used to better capture localised variations on the structure of the data.

In this paper we investigate whether methods from manifold learning theory can be combined with spectral graph theory to develop effective tools for graph structure matching. The idea is to use manifold learning methods to embed the graphs in a low dimensional coordinate space, and to perform spectral analysis on the co-ordinate data for the embedded graph-nodes. We proceed as follows. We commence by using the heat kernel of the graph to compute approximate geodesic distances between nodes. With internode distances to hand, we use a strategy similar to ISOMAP [6] to embed the graphs in a Euclidean pattern space. This is done by computing a matrix of shortest(geodesic) distance between nodes in the graph. We then apply MDS (Multidimensional Scaling) to the distance matrix, and this embeds the individual nodes of the graph in a Euclidean space. Once embedded in this space, we can apply a variant of the Scott and Longuet-Higgins [3] algorithm to match the two sets of embedded points. By doing this we overcome problems of structural difference.

2 Metric Embedding of Graphs

We are interested in the abstract problem of embedding the nodes of a graph into a Euclidean space. Here we use Isomap [6] as a way to solve the low-distortion graph embedding problem. The idea behind Isomap is to apply classical MDS to the matrix of geodesic distances between data-points. In this way the data is mapped from a high-dimensional input space to the low-dimensional space of a nonlinear manifold. Although the method was originally devised for dimensionality reduction, we can use it here for the low-distortion graph embedding problem. Here we use a spectral method to approximate the geodesic distances between nodes.

2.1 Geodesic Distances from Graph Spectra

In this section, we develop a method for approximating the geodesic distance between nodes by exploiting the properties of the heat kernel [2]. To commence, suppose that the graph under study is denoted by $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Since we wish to adopt a graph-spectral approach we introduce the adjacency matrix A for the graph where

$$A(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We also construct the diagonal degree matric D, whose elements are given by $D(u,u) = \sum_{v \in V} A(u,v)$. From the degree matrix and the adjacency matrix we construct the Laplacian matrix $L = D - A$, i.e. the degree matrix minus the adjacency matrix. The normalised Laplacian is given by $\hat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. The spectral decomposition of the normalised Laplacian matrix is $\hat{L} = \Phi \Lambda \Phi^T$ where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{|V|})$ is the diagonal matrix with the ordered eigenvalues as elements and $\Phi = (\phi_1 | \phi_2 | .... | \phi_{|V|})$ is the matrix with the ordered eigenvectors as columns. Since $\hat{L}$ is symmetric and positive semi-definite, the eigenvalues of the normalised Laplacian fall in the interval $[0, 2]$, i.e. they are all positive. The eigenvector assoicated with the smallest non-zero eigenvector is referred to as the Fiedler-vector. We are interested in the heat equation associated with the Laplacian, i.e. $\frac{\partial h_t}{\partial t} = -\hat{L} h_t$ where h_t is the heat kernel and t is time. The solution is found by exponentiating the Laplacian eigenspectrum, i.e.

$$h_t = \sum_{i=1}^{|V|} \exp[-\lambda_i t] \phi_i \phi_i^T = \Phi \exp[-t\Lambda] \Phi^T \tag{2}$$

The heat kernel is a $|V| \times |V|$ matrix, and for the nodes u and v of the graph G the resulting component is

$$h_t(u,v) = \sum_{i=1}^{|V|} \exp[-\lambda_i t] \phi_i(u) \phi_i(v) \tag{3}$$

When t tends to zero, then $h_t \simeq I - \hat{L}t$, i.e. the kernel depends on the local connectivity structure or topology of the graph. If, on the other hand, t is large, then $h_t \simeq \exp[-t\lambda_m]\phi_m\phi_m^T$, where λ_m is the smallest non-zero eigenvalue and ϕ_m is the assoicated eigevector, i.e. the Fiedler vector. Hence, the large time behaviour is governed by the global structure of the graph.

When the graph is embedded on a manifold in Riemannian space then the pattern of geodesic distances between nodes on the manifold is the same as the path length distribution. However, when the manifold is locally Euclidean, then the heat kernel is approximated by the Gaussian

$$h_t(u,v) = [4\pi t]^{-\frac{n}{2}} \exp[-\frac{1}{4t^2}d(u,v)^2] \tag{4}$$

where $d(u,v)$ is the distance between the nodes u and v on the Euclidean manifold and n is the dimensionality of the space. The aim here is to find an approximation to the geodesic distance between nodes in the embbeding, by equating the spectral and Gaussian forms for the kernel. The result is

$$d^2(u,v) = -4t\ln\left\{(4\pi t)^{\frac{n}{2}}\sum_{i=1}^{|V|}\exp[-\lambda_i t]\phi_i(u)\phi_i(v)\right\} \tag{5}$$

Because of the logarithm, the geodesic distance is a slowly varying function of the parameter t. For very small values of t the behaviour is determined by the local topology of the graph, while for large t it is determined by the Fiedler eigenvector, i.e. the cluster-structure of the graph. Here we use a value of $t = 1$.

2.2 Metric Embedding Using Isomap

Our goal is to find a low-distortion or distortion-free embedding from the graph metric space into a normed space. Here we use Isomap [6] as a way to solve the low-distortion graph embedding problem. The idea behind Isomap is to apply classical MDS to map data points from their high-dimensional input space to low-dimensional coordinates of a nonlinear manifold. The key contribution is hence to apply MDS to the pairwise distances not in the input Euclidean space, but in the geodesic space of the manifold.

Although the method was originally devised for dimensionality reduction, we can use it here for the low-distortion graph embedding problem. Viewed as an isometric feature mapping, Isomap is a mapping $f : X \to Y$ from the observation space X to a Euclidean feature space Y that preserves as closely as possible the intrinsic metric structure of the observations, i.e. the distances between observations as measured along geodesic(shortest) paths of X [6]. The distortion c in this embedding is nearly 1.

For graphs, the embedding procedure is straightforward. We first construct the shortest path distance matrix S for each graph. Each element in S is the shortest path distance between the corresponding pair of nodes of the graph. We embed each graph in a Euclidean space by performing MDS on the matrix S.

The pairwise geodesic distances between nodes $d(u,v)$ are used as the elements of an $N \times N$ dissimilarity matrix S, whose elements are defined as follows

$$S(u,v) = \begin{cases} d(u,v) & \text{if } u \neq v \\ 0 & \text{if } u = v \end{cases} \tag{6}$$

The first step of MDS is to calculate a matrix T whose element with row r and column c is given by $T(r,c) = -\frac{1}{2}[d^2(r,c) - \hat{d}^2(r,.) - \hat{d}^2(.,c) + \hat{d}^2(.,.)]$, where $\hat{d}(r,.) = \frac{1}{N}\sum_{c=1}^{N} d(r,c)$ is the average dissimilarity value over the rth row, $\hat{d}(.,c)$ is the similarly defined average value over the cth column and $\hat{d}(.,.) = \frac{1}{N^2}\sum_{r=1}^{N}\sum_{c=1}^{N} d(r,c)$ is the average similarity value over all rows and columns of the similarity matrix T.

We subject the matrix T to an eigenvector analysis to obtain a matrix of embedding co-ordinates X. If the rank of T is $k, k \leq N$, then we will have k non-zero eigenvalues. We arrange these k non-zero eigenvalues in descending order, i.e. $l_1 \geq l_2 \geq \ldots \geq l_k > 0$. The corresponding ordered eigenvectors are denoted by $\boldsymbol{u}_i$ where l_i is the ith eigenvalue. The embedding co-ordinate system for the graphs obtained from different views is $X = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_s]$, where $\boldsymbol{f}_i = \sqrt{l_i}\boldsymbol{u}_i$ are the scaled eigenvectors. For the graph-node indexed u, the embedded vector of co-ordinates is $\boldsymbol{x}_i = (X_{u,1}, X_{u,2}, ..., X_{u,s})^T$.

3 Graph Matching

By applying Isomap to the two graphs, we obtain two point sets I and J, containing m and n features respectively. To match the points we uase a variant of the Scott and Longuet-Higgins method. We regard the points in I and J as lying in the same plane. We then represent the '*proximities*' between the features in I and the features in J. We use the Gaussian form $G_{ij} = exp(-m_{ij}^2/2)$ to compute the matrix of proximity weights. Here $m_{ij}^2/2$ is the Mahalanobis distance $m_{ij}^2(M) = (x_i - x_j)'\Sigma^{-1}(x_i - x_j)$ between nodes where Σ is the point position covariance matrix. The use of the Mahalanobis metric instead of the Euclidean distance has several advantages. First, it automatically accounts for the scaling of the coordinate axes. Second, it corrects for correlation between the different features. And third, it can provide curved as well as linear decision boundaries.

With an inter-graph node distance matrix to hand we follow Scott and Longuet-Higgins [3] and perform the singular value decomposition (SVD) $G = TDU$. The matrices of T and U are orthogonal. The matrix D contains the singular values along its diagonal in descending numerical order. The final step is to compute the correlation between T's rows and U's columns, giving an association matrix $P = TEU$, where E is obtained by replacing each diagonal element in D by a 1. The element P_{ij} indicates the strength of attraction between feature $i \in I$ and $j \in J$. The rows of P, index the features in the first graph, and its columns those in the second graph. If P_{ij} is both the largest element in row i and column j then we regard these features as being in one-to-one correspondence with one-another. If P_{ij} is the greatest element in row i but not

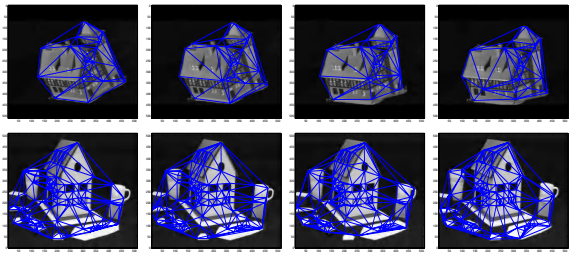


Fig. 1. Delaunay graphs overlayed on the house images

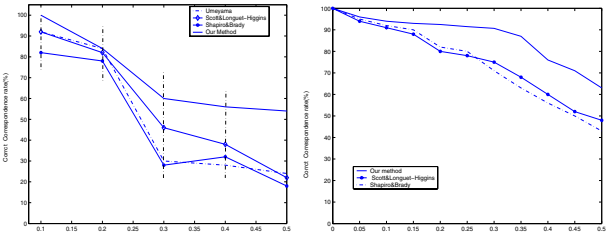


Fig. 2. Synthetic Data Experiments Results

the greatest in column j, then we may regard $i \in I$ competing unsuccessfully for partnership with $j \in J$. Similar remarks apply if P_{ij} is the greatest element in its column but not in its row [3].

4 Experiments

In this section, we provide some experimental evaluation of the new graph-matching method. There are two aspects to this study. First, we compare our method with some alternative methods by using synthetic data. Second, we evaluate our method on real-world data.

We provide a comparison between two methods for weighted graph matching. Here, we have selected the Shapiro and Brady [7] and Scott and Longuet-Higgins [3] feature set matching methods. These two methods use coordinate information for the feature points, and do not incorporate the graph structure information. We also investigated Umeyama's method. In the case of our method we are concerned with matching the Delaunay triangulation of corner-features. Besides the coordinates of the feature points the Delaunay graph incorporates structural information.

Our first experiment is based on synthetic data. We have randomly generated 30 2D points. We use the position of the points to generate a Delaunay graph. We have kept the number of points fixed and have added Gaussian errors to the point positions. The parameter of the noise process is the standard deviation of the positional jitter. In Figure 2(the left plot), we show the fraction of correct

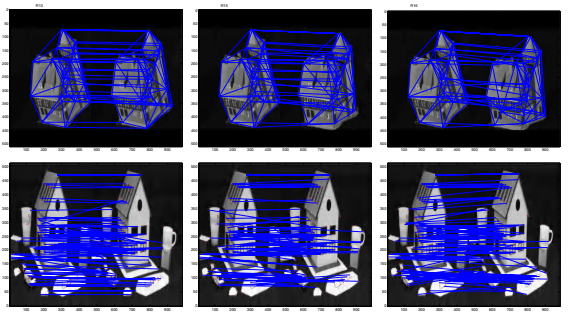


Fig. 3. Our algorithm for CMU and MOVI house sequences

Table 1. Experiments Results for MOVI House Sequence Images

Images	Points	Correct correspondence	False correspondence	No correspondence
house1	140	-	-	-
house2	134	109	9	15
house3	130	96	13	21
house5	137	90	15	29

correspondences as a function of the noise standard deviation for our method, Shapiro and Brady's [7] method, Umeyama's method and Scott and Longuet-Higgins' method [3]. In another synthetic data experiment, we investigate the effect of structural noise. Here we have added a controlled fraction of additional nodes at random positions and have recomputed the Deluanay triangulations. In Figure 2(the right plot), we plot the fraction of correct correspondences as a function of the fraction of added nodes. The plot compares the result of applying our method to the data, and the results obtained using Scott and Longuet-Higgins' method, and Shapiro and Brady's method. Since Umeyama's algorithm can't handle graphs of different size, we have not compared with this method. The main feature to note is that our method outperforms the two alternatives. This means our method can solve the structure matching problem when the graphs are of different size.

To take this study one step further, we perform some real-world data experiments. We apply our matching method to two image sequences(CMU sequences and MOVI sequences). There are rotation, scaling, and perspective distortions present. Example images from these sequences are shown in Fig 1 and correspond to different camera viewing directions. The detected feature points and their Delaunay triangulations are overlayed on the images. The first four images are from the CMU sequence and each contains about 30 nodes, the second four are from the MOVI sequence and each contain about 140 nodes.

In Fig 1, we test our method on some representative image pairs. In Table 1 we summarize the matching results for the MOVI houses. Here we list the number of nodes in the Delaunay graphs, the number of correct correspondence,

the number of correspondence errors, and the number of points without correspondence.

5 Conclusion and Future Work

This paper has presented an efficient approach to structural graph matching. The approach is to first use the Isomap algorithm to embed the graphs in a Euclidean space by using the geodesic distance between nodes. Each node is hence transformed to a point in co-ordinate space. We match points corresponding to nodes in different graphs using the Scott and Longuet-Higgins algorithm.

References

1. A.Ranicki. Algebraic l-theory and topological manifolds. *Cambridge University Press*, 1992.
2. F.R.K.Chung. Spectral graph theory. *CBMS series*, 92, 1997.
3. G.L.Scott and H.C.Longuett-Higgins. An algorithm for associating the features of two images. *Proceedings of the Royal Society of London B*, 244:21–26, 1991.
4. G.R.Hjaltason and H.Samet. Properties of embedding methods for similarity searching in metric spaces. *PAMI*, 25:530–549, 2003.
5. H.Busemann. The geometry of geodesics. *Academic Press*, 1955.
6. J.B.Tenenbaum, V.D.Silva, and J.C.Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:586–591, 2000.
7. L.S.Shapiro and J.M.Brady. Feature-based correspondence: an eigenvector approach. *Image and Vision Computing*, 10:283–288, 1992.
8. B. Luo and E.R. Hancock. Structural graph matching using the em algorithm and singular value decomposition. *IEEE PAMI*, 23:1120–1136, 2001.
9. S.Gold and A.Rangarajan. A graduated assignment algorithm for graph matching. *IEEE PAMI*, 18, 1996.
10. S.Kosinov and T. Caelli. Inexact multisubgraph matching using graph eigenspace and clustering models. *9th International Workshop on Structural and Syntactic Pattern*, LNCS 2396:133–142, 2002.
11. S.Umeyama. An eigen decomposition approach to weighted graph matching problems. *IEEE PAMI*, 10:695–703, 1988.
12. J.R. Ullman. An algorithm for subgraph isomorphism. *J.ACM*, 23:31–42, 1976.
13. W.Christmas, J.Kittler, and M.Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE PAMI*, 17, 1995.

A Matching Algorithm Based on Local Topologic Structure*

Xinjian Chen, Jie Tian**, and Xin Yang

Biometrics Research Group, Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation Chinese Academy of Science, Graduate School of the Chinese Academy of Science, P.O.Box 2728, Beijing, 100080, China
tian@doctor.com; jie.tian@mail.ia.ac.cn
http://www.fingerpass.net

Abstract. How to cope with non-linear distortions in the matching algorithm is a real challenge. In this paper, we proposed a novel fingerprint matching algorithm based on the local topologic structure and a novel method to compute the similarity between two fingerprints. The algorithm firstly aligns the template fingerprint and the input fingerprint. Then local topologic structure matching was introduced to improve the robustness of global alignment. Finally a novel method was introduced to compute the similarity between the template fingerprint and the input fingerprint. The proposed algorithm has been participated in Fingerprint verification competition (FVC2004). The performance was ranked 3rd position in open category in FVC2004.

1 Introduction

Significant improvements in fingerprint recognition have been achieved on the algorithmic side, but a great number of challenging problems still exist. One of the challenging problems is matching of non-linear distorted fingerprints. The acquisition of a fingerprint is a 3D-2D warping process [1][2]. The fingerprint captured with a different contact center will result in a different warping mode. The other reason that will introduce distortion to fingerprint is the non-orthogonal pressure people exert on the sensor. In Fingerprint Verification Competition 2004 (FVC2004)[3], the organizers have insisted on: distortion, dry and wet fingerprints in particular. How to cope with these non-linear distortions in the matching algorithm is a real challenge.

Recently, some algorithms were presented that explicitly deal with the problem of the non-linear distortion in fingerprint images to improve their matching performance. Maio and Maltoni et al. [1] proposed a plastic distortion model to "describe how fin-

* This paper is supported by the Project of National Science Fund for Distinguished Young Scholars of China under Grant No. 60225008, the Key Project of National Natural Science Foundation of China under Grant No. 60332010, the Project for Young Scientists' Fund of National Natural Science Foundation of China under Grant No.60303022.

** Corresponding author: Jie Tian; Telephone: 8610-62532105; Fax: 8610-62527995.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 360–367, 2004.
© Springer-Verlag Berlin Heidelberg 2004

gerprint images are deformed when the user improperly places his/her finger on the sensor plate". This model helps to understand this process. However, due to the insufficiency and uncertainty of information, it is very difficult to automatically and reliably estimate the parameter in that model. Bazen et al. [4] used a thin-plate spline model to describe the non-linear distortions between the two sets of possible matching minutiae pairs. By normalizing the input fingerprint with respect to the template, this method is able to perform a very tight minutiae matching and thus improve the performance. However, the TPS model focuses on smoothly interpolating images over scattered data. When applied this model to fingerprint recognition, it can make two fingerprints, no matter they come from the same finger or not, more similar to each other.

In this paper, we proposed a novel fingerprint matching algorithm based on the local topologic structure. The algorithm firstly aligns the template fingerprint and the input fingerprint using the registration method described in [6]. Then we introduce local topologic structure matching to improve the robustness of global alignment. Finally a novel method is introduced to compute the similarity between the template fingerprint and the input fingerprint.

This paper is organized as follows. Section 2 describes the details of matching using local topologic structure. Section 3 proposes a novel method which uses to compute the similarity between the template fingerprint and the input fingerprint. The performance of the proposed algorithm is shown by experiments in Section 4. Section 5 contains conclusion and discussion.

2 Match Using Local Topologic Structure

The task of the fingerprint registration is to align the two fingerprints and find the corresponding minutiae pairs between the two feature sets. We used the registration method described in [6]. For an input fingerprint image, we use the method described in [5],[7] to enhance the image and detect the minutiae set. The ridge points were sampled at the average inter-ridge distance along the ridge linked with the corresponding minutiae point. In this paper we introduce local topologic structure matching to improve the robustness of global alignment.

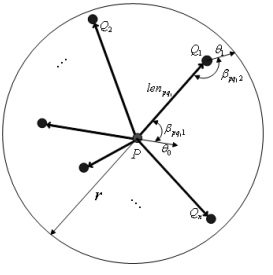


Fig. 1. The local topological structure of P

2.1 Defining of Local Topologic Structure

Let P denote a minutiae in the fingerprint image and $Q1,Q2,...,Qn$ are the minutiaes circle around P within r radius. We use the len_{pq_1}, β_{pq_11}, β_{pq_12} to descript the relationship between P and $Q1$. len_{pq_1} denotes the distance between P and $Q1$., β_{pq_11} denote the angle between the orientation of minutiae P($_0$) and the direction

from P to $Q1$. $\beta_{pq_1 2}$ denotes the angle between the orientation of minutiae $Q1$($_1$) and the direction from $Q1$ to P, Fig. 1 show the meaning of these parameters. And we can find that len_{pq_1}, $\beta_{pq_1 1}$, $\beta_{pq_1 2}$, is no relevant to shift and rotation. Then we can define the local topologic structure of P as following:

$$LoTopo_P = \left((len_{pq_1}, \beta_{pq_1 1}, \beta_{pq_1 2}), (len_{pq_2}, \beta_{pq_2 1}, \beta_{pq_2 2}), \cdots, (len_{pq_n}, \beta_{pq_n 1}, \beta_{pq_n 2})\right)$$

2.2 Local Topologic Structure Match

Suppose P is a minutiae in the template fingerprint, $Q1,Q2,\ldots,Qn$ are n minutiaes circle around P within r radius. Then we can get the local topologic structure of P:

$$LoTopo_P = \left((len_{pq_1}, \beta_{pq_1 1}, \beta_{pq_1 2}), (len_{pq_2}, \beta_{pq_2 1}, \beta_{pq_2 2}), \cdots, (len_{pq_n}, \beta_{pq_n 1}, \beta_{pq_n 2})\right)$$

And suppose R is a minutiae in the input fingerprint. As the fingerprint is deformed, we search minutiaes circle around R within $r + r_s$ radius, r_s is the distance toleance. Suppose $S1,S2,\ldots,Sm$ are m minutiaes circle around R within $r + r_s$ radius. Then we can get the local topologic structure of R:

$$LoTopo_R = \left((len_{rs_1}, \beta_{rs_1 1}, \beta_{rs_1 2}), (len_{rs_2}, \beta_{rs_2 1}, \beta_{rs_2 2}), \cdots, (len_{rs_m}, \beta_{rs_m 1}, \beta_{rs_m 2})\right).$$

We use the following algorithm to determine whether two minutiae P and R is matched:

Step 1: For i (1≤i≤n) and j (1≤j≤m), match PQ_i $(len_{pq_i}, \beta_{pq_i 1}, \beta_{pq_i 2})$ and RS_j $(len_{rs_j}, \beta_{rs_j 1}, \beta_{rs_j 2})$, get the matched number $n_{match\,1}$. Here one pair PQ_i matched one pair RS_j at most. If PQ_i matched several pairs, keep only one pair.

Firstly three parameters len_{diff}, β_{1diff}, β_{2diff} are calculated as following:

$$len_{diff} = \left| len_{pq_i} - len_{rs_j} \right| \tag{1}$$

$$\beta_{1diff} = \left| \beta_{pq_i 1} - \beta_{rs_j 1} \right| \tag{2}$$

$$\beta_{2diff} = \left| \beta_{pq_i 2} - \beta_{rs_j 2} \right| \tag{3}$$

Then we use adaptive matching bounding box $M_{box} = (len_{thr}, \beta_{thr})$ to determine whether PQ_i and RS_j is matched. The size of matching box changes according to the distance len_{pq_i} as follows:

$$len_{thr} = \begin{cases} len_L & if\ len_{pq_i} < Thr_L \\ len_H & if\ len_{pq_i} > Thr_L \\ len_L + \dfrac{len_H - len_L}{Thr_H - Thr_L} \bullet (len_{pq_i} - Thr_L) & otherwise \end{cases} \tag{4}$$

$$\theta_{thr} = \begin{cases} \theta_L & if\ len_{pq_i} < Thr_L \\ \theta_H & if\ len_{pq_i} > Thr_L \\ \theta_L + \dfrac{\theta_H - \theta_L}{Thr_H - Thr_L} \bullet (len_{pq_i} - Thr_L) & otherwise \end{cases} \tag{5}$$

Where $len_L < len_H, \theta_L > \theta_H$. The purpose for using a changeable sized bounding box is to deal with nonlinear deformation more robustly. When the distance of PQ_i is small, a small deformation will mean a large change of the radial angle while the change of radius remains small. Hence in this case the θ_{thr} of the bounding box should be larger and the len_{thr} of the bounding box should be smaller. On the other hand, when the distance of PQ_i is large, a small change in radial angle will cause a large change in the position of the minutia. While the radius can have larger deformation as it is the accumulation of deformation from all the regions between Q_i and P. Hence in this case the len_{thr} of the bounding box should be larger and the θ_{thr} of the bounding box should be smaller.

If Equations (1),(2),(3) satisfy the following conditions: (1) $len_{diff} < len_{thr}$, (2) $\beta_{1diff} < \theta_{thr}$ (3) $\beta_{2diff} < \theta_{thr}$ then we determine that PQ_i and RS_j is matched.

Step 2: If $n_{match1} / n > Thr_{topo}$ then goto Step 3, else we determine that P and R is not matched.

Step 3: Search minutiaes circle around R within $r - r_s$ radius, get the other local topologic structure of R:

$LoTopo_{R2} = \left((len_{rt_1}, \beta_{rt_1 1}, \beta_{rt_1 2}), (len_{rt_2}, \beta_{rt_2 1}, \beta_{rt_2 2}), \cdots, (len_{rt_l}, \beta_{rt_l 1}, \beta_{rt_l 2})\right)$.

Using the same algorithm defined in step 1, match the local topological structure $LoTopo_P$ and $LoTopo_{R2}$, get the matched number n_{match2}

Step 4: If $n_{match2} / n > Thr_{topo}$, then we determine that P and R is matched, else P and R is not matched.

3 Similarity Computing

How to compute the similarity between template fingerprint and input fingerprint for deformed fingerprints is a difficult task. In some algorithms [2][4], they only use the number of matching minutiae to compute the similarity between template fingerprint and input fingerprint. In order to tolerate matching minutiae pairs that are further apart because of plastic distortions, and therefore to decrease the false rejection rate (FRR), most algorithms increase the size of the bounding boxes. However, as a side effect, this gives non-matching minutiae pairs a higher probability to get paired, resulting in a higher false acceptance rate (FAR). Different with above algorithm, we give out a novel method to compute the similarity between two fingerprint images. We considered not only the number of matching minutiae but also the distance difference of the corresponding minutiae pairs.

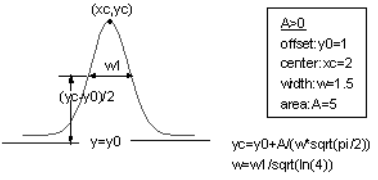


Fig. 2. The illustration of the expression (8).

Using the proposed algorithm, we can get the corresponding minutiae pairs between template fingerprint and input fingerprint. Suppose we get N1 corresponding minutiae pairs, and there are n_i sample points for every minutiae pair. Then we can compute the sum N2 of matched sample points as following.

$$N2 = \sum_{i=1}^{N1} n_i \tag{6}$$

Meanwhile we compute the mean of distance difference between every two minutiaes as following:

$$LenDif = \frac{1}{N1} \sum_{i=1}^{N1} len_{dif} \tag{7}$$

Where the meaning of len_{diff} can be seen expression (1).

After statistical analysis, we find that N2 and LenDif is approximately Guassian distributed. The experiments were done on FVC2002 DB1. It contains 800 fingerprint images captured by optical sensor "Identix TouchView II". Fig. 3,4 show the distribution of N2, LenDif in imposter match and genuine match. From Fig. 3, we can find that the value of N2 in genuine match is much bigger than in imposter match. And From Fig. 4, we can also find that the value of LenDif in genuine match is much smaller than in imposter match. It means that N2 and LenDif have excellent classification performance for match. We use the following Guassian functions to describe the character of N2 and LenDif.

$$y(x) = y0 + \frac{A}{W\sqrt{\pi/2}} e^{-2\frac{(x-x_c)^2}{w^2}} \tag{6}$$

Where x represents N2 or LenDif, the meaning of parameter y0, A, W, Xc can be seen in Fig. 2. Then the similarity between template fingerprint and input fingerprint is computed as following:

$$\text{similarity} = F_{N2} * F_E \tag{7}$$

$$F_{N2} = \begin{cases} \eta_1 \cdot y(N2) \cdot \dfrac{N2 - a1}{a2 - a1} & a1 < N2 < a2 \\ 1 & N2 > a2 \\ 0 & N2 < a1 \end{cases} \tag{8}$$

$$F_{LenDif} = \begin{cases} \eta_2 \cdot y(lenDif) \cdot \dfrac{e2 - lenDif}{e2 - e1} & e1 < lenDif < e2 \\ 0 & lenDif > e2 \\ 1 & lenDif < e1 \end{cases} \tag{9}$$

Where η_1 , η_2 a2,a1,e2,e1 are coefficients and a2>a1, e2>e1.

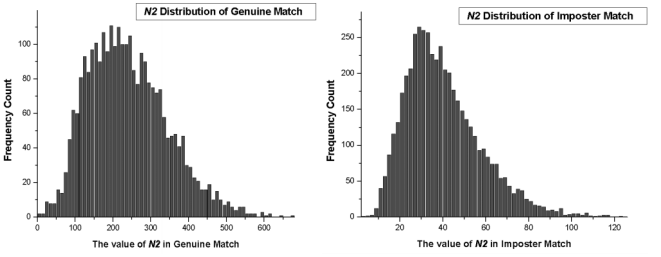


Fig. 3. The distribution of N2 In Genuine and Imposter Match on FVC2002 DB1.

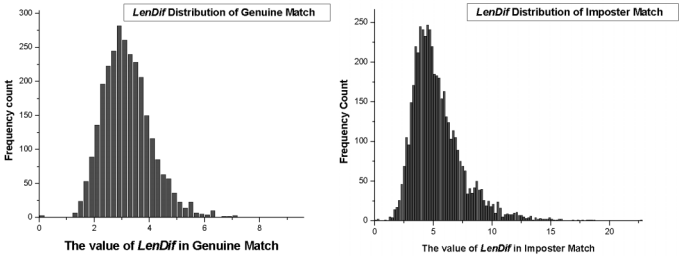


Fig. 4. The distribution of LenDif In Genuine and Imposter Match on FVC2002 DB1.

4 Experimental Results

The proposed algorithm has been participated in FVC2004. The Participant ID is P071 (open). The performance was ranked 3rd position in open category in FVC2004.

The detailed performance of the proposed algorithm can be seen from the website http://bias.csr.unibo.it/fvc2004/default.asp.

In FVC2004, databases are more difficult than FVC2000/FVC2002 ones. In particular in FVC2004, the organizer has insisted on: distortion, dry and wet fingerprints. In fingerprints database DB1 of Fvc2004, the distortion among the fingerprints from the same finger is obviously. The fingerprint images was acquired through CrossMatch V300 (Optical sensor). The size of the image is 640*480 pixels with the resolution about 500 dpi. The fingerprint database set A contains 800 fingerprint images captured from 100 different fingers, 8 images for each finger. Fig. 5 show two examples of big distortion captured from CrossMatch V300 sensor . Using the proposed algorithm, the similarity between these two fingerprints (102_3.tif and 102_5.tif) is 0.420082 (N2 =154, LenDif =8.511742). From Fig. 6, we can judge that these two fingerprints come from the same finger, it is a genuine match.

Fig. 5. The example of big distortion from FVC2004 DB1_B. (a) is 102_3.tif, (b) is 102_5.tif, (c) is the image which (a) (after rotation and translation) was added to (b). In region ◯ , the corresponding minutiae are approximately overlapped. But in region ◯ , the maximal vertical difference of corresponding minutiae is above 100 pixels.

The performance of the proposed algorithm on FVC2004 DB1 was shown in Fig. 6 . The equal error rate (EER) is about 4.37%. The experiments were done on PC AMD Athlon 1600+ (1.41 GHz).The average time for matching two minutiae sets is 0.77 seconds.

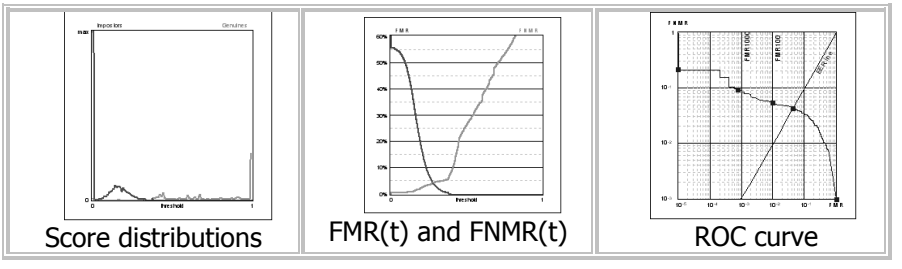


Fig. 6. Experimental results of the proposed algorithm on FVC2004 DB1_A.

5 Conclusion

How to cope with non-linear distortions in the matching algorithm is a real challenge. In this paper, we proposed a novel fingerprint matching algorithm based on the local topologic structure. The algorithm firstly aligns the template fingerprint and the input fingerprint using the registration method described in [6]. Then we introduce local topologic structure matching to improve the robustness of global alignment. Finally

we proposed a novel method to compute the similarity between the template fingerprint and the input fingerprint. The proposed algorithm has been participated in FVC2004. The performance was ranked 3rd position in open category in FVC2004. Experimental results show that the proposed algorithm has good performance on accuracy and processing time.

References

[1] R. Cappelli, D. Maio, and D. Maltoni, "Modelling plastic distortion in fingerprint images", in Proc. ICAPR2001, Rio de Janeiro, Mar. 2001.

[2] Hong Chen, Jie Tian*, Xin Yan,"Fingerprint Matching with Registration Pattern Inspection", Oral Report, AVBPA2003, pp.327-334, Springer, 2003.

[3] Biometric Systems Lab, Pattern Recognition and Image Processing Laboratory, Biometric Test Center, http://bias.csr.unibo.it/fvc2004/

[4] Asker M. Bazen , Sabih H. Gerez, "Fingerprint matching by thin-plate spline modelling of elastic deformations", Pattern Recognition, Volume 36, Issue 8, August 2003, pp.1859-1867

[5] Xiping Luo and Jie Tian, "Knowledge based fingerprint image enhancement", in Proc. 15th ICPR, Barcelona, Spain, Sept. 2000.

[6] Xiping Luo, Jie Tian and Yan Wu, "A Minutia Matching algorithm in Fingerprint Verification",15th ICPR, Vol.4, pp.833-836, Barcelona, 2000

[7] L. Hong, Y. Wan, and A. K. Jain, "Fingerprint image enhancement: algorithms and performance evaluation", IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 8, pp.777-789, 1998.

2-D Shape Matching Using Asymmetric Wavelet-Based Dissimilarity Measure

Ibrahim El Rube'[1], Mohamed Kamel[2], and Maher Ahmed[3]

[1] Systems Design Engineering, University of Waterloo, Canada,
ibrahim@watsup.uwaterloo.ca
[2] Electrical and Computer Engineering, University of Waterloo, Canada,
mkamel@uwaterloo.ca
[3] Physics and Computer Science Department, Wilfrid Laurier University , Canada
mahmed@wlu.ca

Abstract. In this paper, a wavelet-based multiscale asymmetric dissimilarity measure for shape matching is proposed. The wavelet transform is used to decompose the shape boundary into a multiscale representation. Given two shapes, a distance matrix is computed from the moment invariants of the wavelet coefficients at all the scale levels. The asymmetric dissimilarity is then calculated from the minimum values across each row on the distance matrix. The proposed asymmetric dissimilarity is a Hausdorff-like measure and is used for finding globally related shapes. The similarity paths obtained from the locations of the minimum distance values can be used to illustrate these relations.

1 Introduction

Shape matching is a fundamental stage in many areas, including computer vision, pattern recognition, visual information systems, and robotics. In many applications, it is essential for the shape matching to be invariant to geometric transformations such as similarity and affine transformations. Several shape matching techniques and algorithms are reported in the literature. In [1], a survey of shape analysis techniques can be found.

For a pair of patterns, a dissimilarity measure is usually concordant with the notion of a distance, which indicates the degree of the differences between the two patterns. A brief overview to the known dissimilarity measures and their properties for finite sets, curves, and regions is given in [2]. One of the most studied dissimilarity measures in computational geometry, is the Hausdorff distance. Many pattern-matching algorithms have been derived from the Hausdorff metric.

For two different sets $A = \{a_1, a_2, ...a_N\}$ and $B = \{b_1, b_2, ...b_M\}$, the Hausdorff distance is defined as follows:

$$\begin{aligned} H_k(A,B) &= max\{\overrightarrow{h}_k(A,B), \overrightarrow{h}_k(B,A)\}, \\ \text{where} \quad \overrightarrow{h}_k(A,B) &= max_{a\in A}\{min_{b\in B}\{d(a,b)\}\}. \end{aligned} \tag{1}$$

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 368–375, 2004.
© Springer-Verlag Berlin Heidelberg 2004

However, the Hausdorff distance is sensitive to noise and to the outlier points. To solve this problem, a number of Hausdorff-variant distances have been introduced in the literature. The partial Hausdorff distance [3] replaces the maximum operator in Equation (1) with a fixed rank, but this distance is not metric and does not satisfy most of the metric axioms.

Wavelets have been widely applied in many areas, such as computer vision, pattern recognition, image processing, mathematics, and physics. The reason for this is that it provides time-frequency localizations and a hierarchical representation of the signals. Wavelet transform coefficients have been used in [4] and [5], as shape descriptors for shape matching. Other researchers, [6], [7], and [8], have derived invariant functions from the dyadic wavelet transform of the contour of the shape. These functions are sensitive to noise in the first few decomposed scale levels, due to their dependance on the detail coefficients. Other techniques consist of combinations of the WT and Fourier Transform (FT) [9], of the wavelet multiscale features and the Hopfield neural networks [10], and of geometric moments with the wavelet transform [11].

In this paper a new asymmetric dissimilarity measure, based on the multiscale decomposition of the wavelet representation is described. The proposed measure is a Hausdorff-like measure in the sense that it is directionally sensitive. Also, mutual similarity paths between shapes are introduced in this paper. These paths give a succinct representation of the similarities between the shapes.

2 Proposed Method

The directed distance measure (e.g., the Hausdorff distance) is an attractive and a very useful technique, especially in partial shape matching applications. We have extended the idea of the directed measure to derive more meaningful dissimilarities that are related to the wavelet decomposition of the shape boundaries. Consider two similar shapes as shown in Figure 1. The relation between these shapes can be summarized as follows:

- Shape A will be similar to shape B, if some details are removed from shape B.
- Shape B will be similar to shape A, if some details are added to shape A.
- Shapes A and B will be similar to each other if some/all details are removed from both shapes.
- Shapes A and B will be similar to each other if some details are added to both shapes.

Practically, it is easier to remove (smooth) details from the shape than to add them. Consequently, the second and the fourth augments are not considered in this study. Furthermore, the first and the third arguments can be combined if both shapes are decomposed into a hierarchical multiscale scheme (i.e., a wavelet transform). In this hierarchical scheme, all the decomposed levels are related to each other by the filtration process that is carried out by the transform.

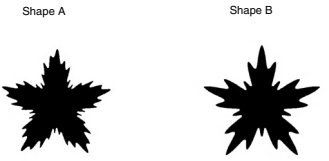


Fig. 1. Two similar shapes

2.1 Feature Extraction

In order to compute the multiscale dissimilarity measures, the features are extracted from the segmented shapes by performing three steps: boundary extraction, wavelet decomposition, and the invariant moments computation of the wavelet coefficients.

Boundary Extraction: The outer boundary of the shape is extracted by using one of the known boundary extractor algorithms (in this study, the bug following technique was used). The extracted 2-D boundary is then converted into two 1-D sequences $x(k)$ and $y(k)$.

Wavelet Decomposition: A 1-D discrete wavelet transform (DWT) is applied to $x(k)$ and $y(k)$ to obtain the different approximation and detail coefficients. The boundary sequences ($x(k)$ and $y(k)$) are decomposed to a certain wavelet scale level L. Figure 2 plots the approximation and the detail coefficients as 2-D contours for $L = 1$ *to* 6. If the shape is subjected to an affine transformation, then the wavelet coefficients of these sequences will be affected by the same transformation (after eliminating the translation parameters).

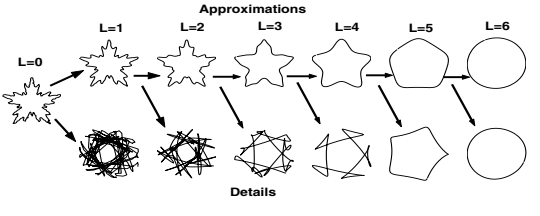


Fig. 2. Multiscale representations of a star shape using wavelet decomposition.

Moment Invariants: The affine invariant curve moments, defined in [12], are computed for the approximation coefficients at all the scale levels. Six moment invariants from [13] are used here with each scale level to obtain the distance matrices between the shapes. Since the moments are invariant to the affine transformation, the dissimilarity measure also becomes invariant to this transformation

group. Due to their sensitivity to noise and to boundary deformations, moments are normalized at each scale by subtracting the mean and dividing by the standard deviation of these moments.

For the detail coefficients, a simple invariant representation is used to represent the detail contours, shown in Figure 2, as 1-D sequences. This representation is computed from the triangle area of each adjacent three points on the coefficient's contour. Only 1-D moments are required for computing the invariant features of the detail coefficient after the area representation is employed.

2.2 Wavelet-Based Multiscale Dissimilarities

In this paper, to calculate the dissimilarity between the shapes, the distance matrix computation is based on the Euclidian distances between the curve moment invariants of the wavelet coefficients at all the scale levels. Two dissimilarities are introduced in this work:

Symmetric Dissimilarity: A symmetric dissimilarity measure (DS1) is calculated by taking the diagonal values of the distance matrix. This dissimilarity measures the distances between the corresponding scale levels of each two shapes. If U_{AB} is a distance matrix between shapes A and B, then

$$DS1_{AB}(L) = \sum_{l=l_o}^{L} U_{AB}(l,l), \tag{2}$$

where L is the highest and l_o is the lowest used scale levels.

Asymmetric Dissimilarity: The asymmetric dissimilarity measure (Hausdorff-like distance measure) is computed by tracking and capturing the minimum values across each row in the distance matrix. This measure is directed and asymmetric which indicates that two shapes can share two different dissimilarity values between them. The forward and the reverse dissimilarities between shapes A and B are,

$$\overrightarrow{V}_{AB}(l) = min_{m\in B}\{U_{AB}(l,m)\} \tag{3a}$$

$$\overrightarrow{V}_{BA}(l) = min_{m\in A}\{U_{BA}(l,m)\}, \tag{3b}$$

respectively, where $U_{AB} = (U_{BA})^T$.

The symmetry is achieved by taking the maximum values of both directions

$$DS2_{AB}(L) = \sum_{l=l_o}^{L} max\{\overrightarrow{V}_{AB}(l), \overrightarrow{V}_{BA}(l)\}. \tag{4}$$

The differences between this measure and the Hausdorff measure are:

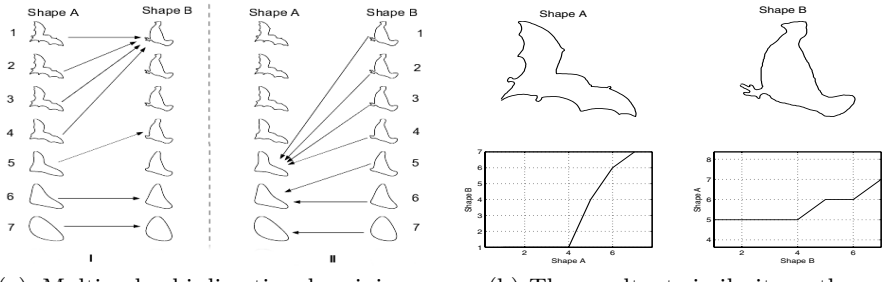
- The Hausdorff distance is sensitive to the outlier noise, whereas the wavelet-based dissimilarity is less sensitive to noise because of the filtration of the noise in the first levels of the wavelet decomposition.
- The hierarchical wavelet decomposition provides more flexibility in selecting the scale levels that are involved in computing the dissimilarity.

The advantage of using the asymmetric dissimilarity, DS2, over the symmetric one, DS1, is that related shapes are easily detected by DS2. Both the DS1 and DS2 are adopted in this paper for shape matching.

The symmetric dissimilarity can be computed for both the approximation and the detail coefficients, whereas the asymmetric dissimilarity is computed for the approximation coefficients only. The reason for this is that the detail coefficients are usually uncorrelated, and in most cases, are independent from one scale level to another, whereas the approximation coefficients are not.

2.3 Similarity Paths

As mentioned earlier in section 2.2, when the minimum distances are tracked from the distance matrix, the minimum values and their locations are recorded. The locations represent a mutual similarity path between the two shapes, as seen from one of them. The two shapes could have two different similarity paths. If the two mutual similarity paths are similar, then these shapes can be classified as globally similar shapes.



(a) Multiscale bi-directional minimum distance locations.

(b) The resultant similarity paths.

Fig. 3. Example of finding and plotting the similarity paths between two shapes.

Fig. 3(a) illustrates the locations of the minimum distances between the scale levels of shapes A and B as seen from both directions. Fig. 3(b) exhibits the resultant unweighted mutual similarity paths between shape A and Shape B. These paths do not indicate the value of the dissimilarity between the shapes, but they do give an indication of the hierarchical relations between the shapes. The first three levels are important for the local closeness between the shapes, and the last levels are important for the global relations between them.

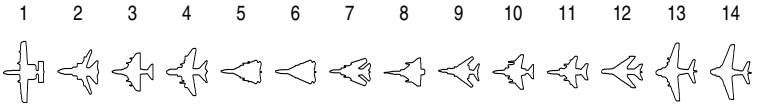


Fig. 4. The shapes used in the first experiment

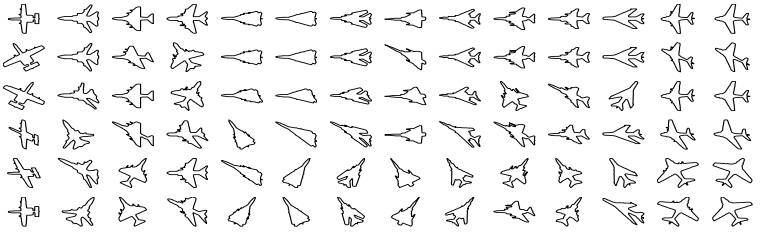
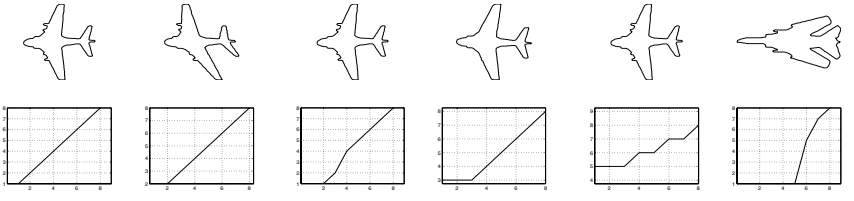


Fig. 5. Matching results of the first experiment. From top to down, the first top row represent the original shapes, the second row is the first match, and so on.



(a) Affine distorted shapes (b) similar shapes (c) Different shapes

Fig. 6. Similarity paths for three type of relations between shapes.

3 Experiential Results

Two experiments are reported in this paper. The first one is to ensure the invariance of the dissimilarity measure to the affine transformation. The second experiment is carried out to measure the efficiency of the dissimilarity in finding similar shapes.

3.1 Affine Invariance Results

The first data set, shown in figure 4, contains 14 airplane shapes which are adopted in many papers (e.g.,[6], [7], and [8]). The affine distorted shapes are obtained by transforming the original 2-D shapes by using affine transformation equations. Figure 5 presents the results of matching each of the original shapes (the first row) with the distorted ones. The results indicate a perfect recovery of all the affine transformed shapes using the $DS2_{AB}$ dissimilarity measure.

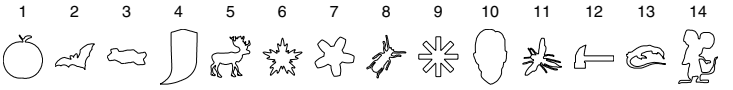


Fig. 7. The shapes used in the second experiment

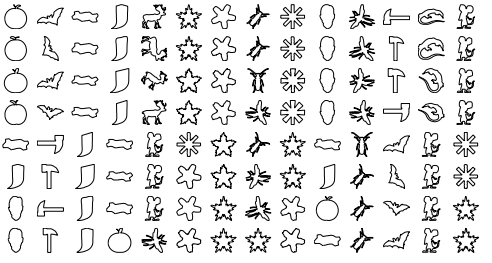


Fig. 8. Matching results from the approximation-based dissimilarity

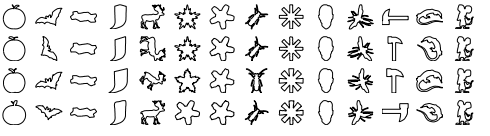


Fig. 9. Matching refined results after using the detail-based dissimilarity

3.2 Similar Shapes Matching Results

The second data set includes 14 groups of shapes, shown in figure 7, with each group contains four similar shapes. These shapes are used in MPEG-7 system [14] and used by many research groups.

The results of the experiment in Figure 8 indicate that the proposed dissimilarity is able to find globally similar shapes. This is because that this dissimilarity is computed from the approximation coefficients of the wavelet transform. Figure 9 reveals the refined matching results after applying the detail-based dissimilarity to the closest 12 shapes resulted from the experiment of figure 8. The co-operation of the approximation-based dissimilarity with the detail-based dissimilarity in this manner, ensures that only the similar shapes are captured even if the shape boundary is subjected to some deformations.

4 Conclusions

In this paper, a wavelet-based asymmetric dissimilarity measure is proposed and tested. The asymmetric dissimilarity is a Hausdorff-like measure with the advantages of less sensitivity to noise and small boundary deformations. The approximation coefficients are more suitable for computing this dissimilarity due to the dependency of each level on the previous one. In the detail coefficients,

the symmetric dissimilarity is more convenient and more stable due to the independency of these coefficients from one scale level to another. As a result of the tracking of the minimum values across the distance matrix between two shapes, the so called mutual similarity paths are found. The mutual similarity paths that are obtained are useful representations for global shape comparisons.

References

1. Loncaric, S.: A survey of shape analysis techniques. Pattern Recognition **31** (1998) 983–1001
2. Veltkamp, R., Hagedoorn, M.: State-of-the-art in shape matching. In (ed.), M.L., ed.: Principles of Visual Information Retrieval, Springer (2001) 87–119
3. Huttenlocher, D., Klanderman, D., Rucklige, A.: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15** (1993) 850–863
4. Chauang, G., Kuo, C.: Wavelet descriptor of planar curves: Theory and applications. IEEE Transaction on Image Processing **5** (1996) 56–70
5. Kashi, R., B-Kavde, P., Nowakowski, R., Papathomas, T.: 2-D shape representation and averaging using normalized wavelet descriptors. Simulation **66** (1996) 164–178
6. Alferez, R., Wang, Y.: Geometric and illumination invariants for object recognition. IEEE Trans. on PAMI **21** (1999) 505–536
7. Tieng, Q., Boles, W.: An application of wavelet based affine invariant representation. Pattern Recognition Letters **16** (1995) 1287–1296
8. Khalil, M., Bayoumi, M.: Affine invariants for object recognition using the wavelet transform. Pattern Recognition Letters **23** (2002) 57–72
9. Chen, G.: Applications of wavelet transformation in pattern recognition and denoising. Master thesis, Concordia University (1999)
10. Lin, W., Chen, C., Sun, Y.: Multiscale object recognition under affine transformation. IEICE Transaction Information and Systems **E82d** (1999) 1474–1482
11. Ohm, J.R., Bunjamin, F., Liebsch, W., Makai, B., M ller, K., A. Smolic, D.Z.: A set of visual descriptors and their combination in a low-level description scheme. Signal Processing: Image Communication **16** (2000) 157–179
12. Zhao, D., Chen, J.: Affine curve moment invariants for shape recognition. Pattern Recognition **30** (1997) 895–901
13. Flusser, J., Suk, T.: A moment-based approach to registration of images with affine geometric distortion. IEEE Transactions on Geoscience and Remote Sensing **32** (1994) 382–387
14. Latecki, P.L.J.: (http://www.cis.temple.edu/ latecki/)

A Real-Time Image Stabilization System Based on Fourier-Mellin Transform

J.R. Martinez-de Dios and A. Ollero

Grupo de Robótica, Visión y Control.
Departamento de Ingenieria de Sistemas y Automatica.
Escuela Superior de Ingenieros. Universidad de Sevilla
Camino de los Descubrimientos, sn, 41092, Sevilla (Spain)
Phone: +34 954487357; Fax: +34 954487340
{jdedios, aollero}@cartuja.us.es

Abstract. The paper presents a robust real-time image stabilization system based on the Fourier-Mellin transform. The system is capable of performing image capture-stabilization-display at a rate of standard video on a general Pentium III at 800 MHz without any specialized hardware and the use of any particular software platforms. This paper describes the theoretical basis of the image matching used and the practical aspects considered to increase its robustness and accuracy as well as the optimizations carried out for its real-time implementation. The system has been submitted to extensive practical experimentation in several applications showing high robustness.

1 Introduction

Image stabilization is a relevant topic in many applications including those in which the video is only used for human visualization and those in which the sequences of images are processed by a computer. In human visualization applications image vibrations introduce stress in the operator, which involves a decrease in the capacity of attention. In computerized image processing applications vibrations have harmful effects and they often include a step devoted to vibrations cancellation.

Two main approaches have been developed for image stabilization. The first one aims to stabilize the camera vibrations. This approach is used by various types of systems from simple mechanical systems for handheld camcorders to inertial gyrostabilized camera systems for gimbals. Mechanical systems for handheld camcorders usually have low accuracy and perform "vibrations reduction" more than "vibrations cancellation". Gyrostabilized camera systems are restricted to only some applications to due their usual high cost, size and weight. Another approach corrects the images by applying image processing techniques. Several image-processing stabilization methods have been proposed. The main limitation of these methods is that they require high time-consuming computations. This paper presents a robust real-time image

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 376–383, 2004.
© Springer-Verlag Berlin Heidelberg 2004

stabilization system based on Fourier-Mellin matching. One of its objectives is to avoid any dependence with hardware or software platforms.

2 Principle for Image Stabilization

The scheme of the processing applied is depicted in Fig.1. *Image Stabilization* performs image matching techniques between the current image, $Im_n(x,y)$, and the stabilized version of the image captured in the previous time instant, $Im^E_{n-1}(x,y)$. Image stabilization corrects $Im_n(x,y)$ by applying geometric relations inverse to those found between $Im_n(x,y)$ and $Im^E_{n-1}(x,y)$.

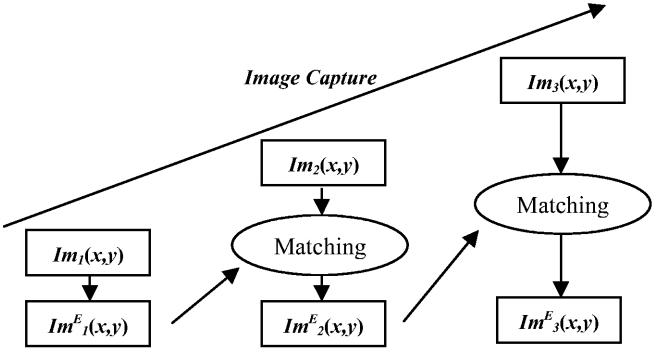


Fig. 1. Operation scheme of the application of image stabilization.

Preliminary simulations showed that image vibrations can be modeled as the combination of image translations and rotations while the scale component could be neglected. The most common approach to image matching is based on cross-correlation, [2]. The straightforward formulation of cross-correlation is not capable of matching rotated images. Besides, its has poor selectivity capacity and does not behave well in presence of noise. Some alternatives have been proposed to cope with rotated images. But, these alternatives involve very time-consuming computations. Matching techniques based on invariant moments are sensitive to noise and have low discriminating capacity, [1]. Other group of techniques is based on matching a number of features present in both images [5]. These techniques require the presence of a considerable number of features which could not always be present in the images. Moreover, the features matching is usually carried out by cross-correlation, with the above-mentioned limitations. Several Fourier transform based matching techniques have been proposed in [7] and [6] but they can not match rotated images. Fourier-Mellin transform is capable of matching translated and rotated images, [3].

2.1 Matching of Two Images Through the Fourier-Mellin Transform

Consider that image $Im_n(x,y)$ is a rotated and translated replica of image $Im_n^E(x,y)$. The stabilization consists of two steps: rotation correction and translation correction. Consider that $s(x,y)$ and $r(x,y)$ are respectively the central rectangular region of $Im_n(x,y)$ and $Im_{n-1}^E(x,y)$. Thus, $s(x,y)$ is a rotated and translated replica of $r(x,y)$:

$$s(x,y) = r\big((x\cos\alpha + v\sin\alpha) - x_0,\ (-x\sin\alpha + v\cos\alpha) - y_0\big), \tag{1}$$

where is the rotation angle and x_0 and y_0 are the translational offsets. The Fourier transforms of $s(x,y)$ and $r(x,y)$ are related by:

$$S(u,v) = e^{-j\phi_S(u,v)} R\big((u\cos\alpha + v\sin\alpha),\ (-u\sin\alpha + v\cos\alpha)\big), \tag{2}$$

where $\phi_S(u,v)$ is the spectral phase of $s(x,y)$ and depends on the translation and rotation. The spectral magnitude of $S(u,v)$ is translation invariant:

$$|S(u,v)| = \big|R\big((u\cos\alpha + v\sin\alpha),\ (-u\sin\alpha + v\cos\alpha)\big)\big| \tag{3}$$

Thus, a rotation of the image involves a rotation of the same angle of the spectral magnitude. Assume that $r_p(\theta,\rho)$ and $s_p(\theta,\rho)$ are the spectral magnitudes of $r(x,y)$ and $s(x,y)$ in the polar co-ordinates (θ,ρ). It is easy to check that:

$$s_p(\theta,\rho) = r_p(\theta - \alpha,\rho) \tag{4}$$

The image rotation is transformed to a translation along the angular axis. It is easy to observe that $S_p(\nu,\varpi)$ and $R_p(\nu,\varpi)$, the Fourier transform of $s_p(\theta,\rho)$ and $r_p(\theta,\rho)$, are related by $S_p(\nu,\varpi) = e^{-j2\pi\alpha\theta} R_p(\nu,\varpi)$. Thus, $s_p(\theta,\rho)$ and $r_p(\theta,\rho)$ have the same spectral magnitude.

2.2 Image Rotation Correction

From $\phi_{S_p}(\nu,\varpi)$ and $\phi_{R_p}(\nu,\varpi)$, phases of $S_p(\nu,\varpi)$ and $R_p(\nu,\varpi)$, we define:

$$Q_r(\nu,\varpi) = exp\Big[-j\Big(\phi_{S_p}(\nu,\varpi) - \phi_{R_p}(\nu,\varpi)\Big)\Big] \tag{5}$$

The rotation angle between $s(x,y)$ and $r(x,y)$ can be obtained by computing the inverse Fourier transform of $Q_r(\nu,\varpi)$, $q_r(\theta,\rho) = F^{-1}\{Q_r(\nu,\varpi)\}$, where F^{-1} stands for the inverse Fourier transform. The peak of $q_r(\theta,\rho)$ is located at $\theta = \theta_{max}$ and $\rho = \rho_{max}$. The rotation angle is given by $\alpha = \theta_{max}$. Once the rotation angle has

been obtained the translation correction is applied by rotating $Im_n(x,y)$ an angle $-\alpha$. The rotation-corrected image is $Im_n^R(x,y)$, its central part is $s^R(x,y)$.

2.3 Image Translation Correction

The $s^R(x,y)$ is a translated version of $r(x,y)$. It is possible to define $Q_t(u,v)$ from the phases of $S^R(u,v)$ and $R(u,v)$, the Fourier transform of $s^R(x,y)$ and $r(x,y)$:

$$Q_t(u,v) = exp\left[-j\left(\phi_{S^R}(u,v) - \phi_R(u,v)\right)\right] \tag{6}$$

The translations between $s^R(x,y)$ and $r(x,y)$ can be obtained by computing the inverse Fourier transform of $Q_t(u,v)$, $q_t(x,y) = Fourier^{-1}\{Q_t(u,v)\}$. The peak of $q_t(x,y)$ is located at $x = x_{max}$ and $y = y_{max}$. The translations between $s^R(x,y)$ and $r(x,y)$ are given by $x_0 = x_{max}$ and $y_0 = y_{max}$. The translation correction is applied by shifting $Im_n^R(x,y)$ by $-x_0$ and $-y_0$.

3 Practical Aspects

3.1 Drift Correction

The accumulation of small errors along the stabilization of a certain number of images produces. A drift correction technique is periodically carried out every *N* of images. The rotation angle between $Im_n(x,y)$ and $Im_{n-1}^E(x,y)$ is $\alpha|_1$ and, between $Im_n(x,y)$ and $Im_{n-N}^E(x,y)$ is $\alpha|_N$. The rotation angle is taken as a combination of them:

$$\alpha = (1-\tau_r)\,\alpha|_1 + \tau_r\,\alpha|_N, \tag{7}$$

where $\tau_r \in [0,1]$ is called *drift correction rotation factor*. If $\tau_r = 0$ no drift correction is applied, while $\tau_r = 1$ usually generates sudden changes in image rotation corrections. Once $Im_n(x,y)$ has been rotation corrected with $-\alpha$, the translational offsets are computed by combining $x_0|_1$ and $y_0|_1$, the translational offsets between $Im_n^R(x,y)$ and $Im_{n-1}^E(x,y)$, together with $x_0|_N$ and $y_0|_N$, translational offsets between $Im_n^R(x,y)$ and $Im_{n-N}^E(x,y)$. The expressions of such combinations are:

$x_0|_1 = (1-\tau_t)\, x_0|_1 + \tau_t\, x_0|_N$ and $y_0|_1 = (1-\tau_t)\, y_0|_1 + \tau_t\, y_0|_N$, where $\tau_t \in [0,1]$ is the *drift translation correction factor*.

3.2 Contrast Correction

The matching method obtains poorer results with low contrasted images. A contrast-enhance method based of histogram stretching has been used to improve the luminance of the images. Luminance of $Im(x,y)$ is often characterized by the image bright (*MI*) and contrast (*C*). The transformation function that should be applied to obtain the desired bright and contrast values (MI_{ref} and C_{ref}) is: $Im^*(x,y) = (C_{ref}/C)(Im(x,y) - MI) + MI_{ref}$.

3.3 Operation Modes

The resolution in the computation of the rotation angle and translational offsets is highly dependent on the size of the matrices that represent the images. Consider that $r(x,y)$ and $s(x,y)$ are represented by square matrices of size *M*x*M* and that $s_p(\theta,\rho)$ and $r_p(\theta,\rho)$ are represented by square matrices of size *W*x*W*.

The resolution of the rotation angle depends on the size of the matrices that represent $S_p(\nu,\varpi)$ and $R_p(\nu,\varpi)$, i.e. *W*x*W*. It is easy to notice that the minimum detectable angle is $\alpha_{min} \approx 2/W$. The value of *W* also has influence on the errors in the computation of the rotation angle. The higher *W* is, the more accuracy in the computation of the rotation angle can be obtained. The value of *W* depends on the number of different radius values considered in the polar conversion, which is constrained by the size of $|S(u,v)|$ and $|R(u,v)|$. The size of the matrices that represent $r(x,y)$ and $s(x,y)$ has straightforward influence on the resolution and errors in the computation of the translations. Lower values of *M* involve poor accuracy since the peak of $q_t(x,y)$ is more broad and more affected by noise. Two operation modes have been selected to cope with the compromise between computer requirements and stabilization accuracy: *Mode1* (low values of *M* and *W* and medium stabilization capability) and *Mode2* (for high magnitude or high frequency vibrations).

3.4 Increase Accuracy Through Sub-pixel Resolution

The position of the peak of $q_r(\theta,\rho)$ and $q_t(x,y)$ determine the value of α, and x_0 and y_0, respectively. The resolutions in the computation of α, x_0 and y_0 are limited by the values of *M* and *W*. Selecting higher values of *M* and *W* increases the computation load. An efficient alternative applied considers a sub-pixel estimation to the peak position. The sub-pixel estimation considers the peak is located at the centroid in a neighborhood of certain size centered at the position of the peak:

$p_m = \left(A(X-1) + B\,X + C(X+1)\right)/(A+B+C)$, where A, B and C are the magnitude at X-1, X and X +1 respectively. For estimating the position of the peak in matrices, the centroid is computed in a 2D neighborhood.

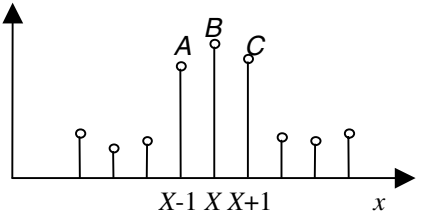


Fig. 2. Sub-pixel estimation of peak position.

4 Computational Aspects

Image matching based on Fourier Mellin requires the computation of six 2D FFTs and two 2D Inverse FFTs. Special care has been put in the optimization of the computation of the FFT. The Cooley-Tukey FFT algorithm [4] was used due to its combination of efficiency and simplicity. The size of the matrices has been selected to be power of two. The row-column approach was used for the 2D FFT.

Two approaches have been considered for the optimization of the Fourier transforms. The first one exploits the symmetry properties of FFTs of real data. If $x(k)$ is a sequence of real data its FFT, $X(k)$, satisfies: $Re\{X(k)\} = Re\{X(-k)\}$ and $Im\{X(k)\} = -Im\{X(-k)\}$, where $Re\{X(k)\}$ and $Im\{X(k)\}$ are the real and imaginary components of $X(k)$. The computation of the 2D FFT also exploits the following symmetry property: if $A(x,y)$ is a matrix of real data its 2D FFT, $A(u,v)$, satisfies: $Re\{A(u,v)\} = Re\{A(-u,-v)\}$ and $Im\{A(u,v)\} = -Im\{A(-u,-v)\}$. Both properties can save up to 50% of the total computation of the 2D FFT of matrices with real data.

The second one computes the twiddle factors of the Cooley-Tukey algorithm at the initialization of the stabilization system. The twiddle factors, $W_i = e^{2\pi i/N}$, are constant values that only depend on N, the length of the vectors which FFT is to be computed, which only depend of the operating mode. The pre-computation of the twiddle factors avoids calculating them each time a FFT is computed. This represents an important reduction (more than 40%) in the operations required for FFT.

Rotation and translation corrections involve the application of image interpolation to deal with non-integer rotation angles and translational offsets. Bilinear interpolation was chosen for its simplicity and efficiency. Further reduction in the computational load (up to 30%.) of bilinear interpolations can be obtained by using 'integer' Mathematical operations which are more efficient than 'floating point'.

5 Experiments

The image stabilization method was implemented with ANSI C on a Pentium III at 800 MHz. It was implemented in Windows NT and Vx-Works to test its portability. The system was submitted to extensive experiments in several different applications. Consider that $Im_{n-1}^{E}(x,y)$ and $Im_n(x,y)$, shown in Fig. 3a-b, are two consecutive images. The first step is the computation of the rotation angle between $Im_{n-1}^{E}(x,y)$ and $Im_n(x,y)$. The peak of $q_r(\theta,\rho)$ takes place at $\theta_{max}=3$, which corresponds to $\alpha = 2'109375°$. Then, $Im_n(x,y)$ is rotated an angle $-\alpha°$. The rotation corrected image, $Im_n^R(x,y)$, is shown in Fig. 3c. In the computation of the translational offsets, the peak of $q_t(x,y)$ takes place at x_0=0.4 and y_0=-585. Translation correction is applied by shifting $Im_n^R(x,y)$ $-x_0$ and $-y_0$. The stabilized image $Im_n^E(x,y)$ is shown in Fig. 4c together with the original image (shown in Fig. 4b) and reference image, in Fig. 4a.

a) b) c)

Fig. 3. a), b) Two consecutive images from a camera under vibrations $Im_{n-1}^{E}(x,y)$ and $Im_n(x,y)$, c) the rotation-corrected version of $Im_n(x,y)$.

Numerous experiments have been carried out to test the robustness of the system. In the experiments carried out *Mode1* uses *M*=*W*=128 and *Mode2*, *M*=*W*=256. The image stabilizing time at *Mode1* is 28.6 ms., which allows real-time stabilization for PAL and NTSC video standards. The stabilizing time at *Mode2* is 102.1 ms.

a) b) c)

Fig. 4. a) $Im_{n-1}^{E}(x,y)$; b) original image with vibrations $Im_n(x,y)$; c) stabilized $Im_n^E(x,y)$.

6 Conclusions

The paper presents a robust real-time image stabilization system based on Fourier-Mellin transform. The stabilization system is based on applying matching based on Fourier-Mellin transforms between consecutive images in a sequence. The stabilization system was optimized to correct the rotations and translations since the scale factor between consecutive images could be neglected in the applications considered. Image matching is applied in two steps: detection and correction of rotations and detection and correction of translations. To increase the robustness, the system includes drift correction techniques and contrast correction. To increase the accuracy of the system, it includes sub-pixel computation of the rotation angle and translational offsets. Special effort has been applied on the minimization of the computer load including -among others- the pre-computing of Cooley-Tukey twiddle factors and the computation of several operations with integer data. The method was implanted in a Pentium III at 800 MHz with 128 Mbytes of RAM. It is capable of performing image capture-stabilization-display at a rate of PAL and NTSC video.

Acknowledgements. The authors would like to thank Joaquin Ferruz and Luis Fernández. The work described in this paper has been developed in the project SEOAN "Sistema Electroóptico de Ayuda a la Navegación". SEOAN project is leaded by "Division de Sistemas" of the Spanish company IZAR and funded by the "Gerencia del Sector Naval". The authors express their gratefulness to Antonio Criado, Francisco López, Alfonso Cardona, Baltasar Cabrera, Juan Manuel Galán and José Manjón from IZAR.

References

1. Abu-Mostafa Y.S. and D. Psaltis. "Recognition aspects of moment invariants". IEEE Trans. Pattern Anal. Mach. Intel., 16(12) (1984). 1156-1168.
2. Barnea, D. I. and H. F. Silverman, "A class of algorithms for fast image registration". IEEE Trans. Computers, C-21, (1972). 179-186.
3. Chen Q., M. Defrise, F. Deconinck, "Symmetric Phase-Only Matched Filtering of Fourier-Mellin Transform for Image Registration and Recognition", IEEE Trans. P.A.M.I., vol. 16, no 12, (1994). 1156-1167.
4. Cooley J.W. and J.W. Tukey, "An algorithm for the machine calculation of complex Fourier series", Math. Comput. 19, (1965). 297–301.
5. Faugeras O., Q. Luong, and T. Papadopoulo, "The Geometry of Multiple Images. MIT Press, 2001. ISBN 0-262-06220-8.
6. Horner J.L. and P.D. Gianino, "Phase-only matched filtering", Applied Optics, vol. 23, no. 6, (1984). 812-816.
7. Oppenheim A.V. and J.S. Lim, "The importance of phase in signals", IEEE Proc. Vol. 69, no. 5, (1981). 529-541.

A Novel Shape Descriptor Based on Interrelation Quadruplet

Dongil Han[1], Bum-Jae You[2], and Sang-Rok Oh[2]

[1] Department of Computer Engineering
Sejong University
98 Gunja-Dong, Gwangjin-Gu, Seoul 143-747, Korea
dihan@sejong.ac.kr
[2] Intelligent Robotics Research Center
Korea Institute of Science and Technology
39-1, Haweolkok-Dong, Seongbuk-Gu, Seoul 136-791, Korea
{ybj, sroh}@kist.re.kr

Abstract. In this paper, we propose a new shape descriptor, which represents the 2-D shape information by using the concept of *interrelation quadruplet*. For this purpose, the polygonal approximation of 2-D shape is applied first. The line segments can be extracted from the polygonal shapes and the definition of interrelation quadruplet between ling segments is introduced. The properties of interrelation quadruplet that is invariant to translation, rotation and scaling of a pair of line segments is described. Several useful properties of the interrelation quadruplet are also derived in relation to efficient partial shape recognition. The shape recognition using the interrelation quadruplet requires only small space of storage and is shown to be computationally simple and efficient.

1 Introduction

Shape descriptors are important tools in many applications of pattern recognition systems, which allow searching and matching images in a database model with respect to the shape information. The goal of the shape descriptors is to uniquely characterize the object shape in a large image database. A vigorous shape descriptor should contain sufficient information to resolve distinct images and compact enough to ignore the redundancies in the shapes. Additionally, it should give results consistent to human visual system.

The shape description methods can be divided into several methods; Fourier descriptors [1-3], invariant moments [4-5], skeleton based descriptors [6-7]. Fourier descriptors are one of the most popular technologies and provide a means for representing the boundary of a two-dimensional shape. The advantages of using Fourier descriptors are that the shape information is concentrated in the low frequency and noise usually affects only the high-frequency parts.

Moments describe shape in terms of its area, position, orientation and other parameters. The set of invariant moments makes useful feature information for the rec-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 384–391, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ognition of objects and matching of the invariant moments feature is computationally inexpensive and is a promising candidate for interactive applications.

But, the shape description features mentioned above are global in nature, dealing with the entire object boundary, silhouette, intensity profile or range map and rely on the entire shape for the determination of the features. Distortion of an isolated region of the shape will result in changes to every feature. This property is undesirable when partial shapes are under consideration.

The motivation of this paper is to find a compact representation of the two-dimensional shape information with the local features to describe and recognize images including partial shapes.

2 Interrelation Quadruplet

Many studies related to the shape analysis use edge information obtained from the boundary of object [8-9]. In this research, the polygonal approximated line segment information is used for describing the invariant property of objects.

Let's consider a polygon that consists of m line segments. Each segment S_i can be expressed as

$$S_i = (x_i, y_i, l_i, \theta_i), \ i = 1,..., m \ . \tag{1}$$

Here x_i and y_i denote the midpoint of the segment, l_i is a length of the segment, and θ_i is an orientation of the segment. Then, new line segment that connects midpoint of two segments can be defined as follows.

Definition 1. For two non-intersecting line segments S_i and S_j, the line segment connecting the midpoint of two line segment from the midpoint of S_i to the midpoint of S_j is called the *interconnection segment* IS_{ij}.

The *interconnection segment* IS_{ij} can be represented as four components and each component of *interconnection segment* IS_{ij} shows values as in equation (1).

$$IS_{ij} = (x_{ij}, y_{ij}, d_{ij}, \gamma_{ij}), \ i = 1,..., m \ . \tag{2}$$

Definition 2. For two non-intersecting line segments $S_i = (x_i, y_i, l_i, \theta_i)$ and $S_j = (x_j, j_j, j_j, \theta_j)$, the *interrelation quadruplet,* IQ_{ij} is a 4-element set each of whose element carries the geometrical relation between the base line segment S_i and the forward line segment S_j as follows:

$$IQ_{ij} = (\theta_{ij}, \phi_{ij}, l_{ij}, ld_{ij}) \, . \tag{3}$$

Where

$$\theta_{ij} = \theta_i - \theta_j, \quad \phi_{ij} = \gamma_{ij} - \theta_j, \quad l_{ij} = \frac{l_j}{l_i}, \quad ld_{ij} = \frac{d_{ij}}{l_i}. \tag{4}$$

The *interrelation quadruplet* IQ_{ij} conveys geometrical relation between two line segments S_i and S_j existing on the polygonal boundary shapes. The following theorems give many useful properties to recognize polygonal objects from the definition of *interrelation quadruplet*.

Theorem 1. *Interrelation quadruplet* is invariant under translation, rotation, and scaling for a pair of line segments – *necessary condition for invariance*.

Proof. Let us consider a pair of non-intersecting line segments S_i and S_j. If we let S_i^{st} and S_j^{st} be the scaled and translated line segment of S_i and S_j, respectively. From the definition of *interrelation quadruplet*, we can easily show that

$$IQ_{ij}{}^{st} = IQ_{ij}. \tag{5}$$

If we let S_i^r and S_j^r be the scaled and translated line segment of S_i and S_j, respectively and if we denote the rotation angle as θ, then the two elements $l_{ij}{}^r$, $ld_{ij}{}^r$ are invariant under rotation and

$$\theta_{ij}{}^r = (\theta_j + \theta) - (\theta_i + \theta) = \theta_{ij}. \tag{6}$$

$$\phi_{ij}{}^r = (\phi_{ij} + \theta) - (\theta_i + \theta) = \phi_{ij}. \tag{7}$$

Thus

$$IQ_{ij}{}^r = IQ_{ij}. \tag{8}$$

Therefore *interrelation quadruplet* is invariant under translation, rotation and scaling of a pair of line segments.

Theorem 2. If there are two pairs of line segments that have the same *interrelation quadruplet*, then one pair of line segment can be represented as a translated, rotated or scaled version of the other pair of remaining line segments – *sufficient condition for invariance.*

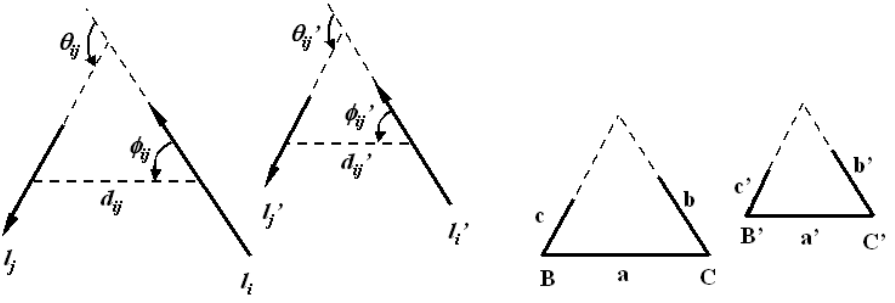
Proof. Let us consider two pairs of line segments that have same *interrelation quadruplet.* Here, we can extract triangular shapes as shown in figure 1. From the hypothesis of the Theorem 2,

$$\theta_{ij} = \theta_{ij}', \quad \phi_{ij} = \phi_{ij}', \quad \frac{l_j}{l_i} = \frac{l_j'}{l_i'}, \quad \frac{d_{ij}}{l_i} = \frac{d_{ij}'}{l_i'}. \tag{9}$$

Thus

$$\angle B = \angle B', \quad \angle C = \angle C', \quad \frac{a'}{a} = \frac{b'}{b} = \frac{c'}{c} \tag{10}$$

can be formed, and it shows that two pairs of segments which have a same *interrelation quadruplet* are alike and a line segment pair can be represented as translated, scaled and rotated version of the other line segment pair.



(a) Two pairs of line segments (b) Two triangular shapes

Fig. 1. Two triangular shapes

There are *m* x *(m-1)* number of *interrelation quadruplet* existing in an *m*-vertexed polygon. But many of the derived *interrelation quadruplets* are dependent on others. Therefore memory capacity and processing time may be reduced, if a polygon is formed with a few of *interrelation quadruplet* after taking a small number of *independent quadruplet* among *m* x *(m-1)* number of *interrelation quadruplet*. Following theorems are induced for this.

Theorem 3. For three non-intersecting line segments S_i, S_j and S_k, the *interrelation quadruplet* IQ_{ik} is uniquely determined as a combination of two *interrelation quadruplets* IQ_{ij} and IQ_{jk} *– Chain relation.*

The proof of Theorem 3 could be appeared in the full version of this paper. If there are three non-intersecting line segments, six *interrelation quadruplets* can be extracted. But, only two of them are independent and others can be calculated by using the Theorem 3. Thus the following theorem can be satisfied for arbitrary number of line segments after expanding the chain relation.

Theorem 4. For *m (m>1)* non-intersecting line segments, there are only *m-1* independent *interrelation quadruplets*.

The proof of Theorem 4 could be appeared in the full version of this paper. In a special case of Theorem 4, the following corollary is satisfied if applied to a simple *m*-vertexed polygon.

Corollary 1. For a simple *m*-vertexed polygon, there are only *m-1* independent *interrelation quadruplets*.

It is possible to describe a simple m-vertex polygon with one base line segment and *m-1* independent *interrelation quadruplet*. And if two polygons are alike, *m-1* independent *interrelation quadruplets* that have same values can be abstracted.

3 Application to Pattern Recognition

As described in previous section, *interrelation quadruplet* IQ_{ij} is a value which represents a geometrical relation between two line segments. This value is irrelevant to translation, rotation, or scaling of a pair of line segments. By using the properties of *interrelation quadruplet*, a very simple pattern matching method can be successfully used for recognizing partially occluded objects.

Many researches focus on recognition methods of a polygonal shape that is approximated from a boundary curve of an object. These can reduce a process time and make an algorithm simple. This research also approximates a boundary shape to a polygon, then *interrelation quadruplet* IQ_{ij} between line segments which form edges of the polygon, use as feature information. We used the polygonal approximation algorithm proposed in [10].

A method to extract *interrelation quadruplet* IQ_{ij} that uses as a feature is as follows.

Step 1) Extract vertex points from boundary curve of an object by using a polygon approximation method in [10].

Step 2) Extract the order sequence of line segments S_i, $i = 1, ..., m$.

Step 3) Extract *m interrelation quadruplet* IQ_{ij} between the current segment S_i and their respective successor S_{i+1}.

From the three-step function, we can get *m* line segment and it is used as auxiliary feature information. The *m interrelation quadruplet* is used as main feature information. Before describing pattern matching method, let us consider the meaning of two terminologies.

Definition 3. A pair of model and scene *interrelation quadruplets* ($IQ_{ij}^{\ m}$, $IQ_{kl}^{\ s}$) is said to be *compatible pair* when $\sqrt{(IQ_{ij}^{\ m} - IQ_{kl}^{\ s})^2} \le d_{th}$. Where d_{th} denotes the predetermined threshold level.

Definition 4. A pair of model and scene *interrelation quadruplets* ($IQ_{ij}^{\ m}$, $IQ_{kl}^{\ s}$) is said to be *matched pair* when each *interrelation quadruplets* are extracted from the corresponding line segments between same objects.

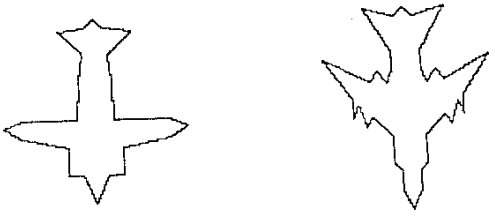
Many kinds of feature searching and pattern matching algorithms can be possible by using an *interrelation quadruplet* as feature information. A very simple pattern matching scheme described below can be a solution for partial shape recognition.

Step 1) Find several *compatible pairs* between the model and the scene until a matched pair is found.

Step 2) Using the *matched pair* obtained in step 1), fine all *matched pairs* between the model and scene.

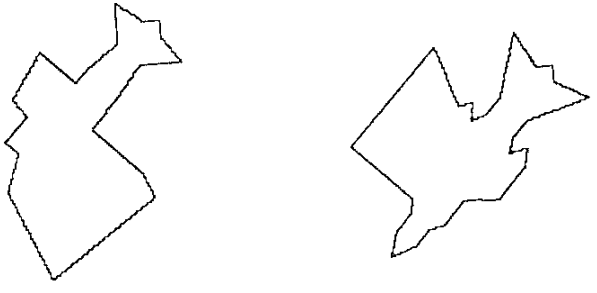
4 Experiments

For the partial shape recognition experiments, libraries of aircraft images are used. To create partial shapes, the unknown contours are chopped, with the chopped portions being replaced by a straight line, an arbitrary angle turn, and another straight line. These are much like the contours used by Liu [10]. Some sample contours are shown in Figure 2. Figure 3 shows contours obtained from chopping at a different orientation and scaling.



Aircraft-A Aircraft-B

Fig. 2. Sample patterns



Unknown shape-A Unknown shape-B

Fig. 3. Unknown object Samples

Table 1 shows sample *interrelation quadruplets* generated from the aircraft-A. The matched segments and its distance between the aircraft-A and unknown shape-A are shown in Table 2. Figure 4 shows several superimposed matching results.

Table 1. Feature list example

segment	θ_{ij}	ϕ_{ij}	l_{ij}	ld_{ij}
0	0.653	0.449	2.010	1.498
1	-2.224	-1.306	1.215	0.500
2	0.918	0.411	0.823	0.818
3	-0.083	-0.047	1.323	1.161
4	1.521	0.948	1.498	0.921
5	-0.381	-0.100	0.359	0.670
6	-2.221	-1.190	1.083	0.464
7	-0.343	-0.236	2.186	1.573
8	1.373	0.477	0.588	0.628
9	-1.571	-0.490	0.533	0.567

Table 2. Matched list example

Model segment id	Scene segment id	Distance
1	19	0.036
2	0	0.034
9	6	0.053
10	7	0.027
17	11	0.017
18	12	0.031
19	13	0.030
20	14	0.028

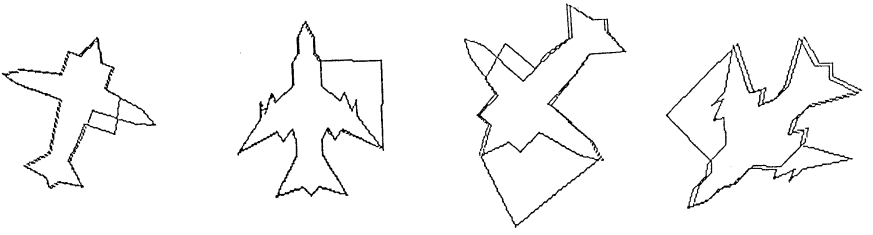


Fig. 4. Superimposed matching result

5 Conclusion

This paper presents a novel shape descriptor for identifying the similar objects in an image database. We developed several useful properties of *interrelation quadruplet* that fulfills the necessary and sufficient requirements for shape recognition. Local features were obtained from *interrelation quadruplet* of contour segments, and simple matching technique was worked successfully. This technique has been shown to recognize unknown shapes which have been rotated, translated and scaled, and which may be occluded or may overlap other objects.

As a future work, we can consider the more complex scenes with noisy environments. And future work is directed towards extending these conclusions using the *interrelation quadruplet* as a same feature information and more powerful pattern matching scheme, and to verify the validity of *interrelation quadruplet* as a shape descriptor. Another challenge is to adopt this descriptor to be used in recognizing object families whose subparts are allowed move with respect one another.

References

1. H. Kauppinen, T. Seppanen and M. Pietikainen: An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. in IEEE Trans. PAMI, vol. 17, Feb. 1995, pp. 201-207.
2. Klaus Arbter, Wesley E. Snyder, Hans Burkhardt and Gerd Hirzinger: Application of Affine-Invariant Fourier Descriptor to Recognition of 3-D Objects. in IEEE Trans. PAMI, vol. 12, no. 7, July 1990, pp. 640-647.
3. Ming-Fang Wu and Hsin-Teng Sheu: Representation of 3D Surfaces by Two-Variable Fourier Descriptors. in IEEE Trans. PAMI, vol. 20, Issue 8, Aug. 1998, pp. 858-863.
4. Xiaolong Dai and Siamak Khorram: A Feature-Based Image Registration Algorithm Using Improved Chain-Code Representation Combined with Invariant Moments. in IEEE Trans. On Geoscience and Remote Sensing, Vol. 37, no. 5, September 1999, pp.2351-2362
5. P. Nassery and K. Faez: Signature Pattern Recognition Using Pseudo Zernike Moments and a Fuzzy Logic Classifier. in Proc. 1996 International Conference on Image Processing, Vol. 1, pp.197-200, September 1996.
6. Juanning Xu: A Generalized Discrete Morphological Skeleton Transform With Multiple Structuring Elements for the Extraction of Structural Shape Components. in IEEE Trans. On Image Processing, Vol. 12, no. 12, December 2003, pp.1677-1686
7. Renato Kresch and David Malah: Skeleton-Based Morphological Coding of Binary Images. in IEEE Trans. On Image Processing, Vol. 7, no. 10, October 1998, pp.1387-1399
8. Latecki, L. J., Lakamper, R.: Shape Similarity Measure Based on Correspondence of Visual Parts. in IEEE Trans. PAMI, vol. 22, no. 10, October 2000, pp. 1185-1190.
9. Arkin, M., Chew L. P., Huttenlocher D. P., Kedem K., and Mitchell J. S. B.: An Efficiently Computable Metric for Comparing Polygonal Shapes. in IEEE Trans. PAMI, vol. 13, 1991, pp. 209-206
10. Hong-Chih Liu and Mandyam D. Srinath: Partial Shape Classification Using Contour Matching in Distance Transformation. in IEEE Trans. PAMI, vol. 12, no. 11, November 1990, pp. 1072-1079.

An Efficient Representation of Hand Sketch Graphic Messages Using Recursive Bezier Curve Approximation

Jaehwa Park and Young-Bin Kwon

Dept. of Computer Science and Engineering, Chung-Ang University,
221 HukSuk-Dong, DongJak-Gu,
Seoul 156-756, Korea
{jaehwa,ybkwon}@cau.ac.kr

Abstract. A practical solution to represent simple hand drawing graphic messages is presented. A freehand-sketch message captured by a digitizing tablet is approximated using the quadratic Bezier curve representation and generated curve control points are adjusted to reduce the dynamic range of the first order difference. The control point data is compressed into a bit-stream to perform efficient graphic representation for use in low bandwidth transmission and data storage applications. A recursive architecture performing a piecewise curve approximation is proposed to maximize the data compression rate. The experimental results show good curve fitting ability and high data compression rate of the proposed method, which can be applicable for practical real-time applications.

1 Introduction

Freehand sketching graphic messages are considered as a natural way to visualize ideas or messages that cannot be efficiently represented by speech or text. These graphic messages are usually composed of several hand-drawn objects expressed in a group of pen strokes such as polygons, lines, arcs and handwritten characters. For an on-line system, the pen movements are typically captured by a digitizing tablet and stored as sampled pen points of their paths, so called as digital ink, while an image for an off-line system is captured by a camera or a scanner and represented as a two dimensional array of pixels.

Recently, small mobile devices are becoming very popular and pen-based user interface is considered as a primary input method that can replace traditional input devices such as keyboards and pointing devices. Moreover short messaging service between mobile devices through digital wireless networks is widespread and the feature to transfer graphic messages between them is highly desirable to overcome the inconvenience of text-only messaging.

However, processing of graphic message data usually requires considerable storage capacity, computing power and transmission bandwidth. Despite recent rapid progress in these areas, the demands still exceed the capabilities of available technologies.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 392–399, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The rapid growth of graphic data in mobile devices has sustained the need for more efficient ways to represent and to compress the graphic data.

The need for efficient representation of hand-drawing object is highly desirable to overcome the resource limitation of mobile devices. The data representing freehand sketching needs not only to be compressed in order to reduce the internal handling size and to transfer in low bandwidth but also to preserve the convenience of easy access of the information for further processing such as handwriting or shape recognition.

In this paper, a method performing efficient graphic representation of on-line hand drawing objects captured in digital ink is introduced. The target is to represent freehand sketching graphic objects in a compact format, achieving high compression rate within a certain error tolerance (lossy compression), while the proposed method is practically adaptable solution for the real mobile devices of today. A piecewise quadratic Bezier curve approximation method is proposed to reduce the burden of computation. And an optimization idea to achieve higher compression rate is represented.

2 Bezier Curve Approximation

A Bezier curve is defined using two anchor points, on-curve control points and at least one shape point, off-curve control point. The on-curve control points are the two end points of the curve actually located on the path of the curve, while the other off-curve control points define the gradient from the two end points, which are usually not located on the curve path. The off points control the shape of the curve. The curve is actually a blend of the off-curve control points [1].

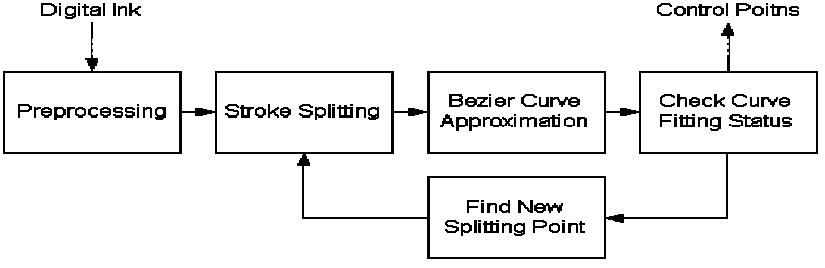


Fig. 1. Block diagram of the recursive approximation process

The more off-curve control points a Bezier curve has, the more complicated shape can be represented; however the order of the mathematical curve equation becomes higher. The number of off-curve points determines the order of the Bezier Curve. The approximation of hand-drawing strokes using high order Bezier curve reduces the number of on-curve points and produces more compact data representation. However,

the approximation of high order curve equation usually requires large amount of computation since we don't have a clear solution but has trail-error approaches.

For the computational efficiency, the quadratic Bezier curve representation is only used in our method since it is relatively simple and the curve coefficients can be easily obtained by the least square error estimation. But the quadratic Bezier curves can only represent simple arc shape curves. Thus, complicate shaped strokes are represented by piecewise approximated quadratic Bezier curves in our approach. The disadvantage generating excessive on-curve points because of using series of low order Bezier curves (compared to approximated by high order Bezier curves) can be somewhat overcome by optimization of control points described in next section.

Fig 1 shows the block diagram of the proposed approximation method. The approximation process accepts the set of strokes represented in digital ink and produces a set of curve control points. The curve control points are produced by piecewise fitting of a series of quadratic Bezier curves using least square error approximation. This process has two independent sub-processing modules, digital ink preprocessing and curve approximation loop. The preprocessing is performed only once per given input set of stroke digital inks, but the curve approximation routine operates recursively till the piecewise fitting result is satisfactory.

2.1 Digital Ink Preprocessing

The purpose of preprocessing is to extract sharp turning pen movement points so-called bending points; for example the sharp peak point of characters M or W. These bending points are generally difficult to handle for low order Bezier curve fitting. To minimize the burden of computation, the bending points are obtained based on curvature (tangent) estimation on each ink point before the curve fitting process. And the strokes that contain bending points are split into a set of smooth sub-strokes.

To minimize the effect of jitter noise erroneously being detected as bending points, a sliding window technique is used. The curvature on each point is estimated using the average of all curvatures obtained in the permissible window. It can minimize the chance of over-splitting caused by jitter of pen movements. The window size is given as a function of the perimeter and bounding box of each stroke.

An inward vector is defined as a transition from previous to current ink point and an outward vector is defined as a transition from the current to next ink point. The curvature angle is defined as the minimum angle between the inward and outward vectors. The difference between the inward and outward angles is assumed to be the curvature angle of the point. The final curvature angle is estimated by averaging of all the obtained curvature values between the corresponding pairs split by current point within the window. If the curvature angle is higher then a threshold, the curvature at the point is assume to be high.

A bending zone is established by a group of series points that have high curvature. Then the highest curvature point within the bending zone is assumed to be the bending point. The ink stroke between any two adjacent bending points is separated as a sub-stroke after all the bending points are detected. The curvature thresholds are de-

termined in several values by the application to minimize the overhead of subsequent recursive curve approximation. If it is too low, the burden of recursive operation become large, otherwise it results in degradation of compression rate because of unnecessary over-splitting.

2.2 Curve Fitting

For an ink stroke (or a sub-stroke), the minimum required curve-fitting condition (such as minimum number of ink points, size of bounding box, etc) is checked. If the ink stroke is determined to be eligible for the further processing, a quadratic Bezier curve approximation method is applied to find out the curve control points.

For a given ink sequence, a quadratic Bezier curve representation coefficients (usually coordinate of control points) are estimated which satisfy following conditions. i) The starting point of the estimated curve should be the same point of the first ink point of the stroke ink sequence, ii) The ending point of the estimated curve should be the same point of the last ink point of the stroke ink sequence, and iii) That has the Least Square Euclidean distance error between the actual ink points. The meditative parameter of Bezier curve representation (usually bounded in [0,1]) is estimated based on the proportional distance of each ink point on the perimeter from staring to end ink points [4].

Using the estimated curve control points, the fitting error between actual ink points and corresponding points of the approximated curve is calculated based on Euclidean distance measurement. If the fitting error is within acceptable range, the approximation is completed. Otherwise, new splitting points are determined using a relaxed curvature threshold. The recursive curve fitting operation is applied to each newly split strokes till the error falls within tolerance or the size of split piece become smaller than the minimum size. Curve approximation function is controlled by two parameters: the minimum ink size and error tolerance. Both parameters control the accuracy of approximation and efficiency of compression.

3 Optimization of Control Points

The key ideas of the optimization are i) recursive optimization of piecewise curve approximation to minimize the dynamic range of data, and ii) data compression stored in bit streams using first order difference of the values, so called as *delta compression* which represents coordinate of curve control points in order to reduce the redundancy. The overall procedure is shown in Fig 2. Adjusting, insertion, and deletion of the initial approximated quadratic Bezier curve control points perform the concept of recursive optimization of Bezier curve approximation.

3.1 Regulate Delta Size

After the initial piecewise Bezier curve approximation, all the estimated control points are examined and adjusted to improve the overall data compression rate. The optimization idea is to reduce the dynamic range of the first order difference of the control points coordinate values by insertion and deletion of control points. It is advantageous to reduce of the overall compressed bit stream size, even if the number of control points is increased than that of initial approximation.

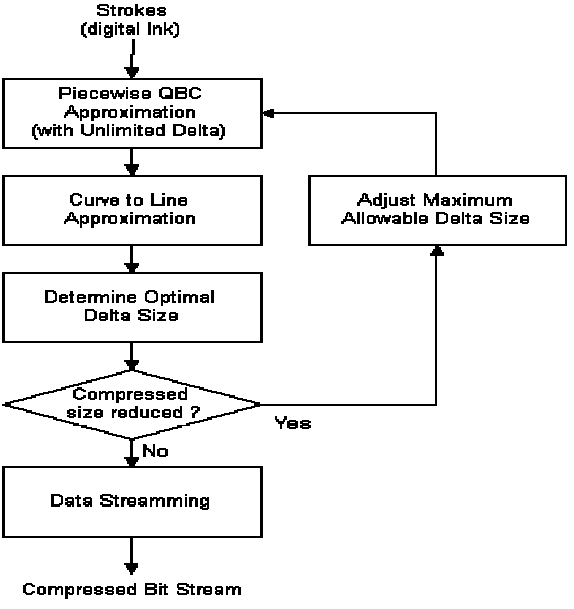


Fig. 2. Block Diagram of Control Point Optimization and Data Compression

Since the final representation of control points is encoded by difference from the current point to the previous control point, the minimum required bit-size to hold the each encoded data for bit streaming is determined by the maximum possible value of the difference. If the dynamic range of the data set is large, i.e. difference between the minimum to the maximum is large, the compression rate become poor since the waste of bits in data of smaller values is increased. For example, if an arc piece that generates a significantly large gap is split into multiple pieces (or insertion of pseudo off-line curve control points), it can reduce the dynamic range of the gaps while it pays the cost of additional storage size due to inserted control points.

If parameters are given as follows,

D	dimension of control points, usually two: (x, y)
V_{max}	the maximum value of difference of same coordinate values between two consecutive control points

b	size of bit field, $[1 \sim b_{max}]$,
b_{max}	smallest integer value larger than $log_2 V_{max}$
N_{total}	total number of control points except the 1st control point.
N_b	number of points in which the minimum bit size of the difference is b
S_b	compressed bit stream size of the difference when b is selected as the unit bit size for the axis (dimension).

S_b is given by

$$S_b = b(N_{total} + \sum_{n=b+1}^{b_{max}} N_n \, D \, (2^{(n-b)} - 1)).$$

The goal is to find the value of b to minimize the S_b for each dimension (x, y and t if applicable). The summation part means overhead caused by the insertion of a control (or bridging) point. Since the summation operation depends on b, the derivation equation becomes relatively complicate. Thus in our approach, an exhaustive search method with pre-determined bit range is used. The possible bit size is practically small and limited by the digitizing tablet device.

3.2 Curve to Line Approximation

Some of curve components can be approximated as a "line" component instead of curve representation within the allowable error range, if the curvature is not large. Since a line component only requires two on-line ending points, the line approximation can reduce the storage space of the off-line curve control point(s). However, the removal of off-line curve control points does not always give advantages since it can increase the dynamic range of the first order difference (gap between two consecutive points) of the data.

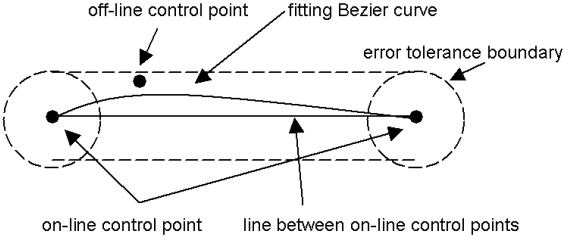


Fig. 3. Curve to line approximation

Since the line approximation from raw digital ink data is a computationally intensive process, the Bezier curve fitting parameters are used directly to examine the possibility of a line approximation. The off-line control point of a quadratic Bezier curve always exists outside of fitting curve, i.e. not within the area of fitting curve and straight line between two on-curve points. So, if the off-line control point is located within the error tolerance boundary as shown in the Fig 3, we can assume that the fitting curve trajectory exists within the error boundary. In this case we can approxi-

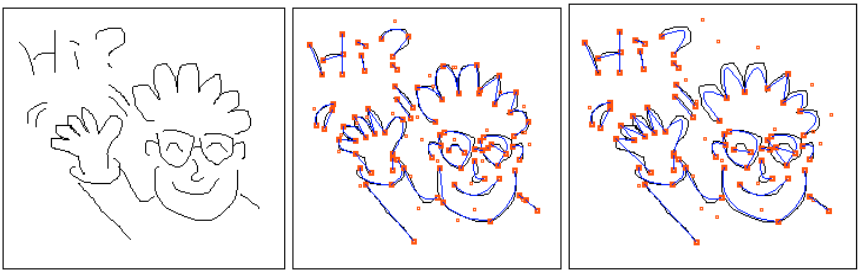
mate the Bezier curve into a straight line between the two on-line control points within the allowable fitting error tolerance.

If two conditions are satisfied, i) the Euclidean distances between the on and off control points are smaller than allowable delta tolerance (means maximum allowable value for the first order difference), and ii) the minimum Euclidean distance between the off-line point and the straight line between the two on-line points is smaller than the error tolerance. Then, the off-line control point is assumed to be within the error tolerance. If the test is successful, the Bezier Curve representation is converted into a line component representation by discarding the off-line control point.

4 Experiments

The proposed method has been tested in two different data sets: object drawings and handwriting message in Chinese characters (such as greeting message etc). The data set collected from various peoples for short messaging system that can be rendered through wireless network.

Fig 4 shows one of the examples. It shows the approximation using the optimization method to reduce the delta size. The proposed algorithm generates the essential Bezier curve control points as shown in Fig 4, allowing the maximum error of four and eight pixel distance (Fig 4-(b) and (c)). It also shows the synthesis curves regenerated by the obtained control points (for comparison, the synthesis curves are displayed with the original drawings). It achieves more than 91% data compression rate before actual bit streaming without significant graphical distortion.



(a) (b) (c)

Fig. 4. An approximation example (a) original drawing, (b) with 4 pixel error and (c) with 8 pixel error tolerance (big dot: on-curve small dot, off-curve points)

Table 1 shows experimental result of overall data compression performance. The fitting error tolerance is given as four-pixel distance as the shown example. The numbers in the table show the ratio of number of control points when the number of control points for the original (raw) ink data is assumed to be 100. In case of "compressed" in the Table 1, which means all the control points are represented by bit stream data, the numbers are calculated by means of actual compacted data size to be

stored. One ink (or control) point is assumed to be 4 byte: 2 bytes for each dimension, x and y (t is not considered here).

The proposed method achieves overall 90% data compression rate. Since hand-written Chinese characters have more short strokes than hand-sketched figures, it requires more on-curve (anchor) points. Thus the approximated size is slightly larger than that of the hand-sketch case.

Table 1. Compression Rate using proposed method with error tolerance: maximum four-pixel distance.

Data Set	Number of Drawings	Original Digital Ink	Not Optimized	Optimized	Com-pressed
Hand Sketch message	45	100	11.52	9.46	7.36
Handwritten Chinese Message	120	100	19.50	17.51	13.94
Overall	165	100	17.32	15.80	10.14

5 Conclusion

In this paper, we present a practical solution for efficient representation of simple hand drawing graphic messages. The goals of the proposed method are to reduce the redundancy of on-line hand drawing graphic data and to achieve high compression rate when the data is packed for message exchange that can be transmitted in low-bandwidth wireless networks. A piecewise quadratic Bezier curve approximation and suppressed delta compression are implemented in recursive architecture. The experimental results show that the computational-efficient implementation using all integerized program code is to be a practical solution. Considering rapid progress in digital wireless network and developing mobile devices, efficient high order curve approximation and optimization idea is highly desired in the near future.

Acknowledgement. This Work is supported by ITRI, Chung-Ang University.

References

1. Farin G.: Curves and Surfaces for CAGD, Fifth Edition. Academic Press, London (2002) 57-74
2. Sederberg T. W., Farouki R. T.: Approximation by interval bezier curves: IEEE Computer Graphics and Applications, Vol. 12(5) . IEEE, September 1992, 87-95
3. Hagglund R., Lowenborg P., Vesterbacka M.: A Polynomial-Based Division Algorithm: Proceedings on IEEE International Symposium 2002
4. Ohno K., Ohno Y.: A Curve Fitting Algorithm for Character Fonts: Electronic Publishing Vol 6(3) September 1993, 195-205
5. Hussain F., Pitteway M.L.V.: Rasterizing the outlines of fonts: Electronic Publishing Vol 6(3) September 1993, 171-181

Contour Description Through Set Operations on Dynamic Reference Shapes

Miroslav Koprnicky[1], Maher Ahmed[2], and Mohamed Kamel[1]

[1]Pattern Analysis and Machine Intelligence Laboratory, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1
mkoprnic@pami.uwaterloo.ca, mkamel@pami.uwaterloo.ca
[2]Department of Physics and Computing, Wilfrid Laurier University, Waterloo, Ontario, Canada, N2L 3C5, mahmed@wlu.ca

Abstract. Eight novel features for irregular shape classification which use simple set operations to compare contours to regularized reference shapes are introduced. The features' intuitive simplicity and computational efficiency make them attractive choices for real time shape analysis problems such as defect inspection. Performance is evaluated through a brute force feature combination search, in which KNN classification rates of the proposed features are compared to several existing features also based on contour comparison. Results indicate that combinations of the proposed features consistently improve classification rates when used to supplement the feature set. Efficacy of the individual features ranges greatly, but results are promising, especially for Outer Elliptical Variation; its strong performance, in particular, calls for further investigation.

1 Introduction

There has been a great deal of effort put into the area of shape analysis. This is largely due to the fact that it is such a key component of computer vision, which has applications in a large number of disparate fields such as engineering, manufacturing, automation, and health science [1]. The majority of the work has been focused into recognizing regularized shapes belonging to distinct categories such as "Triangle" vs. "Circle", and "ScrewDriver" vs. "Hammer" [2]. Somewhat more neglected has been the work on irregular shapes, which can take on an infinite spectrum of stochastic contours, and are often difficult to differentiate, by humans as well as by algorithms.

An interesting set of simple features was proposed for this task by Peura and Iivarinen in [3], in which it was demonstrated that although no single feature in the group was descriptive enough to distinguish between the irregular shapes presented it, they contained sufficient information in tandem to separate the shapes into visually distinct groups. The features selected were made all the more attractive by the fact that they were relatively simple to compute, and thus could be readily incorporated into a real-time shape analysis problem, such as defect classification [4]. The five features investigated were: Convexity, Principal Axis Ratio, Compactness, Circular Variation and Elliptical Variation.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 400–407, 2004.
© Springer-Verlag Berlin Heidelberg 2004

One drawback to the method was that classification was performed by Self-Organizing Maps [5], initially trained by clustering the shapes into the most readily separable groups, which were subsequently labeled by an operator according to the characteristics of the shapes residing within the clusters. This a posteri classification might translate into deceptively effective classification results when the system is subsequently presented with a test dataset.

This work introduces eight novel shape features based on simple set operations on dynamically created reference shapes. They are compared with the aforementioned features, by their performance in a standard KNN classifier system. The irregular shape dataset employed was pre-classified by qualitative inspection prior to exposure to the classifiers, in an attempt at a more unbiased performance analysis. It was found that by supplementing the original features described in [3] with the proposed features, classification performance was consistently improved.

2 Method

The proposed features were mainly inspired by two previously examined features, namely circular and elliptical variance. For clarity, these features as well as the others examined in [3,4] will from herein be referred to as the Classic Features.

2.1 Classic Features

Circular Variance measures how a shape's contour deviates on a point by point basis from a circle with equal area.

$$C\,\mathrm{var} = \frac{1}{N\mu_r{}^2}\sum_i \left(\left\|\vec{p}_i - \vec{\mu}\right\| - \mu_r\right)^2 \tag{1}$$

p_i are points on the shape's contour, N is the number of points in the contour, μ is the shape's centroid, and μ_r is the shape's mean radius. Likewise, Elliptical Variance defines a contour's deviation from an ellipse with equal covariance matrix. It is defined as:

$$E\,\mathrm{var} = \frac{1}{N\mu_{rc}{}^2}\sum_i \left(\sqrt{\left(\vec{p}_i - \vec{\mu}\right)^T C^{-1}\left(\vec{p}_i - \vec{\mu}\right)} - \mu_{rc}\right)^2 \tag{2}$$

The term μ_{rc} is calculated as:

$$\mu_{rc} = \frac{1}{N}\sum_i \sqrt{\left(\vec{p}_i - \vec{\mu}\right)^T C^{-1}\left(\vec{p}_i - \vec{\mu}\right)} \tag{3}$$

Although these shape measures appear computationally intensive, they are in fact O(N) operations, since they only require one pass of a shape's contour for calculation.

2.2 Reference Shapes

The features presented here were initially formulated as potentially less computationally demanding alternatives to Classic Features 1 and 2. Instead of calculating point by point Euclidian distances, the proposed features require only simple area subtraction operations to obtain contour variance estimates.

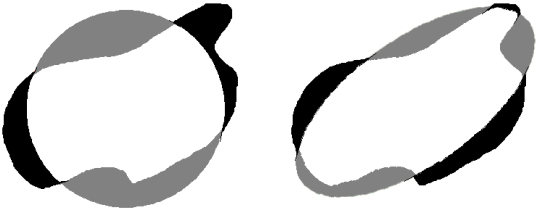


Fig. 1. Irregular shape superimposed over its reference circle and ellipse

Reference sets R for any shape can be easily computed as the set of all pixels r_{xy} which belong to either the reference circle with equal area A, and centroid (p,q):

$$\left\{ r_{xy} \in R_C \mid (x-p)^2 + (y-q)^2 \leq \frac{A}{\pi} \right\} \tag{4}$$

or the reference ellipse with equal contour covariance matrix - that is an ellipse with x axis radius a, and y axis radius b, as well as an angle of inclination θ.

$$\left\{ r_{xy} \in R_E \mid x = x'\cos\theta, y = y'\sin\theta, \frac{(x'-p)^2}{a^2} + \frac{(y'-q)^2}{b^2} \leq 1 \right\} \tag{5}$$

Although these relations, used to generate the reference shapes employed by the proposed features, also appear computationally intensive, it should be stressed that they too, require only one pass over a region of interest - the bounding box surrounding the shape. Also, it should be said that computer graphics is an area that has been thoroughly researched in the past, and there exist many advanced solutions to plotting circles and ellipses, many of which are implemented as very fast hardware solutions. These considerations are why the reference shape calculations are negligible in the consideration of the proposed features' computational complexity.

2.3 Proposed Features

The following expressions describe two features each, one which compares shapes to reference circles, and one which compares shapes to reference ellipses, both represented by the set R of all reference pixels r_{xy}. S is the set of all pixels s_{xy} belonging to the shape. The function *area* denotes a simple pixel counting operation.

Outer Elliptical and Circular Differences are defined as the area of the shape residing outside the contour of the reference ellipse and circle.

$$\Delta_O = \frac{area(S \cap \overline{R})}{area(S)} \tag{6}$$

Inner Elliptical and Circular Differences represent the area of the reference shape not enveloped by the irregular shape.

$$\Delta_I = \frac{area(\overline{S} \cap R)}{area(S)} \tag{7}$$

Relative Elliptical and Circular Differences are defined as the difference between the above features.

$$\Delta_R = \Delta_O - \Delta_I = \frac{area(S \cap \overline{R}) - area(\overline{S} \cap R)}{area(S)} \tag{8}$$

Absolute elliptical and circular difference, are the sums of the first two features. They can be more easily calculated as the absolute difference or exclusive or operation, between the shape and the reference ellipse or circle.

$$\Delta_A = \frac{area(S \oplus R)}{area(S)} = \frac{area|S - R|}{area(S)} \tag{9}$$

All features are normalized over the area of the shape, in order to negate the effect of possibly large differences in area between shapes. It can be shown that all the feature computations are order O(n), because they only require a single pass of the shape's bounding rectangle. In fact, intersection, negation, exclusive or, and subtraction operations can be implemented as concurrent matrix operations, and could easily be performed by array processors, and pixel count operations can be performed in hardware as well.

Fig. 1 illustrates the nature of the proposed features. Outer Differences (6) correspond to the black areas which lie outside of the reference shape, while the Inner Differences (7) are the areas in gray. The Relative Differences (8) show how much more the shape deviates externally as opposed to internally, which can be thought of as subtracting the gray from the black, while the Absolute Differences (9) correspond most to the original Circular and Elliptical Variances (1&2) examined in [3,4], and are seen to be the total shaded area, black + gray.

3 Results

It can be seen from their descriptions (6,7,8,9) that the proposed features are quite mathematically interdependent. It is therefore natural to assume that there exists a large correlation between them. By using a brute-force search all possible permutations of the features can be examined, so that the most effective features can

be determined, without incurring the penalties for correlation that some feature evaluation algorithms employ to shorten search times.

KNN classifier accuracies were noted to ascertain the efficacy of all possible combinations of both the Classic and Proposed shape features. K parameters were set to 5 after it was found experimentally that this was an effective value for classification. Features were extracted from a synthesized dataset of 200 irregular shapes pre-classified qualitatively into 5 classes. 10 trials were performed, with performances averaged over the trials. Random training sets of 140 shapes were created for each trial; testing sets consisted of the remaining 60 shapes.

3.1 Classification Performance

Performance of the feature set as a whole was quite encouraging, with the vast majority of the 8191 total possible permutations of the 13 features resulting in classification error rates of less than 15%, which improves approximately linearly to the optimal feature combination's 8.62% (Fig.2).

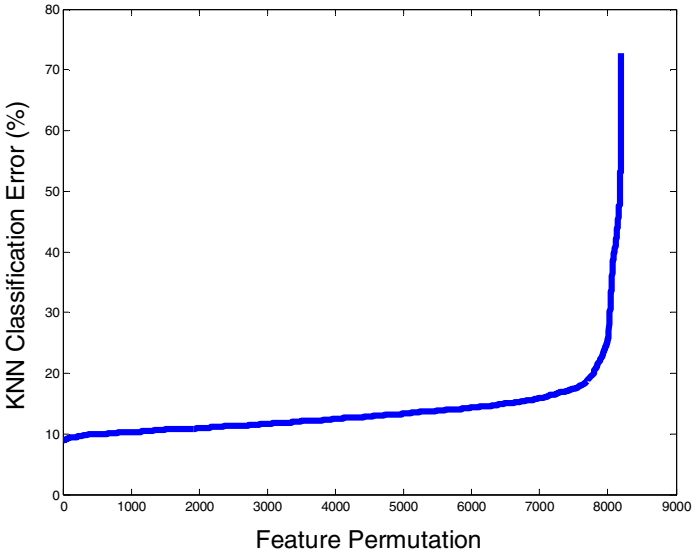


Fig. 2. Feature performance as compared by KNN error rates. All possible feature permutations are displayed, ordered by descending accuracy

This result is positive, in that even if an optimal feature set cannot be unambiguously identified, we can be confident that the features proposed are effective enough to still be used in the majority of the possible permutations with good classification rates as a result.

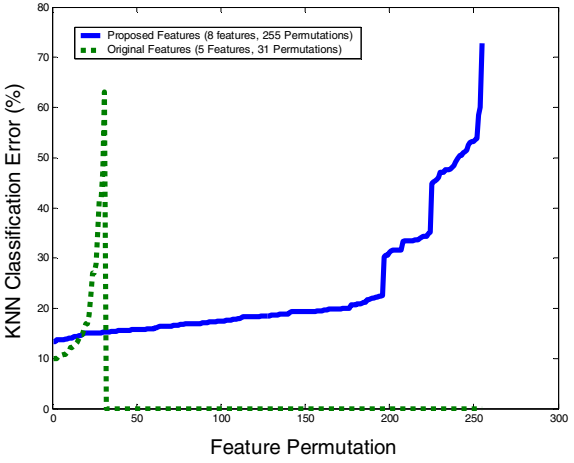


Fig. 3. Classic Feature performance (broken line) as compared to Proposed Features (solid line) in isolation. Feature permutations are ordered by descending accuracy

The proposed features when used in isolation do not perform as well as the classic features (Fig. 3), with the best possible combination of proposed features yielding an error rate of 13.4% as compared to the original features' 9.9%. The best performance, however, is achieved when supplementing the original features with the information of the proposed features. As is clearly illustrated in Fig. 4, classification accuracy is improved for every possible permutation of the classic features, when combined with a permutation of the proposed features, with a peak performance of 8.62% having been reached.

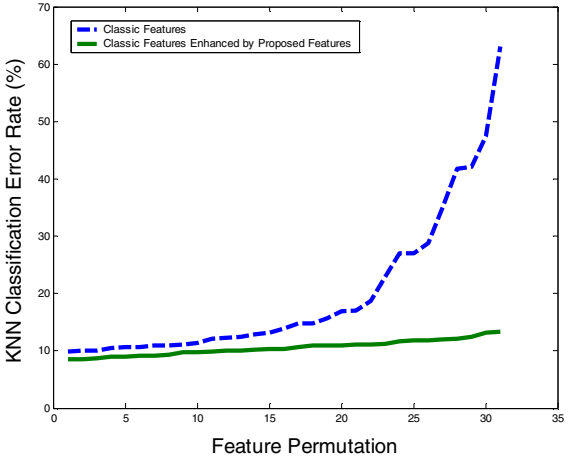


Fig. 4. Classic feature performance (broken line) vs. a mixture of Classic and Proposed features (solid line)

3.2 Individual Feature Efficacy

An estimate of the individual features' importance to classification accuracy can be found through feature utilization percentages. That is, the top 40 feature permutations from Fig. 2 were examined and analyzed to see which features were most prevalent in the highest performing feature spaces.

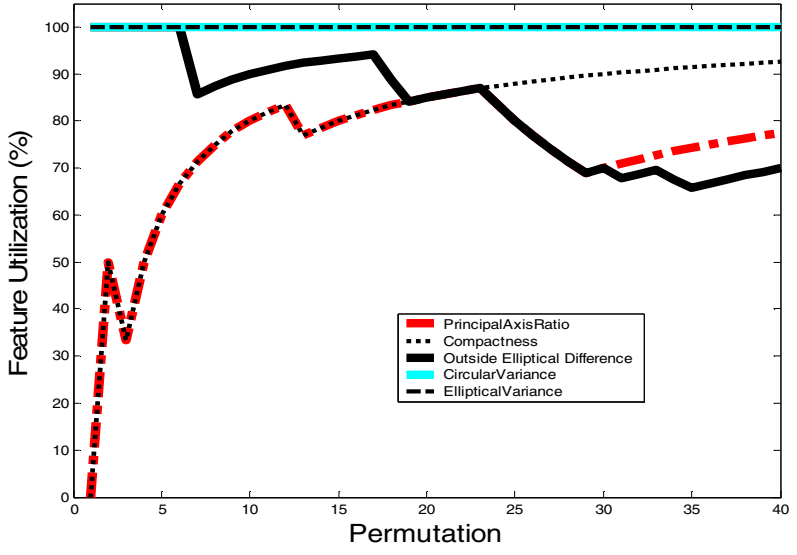


Fig. 5. Individual feature utilization percentages. Shows five best features, as gauged by the number of times they were used in the 40 most successful permutations (from Fig. 2)

Table 1. Computation time comparison (in ms)

Circ. Var.	Ellip. Var.	Absolute Diff.	Relative Diff.	Outer Diff.	Inner Diff.
34.39	52.30	16.52	17.53	8.27	9.21

The proposed feature *Outer Elliptical Difference* is used in more permutations than all other features except *Circular* and *Elliptical Variance* in the most successful feature spaces.

3.3 Computational Performance

The proposed feature groups' algorithmic complexities were compared to the two classical features that they resemble most closely, namely elliptical and circular variance, on an Intel-based P4 1.6 GHz PC running the MATLAB development environment. Times are averaged over 100 shapes. As is shown, all of the proposed features outperform the classical features by approximately a factor of 2; Outer and Inner Differences are actually almost four times as efficient as Circular Variance, while the relatively complex feature Elliptical Variance took the longest to calculate.

4 Conclusion

Eight computationally efficient shape descriptors based on reference shape comparison were presented and compared to various existing shape features for the task of irregular shape classification. All features are worth further investigation, as the vast majority of possible feature combinations yield good classification results (Fig. 2). Although the proposed features are not as effective as the original shape measures in isolation, it is clear that optimum performance is achieved through a mix of the two feature sets (Fig. 4). The proposed feature *Outer Elliptical Difference* proved most promising, as it was present in the majority of the most effective feature permutations (Fig. 5).

Further work is being performed to analyze the features' ability to discriminate between specific shape classes, as well as incorporating the features into selection algorithms in an industrial system for web defect inspection, to evaluate the features' performances with live industrial data.

Acknowledgements. This research is supported in part by research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by the province of Ontario through the Ontario Graduate Scholarship (OGS) program.

Experimental work used functions from the Matlab based pattern recognition toolbox *PRTools Version 3.0*, courtesy of Bob Duin [7].

References

1. B. G. Batchelor & P. F. Whelan "Intelligent Vision Systems for Industry", Springer Verlag, Springer Verlag, London & Berlin, 1997.
2. S. Locaric "A Survey of Shape Analysis Techniques" in Pattern Recognition, Vol. 31, Num. 8, 1998, pp 983-1001.
3. M. Peura, J. Iivarinen "Efficiency of simple shape descriptors" in Aspects of visual form processing, Arcelli C, Cordella LP, Sanniti di Baja G, World Scientific, Singapore, 1997 pp 443-451
4. J. Iivarinen, A. Visa "An Adaptive Texture and Shape Based Defect Classification" In Proc. International Conf. on Pattern Recognition, 117–123, 1998
5. T. Kohonen "Self-Organization and Associative Memory 3rd Ed", Springer Series in Information Sciences, Springer-Verlag, 1989
6. M. James "Pattern Recognition", BSP Professional Books, Oxford, 1987.
7. B. Duin "PRTools Version 3.0", Pattern Recognition Group, Delft University of Technolgy, P.O. Box 5046, 2600 GA Delft, The Netherlands, 2000. http://www.ph.tn.tudelft.nl/~bob/PRTOOLS.html

An Algorithm for Efficient and Exhaustive Template Matching

Luigi Di Stefano[1,2], Stefano Mattoccia[1,2], and Federico Tombari[1,2]

[1] Department of Electronics Computer Science and Systems (DEIS)
University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
{ldistefano, smattoccia}@deis.unibo.it
[2] Advanced Research Center on Electronic Systems 'Ercole De Castro' (ARCES)
University of Bologna, Via Toffano 2/2, 40135 Bologna, Italy
federicot@omniway.sm

Abstract. This paper proposes an algorithm for efficient and exhaustive template matching based on the Zero mean Normalized Cross Correlation (ZNCC) function. The algorithm consists in checking at each position a sufficient condition capable of rapidly skipping most of the expensive calculations involved in the evaluation of ZNCC scores at those points that cannot improve the best score *found so far*. The sufficient condition devised in this paper extends the concept of Bounded Partial Correlation (BPC) from Normalized Cross Correlation (NCC) to the more robust ZNCC function. Experimental results show that the proposed technique is effective in speeding up the standard procedure and that the behavior, in term of computational savings, follows that obtained by the BPC technique in the NCC case.

1 Introduction

Template matching consists in calculating at each position of the image under examination a function that measures the degree of similarity between a template and a portion of the image [1]. Normalized Cross-Correlation (NCC) and Zero mean Normalized Cross Correlation (ZNCC) are widely used similarity functions in template matching (e.g. [1,2,3,4]) as well as in motion analysis, stereo vision, industrial inspections and many other applications, since the normalization process embodied into the NCC and ZNCC allows for handling linear brightness variations. Furthermore, thanks to the subtraction of the mean intensity, the ZNCC function is even a more robust solution than the NCC since it can handle also uniform brightness variations. Since NCC and ZNCC are rather computationally expensive, several non exhaustive techniques aimed at reducing the computational cost have been proposed (e.g. [2,3,4]). Yet, *non-exhaustive* algorithms do not explore the entire search space and hence can be trapped into local maxima, thus yielding a non-optimal solution.

Conversely, in this paper we propose an algorithm that finds exactly the same optimal solution as a *brute force* ZNCC-based template matching process but at a significantly reduced computational cost. The proposed algorithm extends

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 408–415, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the concept of Bounded Partial Correlation (BPC), previously devised only for a template matching process based on the NCC [5,6], to the ZNCC function.

2 A Brief Review of the BPC Technique

Let I be the image under examination, of size $W \times H$ pixels, T the template, of size $M \times N$ pixels, and $I_c(x, y)$ the sub image of I at position (x, y) having the same size as the template (e.g. $I_c(x, y) = \{I(x+i, y+j) \mid i \in [1..M],\ j \in [1..N]\}$).

The Normalized Cross Correlation between the template T and the image I at position (x, y) is defined as:

$$NCC\,(x,y) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M} I\,(x+i, y+j) \cdot T\,(i,j)}{\sqrt{\sum_{j=1}^{N}\sum_{i=1}^{M} I^2\,(x+i, y+j)} \cdot \sqrt{\sum_{j=1}^{N}\sum_{i=1}^{M} T^2\,(i,j)}} \in [0,1] \qquad (1)$$

The numerator of (1) represents the *dot product* between $I_c(x, y)$ and T, while in the remainder the ℓ_2 norms of $I_c\,(x, y)$ and T at the denominator will be denoted as $||I_c\,(x, y)\,||$ and $||T||$.

The BPC technique [5,6] allows for speeding up an exhaustive template matching process based on the NCC by rapidly detecting unsatisfactory matching candidates. Such detection is achieved by evaluating a sufficient condition, obtained at a reduced computational cost, that relies on an upper bound, $\gamma(x, y, n)$, of the dot product term in (1):

$$\gamma\,(x,y,n) \geq \sum_{j=1}^{N}\sum_{i=1}^{M} I\,(x+i, y+j) \cdot T\,(i,j) \qquad (2)$$

Let η_{max} represent the correlation maximum *'found so far'* during the matching process and that the point at coordinates (x, y) is currently under examination. The following inequality,

$$\frac{\gamma\,(x,y,n)}{||I_c\,(x,y)\,|| \cdot ||T||} < \eta_{max} \qquad (3)$$

when holds, provides a sufficient condition for skipping the current position without carrying out the entire calculation of the computationally expensive dot product term since (3) guarantees that the current position cannot improve the η_{max} score.

An effective sufficient condition is obtained splitting both T and I_c into two parts, respectively denoted by rows [0..n] and [n+1..N] as showed in Figure 1,

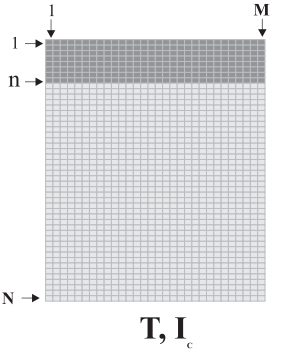


Fig. 1. Splitting of the template T and sub-image I_c.

and, correspondingly, the dot product term into two partial terms:

$$\begin{aligned}&\sum_{j=1}^{N}\sum_{i=1}^{M} I\left(x+i,y+j\right)\cdot T\left(i,j\right) = \\ &\sum_{j=1}^{n}\sum_{i=1}^{M} I\left(x+i,y+j\right)\cdot T\left(i,j\right) + \sum_{j=n+1}^{N}\sum_{i=1}^{M} I\left(x+i,y+j\right)\cdot T\left(i,j\right)\end{aligned} \tag{4}$$

Then, the upper bound $\gamma(x,y,n)$ is obtained by adding the first partial dot product term and an upper bound, $\beta(x,y,n)$, of second partial dot product term:

$$\gamma\left(x,y,n\right) = \sum_{j=1}^{n}\sum_{i=1}^{M} I\left(x+i,y+j\right)\cdot T\left(i,j\right) + \beta\left(x,y,n\right) \tag{5}$$

As shown in Figure 1, the index n in (5) determines the splitting of the dot product term into two partial terms. Applying the Cauchy-Schwarz inequality to the rightmost term of (4) yields to the bounding function $\beta(x,y,n)$:

$$\beta\left(x,y,n\right) = \left|\left|I_c\left(x,y\right)\right|\right|_{n+1}^{N} \cdot \left|\left|T\right|\right|_{n+1}^{N} \tag{6}$$

where $||I_c(x,y)|||_{n+1}^{N}$ and $||T|||_{n+1}^{N}$ represent the *partial* ℓ_2 *norms* of terms I_c and T within rows $(n+1)$ and N.

By plugging (6) into (5) we obtain a sufficient condition (e.g. (3)) that allows for skipping unsatisfactory matching candidates. It relies on a portion of the dot product term and a bounding function $\beta(x,y,n)$ that can be calculated very efficiently using incremental computation schemes (i.e. *box-filtering* [7]), at the cost of a limited and fixed number of operations.

It is worth observing that, in the examined case of a single partition of T and I_c, the splitting procedure of the dot product term is mandatory since if the bounding function is defined over the whole area (e.g. case of n set to zero) then inequality (3) never holds.

3 Extension of the BPC Technique to the ZNCC

This section describes how to extend the BPC technique based on the Cauchy-Schwarz inequality to the more robust and computationally expensive ZNCC function. The novel technique will be referred to as Extended Bounded Partial Correlation (EBPC).

Denoting with $\mu(T)$ and $\mu(I_c(x,y))$ the mean intensity values computed, respectively, on T and on $I_c(x,y)$, the Zero mean Normalized Cross Correlation between T and I at position (x,y) is defined as:

$$\begin{aligned} ZNCC(x,y) &= \frac{\sum_{j=1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right]}{\sqrt{\sum_{j=1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]^2}\cdot\sqrt{\sum_{j=1}^{N}\sum_{i=1}^{M}\left[T(i,j)-\mu(T)\right]^2}} \\ &= \frac{\sum_{j=1}^{N}\sum_{i=1}^{M}I(x+i,y+j)\cdot T(i,j)-M\cdot N\cdot\mu(I_c(x,y))\cdot\mu(T)}{\sqrt{||I_c(x,y)||^2-M\cdot N\cdot\mu^2(I_c(x,y))}\cdot\sqrt{||T||^2-M\cdot N\cdot\mu^2(T)}}\in[-1,1] \end{aligned} \tag{7}$$

Similarly to the NCC case, let's split the template T and the sub-image I_c into two portions, as shown in Figure 1, and correspondingly the numerator of (7) into two terms:

$$\begin{aligned} &\sum_{j=1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right]= \\ &\sum_{j=1}^{n}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right]+ \\ &\sum_{j=n+1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right] \end{aligned} \tag{8}$$

where, as usual, n represents the number of rows determining the two portions of template T and sub-image I_c.

The first term at the second member of (8), referred to as *partial correlation* (e.g. $P_c(x,y,n)$), may be written in a more convenient form as follows:

$$\begin{aligned} P_c(x,y,n) = &\sum_{j=1}^{n}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right]= \\ &\sum_{j=1}^{N}\sum_{i=1}^{M}I(x+i,y+j)\cdot T(i,j)+ \\ &-M\cdot n\cdot\left[\mu(T)\cdot\mu(I_c(x,y))\big|_1^n+\mu(I_c(x,y))\cdot\mu(T)\big|_1^n+\mu(I_c(x,y))\cdot\mu(T)\right] \end{aligned} \tag{9}$$

where $\mu(I_c(x,y))|_1^n$ and $\mu(T)|_1^n$ represent the *partial mean intensity values* between rows 1 and n referred, respectively, to the I_c and T term.

A bounding function of the numerator of the ZNCC function can be devised by applying the Cauchy-Schwarz inequality to the rightmost term in (8):

$$\beta_Z(x,y,n) = \sqrt{\sum_{j=n+1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]^2} \cdot \sqrt{\sum_{j=n+1}^{N}\sum_{i=1}^{M}\left[T(i,j)-\mu(T)\right]^2} \tag{10}$$

Then, by simple algebraical manipulations:

$$\beta_Z(x,y,n) = \sqrt{(||T||\,|_{n+1}^{N})^2 + (N-n)\cdot M\cdot\mu(T)\cdot[\mu(T)-2\cdot\mu(T)|_{n+1}^{N}]}\cdot$$
$$\cdot\sqrt{(||I_c(x,y)\,||\,|_{n+1}^{N})^2 + (N-n)\cdot M\cdot\mu(I_c(x,y))\cdot[\mu(I_c(x,y))-2\cdot\mu(I_c(x,y))|_{n+1}^{N}]} \tag{11}$$

Since function $\beta_Z(x,y,n)$ turns out to be an upper bound of a portion of the dot product term:

$$\beta_Z(x,y,n) \geq \sum_{j=n+1}^{N}\sum_{i=1}^{M}\left[I(x+i,y+j)-\mu(I_c(x,y))\right]\cdot\left[T(i,j)-\mu(T)\right] \tag{12}$$

replacing the latter term of (8) with $\beta_Z(x,y,n)$ leads to the following upper bound of the numerator of the ZNCC function:

$$\gamma_Z(x,y,n) = P_c(x,y,n) + \beta_Z(x,y,n) \tag{13}$$

Finally, denoting as η_{zmax} the maximum ZNCC score *found so far*, (13) allows to obtain the following sufficient condition for safely rejecting unsatisfactory matching candidates:

$$\frac{\gamma_Z(x,y,n)}{\sqrt{||I_c(x,y)\,||^2 - M\cdot N\cdot\mu^2(I_c(x,y))}\cdot\sqrt{||T||^2 - M\cdot N\cdot\mu^2(T)}} \leq \eta_{z_{max}} \tag{14}$$

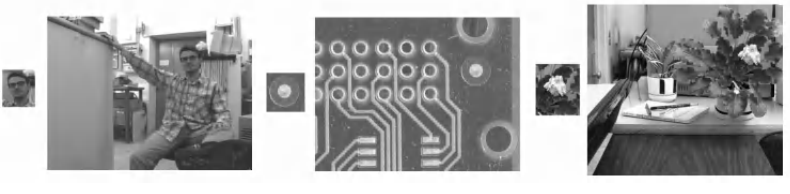


Fig. 2. Data set: (Left) Albert (Center) Pcb3 (Right) Plants

It is worth pointing out that with EBPC only a limited portion of the expensive dot product term needs to be calculated when the sufficient condition (14) holds. Viceversa, if the sufficient condition does not hold the dot product term has to be entirely computed.

Since the strength of the technique relies in avoiding this whole computation, in order to achieve effective performance improvements it is mandatory that the sufficient condition (14) could be calculated very efficiently and its outcome could hold as much as possible.

For this reason it is worth pointing out that the sufficient condition is made of terms (e.g. $||I_c(x,y)|||_{n+1}^{N}$, $\mu(I_c(x,y))$ and $\mu(I_c(x,y))|_{n+1}^{N}$) that can be efficiently computed using well-known incremental calculation techniques (e.g. [7]) requiring reduced and fixed overhead (i.e. 4 elementary operations for each term). This compares favorably with the dot product term since, conversely, this cannot be computed with incremental calculation techniques and hence its complexity grows with the template size resulting in the true bottleneck of the standard ZNCC-based algorithm

Finally, the remaining terms (e.g. $||T|||_{n+1}^{N}$, $\mu(T)$ and $\mu(T)|_{n+1}^{N}$) involved in the evaluation of (14) need to be computed and stored only once, at initialization.

4 Experimental Results

This section provides experimental results concerned with the data set *Albert*, *Pcb3* and *Plants* shown in Figure 2. For each image, Table 1 shows the speed-up of the EBPC algorithm compared to the standard ZNCC-based template matching algorithm with four different initial values of $\eta_{z_{max}}$ (respectively 0%, 90%, 95% and 98% of the actual $\eta_{Z_{max}}$ score). For each test $\frac{n}{N}$ was set to 0.18. All the algorithms were implemented in C and the system used for the experimental results was a Linux PC with an AMD Thunderbird 900 MHz processor.

First column of Table 1 shows that the proposed EBPC technique is effective in increasing the computational efficiency of a ZNCC based template matching process by at least a factor of 1.9. Moreover, better results have been obtained by using a higher initial value of $\eta_{z_{max}}$. In fact, this allows to use a very effective sufficient condition starting from the initial image points examined during the search process.

Table 1. For the three images of Figure 2: measured speed-ups for the EBPC algorithm with four different initial values of $\eta_{z_{max}}$.

Image	**EBPC** [0%]	**EBPC** [90%]	**EBPC** [95%]	**EBPC** [98%]
albert	1.90	2.69	2.74	2.78
pcb3	1.90	2.16	2.47	2.57
plants	2.40	2.56	3.00	3.32

Table 2. For the three images of Figure 2: percentages of points skipped by the EBPC algorithm with four different initial values of $\eta_{z_{max}}$.

Image	**Points**	**EBPC** [0%]	**EBPC** [90%]	**EBPC** [95%]	**EBPC** [98%]
albert	48958	62.15 %	81.86 %	82.58 %	82.79 %
pcb3	97080	62.81 %	70.16 %	77.58 %	79.44 %
plants	113832	73.38 %	76.56 %	83.61 %	87.35 %

Table 3. For the three images of Figure 2: measured speed-ups for the original BPC algorithm with four different initial values of $\eta_{z_{max}}$.

	BPC [0%]	**BPC** [90%]	**BPC** [95%]	**BPC** [98%]
albert	2.67	2.67	2.73	4.32
pcb3	2.09	2.09	2.09	2.10
plants	2.62	2.62	2.62	3.25

Table 4. For the three images of Figure 2: percentages of points skipped by the BPC algorithm with four different initial values of $\eta_{z_{max}}$.

Image	**Points**	**BPC** [0%]	**BPC** [90%]	**BPC** [95%]	**BPC** [98%]
albert	48958	79.64 %	79.64 %	80.48 %	96.71 %
pcb3	67080	66.19 %	66.19 %	66.19 %	66.62 %
plants	113832	76.26 %	76.26 %	76.32 %	85.12 %

Table 2 shows the percentage of skipped points relatively to each algorithm and image presented in Table 1. The table shows that the basic EBPC technique allows for skipping more than 62% of the examined points. Moreover, as expected, when $\eta_{z_{max}}$ gets higher the number of skipped points increases significantly.

Finally, Table 3 and Table 4, show respectively the measured speed-up and the number of skipped points obtained in the case of the standard BPC algorithm compared to the the *brute force* NCC algorithm. It is worth observing that these results are similar to those obtained comparing the EBPC algorithm to the *brute force* ZNCC algorithm (Tables 1 and 2).

5 Conclusions

We have described an efficient and exhaustive template matching algorithm based on direct computation of the ZNCC function. The algorithm extends the principles of the BPC technique, previously devised for the NCC, to the more robust ZNCC function. The proposed algorithm, referred to as EBPC, is capable of rapidly rejecting mismatching positions thanks to a sufficient condition based on the Cauchy-Schwarz inequality. The EBPC algorithm can be implemented very efficiently thanks the use of computational schemes that require limited and fixed numbers of operations.

Experimental results show that the EBPC algorithm compares favorably to the *brute force* ZNCC algorithm and that the behavior, in terms of measured speed-up, is similar to those obtained with the BPC technique in the NCC case.

A further improvement could be achieved using several elimination conditions based on increasing values of the parameter n. Besides, the implementation, currently under development, of the proposed algorithm with the parallel, SIMD-style, multimedia instructions available nowadays in most state-of-the-art microprocessors shall allow for further performance improvements.

References

1. L. Gottesfeld Brown, "A survey of image registration techniques" *ACM Computing Surveys*, Vol. 24, 1992, 325-376
2. W. Krattenthaler, K.J. Mayer, M. Zeiler, "Point correlation: a reduced-cost template matching technique" *1st IEEE Int. Conf. on Image Processing (ICIP 1994)*, Vol. I, September, 1994, Austin, Texas, USA, 208-212
3. A. Rosenfeld, G.J. Vanderburg, "Coarse-Fine template matching", *IEEE Trans. on Sys., Man and Cyb.*, Vol. 7, 1977, 104-197
4. A. Rosenfeld, G.J. Vanderburg, "Two-stage template matching", *IEEE Trans. on Image Processing*, Vol. 26, 1977, 384-393
5. L. Di Stefano, S. Mattoccia, "Fast Template Matching using Bounded Partial Correlation", *Machine Vision and Applications*, Vol. 13, 2003, 213-221
6. L. Di Stefano, S. Mattoccia, "A sufficient condition based on the Cauchy-Schwarz inequality for efficient Template Matching", *IEEE Int. Conf. on Image Processing (ICIP 2003)*, September 14-17, 2003, Barcelona, Spain
7. M. J. Mc Donnell, "Box-Filtering Techniques", *Computer Graphics and Image Processing*, Vol. 17, 1981, 65-70

Modelling of Overlapping Circular Objects Based on Level Set Approach

Eva Dejnozkova and Petr Dokladal

School of Mines of Paris, Centre of Mathematical Morphology
35, Rue Saint Honoré
77 300 Fontainebleau, FRANCE
dokladal@cmm.ensmp.fr

Abstract. The paper focuses on the extraction and modelling of circular objects embedded up to any extent. The proposed method is inspired from the continuous Level Set theory and consists of two stages. First, by using the local curvature and the normal vector on the boundaries are detected the shape parameters of the sought circular objects. Second, an area-based matching procedure detects such cases where several identified circles correspond to only one, partially ocluded object.

Keywords: Shape analysis, computer vision, smoothing, segmentation, curvature, level set, part decomposition

1 Introduction

Many methods dealing with detection and separation of overlapping objects can be found in the literature. The Hough transform (HT) [1] and its extensions [2] represent one popular method for extracting analytic curves. However, its implementation is often memory consuming. Another group of methods is based on classical morphological tools. Meyer [3] proposes a method based on a bisectrice function of the distance to the complement. However, this method can only be used to separate objects, embedded up to a limited extent. A more recent algorithm by Talbot [4] computes the skeleton on the elliptical distance. This method is computationally expensive and the author does not explain up to which extent this algorithm works.

The scale-space approach constitutes another separation technique. As an example, one can cite Lindeberg [5] who focuses on the junction detection by the normalized curvature in automatically selected scale-space. However, this method can lead to poor junction localization and requires an additional correction of the localization. Zhang [6] proposes a direct part decomposition of triangulated surfaces by thresholding the estimated gaussian curvature.

The second problem, shape fitting, usually deals with some optimization method. A widely used approach for shape fitting is the minimization of the squared error (with constraints or not) [7], [8]. The problem of these methods is the robustness and its numerical stability. Another approach using a Bayesian

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 416–423, 2004.
© Springer-Verlag Berlin Heidelberg 2004

formulation is shown in Werman [9]. However, this approach results in optimizing nontrivial cost functions for the fitted model.

This paper proposes a new method to detect and model embedded circular objects by using the local curvature of the object borders and tries to show that a local information can also be used for a global shape analysis. The parameter estimation is based on the continuous Level Set theory [10] which introduces a sub-pixel precision and allows to improve the numerical accuracy. The computation cost is reduced by measuremements performed only on a narrow band around the contours. The second goal is to minimize the optimization effort of the shape fitting task by using a clustering method on a set with a reduced number of elements.

The paper is organized as follows: the basic notions and principles are introduced first, followed by the algorithm description. Finally an application example used for an analysis of microscope photographs of polymer crystals is presented.

2 Basic Notions

Below we use the following notations and definitions.

Let X be a discrete, binary object $X\colon \mathbb{Z}^2 \to \{0, 1\}$, where 1 denotes the objets. Let $\mathcal{C}$ be a closed curve placed in $\mathbb{R}^2$ such that $\mathcal{C}$ is a continuous representation of the boundary of X. $\mathcal{C}$ is obtained from X by some interpolation method, see e.g. Siddiqi [11] or Shu [12]. For an easy manipulation, the curve $\mathcal{C}$ is defined implicitly by the signed distance function $u : \mathbb{Z}^2 \to \mathbb{R}$ such that

$$u(x) = d(x, \mathcal{C}) \tag{1}$$

The distance u is assumed positive outside and negative inside the object X. $\mathcal{C}$ is therefore the zero-level set of u.

To describe the curve $\mathcal{C}$, the distance u is calculated only on a narrow band NB close to the curve

$$\mathrm{NB} = \left\{ x \middle| \, x \in \mathbb{Z}^2,\ |u(x)| < \mathrm{NB}_{\mathrm{width}} \right\} \tag{2}$$

where $\mathrm{NB}_{\mathrm{width}} > 0$. Outside NB, the values of u are limited to $\pm\, \mathrm{NB}_{\mathrm{width}}$.

Let X be some set, and $\mathcal{P}(X)$ the set of all subsets of X. The mapping $\mathrm{CC}^{\gamma} : X \to \mathcal{P}(X)$ denotes the set of connected components of X, with a given neighbourhood connectivity γ (i.e. 4- or 8-neighbourhood). A γ-connected component of X is the equivalence class for γ-connected path of points in X.

A circle is defined by a pair $\{c, r\}$, where $c \in \mathbb{R}^2$ and $r \in \mathbb{R}^+$ stand for the centre and radius.

2.1 Local Curvature

The curvature of a curve is defined as the inverse of the radius r of the osculating circle. It is proportional to the angular speed of the normal vector n_0 travelling alongside the curve:

$$\kappa = \operatorname{div} n_0 \tag{3}$$

An exhaustive discussion of the curvature representations can be found in Sapiro [13]. In terms of implicit description by the distance function u, the curvature κ is given by:

$$\kappa = \operatorname{div} \frac{\nabla u}{|\nabla u|} \tag{4}$$

In this paper the curvature κ is used for two objectives: 1) smoothing and 2) radius estimation.

The smoothing makes use of the traditional discretization scheme which is obtained by using central differences in Eq. (4).

$$\kappa = \frac{u_{xx}u_y^2 - 2u_x u_y u_{xy} + u_{yy}u_x^2}{(u_x^2 + u_y^2)^{3/2}} \tag{5}$$

The main feature of this discretization scheme is that it allows to estimate the curvature even in singular points, i.e. the points where $u_x = u_y = 0$, by solving the limit case. However this scheme is less robust for the radius estimation because the surface deformations can influence the resulting level-set curvature.

The radius estimation requires a better accuracy of the curvature measurement.We propose to compute first the normal vector n_0 with a more sophisticated scheme taking the mean of four normal vectors obtained by one-sided differences as proposed by Sethian [10].

$$\begin{aligned} n_0' = (n_{0x}', n_{0y}') = &\frac{(u_x^+,\, u_y^+)}{\left((u_x^+)^2 + (u_y^+)^2\right)^{1/2}} + \frac{(u_x^+,\, u_y^-)}{\left((u_x^+)^2 + (u_y^-)^2\right)^{1/2}} + \\ &+ \frac{(u_x^-,\, u_y^+)}{\left((u_x^-)^2 + (u_y^+)^2\right)^{1/2}} + \frac{(u_x^-,\, u_y^-)}{\left((u_x^-)^2 + (u_y^-)^2\right)^{1/2}} \end{aligned}$$

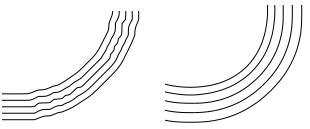
and after normalization :

$$n_0 = \frac{n_0'}{|n_0'|} \tag{6}$$

The divergence operator Eq. (3) is only applied after having numerically obtained n_0. In this case, the curvature estimation is based on the fine measurements of the changes in the normal vector direction and the resulting curvature estimation is more homogeneous. Nevertheless, this directional approach to the approximation is to be used carefully. The divergence operator can only be applied directly on such points where the normal vector n_0 exists both in the point itself and in its neighbourhood. Otherwise the obtained numerical value is incorrect.

3 Algorithm

Before the actual identification, the contours are decomposed in parts to separate the fused objects. The algorithm consists in several steps, described in this section in the order they are applied.



(a) (b)

Fig. 1. Iso-distance lines before (a), and after smoothing (b).

3.1 Construction of the Level-Set Function

The initial objects are defined as $X\colon \mathbb{Z}^2 \to \{0, 1\}$. The boundary level-set function is the distance function u calculated by using Eq. (1). Next, the distance u has to be smoothed to eliminate the discretization effect and the segmentation artefacts (cf. Fig. 2). Recall that u is calculated with a sub-pixel accuracy. Unless one uses some higher-order interpolation to obtain the initial line $\mathcal{C}$, the iso-distance lines will have a staircase-like aspect, unusable for global shape estimation based on the local curvature.

The level-set smoothing can be considered as a choice of the appropriated scale-space used for the part decomposition. There exist many smoothing methods; see e.g. [14], [15] or [5]. We have adopted the smoothing governed by the equation called *geometric heat flow* [15]:

$$\frac{\partial u}{\partial t} = \kappa |\nabla u| \tag{7}$$

This choice has been made for the following reasons. It has been shown that any closed curve will be deformed to become convex and will converge to a circular form without developping self-intersections. The expected circular shape of the contours is then naturally preserved. Eq. (7) applies the stronger smoothing factor (see Fig. 1), the higher the local curvature κ is. The smoothing stops when some norm of the difference of the successive iterations is smaller than some arbitrary limit.

3.2 Boundary Decomposition

The overlapping objects are separated by detecting the cusp points in $\mathcal{C}$. After the segmentation of $\mathcal{C}$ in these points one obtains a set of smooth circular arcs.

The segmentation is based on sign(κ) similarly to the methods proposed by [6] and [16]. The contour is decomposed in convex parts ($\kappa > 0$) by splitting it where κ changes sign and by dropping the concave parts. The concave parts correspond to the cusp points before the smoothing.

Recall, that here the set of arcs is described implicitly by the set $\mathcal{A}$ of portions of the narrow band around the decomposed curve $\mathcal{C}$. The set of arcs $\mathcal{A} = \{A_1, A_2, \dots\}$, $\mathcal{A} \subset \mathbb{Z}^2$, is obtained as : $\mathcal{A}' = \mathrm{CC}^{N_4}\{x \mid x \in \text{ NB and } \kappa(x) > 0\}$. The 4-connectivity N_4 was used for this application. Optionally, the arc segments that are too short are filtered out: $\mathcal{A} = \{A_i \mid A_i \in \mathcal{A}', \#A_i > const.\}$, where # stands for the cardinal of a set.

3.3 Detection of Centres and Radii of the Circles

The centre coordinates of the circle approximating a given arc $A_k \in \mathcal{A}$ are calculated in all points of A_k where the curvature is not biased by the border effect (introduced by points close to the border of NB where the derivatives of u use neighbors from outside the narrow band). Let $x_i \in A_k$ be any such point. The circle approximating the iso-distance line in x_i is given by: the radius $r_i = 1/\kappa(x_i)$ and the centre $c_i = x_i + n_0(x_i)/\kappa(x_i)$, $c_i \in \mathbb{R}^2$, where $n_0(x_i)$ is the normal vector (Eq. (6)) of the iso-distance line in x_i, and $\kappa(x_i)$ the curvature. Recall that negative values of κ were filtered out (cf. section "Boundary decomposition").

Let $C(A_k) = \{c_i\}$ denote the set of the circle centre coordinates, obtained for all $x_i \in A_k$. The centre $\hat{c}_k$ of the circle, approximating a given arc A_k, is obtained by taking the median: $\hat{c}_k = \text{med}(C(A_k))$. Finally, the radius is $\hat{r}_k = \text{med}(d_e(\hat{c}_k, x_i))$, for all $x_i \in A_k$, where d_e: $\mathbb{R}^2 \to \mathbb{R}^+$, denotes the euclidean distance from a to b calculated $d_e(a, b) = [(a_x - b_x)^2 + (a_y - b_y)^2]^{1/2}$.

Note that the computationaly expensive voting and following maxima searching processes of the standard HT are replaced by simple statistical measure.

3.4 Circles Matching

Obviously, several arcs may form one circular object and have to be matched. We need some arbitrary condition authorizing to couple circles even if their centres and radii are not perfectly identical. This condition is a trade-off between the accuracy and the capacity to model circles fused to an unlimited extent. We use a classical clustering technique, proceeding iteratively by agglomerating the most similar pairs.

The procedure uses the similarity matrix $\Xi = \{\xi_{ij}\}$ where ξ_{ij} denotes the similarity of circles i and j. The similarity criterion ξ of two circles is the area of their intersection divided by the area of their union. The strictly positive values of ξ are allowed only for the circles of similar radii and position (i.e. low d_e).

$$\xi_{ij} = \begin{cases} \frac{S(A)}{S(A_i)+S(A_j)-S(A)} & \text{if } d_e(\hat{c}_i, \hat{c}_j) < \frac{\min\{\hat{r}_i, \hat{r}_j\}}{2} \text{ and } i > j \\ 0 & \text{elsewhere} \end{cases} \tag{8}$$

where $S(A_i)$ and $S(A_j)$ are the areas of the i-th and j-th circle and $S(A)$ the area of their intersection. The values ξ_{ij} are calculated for $i > j$ only to have a triangular matrix under the main diagonal. The matching algorithm reads:

```
repeat while (max_{i>j}(ξ_ij) > 0)
   [k, l] := arg max_{i>j}(ξ_ij)  // get the most similar pair k,l of all i,j
   A_k := A_k ∪ A_l               // merge the circle l to k
   ĉ_k := med(C(A_k))             // recompute the centre and radius
   r̂_k := med(d_e(ĉ_k, x_i)) for all x_i ∈ A_k
   delete A_l, line and column l from Ξ and recompute Ξ
```

This algorithm is run on a reduced population set and is not computationally expensive. It stops as soon as there are no more pairs of circles verifying $\xi_{ij} > 0$.

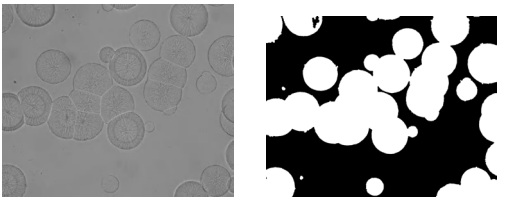


Fig. 2. Original image: (a) A microscope photograph of polymer crystals, (b) Segmentation of the original image.

4 Experiment Results

The motivation of the study presented above was an automatic crystal growth analysis [1]. In the early stage of the growth, the crystals are circular. The goal is to successively study crystal properties such as the size or the degree of incrustation of artificial polymer crystals (Fig. 2).

4.1 Segmentation

Although the segmentation is not the objective of this paper, we briefly describe how the initial contours are obtained. Recall the basic operators from the mathematical morphology. Let f denote an image and $\mathcal{F}$ the family of images. Then ε_X, δ_X: $\mathcal{F} \to \mathcal{F}$ denote respectively the erosion and dilation by X. If no structuring element is given then the unitary disk is used.

The segmentation strategy has been chosen according to the observations made on the original image (Fig. 2). One can see that the image background is almost flat. On the other hand, the range of intensity in the crystal interior can be high. The gray levels on the crystal borders can reach both low and high values. In order to overcome this problem we start the segmentation procedure by computing the morphological gradient which allows to extract the information about the contrast changes alongside the borders. Let f denote the original grey-scale image and $f_g = \delta(f)$ - $\varepsilon(f)$ the morphological gradient of f.

Since the next objective is to segment the image by a threshold we first have to equalize the values in the crystal interior. For this purpose we use the hole filling operator HoleFill. HoleFill(f_g) fills the attraction basins of f_g (corresponding to the agglomerated crystals): f_1 = HoleFill(f_g). Note, that the HoleFill operator can be applied either to binary or grey-scale images. More details can be found in [17], for example.

Finally, the following thresholding extracts the gradient crests delimiting the circular objects:

$$f_2 = \begin{cases} 1, & \text{if } f_1 > Th \\ 0, & \text{otherwise} \end{cases}$$

[1] The microscope images of polymer crystals were used with kind permission of the Centre for Material Forming (CEMEF), School of Mines of Paris, France

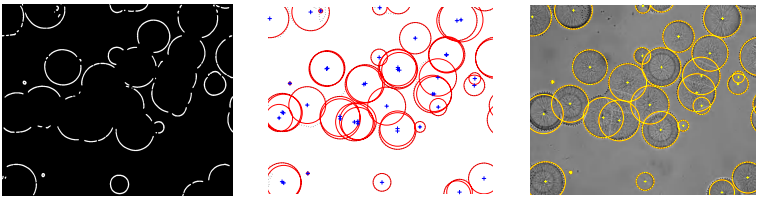


Fig. 3. (a) Arcs segmented by the boundary decomposition, (b) Circles detected from the arcs. (c) Result of the circles matching superposed on the original gray-scale image.

where $Th > 0$ (a convenient value for this application is $Th = 15$). In the next stage a morphological closing filter will smooth noisy borders: $f_3 = \varepsilon_X(\delta_X(f_2))$, where X is a disk of a four point radius. In order to eliminate the possible holes inside these objects, a following hole-filling operator is applied to fill the interior of the objetcs: $f_4 = \text{HoleFill}(f_3)$. Finally, an area opening suppresses small, noisy objects in the background: $f_5 = \text{AreaOpen}_N(f_4)$, with $N = 300$ (to eliminate objects smaller than 300 points). f_5 is the result of the segmentation (see Fig. 2 (b)).

The binary objects are then submitted to boundary decomposition to extract the smooth arcs, see Fig. 4.1. For every extracted arc one osculating circle is found (Fig. 4.1). Finally, if several circles correspond to only one circular object they are matched and replaced by only one circle. The parameters of the new circle are accordingly adjusted. See the results of the matching (superposed to the original image) at Fig. 4.1.

5 Conclusions

The paper shows the use of local curvature measure for a global shape analysis and gives a specific application example, where the curvature is used to separate circular objects fused theoretically up to any extent.

The proposed method uses a distance-based, implicit description of the contours for the estimation of the radii of circular objects. The measurements are performed only on a narrow band around the contours. Obviously, the curvature varies on different levels of the level set, even if measured in the normal direction. Nevertheless, the osculatory circles tangent to points laying on the normal are cocentric. Using larger narrow band gives birth to a more numerous population of candidates, and consequently to an increase of the accuracy.

Concerning the radius estimation, it has been observed that higher curvature offers better accuracy for the radius estimation. One improvement consists in giving stronger weights to points belonging to level sets closer to the circle centre, and having therefore a higher curvature. The second one consists in using an asymmetric narrow band, larger towards the centre of the circles.

In addition, this paper presents a matching procedure based on the analysis of the circle intersection area. The proposed matching condition allows to couple

circles representing one binary object and, at the same time, to separate crystals embedded up to high degree of incrustation.

The computation complexity of this technique remains quite low. The only computationally extensive step is the preprocessing where the distance function is iteratively smoothed. The proposed method have a high degree of parallelism and can be efficiently implemented on a specific parallel hardware [18] without any constraint.

References

1. Hough, P.V.C.: Methods and means for recognizing complex patterns. US Pattent 3069654 (1962)
2. Atiquzzaman, M.: Coarse-to-fine search technique to detect circles in images. International Journal of Advanced Manufacturing Technology **15** (1999) 96–102
3. Meyer, F.: Cytologie quantitative et morphologie mathématique. PhD thesis, Ecole des Mines de Paris (1979)
4. Talbot, H., Appleton, B.: Elliptical distance transform and the object splitting problem. In: ISMM, Australia. (2002)
5. Lindeberg, T.: Scale-Space Theory In Computer Vision. Kluwer Academic Publishers, Monograph 1994 (1994)
6. Zhang, Y., Paik, J., Koschan, A., Abidi, M.A.: A simple and efficient algorithm for part decomposition of 3-d triangulated models based on curvature analysis. In: ICIP02. Volume III., Rochester, N.Y. USA (2002) 273–276
7. Fitzgibbon, A., Pilu, M., Fisher, R.: Direct least square fitting of ellipses. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 476–480
8. Gander, W., Golub, G.H., Strebel, R.: Least-squares fitting of circles and ellipses. In editorial board Bulletin Belgian Mathematical Society, ed.: Numerical analysis (in honour of Jean Meinguet). (1996) 63–84
9. Werman, M., Keren, D.: A bayesian method for fitting parametric and nonparametric models to noisy data. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 528–534
10. Sethian, J.: Level Set Methods. Cambridge University Press (1996)
11. Siddiqi, K., Kimia, B., Shu, C.W.: Geometric shock-capturing ENO schemes for subpixel interpolation, computation and curve evolution. Graphical models and image processing: GMIP **59** (1997) 278–301
12. Osher, S., Shu, C.W.: High-order Essentially Non-oscillatory schemes for Hamilton-Jacobi equations. SIAM Journal of Numerical Analysis **28** (1991) 907–922
13. Sapiro, G.: Geometric Partial Differential Equations and Image Analysis. Cambridge University Press (2000)
14. Leymarie, F., Levine, M.D.: Curvature morphology. Technical Report TR-CIM-89-1, Computer Vision and Robotics Laboratory, McGill University, Montreal, Quebec, Canada (1989)
15. Kimia, B., Siddiqi, K.: Geometric heat equation and nonlinear diffusion of shapes and images. Computer Vision and Image Understanding: CVIU **64** (1996) 305–322
16. Siddiqi, K., Kimia, B.: Parts of visual form: Computational aspect. IEEE Transactions on Pattern Analysis and Machine Intelligence (1995)
17. Serra, J.: Image analysis and mathematical morphology. Academic Press, London (1982) ENSMP - CMM Fontainebleau.
18. Dejnožková, E., Dokládal, P.: Asynchronous multi-core architecture for level set methods. In: ICASSP, IEEE (2004) Proceedings.

A Method for Dominant Points Detection and Matching 2D Object Identification*

A. Carmona-Poyato, N.L. Fernández-García, R. Medina-Carnicer, and F.J. Madrid-Cuevas

Departament of Computing and Numerical Analisis, Córdoba University. Spain,
ma1capoa@uco.es

Abstract. A method for dominant points detection and matching 2D object identification using a new procedure for selecting dominant points is presented. This method can be classified as searching corner detection using some significant measure other than curvature category. For matching 2D object identification, an easy extension of the Gu–Tjahjadi method is proposed, modifying the comparison criterion in order to obtain a symmetrical and normalized function which allows metrics to be defined between contours. The experimental results show that this method is efficient and effective and significantly reduces the number of dominant points as compared to other proposed methods.

Keywords: Contour description, dominant points, matching 2D object.

1 Introduction

Dominant points detection is an important research area in computer vision given that information on a curve is concentrated at the corner (dominant) point. Many algorithms are used to detect dominant points. These methods can be classified into three categories [3]:those which search for dominant points using some significant measure other than curvature [2,4,7,1,12,9], those which evaluate the curvature by transforming the contour to the Gaussian scale space [10,11,5] and those which search for dominant points by estimating the curvature [6,13,14].

Contour representation is another problem in shape analysis. Many contour representations have been proposed in the literature. For example, Zhang et al.[15] reviewed and clasified shape representation and description techniques.

Dominant points detection consists of two steps: estimate a measure of curvature or its equivalent and locating the local maximum of this measure.

The present paper proposes a new method for dominant points detection using significant measure other than curvature. The detection procedure is applied to find the dominant points in the curve and compared with Mokhtarian's

* This work has been carried out with the support of the Research Project "DPI2002-01013" financed by the Spanish Ministry of Science and Technology Ministry and FEDER.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 424–431, 2004.
© Springer-Verlag Berlin Heidelberg 2004

method and Gu's method [10,6]. An easy extension of [6] has been proposed for the matching process. Our proposal uses the polar coordinate system in relation to the centroid as contour representation and dominant points obtained as contour descriptors.

In section 2 the proposed method, Gu's method [6] and Mokhtarian's method [11] for dominant points detection are described. An extension of the matching process [6] is proposed in section 3. The experiments are described in section 4 and, finally, the main conclusions are summarized in section 5.

2 Dominant Points Detection

2.1 Proposed Method

The set of n consecutive points define a digital curve C. That is $C = \{P_i(x_i, y_i)|i = 1, 2, 3, ..., n\}$ where n is the number of points, p_i is the ith point with coordinate (x_i, y_i). Let $\boldsymbol{a}_{ik} = (x_{i-k} - xi, y_{i-k} - yi)$ be the vector $\boldsymbol{P_iP_{i-k}}$ and $\boldsymbol{b}_{ik} = (x_{i+k} - xi, y_{i+k} - yi)$ be the vector $\boldsymbol{P_iP_{i+k}}$.

The meausure used for dominant points detection is the ratio r_i defined by

$$r_i = \frac{|\boldsymbol{u}_a| + |\boldsymbol{u}_b|}{2} \tag{1}$$

where $\boldsymbol{u}_a$ and $\boldsymbol{u}_b$ are unitary vectors respective to $\boldsymbol{a}_{ik}$ and $\boldsymbol{b}_{ik}$. Given that $\boldsymbol{u}_a$ and $\boldsymbol{u}_b$ are unitary vectors $r_i \in [0, 1]$. If the value of r_i is 0 or close to 0, the i-th point is on a straight line and cannot be considered a dominant point. If the value of r_i is close to 1, the i-th point can be considered a corner. The corner points are the candidates of dominant points. The candidates points can be reduced using a support region in the ratio r_i estimation and a prederterminate threshold for the r_i value. This reduction requires one or more input parameters to define the length of the support region and the threshold value, and can drastically influence the dominant points detection method. Because of this and due to the comparison proccess with other methods, a variable support region is not used; it is posible than a dominant points detection method can benefit from the selection of the support region estimation method. For this purpose, point P_i is considered a dominant point if r_i is a local maximum, and all the candidates will be dominant points.

2.2 Gu and Mokhtarian Methods

Gu's method [6] defines an angle φ_i associated with the i-th point as the angle between the two vectors $\boldsymbol{a}_{ik}$ and $\boldsymbol{b}_{ik}$:

$$\varphi_i = \cos^{-1}\left(\frac{a^2 + b^2 - c^2}{2ab}\right) \tag{2}$$

where $a = |\boldsymbol{a_{ik}}|$, $b = |\boldsymbol{b_{ik}}|$, and $c = |\boldsymbol{a_{ik}} + \boldsymbol{b_{ik}}|$. To prevent artifitial variation in angular values due to discrete points, the curve is smoothed before the angles

are calculated with a mean filter. The local minimum values of φi correspond to dominant points.

Mokhtarian's method [11] uses a curvature scale space representation to obtain the curvature in a point. The curvature c_i in a point p_i is calculated as

$$c_i = \frac{x'y'' - x''y'}{\sqrt{(x'^2 + y'^2)^3}} \tag{3}$$

where x' and y' are the convolution of x_i and y_i with the first derivative of Gaussian of width σ; and x'' and y'' are the convolution of x_i and y_i with the second derivative of Gaussian of width σ. The local maximum values of c_i correspond to dominant points.

3 Extension of Gu–Tjahjadi Method for the Matching Process

A contour comparison algorithm which works in two different stages is proposed in [6]. In the first stage, called "coarse object identification", the shapes that differ from the desired shape are discarded. In the second stage, called "the fine object identification", the desired and candidate shapes are compared using a fine-matching algorithm. In the present paper we attempt to improve this second stage of algorithm in [6], which consists of the following phases: determination of the first characteristic point of every contour and application of the contour comparison function.

Given that contours must have identical number of characteristic points in order to be compared, the following criterion is used to choose the dominant points of every contour: original dominant points, interpolated points in positions corresponding to dominant points of the other contour, and extra points obtained by means of uniform interpolation.

3.1 Extension to Obtain the First Dominant Point

In [6] the method considers the first dominant point of every contour to be the point with the global minimun angular value. This criteria depends on contour orientation and the applied level of smoothing and is very sensitive to noise.

An invariable to rotations and translations contour representation method has been used. In order to achieve the translational invariance, the polar coordinate system in relation to centroid has been chosen. Rotational invariance has been obtained by means of the following method:

- Computation of the contour centroid.
- Computation of the contour minimun inertia axis.
- Obtaining the contour points which belong to the contour minimun inertia axis and considering the point which is farthest from the centroid as the first contour point. If the intersection between the contour and minimun inertia axis is empty, the third criterion must be modified by obtaining the contour points closest to the minimun inertia axis and considering the point farthest from the centroid as first contour point.

– Rotation of the contour points so that the first contour point has an angle (in polar coordinates in relation to centroid) equal to 0^o.

Given that two diametrical opposed points can be equidistant from the centroid, both points have been chosen as first point. This criterion produces two different parametrized curves for every contour, thus implying a double comparison between objects in return for a more robust procedure.

The following characteristic points are the contour descriptors of the proposed algorithm: the dominant points obtained with the algorithm proposed above and the "equivalent" points in positions corresponding to the dominant points of the curve of the other contour.

3.2 Extension to Obtain the Comparison Function

Once the characteristic points are obtained, the contour comparison function is applied based on the square minimun algorithm. In [6], the comparison function has two drawbacks: a) the function is not symmetrical as the second contour has to be scaled and rotated in order to adjust to the size of the first contour and, therefore, cannot be used to define a metric and b) the function value is not normalized because it depends on the first contour size.

To avoid the drawbacks to the Gu–Tjahjadi comparison function, namely lack of symmetry and normalization, the following modifications are proposed:

– Normalization of dominant points radii:

$$P'_j(i) = (r'_j(i), \alpha_j(i)) = (\frac{c}{rmax_j} r_j(i), \alpha_j(i)) \tag{4}$$

where $P_j(i) = (r_j(i), \alpha_j(i))$ denotes the dominant point number i in coordinate polar in relation to the centroid in the $contour_j$, where $j \in \{1, 2\}$, $c > 0$ and $rmax_j = max\{r_j(i) | i \in \{1, .., N\}\}$ $contour'_j$ is normalized since $r_j(i) \in [0, c], \forall i$.

– Application of the following comparison measure:

$$\varepsilon_1^2(contour'_1(\theta_1), contour'_2(\theta_2)) = \frac{100}{4c^2 N} \sum_{i=0}^{N-1} d(P'_{1(\theta_1)}(i), P'_{2(\theta_2)}(i))^2 \tag{5}$$

where $P'_{j(\theta_j)}(i) = (r'_j(i), \alpha_j(i) + \theta_j)$ and $j \in \{1, 2\}$.
Obviously, the proposed measure is symmetrical and, accordingly, it can be used to define a new metric whose values are normalized between 0 and 100.

4 Experimental Results

4.1 First Experiment

To prove the extension of Gu's method to object recognition, the developed algorithm has been applied to forty contours obtained from four objects included

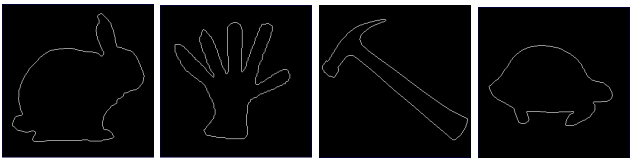


Fig. 1. Contours used in first experiment.

in binary images. The original objects were rotated 0^o, 30^o, 45^o, 60^o and 90^o and scaled with factors equals to 1.0 and 1.5. The external contour was then extracted from every rotated and scaled object. The contours of the original objects are showed in figure 1.

Two parametrized curves were obtained from every contour and compared with the parametrized curves pertaining to the other contours. To obtain dominant points [11] has been used given that the recognition algorithm does not depend on the dominant points detection algorithm. The values σ_1 and σ_2, used by the algorithm to evaluate dominant points, varied from 1 to 5 with a 0.5 increment. The results allow the following observations to be made:

- The rotational angles θ_{1_k} and θ_{2_k}, obtained from the minimization process of function ε_1^2 (equation 5), are correlated: $Pearson(\theta_{1_k}, \theta_{2_k}) = -1.000$ This correlation is significant to level 0.01 (bilateral). Therefore, it is only necessary to compute the rotational angle of a normalized contour, because the angle of the other contour will be the opposite one.
- As shown in table 1, the value of θ_{1_k} is very close to zero, that is, it is only necessary that the contour rotate to reach the minimum value. Consistently, the contour parametrization using the minimun inertia axis and the first point angle for the initial rotation is sufficient to provide a good approximation to the minimun value of the comparison function. This results shows that the minimization process is not necessary when only a good aproximation to the minimum value is required.
- The proposed algorithm can correctly classify every object (figure 2).

Table 1. Statistical descriptors of the θ_{1_k} parameter corresponding to the rotational angle of the first contour where the comparison function reaches the minimun value.

Mean	Standard Deviation	Minimum	Maximum
4.85E-19	1.47E-02	-0.042500	0.042500

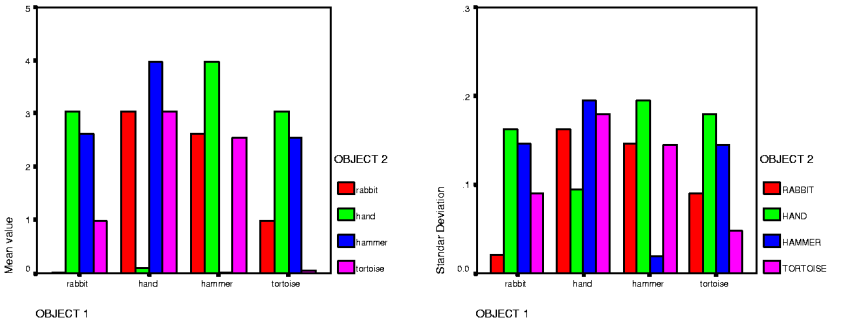


Fig. 2. Mean and standard deviation of the comparison function values.

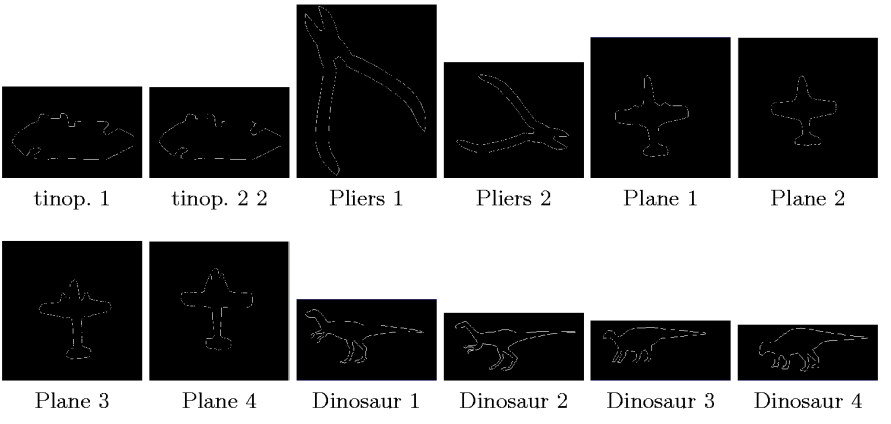


Fig. 3. Contours used in the second experiment.

4.2 Second Experiment

The object recognition method developed here has been applied to twelve contours to evaluate dominant points detection. The contours of the original objects are shown in figure 1. In this paper we have compared:

- Mokhtarian's method for obtaining dominant points and the proposed extension of Gu's method for the matching process.
- Gu's original method [6].
- Gu's original method for obtaining dominant points and the proposed extension of Gu's method for the matching process.
- The proposed point detection method and the proposed extension of Gu's method for the matching process.

Ten neighbours were used in dominant points detection to estimate curvature measure or its equivalent.

The results are shown in the table 2. When similar objects or different images of the same object are compared, the comparison function values are lower than

Table 2. Comparison function values and dominant points/contour points ratio.

Contours	Mokhtarian	Gu	Gu's extension	proposed method
tin op.1-tin op.2	0.025	0.035	0.024	0.017
pliers1-pliers2	1.79	2.30	1.60	1.41
dinosaur1-dinosaur2	0.78	0.85	0.69	0.96
dinosaur1-dinosaur3	4.25	6.36	3.36	4.94
dinosaur1-dinosaur4	3.50	4.16	4.29	3.82
dinosaur2-dinosaur3	3.71	5.30	3.96	4.25
dinosaur2-dinosaur4	3.46	3.50	3.65	3.88
dinosaur3-dinosaur4	1.61	1.80	1.65	1.74
plane1-plane2	1.50	0.9	1.80	2.74
plane1-plane3	10.9	16.3	11.2	9.60
plane1-plane4	9.80	7.30	10.6	9.50
plane2-plane3	13.0	13.8	13.4	11.5
plane2-plane4	12.2	7.70	13.0	12.1
plane3-plane4	0.72	23.02	0.69	0.67
tin op.1-pliers1	28.3	26.3	31.2	22.9
tin op.1-dinosaurs	38.1	38.4	38.6	35.8
tin op.1-planes	47.4	65.0	51.4	34.4
pliers1-dinosaurs	63.0	61.4	67.8	58.4
pliers1-planes	87.5	92.5	99.0	74.9
dinosaur1-planes	15.0	19.1	18.2	13.5
dinosaur3-planes	17.0	17.3	14.0	13.4
ratio	0.109	0.142	0.127	0.098

4, showing that all methods are good. When differents objects are compared, the comparison function values are greater than 5. When using Gu's method in the plane3-plane4 comparison, the comparison function value is high due to fact that the recognition algorithm is sensitive to noise when determining the first contour point. The results show than Mokhtarian's method and the proposed method are better than Gu's methods (original and extension) while Mokhtarian's method and the proposed methods are similar.

The ratio between the number of dominant points and the number of contour points [13,14] is important when evaluating dominant points detection algorithms. The mean values of this ratio are shown in table 2. As seen from the results,the proposed method is better than the other methods given that the number of dominant points obtained is 10 per cent lower than with Mokhtarian's method, 30 percent lower than with Gu's extension and 45 percent lower than with Gu's method.

5 Conclusions

A new method for detecting dominant points and a modified method for 2D object recognition has been proposed in this paper. The new method for dominant points obtains similar results to the Mokhtarian's method using a lower number

of dominant points. The 2D object recognition method is a modification of the fine-matching algorithm in [6]. The main features of the new method are:

- The contour representation is invariable to rotations and translations.
- The comparison function is symmetrical and normalized.

These modifications improve the robustness of the algorithm and permit a new metrics to be defined for classifying bidimensional objects. Furthermore, if only a good approximation to the minimun value is required the minimization process is unnecessary.

References

1. Bandera A., Urdiales C., Arrebola F., Sandoval F.: 2D object recognition based on curvature functions obtained from local histograms of the contour chaincode. Pattern Recognition Letters **20** (1999) 49–55.
2. Cornic P.: Another look at dominant point detection of digital curves. Pattern Recognition Letters **18** (1997) 13–25.
3. Fu A.M.N., Yan H.: Effective classification of planar shapes based on curve segment properties. Pattern Recognition Letters **18** (1997) 55–61.
4. Fu A.M.N., Yan H.: A contour bent function based method to characterize contour shapes. Pattern Recognition **30** (1997) 1661–1671.
5. Garrido A., Pérez N., García-Silvente M.: Boundary simplification using amultiescale dominant-point detection algorithm. Pattern Recognition **31** (1998) 791–804.
6. Gu Y.H., Tjahjadi T.: Coarse-to-fine planar object identification using invariant curve features and B-spline modeling. Pattern Recognition **33** (2000) 1411–1422.
7. Huang P.W., Dai S.K., Lin P.L.: Planar shape recognition by directional flow-change method. Pattern Recognition Letters **20** (1999) 163–170.
8. Loncaric S.: A survey of shape analysis techniques. Pattern Recognition **31** (1998) 983–1001.
9. Marji M., Siy P.: A new algorithm for dominant points detection and polygonization of digital curves. Pattern Recognition **36** (2003) 2239–2251.
10. Mokhtarian F., Mackworth A.K.: A theory of multiscale-based shape representation for planar curves. IEEE Transactions on Pattern Analysis and Machine Intelligence, **14** (1992) 789–805.
11. Mokhtarian F.: Silhouette-based isolated object recognition through curvature scale space. IEEE Transactions on Pattern Analysis and Machine Intelligence, **17** (1995) 539–544.
12. Urdiales C., Bandera A., Sandoval F.: Non-parametric planar shape representation based on adaptative curvature functions. Pattern Recognition **35** (2002) 43–53.
13. Wu W.Y.: Dominant point detection using adaptative bending value. Image and Vision Computing **21** (2003) 517–525.
14. Wu W.Y.: An adaptative method for detecting dominant points. Pattern Recognition **36** (2003) 2231–2237.
15. Zhang D., Lu G.: Review of shape representation and description techniques. Pattern Recognition **37** (2004) 1–19.

Character Recognition Using Canonical Invariants

Sema Doguscu and Mustafa Unel

Department of Computer Engineering, Gebze Instute of Technology
Cayirova Campus
41400 Gebze/Kocaeli Turkey
{doguscu, munel}bilmuh@gyte.edu.tr

Abstract. This paper presents a new insight into character recognition problem. Implicit polynomial (IP) curves have been used for modelling characters. A unique decomposition theorem is employed to decompose these curves into simple line primitives. For the comparison of the characters, canonical invariants have been computed using so called "related points" of the curves, which are the real intersections of the lines. Experimental results are presented to asses discrimination power of proposed invariants and their robustness under data perturbations. The method has also been compared with fourier descriptors.

1 Introduction

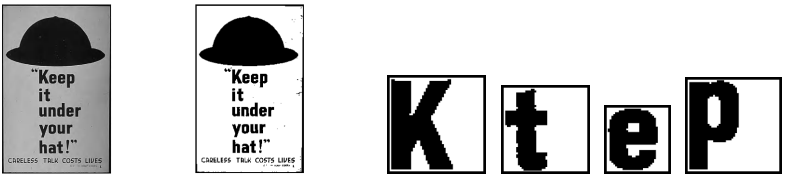
Automatic recognition of characters is an important problem in pattern analysis, and it has been the subject of research for many years. This paper presents a new insight into character recognition problem using IP or so called algebraic curves. The problem is to assign a digitized character into its symbolic class. In this work, IP curves have been used to model characters. Implicit polynomials are one of the most effective representations for complex free-form object boundaries and have certain advantages over other representations [4,5,6,7,8,10]. Character recognition follows three major steps in our approach. These are [1]: Preprocessing; Representation; Recognition and Classification of Characters.

In the preprocessing part, analog documents are converted into digital form and then thresholded. The connected component analysis [2] is performed on digitized image and each character is extracted from the text line. Then boundaries of segmented characters are obtained by eight-neighbor method [3]. In the representation part, characters are modelled using IP curves which are fitted to the boundaries of the characters by a fitting procedure [4]. A unique decomposition theorem [6] is then used to decompose algebraic curves into lines. Line factor intersections are related- points which map to one another under affine transformations. These related-points are used to construct canonical invariants [7], which will then be used in recognition and classification part. In recognition and classification part, characters are recognized by comparing their canonical invariants. To compare invariant vectors, a similarity ratio is employed.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 432–439, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Preprocessing

The raw character data are subjected to a number of preliminary processing steps. These preprocessing algorithms smooth the character images, segment the characters from each other and from the background, remove the noise and calculate the boundaries of characters. The scanned input character images are gray-scaled images. These images should be converted into binary images by *thresholding* [3]. See Fig. 1.b. For representation, characters should be isolated from the document and each other. *Segmentation* is division or separation of the image into regions of similar attribute. Connected component segmentation method [2] is used in this work. Fig. 1.c depicts some segmented characters of the document shown in Fig. 1.a.



(a) (b) (c)

Fig. 1. (a)Original image (b)Binarized Image (c)Segmented characters

Since each character can be represented by a closed curve contour of line segments, tracing the boundary of the character can yield useful information to distinguish characters from one another [9]. *Contour detection* algorithm [3] extracts the information of the boundary of a segmented character and presents it in a more compact form. See the boundaries of some characters in Fig. 2.

(a) (b)

Fig. 2. (a)Original Image (b)Character Contours

3 Representation

3.1 Algebraic Curve Fitting and Implicit Polynomials

Image representation plays one of the most important roles in a recognition system. In order to avoid extra complexity and to increase the accuracy of the

algorithms, representation is required. In this work, algebraic curves have been used for image representation. To get the best fitting polynomial, we have used a fitting algorithm detailed in [4]. This algorithm is linear, computationally fast, Euclidean invariant and robust. Examples of algebraic curve fits of 6^{th} degree to data sets are shown in Fig. 3. Our experiments have shown that virtually all of the characters can be fit well by sixth degree IP curves.

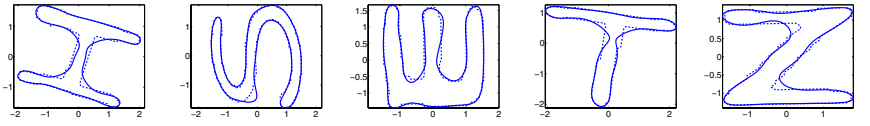


Fig. 3. Algebraic curve fitted characters. Solid curves represent algebraic curves fitted to the characters.

IP curves and surfaces are mathematical models for the representation of 2D curves and 3D surfaces. Algebraic curves are defined implicitly by equations of the form $f(x, y) = 0$, where $f(x, y)$ is a polynomial in the variables x, y, i.e.

$$f(x,y) = \sum_{0 \le i+j \le n} a_{ij} x^i y^j$$

Alternatively, the intersection of an explicit surface $z = f(x, y)$ with the $z = 0$ plane yields an algebraic curve if $f(x, y)$ is a polynomial [6]. The analysis and representation of algebraic curves can be simplified through *a unique decomposition theorem*. Decomposition theorem provides a new expression for the curve as a unique sum of the products of (possibly complex) lines.

Theorem 1. *[6] A non-degenerate (monic) $f_n(x, y)$ can be uniquely expressed as a finite sum of real and complex line products or real conic-line products, namely*

$$f_n(x,y) = \Pi_n(x,y) + \gamma_{n-2}[\Pi_{n-2}(x,y) + \gamma_{n-4}[\Pi_{n-4}(x,y) + \dots]] \qquad (1)$$

where each Π_r is a product of r real and/or complex lines,i.e.

$$\Pi_r(x,y) = \Pi_{i=1}^{r}(x + l_{ri}y + k_{ri}) \;\; with \;\; \Pi_0(x,y) = 1$$

3.2 Affine Equivalence and Related Points

Any two curves defined by a monic $f_n(x, y) = 0$ and a monic $\bar{f}_n(\bar{x}, \bar{y}) = 0$ will be affine equivalent if for some scalar s_n,

$$f_n(x,y) = 0 \overset{A}{\mapsto} f_n(m_1\bar{x} + m_2\bar{y} + p_x, m_3\bar{x} + m_4\bar{y} + p_y) \overset{def}{=} s_n\bar{f}_n(\bar{x},\bar{y}) = 0 \qquad (2)$$

Two corresponding related-points of the affine equivalent curves defined by $f_n(x,y)=0$ and $\bar{f}_n(\bar{x},\bar{y})=0$, such as $\{x_i,y_i\}$ and $\{\bar{x}_i,\bar{y}_i\}$ will be defined by the condition that

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} m_1 & m_2 & p_x \\ m_3 & m_4 & p_y \\ 0 & 0 & 1 \end{pmatrix}}_{A} \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \\ 1 \end{pmatrix} \longrightarrow \{x_i,y_i\} \overset{A}{\mapsto} \{\bar{x}_i,\bar{y}_i\} \tag{3}$$

Any two corresponding related-points will satisfy the relation

$$f_n(x_i,y_i) = s_n \bar{f}_n(\bar{x}_i,\bar{y}_i) \tag{4}$$

In the case of affine transformations, bitangent points, inflection points, centroids and line factor intersections all represent related points which can be determined from knowledge of the curves [6]. Line factor intersections have been used as related-points in this work. To establish the *correct correspondence* between the points in two sets of k corresponding real, distinct related-points, such as $\{x_i,y_i\}$ and $\{\bar{x}_i,\bar{y}_i\}$, we next note that if $f_n(x_i,y_i)=z_i$ and $\bar{f}_n(\bar{x}_i,\bar{y}_i)=\bar{z}_i$, then $z_i = s_n\bar{z}_i$ and

$$s_n = \frac{z_i}{\bar{z}_i} = \frac{\sum_{i=1}^{k} z_i}{\sum_{i=1}^{k} \bar{z}_i} \tag{5}$$

Therefore, we will always order the related points so that $z_1 < z_2 < ... < z_k$, and

$$\bar{z}_1 < \bar{z}_2 < ... < \bar{z}_k \quad if \quad s_n > 0$$

$$\bar{z}_1 > \bar{z}_2 > ... > \bar{z}_k \quad if \quad s_n < 0$$

4 Recognition

To distinguish characters from one another, a set of features should be extracted for each class and these features should be invariant to characteristic differences within the class. In this work *canonical invariants* have been used for recognition. Let $f_n(x,y)=0$ and $\bar{f}_n(\bar{x},\bar{y})$ be affine equivalent IP curves. Any three related-points of $f_n(x,y)=0$ to any three corresponding related-points of $\bar{f}_n(\bar{x},\bar{y})=0$ will define the affine transformation matrix A via the relation (3).

Any three such related-points of $f_n(x,y)=0$ will define a canonical transformation matrix [7]

$$A_c = \begin{pmatrix} x_1-x_3 & x_2-x_3 & x_3 \\ y_1-y_3 & y_2-y_3 & y_3 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}}_{E}, \tag{6}$$

and a monic canonical curve $f_n^c(x,y)=0$ of $f_n(x,y)=0$ defined by the relation

$$f_n(x,y)=0 \overset{A_c}{\mapsto} s_c\, f_n^c(x,y)=0 \tag{7}$$

Three corresponding related points of $\bar{f}_n(\bar{x},\bar{y})=0$ will define a corresponding canonical transformation matrix $\bar{A}_c=\bar{T}E$, and a corresponding monic canonical curve $\bar{f}_n^c(\bar{x},\bar{y})=0$ of $\bar{f}_n(\bar{x},\bar{y})=0$ defined by the relation

$$\bar{f}_n(\bar{x},\bar{y})=0 \overset{\bar{A}_c}{\mapsto} \bar{s}_c\, \bar{f}_n^c(\bar{x},\bar{y})=0 \tag{8}$$

for some scalar $\bar{s}_c$

We will call the coefficient vectors of canonical curves *canonical invariants*. Our strategy will be to associate a canonical curve with each character and compare characters based on their canonical curves. In practice, the coefficients of the canonical curves will not be the same for the affine equivalent characters because of noise. Therefore we have to introduce some measure of "closeness" of canonical invariant vectors. Comparison of two characters is realized by comparing the similarity of the canonical invariants. Characters have been compared with each other under a similarity ratio. The similarity ratio employed in this work is

$$Similarity = r = \frac{C_1^T C_2}{\|C_1\|\|C_2\|} \tag{9}$$

Here C_1 and C_2 are the canonical invariants of the curves. If two vectors are close to each other, similarity value gets closer to 1, otherwise similarity value gets closer to -1. Characters with the highest similarity will be considered to be equivalent, and therefore to be the same.

$$-1 \le r \le 1 \tag{10}$$

5 Experimental Results

We now present some experimental results which illustrate our procedures. Characters have first been thresholded, segmented and their boundaries have been extracted. Then IP curves have been fit to data sets. After obtaining three related-points from IP curves, the (monic) canonical curves $f_6^c(x,y)=0$ have been determined. Using canonical invariant vectors, similarity ratio between the characters has been computed.

Recognition is performed by comparing the input characters with various model characters in the database using the computed similarity ratios. Characters can be classified into 3 groups by the number of their contours. Each character in the first group has one contour (See fig. 4a). The ones in the second group has two contours (See fig. 4b). Those in the third group has three contours as shown in fig. 4c. Several characters have been tested and their similarities to the model characters have been computed . The character model which has the

Fig. 4. (a) First group (b) Second group (c) Third group

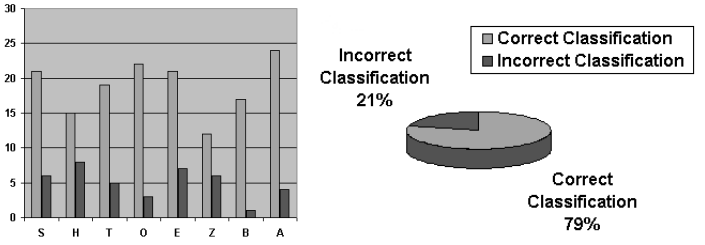


Fig. 5. Recognition rates using implicit polynomials

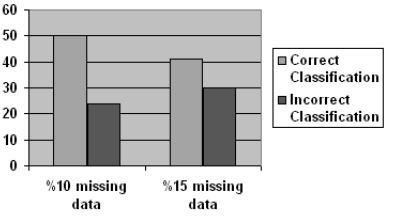


Fig. 6. Correct and incorrect classification rate under 10% and 15% missing data using implicit polynomials. Canonical invariants have yielded 68% recognition rate under10% missing data and 58% recognition rate under 15% missing data.

largest similarity ratio has been declared as the input character. We have used 191 character data and the recognition rate was 79%. See Fig. 5.

A character recognition system usually doesn't have all the boundary information of the characters. Only partial information might be available. To test the robustness and the discrimination power of our canonical invariants with respect to missing data, character data points were chopped at different boundary locations. The similarity ratios based on the canonical invariants of characters under 10% and 15% missing data are computed and shown in Fig. 6. Canonical invariants have yielded 68% recognition rate under10% missing data, and 58% recognition rate under 15% missing data.

We have also compared our method with fourier descriptors using the same characters, same models and the same conditions. Characters have been thresholded, segmented and their boundaries have been extracted. Then fourier de-

scriptors have been computed. From these descriptors, similarity ratios have been computed. Recognition rate has been found to be 69%.

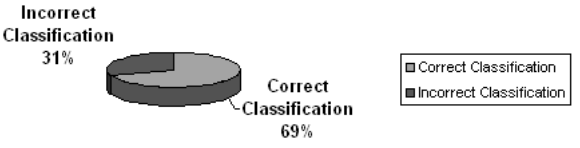


Fig. 7. Recognition rate for fourier descriptors using the same characters, the same models and the same conditions.

To test the robustness and the discrimination power of fourier descriptors with respect to missing data, character data points were chopped at different boundary locations. The similarity ratios of characters under 10% and 15% missing data have been computed and shown in Fig. 8.

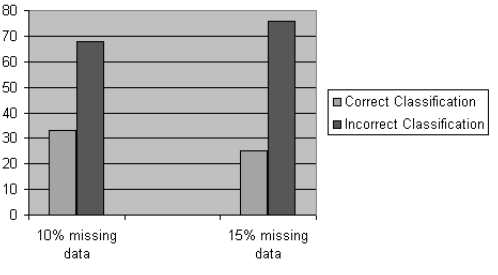


Fig. 8. Correct and incorrect classification rate under 10% and 15% missing data using fourier descriptors. Fourier descriptor based invariants have yielded 33% recognition rate under10% missing data and 25% recognition rate under 15% missing data.

6 Conclusion

We have now outlined a new method for character recognition problem. Algebraic curves are used for modelling characters. Most of the characters can be represented by 6^{th} degree algebraic curves. Since the quality of the fitting algorithm has substantial impact on the recognition performance, a stable and repeatable curve fitting method has been used. Decomposition theorem is employed to decompose these curves into lines. Line factor intersections have been used as related-points. By using related-points, canonical invariants have been computed.

Experiments have been conducted to compare characters based on the similarity between canonical invariant vectors. Robustness and the discrimination capabilities of canonical invariants have been tested on different characters. Experiments have shown that canonical invariants are stable with respect to modest amount of missing data. We have also compared our method with fourier descriptor using the same characters, the same models and the same conditions. Experimental results are promising, and much work must be done to fully exploit advantages of using IP curves as a representation in character recognition problems.

Acknowledgment. This research was supported from GYTE research grant BAP #2003*A*23

References

1. N. Arica & F. Yarman-Vural, An Overview of Character Recognition Focused on Off-Line Handwritting, IEEE Transactions on Systems,Man and Cybernetics-Part C:Applications and Reviews,Vol.31,No.2, May 2001.
2. H. Kuo & J. Wang, A New Method for the Segmentation of Mixed Handprinted Chinese/English Characters, Proceedings of the Second International Conference on Document Analysis and Recognition, pages 810-813, October 1993.
3. C. Jeong & D. Jeong, Handwritten Digit Recogntion Using Fourier Descriptors and Contour Information, IEEE TENCON, vol.6, No.99, 1999.
4. T.Tasdizen & J.P. Tarel & D.B. Cooper, Improving the Stability of Algebraic Curves for Applications, IEEE Transactions in Image Processing, vol.9, No.3, March 2000.
5. M. Unel & W. A. Wolovich, A new representation for quartic curves and complete sets of geometric invariants, International Journal of Pattern Recognition and Artificial Intelligence, December 1999.
6. M. Unel & W. A. Wolovich, On the Construction of Complete Sets of Geometric Invariants for Algebraic Curves, Advances in Applied Mathematics, Vol. 24, No. 1, pp. 65-187, January 2000.
7. W. A. Wolovich & M. Unel, The Determination of Implicit Polynomial Canonical Curves, IEEE Transactions on Pattern Analysis and Machine Intelligence, October 1998.
8. M.Blane, Z.Lei et al., The 3L algorithm for Fitting Implicit Polynomial Curves and Surfaces to Data, IEEE Transaction on Pattern Analysis and Machine Intelligence, Bol.22, No.3, March 2000.
9. Y. Chung & M. Wong, Handwritten Character Recognition by Fourier Descriptors and Neural Network, IEEE TENCON, Speech and Image Technologies for Computing and Telecommunications, 1997.
10. D. Keren & D. Cooper, Describing Complicated Objects by Implicit Polynomials, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.16, No.1, 1994.

Finding Significant Points for a Handwritten Classification Task

Juan Ramón Rico-Juan and Luisa Micó *

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, E-03071 Alicante, Spain,
{juanra, mico}@dlsi.ua.es

Abstract. When objects are represented by curves in a plane, highly useful information is conveyed by significant points. In this paper, we compare the use of different mobile windows to extract dominant points of handwritten characters. The error rate and classification time using an edit distance based nearest neighbour search algorithm are compared for two different cases: string and tree representation.

Keywords: Feature extraction, nearest neighbour, handwritten character recognition, metric space.

1 Introduction

One of the most useful and simplest techniques in Statistical Pattern Recognition that can be used in a wide range of applications of computer science and technology is the Nearest Neighbour (NN) rule. In this rule, an input pattern is assigned to the class of the nearest prototype pattern. Examples of application of NN rule include handwritten/speech recognition, data compression [1], data mining [2] and information retrieval [3].

If patterns can be coded in a vector space, methods based on the coordinates of the representation can be applied. However, this is not the general case and often, only methods that use a distance (and the metric properties of the distance) can be applied to perform the classification.

A popular distance used in general metric spaces is the edit distance. The edit distance between two objects is defined as the number of basic operations (insertion, deletion and substitution) needed to transform one representation into another. Depending on the type of representation (for instance, strings or trees) basic operations are differently defined. Each basic operation has associated a weight, usually identical for insertion and deletion ($W_I = W_D$), and a third weight for substitution (W_S) that fulfils the following relationship:

$$W_I + W_D \geq W_S$$

Different algorithms allow to obtain a good code representation of planar objects [4, 5,6,7]. These algorithms extract points from a figure that help us to obtain the features

* Work partially supported by the Spanish CICYT under contract TIC2003-08496-CO4

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 440–446, 2004.
© Springer-Verlag Berlin Heidelberg 2004

to represent it. Some of them obtain a set of approximately equidistant points (EP) [8]. Others algorithms obtain a set of dominant points with an irregular distribution (IP) [4, 5,7].

Classification using the edit distance and IP methods gives poor results (tested in [9]). Improving classification requires the definition of a more complex distance that takes into account the geometric distance between adjacent significant points.

In [8] two different representations of handwritten characters have been used: the contour string and a tree of significant points obtained after using a modified thinning algorithm [10]. The set of points obtained with the second method are almost equidistant because each new point is obtained when a square window is moved a fixed distance in the original image.

In this work we use a mobile circular window to obtain equidistant points for strings and trees as representation of handwritten characters. The number of points in the representation depends on the radius of the window. In section 2 we describe the method to obtain the string and tree code. The edit distance used to compare these representations is introduced in section 3. In section 4, the obtained results are listed while, in section 5, the concluding remarks are offered.

2 String and Tree Representation of Characters

Here, two different representations of handwritten characters have been used. In both cases, the mathematical morphology opening transformation reference are used to avoid noisy pixels and to smooth the shapes of the characters.

2.1 Tree Code

The Nagendraprasad-Wang-Gupta thinning algorithm modified as in [10] was applied (figure 1b). The result image is transformed into a tree representation using the following steps:

1. The radius R is selected.
2. The first up and left pixel, r, is marked and assigned the tree root with a special label "0". Two empty pixel sets C, G and a pair pixel set T are created.
3. $C \leftarrow \{r\}$
4. While $C \neq \emptyset$ repeat steps 5-7.
5. For all elements $t \in C$, collect in set G every unmarked pixels in the circumference of radius R centred in the pixel associate to t. Follow and mark connected pixels until a pixel, g, is found with one of the following properties:
   a) the branch has the maximum radius R (see figure 4);
   b) the pixel has no unmarked neighbours (terminal pixel);
   c) the pixel has more than one unmarked neighbour (intersection pixel).
6. Add to T the new branches $(t, g) : g \in G$. A label is assigned to the branch depending on the final pixel (g), relative position to the starting one[1] (t).
7. $C \leftarrow \{g\} : (t, g) \in G$. Erase all elements from G.
8. end repeat.
9. return T.

An example showing this feature extraction with character 'F' is presented in figure 1.

[1] The 2D space is divided in 8 regions (figure 3).

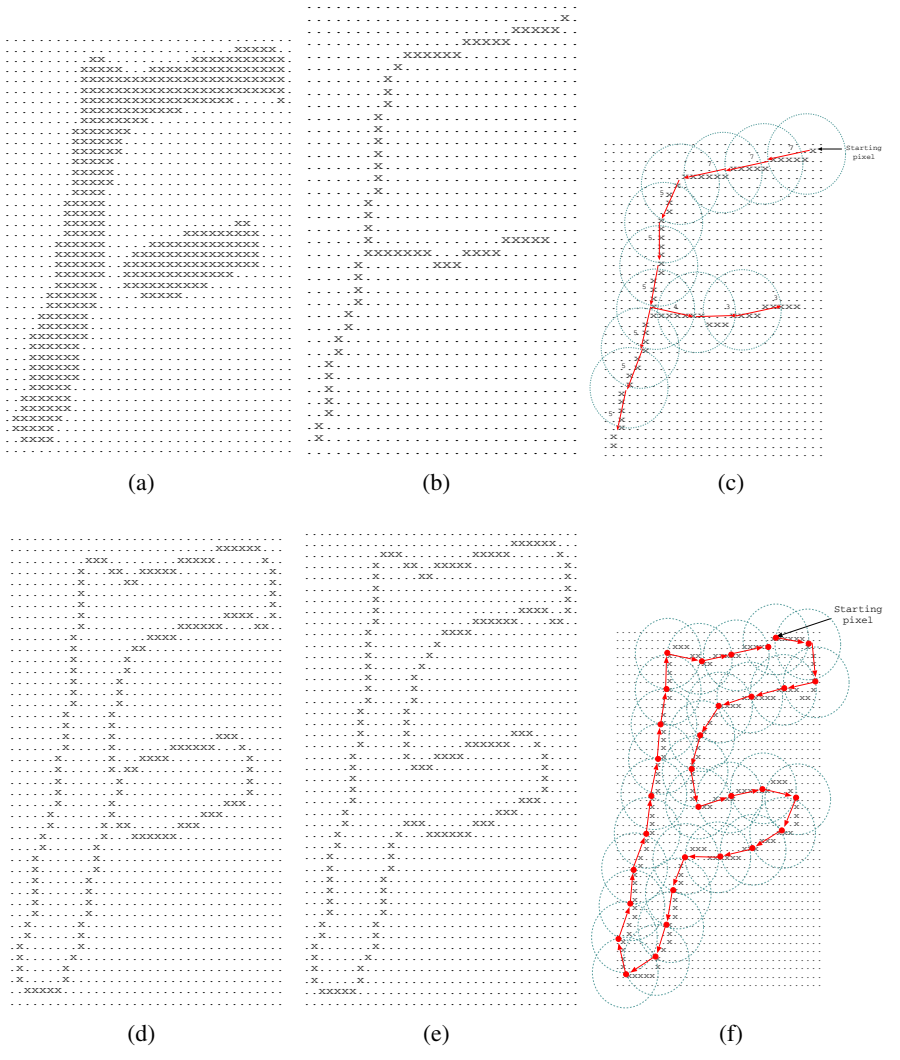


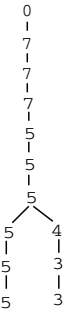
Fig. 1. Example feature extraction (a) original image; (b) thinned image; (c) tree labelling process; (d) image with problems to extract a contour string; (e) image right formed to extract a contour string; (f) string labelled process.

2.2 String Code

After the mathematical morphology opening transform using n pixels [2] is applied the following algorithm is used to extract the external contour of the character .

1. A radius R is selected. String s is empty.

[2] n is the smallest positive integer allow to have a external contour where all the pixels have two neighbours.



"F"=3577765523
46677555681111
1111333
(b)

Fig. 2. Example of extracted features from character in figure 1 (a) tree; (b) string.

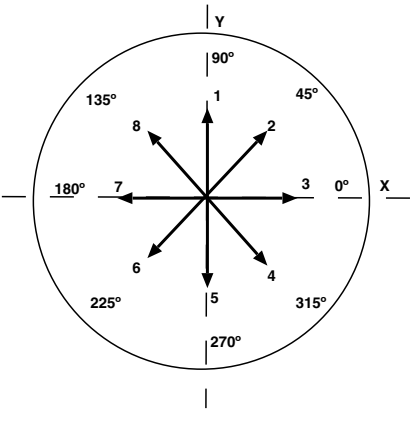


Fig. 3. 2D labelled regions

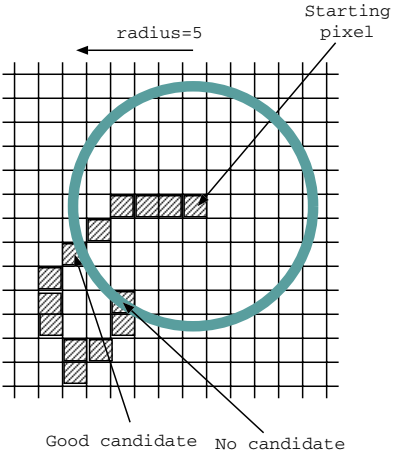


Fig. 4. Example to get next candidates to create branches in structured tree extraction.

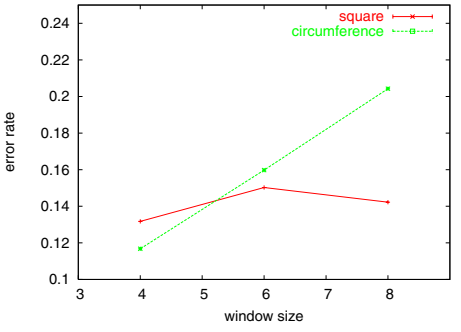
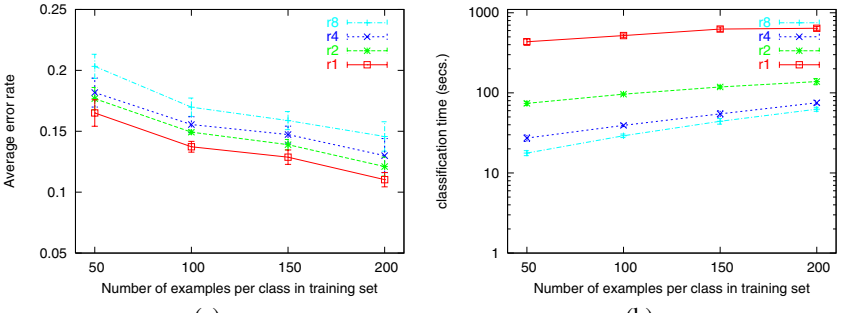


Fig. 5. Results applying NN classification algorithm with different types of window approximation



(a) (b)

Fig. 6. Results for NN classification with `AESA` search with tree representations of character obtained with different sizes of the window as a function of different training examples size belonging to 26 character classes: (a) average error rate; (b) average classification time.

2. The first black pixel, r, is searched with a left-to-right scan starting from the top.
3. From r and clockwise, the contour of the character is followed until a new black pixel, t, is found. This pixel is the intersection between the contour and the circumference centred in r with radius R. Add to the string s the code of the direction[3] (r, t).
4. If $(t \neq \text{last pixel}) r \leftarrow t$ and go step 3.
5. return s.

3 Edit Distances

A general tree edit distance is described in [11]. A dynamic programming algorithm is implemented to compute the distance between two trees, T_1 and T_2 whose complexity is in time $O\left(|T_1| \times |T_2| \times \min\left(\text{depth}\left(T_1\right), \text{leaves}\left(T_1\right)\right) \times \min\left(\text{depth}\left(T_2\right), \text{leaves}\left(T_2\right)\right)\right)$ and in space $O\left(|T_1| \times |T_2|\right)$.

[3] There are eight neighbouring pixels that can be found (figure 1f and 1g), therefore, only eight symbols can appear in this chain-code (see figure 3a)

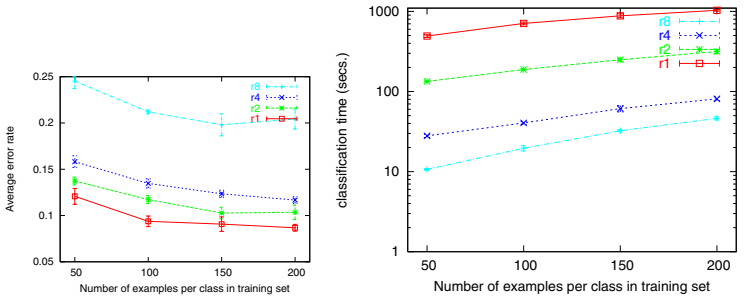


Fig. 7. Results for NN classification with `AESA` search with string representations of character obtained with different sizes of the window as a function of different training examples size belonging to 26 character classes: (a) average error rate; (b) average classification time.

Each basic operation has an associated weight with the following values, used in [6]: substitution $w_{ij} = \min(|i-j|, 8-|i-j|)$ and insertion and deletion $w_I = w_D 2$. This distance is finally normalised with the sum of the number of nodes in each tree.

The cost values on the string edit distance are those used in tree edit distance. The string edit distance can be computed in time in $O(|x|,|y|)$ using a standard dynamic-programming technique [12]. As in the tree edit distance, this final measure is normalised, in this case by the sum of the lengths of the two strings.

4 Experiments

A classification task using the NIST SPECIAL DATABASE 3 of the National Institute of Standards and Technology has been done in this work. Only the 26 uppercase handwritten characters were used. The increasing-size training samples for the experiments were built by taking 500 writers and selecting the samples randomly. To perform the NN search, the Approximating Eliminating Search Algorithm, `AESA`, has been used in this work.

Figure 5 shows the comparison between the error rate in a classification task is evaluated for different sizes, R, of the two types of windows: the square window used in previous work[8] and a circular window. The figure shows the average error rate using a training set of 200 samples per class. This experiment shows that the error rate grows linearly with the radius of the circular window. However, for relatively small windows, the error rate is smaller using a circular window than a square window.

Figures 6 and 7 shows the behaviour of the error rate and the classification time when a circular window is used. In this case, different radius of the window ($R = 1, 2, 4, 8$) with different sizes of the training set have been used for handwritten characters represented as strings and trees.

In all cases the use of strings generates a lower error rate in the recognition task than the use of a tree representation, although the classification time is higher. However, as shown in figures 6 and 7 larger values of the radius of the window allow to reduce the classification time at a little increase in the error rate.

On the one hand, the use of a circular window with the string representation improves the classification error rate (compared to the tree representation) with a radius of the window less or equal than 4.

On the other hand, when the radius grows using a string code, the classification time tends to be similar than that using a tree code.

5 Conclusions

In this paper we have compared the performance and the accuracy of a handwritten recognition task using two different representations (strings and trees) obtained with a circular window. Our experiments show that better results in a classification task are obtained when a circular window with radius higher to one are used for a string representation of the handwritten characters.

References

1. Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1991.
2. T. Hastie and 1996. R. Tibshirani. Classification by pairwise coupling. Technical report, Stanford University and University of Toronto, 1996.
3. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, 1983.
4. X. Li and D. Yeung. On-line handwritten alphanumeric character recognition using dominant points in strokes. *Pattern Recognition*, 30:31–34, 1997.
5. J. M. Iñesta, B. Mateo, and M. A. Sarti. Reliable polygonal approximations of imaged read objects though dominant point detection. *Pattern Recognition*, 31:685–697, 1998.
6. J. R. Rico-Juan and L. Micó. Comparison of AESA and LAESA search algorithms using string and tree edit distances. *Pattern Recognition Letters*, 24(9):1427–1436, 2003.
7. B. Sarkar, S. Roy, and D. Sarkar. Hierarchical representation of digitized curves though dominant point detection. *Patter Recognition Letters*, 24:2869–2882, December 2003.
8. J. R. Rico-Juan and L. Micó. Some results about the use of tree/string edit distances in a nearest neighbour classification task. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Pattern Recognition and Image Analysis*, number 2652 in Lecture Notes in Computer Science, pages 821–828, Puerto Andratx, Mallorca, Spain, june 2003. Springer.
9. J. R. Rico-Juan. Off-line cursive handwritten word recognition based on tree extraction and an optimized classification distance. In M. I. Torres and A. Sanfeliu, editors, *Pattern Recognition and Image Analysis: Proceedings of the VII Symposium Nacional de Reconocimiento de Formas y Análisis de Imágenes*, volume 3, pages 15–16, Bilbao (Spain), May 1999.
10. R. C. Carrasco and M. L. Forcada. A note on the Nagendraprasad-Wang-Gupta thinning algorithm. *Pattern Recognition Letters*, 16:539–541, 1995.
11. K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing*, 18:1245–1262, 1989.
12. R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21:168–173, 1974.

The System for Handwritten Symbol and Signature Recognition Using FPGA Computing

Rauf K. Sadykhov[1], Leonid P. Podenok[2], Vladimir A. Samokhval[2], and Andrey A. Uvarov[1]

[1] Belorussian State University of Informatics and Radioelectronics,
Computer Systems Department,
6 Brovka str., Minsk, Belarus, 220013
rsadykhov@gw.bsuir.unibel.by
[2] National Academy of Science United Institute of Information Problems
NAS UIIP, 6 Surganov st, Minsk, Belarus, 220012
{podenok,sam}@lsi.bas-net.by

Abstract. The system for handwritten symbol and signature recognition with the use of shape descriptors in orthogonal basis of Hadamard transform as informative features is presented. For a construction of the classifier discriminant methods have been used that have supplied acceptable separation between classes. To raise processing speed a part of algorithmic scheme was implemented by hardware facilities. There was used specialized integer-valued co-processor implemented as PCI bus extension board based on FPGA. Proposed hardware/software architecture resulted in computational acceleration factor of 10 with recognition rates averaged about 97-99% in the task of handwritten symbol processing.

1 Introduction

Per the last years the attention of researchers is attracted with the developement methods and algorithms for handwritten symbols and signatures, effective both on a productivity and accuracy. One of the main problems in this task is the selection of informative features that are invariant to a set of affine transforms. In [1] the approach to recognition of handwritten symbols on the base of different kinds of neural networks is presented. Another direction to handwritten symbols recognition is application of the approach on the base of shape descriptors and profile projections [2], [3], [4]. The investigations related to Fourier-Mellin, Legendre and Zernike moment functions and Hough transform also represent significant interest [5], [6]. In [7] the approach to identification of handwritten signatures is considered. To improve the outcomes, represented in [1-7] the system for recognition of handwritten characters and signatures on the base of shape descriptors in the space of Hadamard transform and discriminant methods with hardware implementation was developed.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 447–454, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Preprocessing Stage

The main advantage of preprocessing a handwritten character image is to organize the information so as to make the task of recognition simpler. The first step in this stage is to eliminate noises of original image. The second part of preprocessing is image normalization, which attempts to remove some of those variations in the image which do not affect the identity of the character. To enhance the recognition rate, all patterns must have the same size, so size normalization procedure is important and was performed at preprocessing stage. This procedure determines object boundaries followed by scaling transform. Performed normalization leads to spatial independency and can highly improve classification results.

3 Formalization of Feature Extraction

As it has been pointed out, we intend to investigate methods and algorithms for handwritten character recognition oriented mainly for hardware implementation. This fact tends to select appropriate technics which allow almoust absolute mapping to the hardware platform.

The information below describes the use of Hadamard technique of spectral expansion [8] aimed at image compression and application of synthetic discriminant functions (SDF) for classification. Another approch uses connectivity spectrum combined with multi-class correlation classifier. The basis for SDF synthesis is the use of image training sets for every object class. SDF filter is constructed from image correlation matrix and depending on special conditions the final SDF appears in different forms.

For two-class recognition problem (original and forged signatures) two SDFs $\boldsymbol{h}_i$ are constructed as follows

$$\boldsymbol{f}^{(k)} \otimes \boldsymbol{h}_i = \delta_{ki}, \tag{1}$$

where δ_{ki} is Kronecker symbol, $\otimes$ denotes correlation operator. In this case two orthogonal SDFs from (1) are calculated using N_1 training images $\boldsymbol{f}_i^{(1)}$ of first class objects and N_2 training images $\boldsymbol{f}_j^{(2)}$ of second object class as

$$\boldsymbol{h}_1 = \sum_m a_m \boldsymbol{f}_m^{(1)}, \quad \boldsymbol{h}_2 = \sum_m b_m \boldsymbol{f}_m^{(2)}, \tag{2}$$

where $\boldsymbol{f}_m^{(1)} \boldsymbol{h}_1 = 1$ for images of class 1, $\boldsymbol{f}_m^{(2)} \boldsymbol{h}_2 = 1$ for objects of class 2, and $1 \le m \le N_1 + N_2$. Matrix-vector solution for a_m and b_m of eq. (2) is determined as

$$\boldsymbol{a} = R_{12}^{-1} \boldsymbol{u}_1, \quad \boldsymbol{b} = R_{12}^{-1} \boldsymbol{u}_2, \tag{3}$$

where R_{12} is correlation matrix with dimensions $(N_1 + N_2) \times (N_1 + N_2)$ of total training set and

$$\boldsymbol{u}_1 = [1, \ldots, 1; 0, \ldots, 0]$$
$$\boldsymbol{u}_2 = [0, \ldots, 0; 1, \ldots, 1]$$

Equal correlation peak synthetic discriminant function ECP SDF) [9] $\boldsymbol{h}$ satisfies matrix equation

$$\mathbf{A}\boldsymbol{h} = \boldsymbol{c}, \tag{4}$$

where $\mathbf{A} \in R^{m\times n}$ is matrix with rows $\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_m^T$, and $\boldsymbol{c}$ is a column vector with each entry of desired value c_0, $\boldsymbol{h} \in R^{n\times 1}$ is a linear filter. Mostly in practice equation (4) is underdetermined and has multiple solution. Pseudoinverse solution of (4), given by

$$\boldsymbol{h} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\boldsymbol{c} \tag{5}$$

exists and is unique, when rows of $\mathbf{A}$ are linear independent and filter $\boldsymbol{h}$ is constructed as linear combination of input images. ECP SDF can be adapted for two-class recognition problem (signatures-forgeries) by incorporating vectors $\tilde{\boldsymbol{a}}_j,\ j = 1, \ldots, \tilde{m}$ from forgery training set into matrix $\mathbf{A}$, each correlates with filter $\boldsymbol{h}$ producing new desired value $\tilde{c}_0$.

Now consider the *fast Hadamard transform (HT)* and show its effectiveness for data compression and selection of image features. Two-dimensional HT can be performed as two-cascade one-dimensional transform

$$F(u,y) = \sum_{x=0}^{N-1} f(x,y)(-1)^{q(x,u)} \tag{6}$$

$$F(u,v) = \sum_{y=0}^{N-1} F(u,y)(-1)^{q(y,v)} \tag{7}$$

The Hadamard transform just as the Fourier transform preserve the energy of initial image and tends to decorrelation of image features. The fast Hadamard transform is less effective than FFT or Karhunen-Loeve Transform (KLT) in the sense of mean square error, nevertheless it is more effective in speed.

The procedure below represents the modification of HT called BIFORE transform [8], which results in translation invariant compressed power spectrum. In the case of truncated Hadamard transform spectral coefficients can be calculated as follows:

$$P(0) = F^2(0), \tag{8}$$

$$P(i) = \sum_{j=2^i-1}^{N-1} [F^2(j) + F^2(j+1)], \quad i = 1, 2, \ldots, n, \tag{9}$$

where j is changed by steps of 2^{i+1}, $n = \log_2 N$.

The techniques described above were used in handwritten signature identification system to perform the most important stages of recognition problem, i.e. data compression and classification.

Positional spectrum coefficients of two-dimensional BIFORE transform were considered as image features and were used as input data for signature recognition system.

Reduction of dimensionality was carried out by computing basic spectral coefficients of 2-D BIFORE transform. Compressing in series rows and columns of input image matrix X_i, new matrix Y_i of reduced size was obtained as follows

$$X_i \xrightarrow{T} Y_i \tag{10}$$

where $X_i \in R^{N\times N}$ is $N \times N$ image matrix, Y_i belongs to $R^{n\times n}$ is matrix of spectral coefficients, $n = \log_2 N + 1$, $i = 1, 2, \ldots, m$ and T denotes transform operator. Matrix Y_i is then converted into $(n \times n) \times 1$ column vector $\boldsymbol{a}_i$. Obtained feature vectors $\boldsymbol{a}_i$ were then used for training and (or) verification. Another type of transform we offer is the connectivity spectrum [10] because of its simple realization and ability to reduce original image. Value of connectivity for every image pixel can be calculated by analyzing its neighbourhood. In rectangular image raster every point is rounded by 8 neighbours so that one byte is necessary to encode the value of connectivity. Every neighbour puts its own value in corresponding bit position of connectivity byte, resulted in final value in the range from 0 to 255. This value obviously depends on starting point and the direction of the path. Calculating the connectivity value for all units we can form the distribution (histogram) of image pixels, which represents the connectivity spectrum of input image. The spectra obtained are then considered as 256×1 feature vectors to be used as input data for pattern recognition system.

Minimum distance classifiers based on Euclidean metrics and trained for one- and two-class problem in the task "originals-forgeries", and multi-class problem in the task of handwritten symbols (numerals) recognition problem have been used in experiments.

To construct classifier with one training set the mean vector $\boldsymbol{a}$ of a cluster center and maximal radius R_{max} of decision boundary are calculated as follows

$$\boldsymbol{a} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{a}_i, \tag{11}$$

$$R_{max} = \max\left\{\sqrt{(\boldsymbol{a}_i - \boldsymbol{a})^2}, \quad i = 1, \ldots, m\right\} \tag{12}$$

In the case of multiple training sets minimum distance criterion was used. m mean vectors for m classes were calculated using formula (11). m distances to cluster centers were then computed for every testing vectors $\boldsymbol{a}_t$ as

$$R_{it} = \sqrt{(\boldsymbol{a}_t - \boldsymbol{a}^{(i)})^2}, \quad i = 1, 2, \ldots, m. \tag{13}$$

Testing image is considered to be a member of class i if

$$R_{it} < R_{jt}, \quad i \neq j. \tag{14}$$

4 Implementation

The base volume of computations consists of calculation the scalar product of two vectors (eqs. 2, 6, 7). Besides, the preliminary processing the input images such the spatial filtering can also be decomposed on series of the scalar product calculation. So we significantly improve the performance of the whole system when the parallel hardware to fulfil these calculations is used.

We have applied the multilevel approach to implement all the algorithms described above. The high level of the computations is represented by soft programmed algorithm logic and effectively organized with widely used PC-based platform. The low level of the computations is implemented using programmable hardware such the FPGA. Our FPGA equipped device is capable to interact with host using slave PCI 32/33 bus interface and we have used it to implement low level homogeneous computaions.

The spatial filtering algorithm is the convolution of the source image with the spatial mask h in the vicinity of every point x_i of result image using formula $(1/w_h)\sum_k x_{ik}h_k$. Here w_h is sum of all the values h_k or square root from sum of squares. Index i runs throw all the points of result image.

The spatial scaling is also important process to reduce computational costs and it is formally the same computational process except for different dimensions of source and result images.

The division by w_h used to normalize result is sequential operations in general sense but divisor w_h is the constant for the filter mask and we can take rational representation of inverse value $1/w_h$ to perform result divisions using multiplication. For this purpose the Barret fast division algorithm [11] has been used. To calculate the correlation it is necessary to fulfil division by non-constant values formed from source data flow. To solve the problem we have used fast algorithm to implement division using addition and multiplication [12].

The computational core of hardware to implement spatial $n \times n$ filtering algorithms is organized as the parallel set of pipelines each one consists of n processing element (PE). Every PE (Fig. 1) contains a multiplier, bypath channel, memory cell array, control logic, and is capable to fulfil small set of operations: set memory cell to 1, transit data from input O to W output, load data from input O to memory cell M and transit it to output W, load data from input O, transit it to output W, multiply that data with contents of memory cell M, and set result on S output lines, transit data from input O both to W and S outputs. Chain of the PE forms the pipeline as shown on Fig. 1.

Image row sequentially enters into pipeline with memory cells loaded with filter mask values. All the PEs make mutiplications with the every clock pulse, then products are added in adder S and the result can be used as partial scalar product right now. Additional group of adders was used to complete scalar product for spatial filtering using 3×3 mask (Fig. 2).

Result data presents the strip with noncomplete boundary rows. To correct boundary effects the special input and complementary output pathes A, B are made.

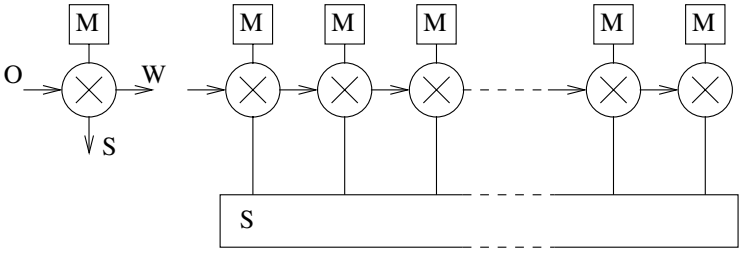


Fig. 1. Processing element and pipeline core organization

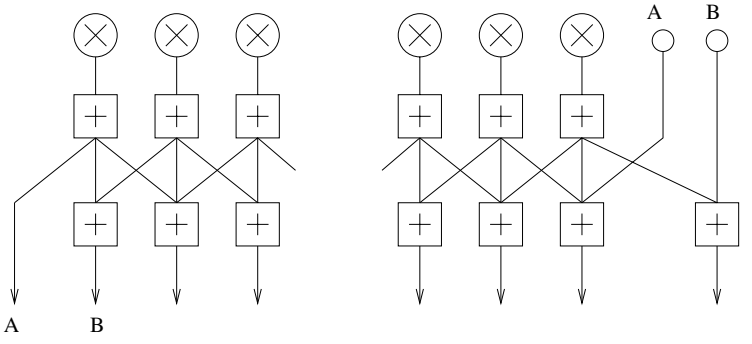


Fig. 2. Pipeline interconnections scehme to provide parallel filtering with $3 \times s$ mask.

The software consists of hardware driver to provide all the communications between host kernel and PCI co-processor device. The next software level is formed by functional call library based on raw driver calls. That library provides application interface and can be replaced with another one implementing software core to make software development and debugging more powerful and convenient.

All of these was used to create high performance classifier. The add-in board based on FPGA with 5 maximum pipelines supports calculation speeds of up to 20 million patterns per second. The board accepts input vectors with a maximum of 32 feature dimensions, each of 12 levels of binary resolution, and it outputs up to 256 classes or distances. High-speed parallel processing unit computes Euclidean/CityBlock distance between an input vector and up to $256 \times 32 = 8194$ stored prototypical examples.

Pattern recognition is a process of sorting input data into categories or classes that are significant to the user. The differences or underlying traits of each class must first be loaded into the chip's memory. The content of the chip's memory can be developed manually or extracted from examples of data typical to the problem, using a learning algorithm. Input data is problem-specific and may consist partially or completely of stored data. Once learning is complete, the system is ready to classify input data. Coprocessor algorithm consists of two stages. At first, initialization and tuning procedures are required, providing

classifier data upload and special parameter setup. On the second stage input object classification procedures can be carried out.

During coprocessor design we took account of modern tendencies in the area of developing high performance digital equipment. An approach with extensive use of high parallelism and pipelining was employed to increase processing speed. Calculation unit of the co-processor is composed of five parallel pipelines. It enables to classify up to five different vectors simultaneously. Classification time depends on the classifier extent, i.e. on the number of classes in current recognition task and the number of vectors which represent these classes. Each pipe is able to calculate the distance between two 64-dimensional vectors by 16 clocks (cycles). Thus, single pipe with 66 MHz can calculate 4 million distances per second. All the five pipes produce 20 million distances per second. The benefit of the proposed approach is that it is possible to further increase the number of vectors to be processed. The amount of processing data depends on the extent of FPGA in use. Another advantage is the ability to increase the dimensionality of classifier (number of classes or class description or data resolution) without reduction of system performance.

5 Experimental Results and Conclusion

Eleven valid signatures of 37 persons each represented in database by the collection of 21 originals were used for training in both 1- and 2-class formulations for signature verification problem. Forgeries, 25-30 per object class were carefully prepared by students with definite portion of patience. Eleven forgeries from every object class were used for training in 2- class formulation. Full database of originals and forgeries was subjected to verification. A single bipolar SDF was used, so that classification results were mapped onto one numeric axis, which represents the inner product of SDF filter and image feature vector to be classified. The system demonstrates reliable classification results providing the equal error rate in the range from 3.9% to 9.9%. In the case of handwritten character recognition (arabic numerals) data base of 4000 symbols was prepared by 14 students with some limiting rules - we offer to write digits using 10 standardized patterns. 50 symbols of every class were used as training. The total error rate of 3.3% was obtained, varying between 0.6% and 4.9%.

Acceleration achieved with the use of FPGA is more clearly demonstrated in the task of handwritten symbol recognition when database is complete enough and processing time is one of the major parameters of system performance.

Classifier has been hardware implemented as co-processor on PCI card containing FPGA Xilinx Virtex XCV400-4. Because of FPGA-based implementation of PCI controller and its functional restrictions (slave 32 bit/33 MHz) real data transfer rate was limited with 50 Mb/s to and 12Mb/s from device. Therefore to raise classification performance using our weak FPGA we have put 4 pipeline each calculates up to $4 \cdot 10^6$ scalar multiplications of 64-component vectors per second. Data transfer to and from device has fully combined in time with data processing within co-processor unit. Thereby classification performance using

parallel pipelined co-processor at application layer has achieved up to $7.5 \cdot 10^3$ symbols per second. Number of classes is equal to 256 with cluster size of 4 vectors. Thereagainst software implementation of classification has realised about 750 symbols per second at two-processor 2.4 GHz Xeon PC. To achieve that results amount of data to be returned from device has been reduced down to 4 estimations per feature vector.

Employing more powerfull up-to-date FPGA one can raise performance of classification up to $2.5 \cdot 10^5$ symbols per second using PCI 32/33 slave communication. In principle highest possible performance of classification using FPGA is limited with interface data transfer rate only.

References

1. Kussul, E., Baidyk, T.: Improved Method of Handwritten digit recognition. In: Proceedings of the 15-th International Conference on Vision Interface, Canada, Calgary, (2002) 192-197.
2. Akiyama, T., Nagita, N.: Automated entry system for printed documents. Pattern Recognition, v.23, N 11, (1990) 1130-1141.
3. Yousefi, H.Al., Udupa, S.S.: Recognition of Arabic Characters. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.14, N 8, (1992) 853-860.
4. Aguado, A.: Parametrizing arbitrary shapes via Fourier descriptors for evidence-gathering extraction. Computer Vision and Image Understanding, v.69, N 2, (1998) 202-219.
5. Mukundan, R.: Fast Computation of Legendre and Zernike moments. Pattern Recognition, v.28, N 9, (1995) 1433-1442.
6. Wallin, A.: Complete sets of a complex Zernike moments invariants and the role of the pseudoinvariants. IEEE Transactions on Pattern Analysis and Machine Intelligence, v.17, N 11, (1995) 1106-1114.
7. Parker, J.R.: Simple Distances between handwritten signature. Procedeengs of the 15-th International Conference on Vision Interface, Canada, Calgary, (2002) 218-222.
8. Ahmed, N., Rao, K.: Orthogonal Transforms for Digital Image Processing. Springer-Verlag, Berlin-Hedelberg-New-York, (1975).
9. Casasent, D.: Unified synthetic discriminant function computational formulation. Appl. Opt., 23, (1984) 1620-1627.
10. Samokhval, V.A., Sadykhov, R.H.: The Use of Connectivity Spectrum for Signature Verification. In: Proceedings of the 3-d International Conference "Pattern Recognition and Image Analysis", Minsk, **3**, (1995) 43-46.
11. Barrett, P.: Implementing the Rivest, Shamir and Adleman public-key encryption algorithm on a standard digital signal processor. // In A. M. Odlyzko, editor, Advances in cryptology: CRYPTO'86: proceedings, volume 263 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, Germany, (1987) 311-323.
12. Hitz, M.A., Kaltofen, E.: Integer Division in Residue Number Systems. In: IEEE Trans. on Computers, Vol.44 N 8, (1995) 983-989.

Reconstruction of Order Parameters Based on Immunity Clonal Strategy for Image Classification

Xiuli Ma and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University,
710071 Xi'an, China
Key Lab for Radar Signal Processing, Xidian University,
710071 Xi'an, China
Lijianer@163.com, Lchjiao@Xidian.edu.cn

Abstract. A novel reconstruction algorithm of order parameters based on Immunity Clonal Strategy (ICS) is presented in this paper, which combines the self-learning ability of Synergetic Neural Network (SNN) with the global searching performance of ICS to construct linear transform and then realize reconstruction. Compared with the reconstruction method based on Genetic Algorithm (GA), the new method not only overcomes the aimless and random searching of GA at the later time of searching but also improves its searching efficiency greatly. The tests on IRIS data and Brodatz texture show that the proposed method can positively find a new set of reconstruction parameters and enhance the classification accuracy rate remarkably.

1 Introduction

Synergetics studies features of spontaneous variation based on spatial structure, temporal structure and functional structure generated by self-organization in complicated systems. In the late 1980s, Haken proposed to put synergetic theory into the area of pattern recognition and introduced a new viewpoint that synergetics takes the recognition process as a pattern formation process [1]. Hence, the application of synergetics in image processing and recognition is a rising field. In the 1990s, he presented a new theory on neural networks, namely, Synergetic Neural Network (SNN). Compared with other traditional neural networks, SNN is constructed from top to down and its remarkable characteristic is not having pseudo-state. In the recent years, its learning algorithms are widely studied especially on the selection of prototype pattern vector, the setting of attention parameter and its invariant properties and so on.

On the selection of prototype pattern vector, Haken proposed to select any sample from each class as prototype pattern vector. Wanger and Boebel made use of SCAP algorithm [2] by averaging training samples simply. Wang *et al* took cluster center obtained from C-Means clustering algorithm as prototype pattern vector. Then he proposed to the learning algorithm based on LAIS [3]. The above methods improve the classification performance of SNN to a certain extent but order parameters got by

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 455–462, 2004.
© Springer-Verlag Berlin Heidelberg 2004

these methods are unreasonable at some degree. Hu discovered this first and proposed the reconstruction of order parameters [4] in concept. He expected to control order parameters' behavior by linear transform on status vector q, namely, by changing the proportion of order parameters. It is a pity that he did not propose a concrete and effective method to construct this linear transform. Wang introduced the reconstruction algorithm based on award-penalty learning mechanism [3]. This method classifies every sample at each iteration thus the training time is long. Then he put forward to realize this linear transform by combining SNN with GA, but it is unreasonable to confine reconstruction parameter γ_k to [0,5] or [0,8] in paper [3]. Moreover, GA is random and aimless at the later time of searching.

Considering that, the reconstruction algorithm based on Immunity Clonal Strategy (ICS) is presented in this paper. The new method combines the self-learning ability of SNN with the global searching performance of ICS to train reconstruction parameters and improves the searching efficiency greatly.

2 Review of Synergetic Pattern Recognition

Haken's synergetic approach [1] to pattern recognition is based on a parallel between pattern recognition and pattern formation. The recognition procedure can be viewed as the competition process of many order parameters. Supposed that the prototype pattern vectors are M and the status vectors N, M is less than N for the sake of linear independence of prototype pattern vectors. A dynamic equation proposed by Haken can be given as follows:

$$\dot{q} = \sum_{k=1}^{M} \lambda_k (v_k^+ q) v_k - B \sum_{k \neq k'} v_k (v_{k'}^+ q)^2 (v_k^+ q) - Cq(q^+ q) + F(t) \ , \tag{1}$$

which is applicable to pattern recognition. Where q is the status vector of input pattern with initial value q_0, λ_k is attention parameter, v_k is prototype pattern vector, and v_k^+ is the adjoint vector of v_k that satisfies:

$$(v_k^+, v_{k'}) = v_k^+ v_{k'} = \delta_{kk'} \ . \tag{2}$$

Order parameter is

$$\xi_k = (v_k^+, q) = v_k^+ q \ . \tag{3}$$

Corresponding dynamic equation of order parameters is

$$\dot{\xi} = \lambda_k \xi_k - B \sum_{k' \neq k} \xi_{k'}^2 \xi_k - C(\sum_{k'=1}^{M} \xi_{k'}^2) \xi_k \ . \tag{4}$$

The strongest order parameter will win by competition and desired pattern will be recognized.

3 Reconstruction of Order Parameters Based on Immunity Clonal Strategy

3.1 Immunity Clonal Strategy

In order to enhance the diversity of the population in GA and avoid prematurity, Immunity Clonal Strategy [5] is proposed by Du *et al*. The clonal operator is an antibody random map induced by the affinity including: clone, clonal mutation and clonal selection. The state transfer of antibody population is denoted as

$$C_{MA}: A(k) \xrightarrow{clone} A'(k) \xrightarrow{mutation} A''(k) \xrightarrow{selection} A(k+1) .$$

Here antibody, antigen, the affinity between antibody and antigen are similar to the definitions of the objective function and restrictive condition, the possible solution, match between solution and the fitting function in AIS respectively. According to the affinity function $f(*)$, a point $a_i = \{x_1, x_2, \cdots, x_m\}$, $a_i(k) \in A(k)$ in the solution space will be divided into q_i different points $a_i'(k) \in A'(k)$, by using clonal operator, a new antibody population is attained after performing clonal mutation and clonal selection. It is easy to find that the essential of the clonal operator is producing a variation population around the parents according to their affinity. Then the searching area is enlarged. The ICS algorithm is shown as follows. In the following description, Φ is a function of f generally and N is the scale of the antibody population.

Immunity Clonal Strategy Algorithm

$k = 0$;
Initialize the antibody population: $A(0) = \{a_1(0), a_2(0), \Lambda, a_N(0)\} \in I^N$;
Calculate the affinity: $A(0): \{f(A(0))\} = \{\Phi(a_1(0)), \Phi(a_2(0)), \Lambda, \Phi(a_N(0))\}$;
When the halt condition is not satisfied, do
 Clone T_c^C: $A'(k) = T_c^C(A(k)) = [T_c^C(a_1(k)), T_c^C(a_2(k)), \Lambda, T_c^C(a_N(k))]^T$;
 Clonal mutation T_m^C: $A''(k) = T_m^C(A'(k))$;
 Calculate the affinity: $A''(k): \{\Phi(A''(k))\}$;
 Clonal selection T_s^C: $A(k+1) = T_s^C(A''(k))$;
 Calculate the affinity:
$A(k+1): \{\Phi(A(k+1))\} = \{\Phi(a_1(k+1), \Phi(a_2(k+1)), \cdots, \Phi(a_N(k+1)\}$;
 $k = k+1$;
End.

3.2 Reconstruction of Order Parameters

The construction of order parameters is unreasonable at some degree in Haken model, which had been proved in paper [3]. In order to make recognition process reasonably

reflect the relation among patterns, it is necessary to reconstruct order parameters. First, supposing $T = P\Gamma P^{+}$, where $\Gamma = diag\{\gamma_1, \gamma_2, \cdots, \gamma_M\}$, we can get the following equations by the image of linear transform T and status vector q:

$$Tq = \sum_{k=1}^{M} \xi_k \gamma_k v_k \ , \ v_k^{+} \bullet Tq = \xi_k \gamma_k \,. \tag{5}$$

In fact, the linear transform to q changes the proportion of patterns, therefore it can provide the means to control order parameters by selecting $(\gamma_1, \gamma_2, \cdots, \gamma_M)$. At this time, we name the status vector transformed as $\overline{q}$ and order parameter as $\overline{\xi}_k$, where $\overline{q} = Tq$ and $\overline{\xi_k} = v_k^{+} \bullet Tq$. In this way, the dynamic equation defined by $\overline{q}$ and $\overline{\xi}_k$ is the same as by q and ξ.

3.3 Generalization of the Reconstruction of Order Parameters

The reconstruction of order parameters has been introduced above, but its form is not universal, which is defined by the formulae $T = P\Gamma P^{+}$ and $\Gamma = diag\{\gamma_1, \gamma_2, \cdots, \gamma_M\}$. When Γ is

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{bmatrix}, \tag{6}$$

which is $M \times M$, the reconstruction of order parameters defined by $T = P\Gamma P^{+}$ is most general. In case of $\overline{q} = Tq$ and $\overline{\xi_k} = v_k^{+} \bullet Tq$, we can get the same dynamic equation as the original one, which has been proved in paper [3].

It is equal to add a new order parameter layer to SNN by introducing the reconstruction of order parameters. By the generalization of Γ, the single joint from order parameter layer to new order parameter layer is transformed into full joint. From neural network theory, we can see that the performance of SNN generalized is better than before. So the generalization of the reconstruction of order parameters is reasonable and effective to improve the performance of SNN.

The most necessary and important problem to the reconstruction of order parameters is how to get reconstruction parameters $(\gamma_{11}, \cdots \gamma_{1M}, \gamma_{21}, \cdots, \gamma_{2M}, \cdots \gamma_{M1}, \cdots, \gamma_{MM})$ and $(\gamma_1, \gamma_2, \cdots, \gamma_M)$. The reconstruction method based on award-penalty learning mechanism [3] has a long training time because it needs to classify every sample at each iteration. In the worse case, it is very difficult to get the reconstruction parameters to classify every training sample correctly if training samples of different kinds are classified mutually incorrectly. The reconstruction method based on GA [3] uses the global searching of GA to get better parameters in reconstruction parameters

space. But GA is easily premature and is aimless and random at the latter time of searching. Moreover, it is local and unreasonable to confine searching space to $[0,5]$ or $[0,8]$.

Considering that, a new reconstruction method based on ICS algorithm is presented in this paper. ICS has the performance of global convergence and local optimization so it is better than GA no matter on convergent speed or on searching efficiency.

3.4 Reconstruction of Order Parameters Based on Immunity Clonal Strategy

The reconstruction of order parameters based on ICS realizes reconstruction by ICS training reconstruction parameters $(\gamma_{11},\cdots\gamma_{1M},\gamma_{21},\cdots,\gamma_{2M},\cdots\gamma_{M1},\cdots,\gamma_{MM})$ and $(\gamma_1,\gamma_2,\cdots,\gamma_M)$. Here, the affinity between antibody and antigen is denoted by the classification accuracy rate of training samples, clonal scale is determined by the affinity or a constant and the optimal individual is preserved during training process. The halt conditions are defined as the restricted iterative number or the classification accuracy rate or the two methods blending. At this time, the preserved individual during training process is the best reconstruction parameters. The detailed algorithm is described as follows.

Step 1: Initialize the antibody population.

Initialize the antibody population $A(0)$ size of N randomly. Every individual represents a set of reconstruction parameters.

Step 2: Calculate the affinity.

The affinity is donated by the classification accuracy rate of training samples.

Step 3: Clone.

Clone every individual in the *k*th parent $A(k)$ to produce $A'(k)$. The clonal scale is a constant or determined by the affinity.

Step 4: Clonal mutation.

Mutate $A'(k)$ to get $A''(k)$ in probability of p_m.

Step 5: Calculate the affinity.

Calculate the affinity of every individual in new population $A''(k)$.

Step 6: Clonal selection.

Select the best individual, which has better affinity than its parent, into the new parent population $A(k+1)$.

Step 7: Calculate the affinity.

Calculate the affinity of every individual in new population $A(k+1)$.

Step 8: Judge the halt conditions reached or not.

The halt conditions can be restricted iterative number or the classification accuracy rate of training samples or the two methods blending. If the searching algorithm reaches the halt conditions, terminate the iteration and make the preserved individual as the best reconstruction parameters, or else preserve the best individual in current iteration and then turn its steps to Step 3.

4 Experiments

4.1 Classification of IRIS Data

IRIS data is selected to test the performance of the method proposed in this paper. This data set has 150 data, which is made up of 3 classes, and each datum consists of 4 attributes. We randomly select 16 data from every class as training set and others as testing set.

In the experiment, SCAP algorithm [2] is used to get prototype pattern vector. The parameters in ICS are defined empirically as: the size of initial population is 6 and mutation probability is 0.2. The clonal scale is determined by the affinity. In GA, the size of initial population is 20, crossover probability 0.9 and mutation probability 0.1 respectively. GA operators are fitness proportional model selection operator, arithmetic crossover operator and un-uniform mutation operator. The halt condition is the classification accuracy rate of training samples up to 100%. The average statistical results of 20 trails are shown in Table 1.

Table 1. Comparative results between the reconstruction method based on GA and on ICS

	Before reconstruction	Reconstruction based on GA	Reconstruction method proposed
Training time (s)	0	26.6141	6.5058
Testing time (s)	0.016	0.01	0.014
Classification rate of training set (%)	81.25	100	100
Classification rate of testing set (%)	78.431	93.532	95.2939

From Table 1 we can see that the reconstruction method based on ICS has not only a shorter training time but also a higher classification accuracy rate than on GA. Moreover, it is very important to select prototype pattern vector, which is crucial to the classification performance of SNN. Because we put emphasis particularly on the study of the reconstruction of order parameters, SCAP is selected. SCAP is simple and rapid for it gets prototype pattern vector by averaging the training samples simply, so it has an effect on the classification accuracy.

4.2 Classification of Brodatz Texture by Brushlets Features

Edges and textures in an image can exist at all possible locations, orientations and scales. In order to obtain a better angular resolution we expand the Fourier plane into windowed Fourier bases thus resulting in an expansion of the image into a set of brushlets [6]. A brushlet is a function reasonably well localized with only one peak in frequency. Furthermore, it is a complex valued function with a phase. The phase of

the bi-dimensional brushlet provides valuable information about its orientation. We can adaptively select the size and location of the brushlets in order to obtain the most concise and precise representation of an image in terms of oriented textures with all possible directions, frequencies, and locations.

16 similar textural images chosen from Brodatz are shown in Fig. 1. Every image size of 640×640 is segmented into 25 nonoverlapping images as a class. Then the samples are 16 kinds in all and each class has 25 images. We randomly select 8 training data and 17 testing of every class respectively. At the same time, Burshlets is used to decompose each image into three layers.

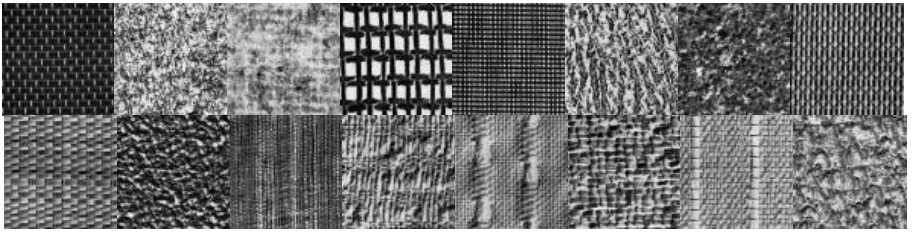


Fig. 1. 16 kinds of Brodatz textural images. These images from left to right and up to down in sequence are D006, D009, D019, D020, D021, D024, D029, D053, D055, D057, D078, D080, D083, D084, D085, and D092 the same as in paper [7].

In this experiment, SCAP [2] is selected to get prototype pattern vector. Supposing empirically the initial population is 5 and mutation probability is 0.1. The clonal scale is determined by the affinity in ICS. In GA, the initial population is 10, crossover probability 0.9 and mutation probability 0.1. The GA operators are fitness proportional model selection operator, arithmetic crossover operator and un-uniform mutation operator. The halt condition is the iteration number up to 200 or the classification accuracy rate more than 99%. The average statistical results of 20 trials are shown in Table 2.

Table 2. Comparative results between the reconstruction method based on GA and on ICS

	Before reconstruction	Reconstruction based on GA	Reconstruction method proposed
Training time(s)	0	4.8827	1.8894
Test time(s)	0.062	0.0655	0.0686
Classification rate of training samples (%)	92.969	98.9065	99.219
Classification rate of testing samples (%)	94.853	96.8017	97.1691

From Table 2 we can see the reconstruction method based on ICS has not only a shorter training time but also a higher classification accuracy rate. The *optimal* recon-

struction parameters obtained from experiments of reconstruction based on ICS are shown in Table 3. At this time, the classification accuracy rate is as high as *98.897%*.

Table 3. Optimal reconstruction parameters

R1	R2	R3	R4	R5	R6	R7	R8
2.6164	2.1079	1.6545	2.9523	3.0517	1.4960	0.7879	2.0735
R9	R10	R11	R12	R13	R14	R15	R16
2.1267	1.7913	1.8909	1.8941	1.1282	1.5021	2.5771	0.9773

5 Conclusions

A novel reconstruction algorithm of order parameters based on Immunity Clonal Strategy (ICS) is presented in this paper, which combines the self-learning ability of Synergetic Neural Network (SNN) with the global searching performance of ICS to train reconstruction parameters. In comparison with the reconstruction method based on Genetic Algorithm (GA), the new method not only overcomes the aimless and random searching of GA at the later time of searching but also improves its searching efficiency greatly. The tests on IRIS data set and Brodatz texture show that the proposed method can positively find a new set of reconstruction parameters and enhance the classification accuracy rate remarkably. Moreover, it is very important to select prototype pattern vector, which is crucial to the classification performance of SNN. Because we put emphasis particularly on the study of the reconstruction of order parameters, SCAP is selected. SCAP is simple and rapid for it gets prototype pattern vector by averaging the training samples simply. As a result, it has an effect on the classification accuracy.

References

1. Haken, H.: Synergetic Computers and Cognition–A Top-Down Approach to Neural Nets. Springer-Verlag, Berlin (1991)
2. Wagner, T., Boebel, F.G.: Testing Synergetic Algorithms with Industrial Classification Problems. Neural Networks. 7 (1994) 1313–1321
3. Wang, H.L.: The Research of Application of Image Recognition Using Synergetic Neural Network. Ph.D. Dissertation. Shanghai Jiao Tong University, China (2000)
4. Hu, D.L., Qi, F.H.: Reconstruction of Order Parameters in Synergetics Approach to Pattern Recognition. J. Infrared Millim. Waves. 7 (1998) 177–181
5. Jiao, L.C., Du, H.F.: Development and Prospect of The Artificial Immune System. Acta Electronica Sinica. 31 (2003) 1540–1548
6. Meyer, F.G., Coifman, R.R.: Brushlets: A Tool for Directional Image Analysis and Image Compression. Applied and Computational Harmonic Analysis. 4 (1997) 147–187
7. Huang, Y., Chan, K.L.: Multi-model Feature Integration for Texture Classification. The 5th Asian Conference on Computer Vision. Melbourne, Australia (2002) 23–25

Visual Object Recognition Through One-Class Learning

QingHua Wang[1], Luís Seabra Lopes[1], and David M. J. Tax[2]

[1] IEETA/Department of Electronics & Telecommunication,
University of Aveiro,
Campus Santiago, 3810-153, Aveiro, Portugal
qhwang@ieeta.pt, lsl@det.ua.pt
[2]Faculty of Information Technology and Systems
Delft University of Technology
P.O. Box 5031, 2600GA,Delft, The Netherlands
d.m.j.tax@ewi.tudelft.nl

Abstract. In this paper, several one-class classification methods are investigated in pixel space and PCA (Principal component Analysis) subspace having in mind the need of finding suitable learning and classification methods to support natural language grounding in the context of Human-Robot Interaction. Face and non-face classification is used as an example to demonstrate effectiveness of these one-class classifiers. The idea is to train target class models with only target (face) patterns, but still keeping good discrimination over outlier (never seen non-target) patterns. Some discussion is given and promising results are reported.

1 Introduction

Let's consider the task of teaching a robot to recognize an object, say, "apple", through its camera, in the context of Human-Robot Interaction (HRI). How can the teaching be conducted? To apply state-of-the-art statistical approaches, e.g., Hidden Markov models [6, 22], Bayesian networks [11], naïve Bayes classifier [14], PCA [18], and other methods described in [20], basically it's necessary to find quite a lot of apples, and to find enough non-apples, which is itself an ambiguous concept, to estimate the class distributions precisely. One might wonder whether these requirements are realistic in the context of HRI. The fact that learning is supervised and the teaching is interactive typically leads to the availability of only a small number of samples. This makes the conventional methods mentioned above not applicable as they require to prepare both target and non-target patterns. Thus, it might be useful to construct classifiers based on only target class patterns but still having good discrimination for never seen non-target patterns.

Following this idea, a method based on the combination of the wavelet domain Hidden Markov Trees (HMTs) and Kullback-Leibler distance (KLD) was proposed in [19]. In that method, only target (face) samples were used to train an object model in terms of parameters of HMTs. Then for each unknown pattern, its KLD to this model

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 463–470, 2004.
© Springer-Verlag Berlin Heidelberg 2004

is computed. If its KLD is smaller than a certain threshold, obtained from training session, it is recognized as a target pattern; otherwise, it is rejected. One problem of this HMT/KLD based approach is that it can't derive robust class models if there are big in-class variations among the training patterns. One cause is that simply the average of individual HMTs is used to attain the overall class model. In that way, if individual HMTs vary greatly from each other, the average method loses precision of HMT parameter estimation. In this paper, several one-class classification methods, previously described in [17], are investigated to solve this problem.

The rest of this paper is organized as follows. The brief review of one-class classification is provided in section 2. In section 3, the experimental setup and results are presented. Conclusion is given in section 4 with some discussion and future work.

2 One-Class Classifiers

The design of one-class classifiers is motivated by the fact that patterns from a same class usually cluster regularly together, while patterns from other classes scatter in feature space. One-class learning and classification was first presented in [7], but similar ideas had also appeared, including outlier detection [12], novelty detection [2], concept learning in the absence of counter-examples [5] and positive-only learning [9]. Generally, in multi-class approaches, one can precisely capture class descriptions through the availability of samples from all classes. In contrast, in one-class approaches, only samples of the target class are required. A very natural method for decision-making under this condition is to use some distance-based criterion. If the measurement of an unknown pattern x is smaller than the learned threshold, it can be accepted as the target class pattern; otherwise, it should be rejected. This can be formulated as follows.

$$Class(x) = \begin{cases} target, & if \quad Measurement(x) \le threshold \ ; \\ non-target, & otherwise \ . \end{cases} \tag{1}$$

It's comparable to the Bayesian decision rule. The main difference is that, here, the threshold is learned only from target class patterns, while in Bayesian decision rule it's determined by both target and non-target class patterns. If an appropriate model of the target class (and thus a proper threshold) is found, one can find that most patterns from this target class are accepted and most non-target class patterns are rejected. Surely the ideal model is one that can accept all target patterns and reject all non-target patterns. But this is usually not easy to find realistically. Common practice is to define *a priori* the fraction of training target patterns that should be discarded (known as reject rate), in order to obtain a compact data description and minimize false positives. In many cases 5% or 1% is used.

Several methods were proposed to construct one-class classification models. A simple method is to generate artificial outlier data [13], and conventional two-class approaches are thus applicable. This method severely depends on the quality of artificial data and often does not work well. Some statistical methods were also proposed. One can estimate the density or distribution of the target class, e.g., using Parzen den-

sity estimator [2], Gaussian [9], multimodal density models [21] or wavelet-domain HMTs [19]. The requirement of well-sampled training data to precisely capture the density distribution makes this type of methods problematic. In [7, 17] some boundary-based methods were proposed to avoid density estimation of small or not well-sampled training data. But a well-chosen distance or threshold is needed. Tax provides a systematic description of one-class classification in [17], where the decision criteria are mostly based on the Euclidean distance. Below is a brief description of seven one-class classifiers previously described in [17] and [10].

The Support Vector Data Description (SV-DD) method, proposed in [17], basically finds a hypersphere boundary around the target class with minimal volume containing all or most of the target class patterns. It can provide excellent results when a suitable kernel is used. Currently, the Gaussian kernel is chosen. It is possible to optimize the method to reject a pre-defined fraction of the target data in order to obtain a good and compact data description of it (thus some remote target data points may be discarded). Thus for different rejection rates, the shape of the boundary changes. For classification, objects outside this sphere decision boundary are regarded as outliers (objects from other classes). The main drawback of the method is that it requires a difficult quadratic optimization.

Another method, GAUSS-DD, models the target class as a simple Gaussian distribution. To avoid numerical instabilities, the density estimate is avoided, and just the Mahalanobis distance $f(x) = (x-\mu)^T \Sigma^{-1} (x-\mu)$ is used, where mean μ and covariance matrix Σ are sample estimates. The classifier can be defined by (1).

In KMEANS-DD, a class is described by k clusters, placed such that the average distance to a cluster center is minimized. The cluster centers $\mathbf{c}_i$ are placed using the standard k-means clustering procedure [2]. The target class is then characterized by $f(x) = \min_i (x - c_i)^2$. The classifier then is defined as in (1).

The PCA-DD method, based on Principal Component Analysis, describes the target data by a linear subspace. This subspace is defined by the eigenvectors of the data covariance matrix Σ. Only k eigenvectors are used, which are stored in a $d \times k$ matrix $\mathbf{W}$ (where d is the dimensionality of the original feature space). To check if a new object fits the target subspace, the reconstruction error is computed. The reconstruction error is the difference between the original object and the projection of that object onto the subspace (in the original data). This projection is computed by:

$$x_{proj} = W(W^T W)^{-1} W\mathbf{x} \tag{2}$$

The reconstruction error is then given by $f(x) = \| x - x_{proj} \|^2$.

The NN-DD method is a simple nearest neighbor method. Here, a new object x is evaluated by computing the distance to its nearest neighbor $NN(x)$ in the training set. This distance is normalized by the distance between its nearest neighbor, $NN(x)$, and the nearest neighbor of $NN(x)$ in training set, $NN(NN(x))$.

The KNN-DD is a k-nearest neighbor method. In its most simple version, just the distance to the k-th nearest neighbor is used. Slightly advanced methods use averaged

distances, which works somewhat better. This simple method is often very good in high dimensional feature spaces.

The LP-DD is a linear programming method [10]. This data descriptor is specifically constructed to describe target classes which are represented in terms of distances to a set of support objects. In some cases it might be much easier to define distances between objects than informative features (for instance when shapes have to be distinguished). This classifier uses the Euclidean distance by default. The classifier has basically the following form $f(x) = \sum w_i d(x, x_i)$. The weights $\boldsymbol{w}_i$ are optimized such that just a few weights stay non-zero, and the boundary is as tight as possible around the data.

3 Experiments and Results

3.1 Experimental Setup

All the seven one-class classifiers are investigated using the dataset in [19]. This dataset contains two parts. There are 400 pictures from AT&T/ORL face database [1] and 402 non-face pictures from our previous work [15, 16]. There are some examples from each part shown in Figure 1 and 2 respectively. It should note that all patterns were resized to 32×32. The reported experiments are all carried out based on the PRTOOLS [4] and DDTOOLS [17] packages, from Delft University of Technology. And face is the target class.

Fig. 1. Some face patterns

Fig. 2. Some non-face patterns

Currently two feature schemes are used in experiments reported in this paper. First experiments are directly conducted in full pixel space (1024 dimensions). Then similar experiments are repeated in PCA subspace. For all the seven methods, the reject rates for target class are set to 0.01. For PCA-DD, its dimension will be 10 if that can't be clearly found from context. For SV-DD, σ=1128 is used. For KMEANS-DD, *k* is 5. For KNN-DD, *k* is 2.

3.2 Results and Discussion

To know how the amount of training patterns affects the performance of each classifier, a fraction, from 10% to 90%, of face data (randomly selected in the whole face database each time) is used for training, and the rest of face data and all non-face data for independent testing. For a certain experiment, it is repeated ten times and average error rate is used as the final score. The first series of experiments are conducted directly in pixel space. PCA is used to reduce the dimension for another series of experiments. Results are demonstrated through Fig. 3.

Over pixel space, SV-DD shows decrease of overall error rate (OA) from about 40% to 5%, and false negatives (FN) from 80% to 30%. Its false positive (FP) rates are very steadily less than 5%. No other methods show similar trend. Two methods LP-DD and GAUSS-DD don't work well over pixel space. Both of them have 100% FN and 0% FP in all experiments. Therefore they are not included in Figure 3.a, 3.b and 3.c. In PCA subspace (10 Principle Components), SV-DD shows similar trend on FNs as it does over pixel space, but the decreases are relatively slight. It shows a very steady performance less than 10% in overall error rate and FPs. Similarly, no other methods work well like SV-DD. This time, the LP-DD method works as with pixel space. Methods like NN-DD, KMEANS-DD and KNN-DD have very low FNs, but very high FPs, both over pixel space and PCA subspace. The relatively good performance of SV-DD in comparison to the other six methods can be contributed by its flexibility. The other methods are mainly using very strict models, such as plane-like shapes or nearest neighbor type of models. They tend to capture large areas in feature space since the reject rates for target class were set relatively low at 0.01, and therefore large FPs and low FNs.

Table 1. Error rates of SV-DD over PCA subspcace (FN = false negatives, FP = false positives, OA = overall error rate)

Data size	Error	10 PCs	15 PCs	20 PCs	30 PCs	Average
	FN	12.97	12.14	10.31	9.36	11.20
10%	FP	6.57	4.18	4.7	9.36	6.20
	OA	9.59	7.94	6.69	6.59	7.70
	FN	10.03	10.07	10.85	10.16	10.28
20%	FP	4.78	4.53	4.13	3.63	4.27
	OA	7.10	7.27	7.13	6.53	7.01
	FN	10.07	11.39	10.04	10.89	10.60
30%	FP	4.95	4.78	4.58	4.18	4.62
	OA	7.05	7.49	6.82	6.94	7.08
	FN	10.67	10.15	10.58	10.89	10.57
40%	FP	5.95	5.65	5.32	5.05	5.49
	OA	7.71	7.31	7.29	7.20	7.38
Average (OA)		7.91	7.50	6.98	6.82	7.29

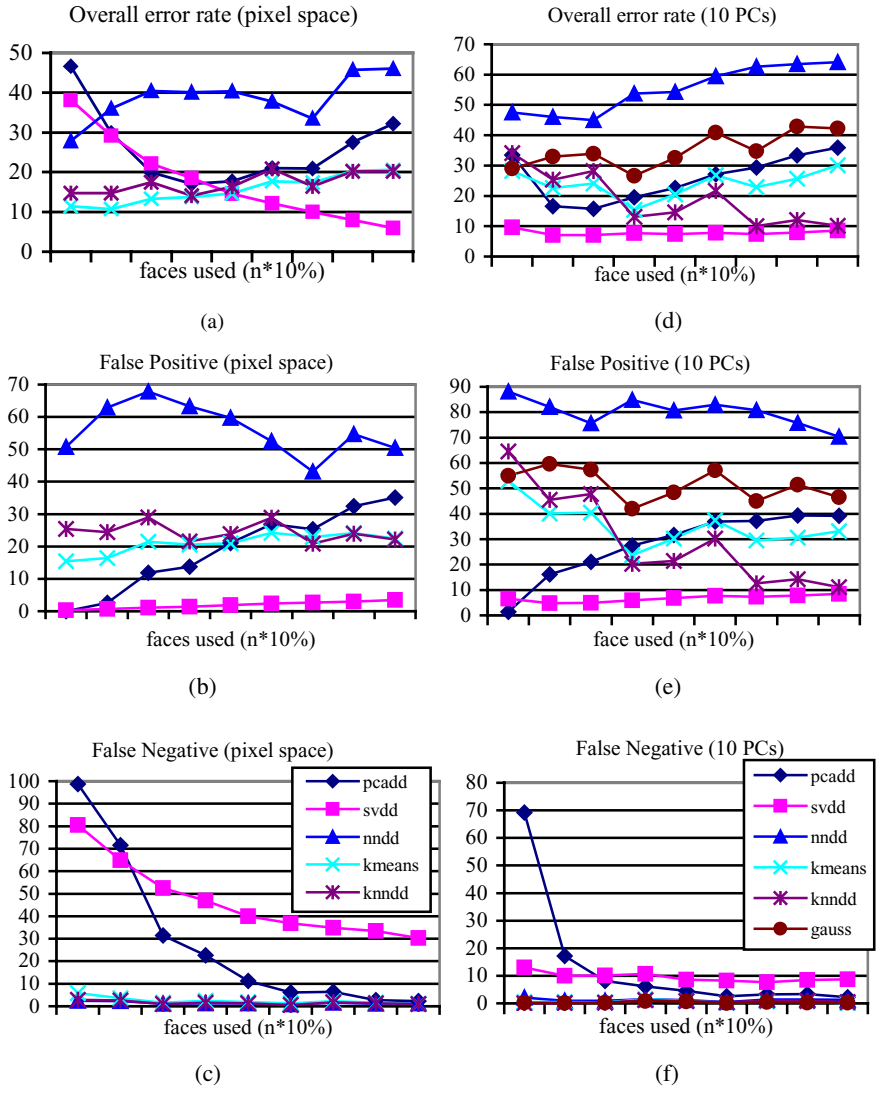


Fig. 3. Some results: diagrams a, b, c show overall classification error, false positives and false negatives of five methods in full pixel space; diagrams d, e, f show overall classification error, false positives and false negatives of six methods in PCA subspace (10 Principal Components). The Y-axis is error rate score (%), and the X-axis is the percentage of faces used in training.

How the number of features used may affect these classifiers is also investigated. For the specific case of SV-DD, 10, 15, 20 and 30 PCs are used. In table 1, a decrease of error rates (OA, FP, FN) can be found when more training patterns are used. There is also a more or less similar trend when more features are used (last row in the table). But when the main variation is captured over a specific training set, more features

don't always guarantee better results. It is because when more features are used, generally more training data are needed to estimate reliably the class models. Thus with a certain training set used above, more features may directly cause that the class models can't be estimated reliably, and the performance dangles a little bit (the "*curse-of-dimensionality*" [3]). This is also why SV-DD performs better in PCA subspace than it does in full pixel space.

4 Concluding Remarks

In this paper, face and non-face classification is used as an example in investigating several one-class classification methods. It's preliminary work towards finding suitable learning and classification methods for natural language concept grounding in the context of Human-Robot Interaction. In the reported experiments, it's intentional to learn target class models with only target patterns, but still keeping good discrimination with respect to outlier patterns. It's found that some of such one-class classifiers, particularly SV-DD, can attain very nice performance (overall error rate, false negative and false positive all less than 10%) on our data set. All other one-class classifiers perform less well in our experiments. Some of them work well to accept target patterns. Some of them work well to reject outlier patterns. Only SV-DD performs very steadily, especially when discriminant features such as PCA subspace is used. It can be concluded that SV-DD can form a good foundation for developing a learning and classification method suitable for HRI, since not only can it obtain reasonable performance with a (relative) small amount of training patterns, but also it can achieve very nice results when more training patterns are available. From a viewpoint of lifelong learning for a robot, this potential of SV-DD can be further utilized.

Obviously further study on these one-class classifiers should be conducted, for example, using other larger data set and/or feature extraction methods. More importantly, it's interesting to apply some of these methods on to Carl, a service robot prototype previously developed by our group [16].

Acknowledgement. Q. H. Wang is supported by IEETA (Instituto de Engenharia Electrónica e Telemática de Aveiro), Universidade de Aveiro, Portugal, under a PhD research grant.

References

1. AT & T Face Database, formerly "The ORL Database of Faces", at http://www.uk.research.att.com/ facedatabase.html
2. Bishop, C.: Novelty detection and neural network validation. In: IEE Proc. Vision, Image and Signal Processing, 141 (1994) 217-222
3. Bishop, C.: Neural Networks for Pattern recognition. Oxford University Press (1995)
4. Duin, R.: PRTOOLS 4.0. Delft University of Technology, The Netherlands (2004)
5. Japkowicz, N.: Concept-Learning in the absence of counter-examples: an autoassociation-based approach to classification. Ph D thesis, The State Univ. of New Jersy (1999)

6. Meng, L. M.: An Image-based Bayesian Framework for Face detection. In: Proc. of IEEE Intl. Conf. On Computer Vision and Pattern Recognition (2000)
7. Moya, M., Koch, M. and Hostetler, L.: One-class classifier networks for target recognition applications. In: Proc. World congress on neural networks (1993) 797-801
8. Muggleton, S. and J. Firth. CProgol4.4: a tutorial introduction. In S. Dzeroski and N. Lavrac (eds.): Relational Data Mining. Springer-Verlag (2001) 160-188
9. Parra, L., Deco, G. And Miesbach, S.: Statistical independence and novelty detection with information preserving nonlinear maps. In: Neural Compiutation 8 (1996) 260-269
10. Pekalska, E., Tax, D. M.J. and Duin, R. P. W.: One-Class LP Classifiers for dissimilarity Representations. In: Advances in Neural Info. Processing Systems, vol. 15. MIT Press (2003) 761-768
11. Pham, T. V., Arnold, M. W. and Smeulders, W. M.: Face Detection by aggregated Bayesian network classifiers. In: Pattern Recognition Letters, 23(4) (2002) 451-461
12. Ritter, G. and Gallegos, M.: Outliers in statistical pattern recognition and an application to automatic chromosome classification. In: Pattern Recognition Letters 18. 525-539 (1997)
13. Roberts, S. and Penny, W.: Novelty, confidence and errors in connectionist systems. Technology report, Imperial College, London, TR-96-1 (1996)
14. Schneiderman, H. and Kanade, K.: A Statistical Method for 3D Object Detection Applied to Faces and Cars. In: Proc. CVPR 2000 (2000) 746-751
15. Seabra Lopes, L.: Carl: from Situated Activity to Language-Level Interaction and Learning. In: Proc. IEEE Intl. Conf. on Intelligent Robotics & Systems (2002) 890-896
16. Seabra Lopes, L. and Wang, Q. H.: Towards Grounded Human-Robot Communication. In: Proc. IEEE Intl. Workshop RO-MAN (2002) 312-318
17. Tax, David M. J.: One-class classification. Ph D dissertation, Delft University of Technology, The Netherlands (2001)
18. Turk, M. and Pentland, A.: Eigenfaces for recognition. In: Journal of Cognitive Neuroscience 3 (1994) 71-86
19. Wang, Q. H. and Seabra Lopes, L.: An Object Recognition Framework Based on Hidden Markov Trees and Kullback-Leibler Distance. In: Proc. ACCV 2004 (2004) 276-281
20. Yang, M. H., Kriegman, D. and Ahuja, N.: Detecting Faces in Images: A Survey. IEEE Trans. PAMI 24 (2002) 34-58
21. Yang, M. H., Kriegman, D. and Ahuja, N.: Face Detection Using Multimodal Density Models. In: Computer Vision and Image Understanding 84 (2001) 264-284
22. Zhu, Y. Schwartz, S.: Efficient Face Detection with Multiscale Sequential Classification. In: Proc. IEEE Intl. Conf. Image Processing '02 (2002) 121-124

Semantic Image Analysis Based on the Representation of the Spatial Relations Between Objects in Images

Hyunjang Kong[1], Miyoung Cho[1], Kwanho Jung[1], Sunkyoung Baek[1], and Pankoo Kim[2]

[1] Dept. of Computer Science, Chosun University, Gwangju 501-759 Korea
{kisofire,irune80,khjung,zamilla100}@mina.chosun.ac.kr
[2] Corresponding Author, Dept. of CSE, Chosun University, Korea
pkkim@chosun.ac.kr

Abstract. The number of images available on the world wide web has grown enormously, because of the increasing use of scanners, digital cameras and camera-phones. Consequently, the efficient retrieval of images from the web is necessary. Most existing image retrieval systems are based on the text or content associated with the image. In this paper, we propose a semantic image analysis for the semantic web. We use the description about the image and try to represent it using OWL. We also define new axioms for representing the spatial relationships based on the spatial description logics.

1 Introduction

The use of image acquisition devices such as scanners, digital cameras, etc., has grown rapidly in recent times. Consequently, the number of images available on the web is increasing. Because of this, it has become necessary to develop a system for the storage and retrieval of these images. The existing image retrieval systems are based on the text annotations associated with the images. However, the precision of these image retrieval systems is very low, because the annotations often have ambiguous meanings. As a result, most of the studies that have been conducted so far about image retrieval focused on content based image retrieval.

This paper focuses on semantic image storage and retrieval for the semantic web. Especially, we attempt to store images using the standard metadata representation regarding the description of the image information and to use the metadata when we retrieve the images from the semantic web. Among the possible concepts that could be used, most images are described using spatial relationships. In this paper, we define the basic spatial relationships based on Egenhofer's spatial relationships. We also design new axioms for these basic spatial relationships based on the description logics employed on the semantic web. In this paper, we define the complex roles in the TBox of the description logics and we represent the individual properties in the ABox of the description logics by using the new axioms. Finally, we apply this image knowledge base, consisting of the TBox and ABox, to the image retrieval system on the semantic web.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 471–478, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Related Works

2.1 Semantic Web and Web Ontology

The current web is aimed at expanding the quantity of information, while the semantic web, or trust web, refers to the expansion of quality. The semantic web is intended to be 'an evolution of the web' rather than a reformation of the web. Tim Berners-Lee, who proposed the semantic web, stated that, "*The Semantic Web is an extension of the current web, in which information is given a well-defined meaning, which will better enable computers and people to work together in cooperation with each other*"[3]. The Semantic Web will have intelligent services such as information brokers, search agents, information filters, etc. Such intelligent services, which are destined to be available on the knowledgeable web, should supersede the currently available versions of these services, which are limited in their functionality, in that they only work as stand-alone services that do not interoperate. The two components constituting the semantic web are ontology, representing the semantic constitution of information, and the markup language, representing the well-defined information. Humans and machines need to communicate with each other, in order to realize the semantic web, by processing and interpreting information. The languages used to represent the information, XML which is used to represent the information structures and RDF, DAML, and OWL which are used to represent the information meaning, have been developed and standardized in various ways by the W3C [1][2][3][15]. In order to represent the ontology that makes up the spatial relationships, in this study, we made use of the OWL language, in which hierarchies of classes can be used to represent web documents and applications. In addition, OWL can be used to define each class's properties and to represent the domain that is used to define the individual properties.

2.2 9-Intersection Model for Line-Region Relations

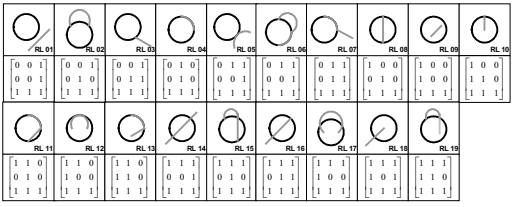
The 9-intersection model is a comprehensive model for binary topological spatial relations and applies to objects of type area, line and point. It characterizes the topological relation between two point sets, A and B, by the set intersections of A's interior (A^o), boundary(∂A) and exterior(A^-) with the interior, boundary and exterior of B, called the 9-intersection(Equation 1)

$$I(A,B)=\begin{pmatrix} A^o \cap B^o & A^o \cap \partial B & A^o \cap B^- \\ \partial A \cap B^o & \partial A \cap \partial B & \partial A \cap B^- \\ A^- \cap B^o & A^- \cap \partial B & A^- \cap B^- \end{pmatrix} \quad (1)$$

The interior, boundary, and exterior of a line are defined according to algebraic topology: the boundary of a simple line comprises the two end points, the interior is the closure of the line minus the boundary, and the exterior is the complement of the closure. Given that each of these 9 intersections can be either empty(0) or non-empty (1), the model distinguishes a total of 512 different topological relations between two

point sets, some of which cannot be realized. Between simple line(1-dimensional, non-branching, without self-intersections) and the region(2-dimensional, simply connected, no holes)embedded in R^2, 19 different situations are found with the 9-intersection model(table 1).

Table 1. 19 line-region topological relations



2.3 Spatial Terms

To obtain the topological relations for the spatial terms, we needed a set of spatial terms. So, we used the results of an experiment involving non-trained college students conducted by Mark and Egenhofer[14,16,20,23]. In this experiment, the subjects were presented with outlines of a park accompanied in each case by an English-language sentence printed underneath it, describing a particular spatial relation between the road and the park. Table 2 lists some of the results concerning the representation of the topological relations for each spatial term. The numbers 1 to 19 indicate the topological relations LR1 to LR19. The total represents the total number of answers which included this topological relation for the representation of the spatial term.

Table 2. Topological relations for each spatial term

Spatial term	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	total
Along edge	20				7				1					1						29
bisects										1	2			27		1		1		32
Bypasses	31									1				1						33
Connected to	5		7		7				1					5				6		31
Crosses										1				28				2		31
Cuts across								2	1					29				2		34
Cuts through										1				29		2				32
Divides										1				32				1		34
Ends in			1						1	1								27		30
Ends near	31		1							1								1		34
Ends outside	14		1											2		2		14		33
exits	2		2							1				1	1			27		34
in									19	3				1		1				24
Near	29				4					1										34
Outside	31									1										32

In total, 15 out of the 19 possible topological relations occurred among the spatial terms. That is, not all of the topological relations were used to represent the spatial terms. Moreover, most of the cases are represented by just 4 relations : LR18 (the line goes from the region's interior to its exterior), LR1 (the line is completely contained in the region's exterior), LR14 (the line goes from the region's exterior through the region's interior to the region's exterior again), LR9 (the line is completely contained in the region's interior).

To identify certain representative relations among the topological relations, we organized the terms into groups. This grouping was done using a statistical technique for detecting natural groupings in data. We focused on the most frequent topological cases and eliminated those with small counts(i.e., equal to or less than 1 inclusion in a response), because the analysis of the latter would not have led to statistically significant results. After grouping, we obtained 3 main groups(the representatives of these three groups are LR1, LR14 and LR18, respectively).

3 Defining New Axioms for Representing the Spatial Relations Based on Description Logics

A variety of relationships should be employed when building the domain ontology. However, the existing capabilities of OWL are not sufficient to completely represent the concepts and define the relationships among them. In this paper, we define the basic spatial relationships for the spatial terms based on Egenhofer's spatial relationships. We also design new axioms for these basic relationships based on the description logics allowing for its use on the semantic web. We define the complex roles in the TBox of the description logics and we represent the individual properties in the ABox of the description logics by using the new axioms. Finally, we apply this image knowledge base, consisting of the TBox and ABox, to the problem of image retrieval on the semantic web.

3.1 New Axioms for Representing the Spatial Relations Between Objects

In order to understand the relationships between the spatial regions, ***ALCRP(D)*** was developed as an extension of the DL, ***ALC(D).*** This DL provides a foundation to support the spatial logical reasoning with DLs[9][10].

In this section, the new axioms used for representing the spatial relationships based on the ***ALCRP(D),*** are defined and used it to construct the web ontology. We introduce the concept that the ***ALCRP(D),*** is more suitable for describing the spatial relationships than the other DLs. Using ***ALCRP(D)'***s role-forming predicate-based operator, a set of complex roles can be defined based on the aforementioned ***RCC-8*** predicates.

In section 2.3, we clustered the spatial terms into 3 groups. Therefore, we try to define three basic axioms to represent the 3 groups. Thus, we define the predicates – 'tpp', 'dc' and 'co' for the three relationships. Here, 'co' means 'ec $\vee$ po $\vee$ eq' or in other words, "connected" and 'tpp' means 'tpp $\vee$ ntpp' or in other words, "inside".

The formal semantics of ‘dc’, ‘tpp’ and ‘co’ are as follows:

$$\begin{array}{ll} co\ (X_1, X_2) & \exists x\ \ x \in X_1 \cap X_2 \\ tpp(X_1, X_2) & \forall x\ \ x \in X_1 \cap X_2 \\ dc\ (X_1, X_2) & \neg\exists x\ \ x \in X_1 \cap X_2 \end{array}$$

Table 3 shows how to add these new spatial predicates to the OWL axioms.

Table 3. Adding the new axioms to the OWL axioms

Axiom	DL syntax	Example
subClassOf	C1 ⊆ C2	Human⊆Animal∩Biped
sameClassAs	C1 ≡ C2	Man≡Human∩Male
disjointWith	C1 ⊆¬C2	Male⊆¬Female
…….	……	…….
connectedWith	C_1 co C_2	Banana co Pineapple
disconnectedWith	C_1 dc C_2	France dc England
insideInto	C_1 tpp C_2	Seoul tpp Korea

Using the spatial predicates, a set of complex roles can be defined for the TBox. In the image information that is contained in the expression “bananas overlap the pineapple”, the complex role ‘overlap’ could be defined as follows:

overlap ≡ ∃(has_area)(has_area).*co*

4 Example of the Semantic Image Retrieval Process

4.1 Example of the Semantic Image Retrieval Process on the Semantic Web

We apply the new axioms to the problem of image retrieval on the semantic web. The images consist of many objects. We use the description of the image information. Most of the descriptions of the images include spatial relationships among the objects. Table 4 shows simple images about fruit.

Table 4. An example of OWL regarding fruits

Image 1

Image 2

Image 3

Image 4

Image 5

Image 6

Image 7

Image 8

To search for the images, the retrieval system invokes a sequence of internal processing steps, similar to those depicted in figure 1.

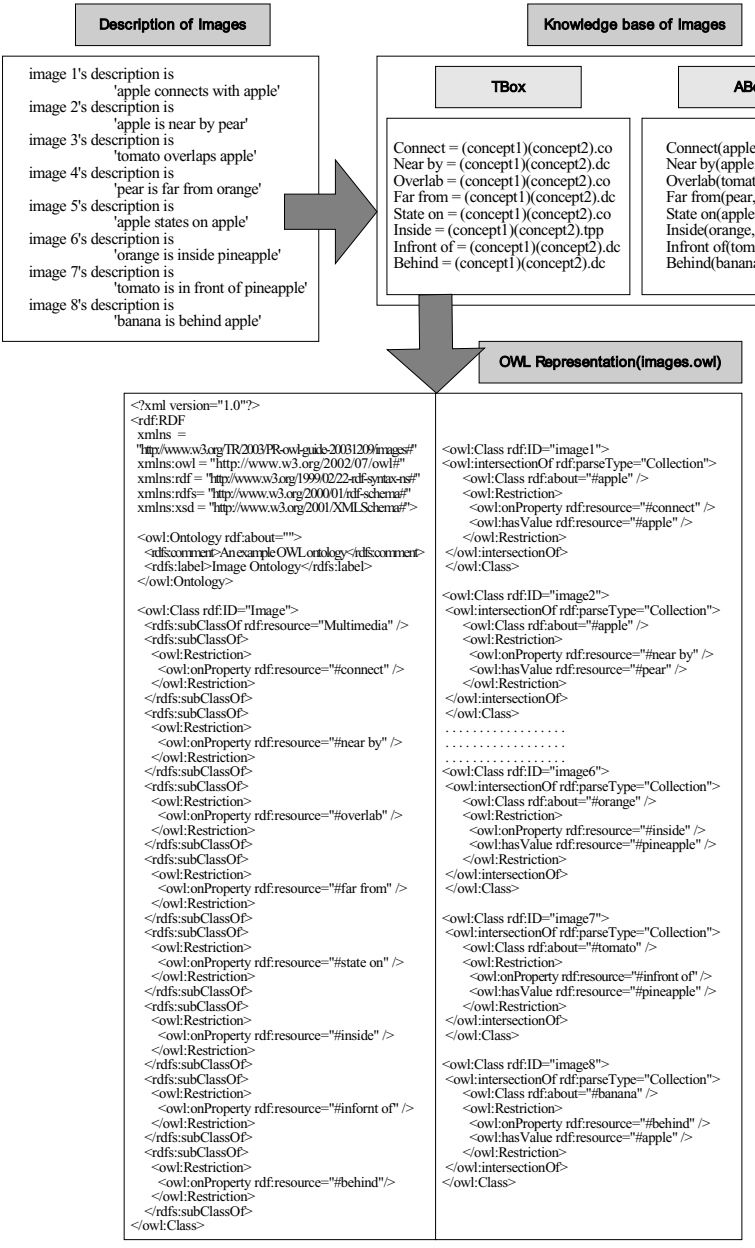


Fig. 1. Internal processing steps of the image retrieval system for the semantic web

In figure 1, the retrieval system constructs a knowledge base of the images by using the descriptions of each image. The knowledge base consists of a TBox and ABox, each containing description logics, and we use the new axioms to define the complex roles used to describe the spatial relationships. Finally, we construct the domain ontology(images.owl) using the OWL language, based on this knowledge base.

In the TBOX in figure 1, we know that some of the properties are defined using the same semantics. That is, those properties that have the same semantics have similar spatial relationships. So, we can resolve the ambiguity among the terms and, consequently, the retrieval of the images can be accomplished in a more semantically oriented fashion on the semantic web.

5 Conclusion and Future Works

In the present study, we represent the spatial relationships based on spatial description logic and use new axioms to construct a semantic image retrieval system, which is used for representing spatial relationships in the description of images. The use of these new axioms in the image retrieval system, allows for more semantic results to be obtained. In future works, the representation of the spatial relations will be extended, in order to develop the more accurate web ontology.

References

1. J. P. Eakins. Automatic image content retrieval | are we getting anywhere? Pages 123-135. De Montfort University, May 1996.
2. M. Koskela, J. Laaksonen, S. Laakso, and E. Oja. The PicSOM retrieval system: description and evaluations. In The challenge of image retrieval, Brighton, UK, May 2000. http://www.cis.hut.~/picsom/publications.html.
3. M. Agosti and A. Smeaton, editors. *Information retrieval and hypertext*. Kluwer, New York, 1996.
4. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, New York, 1999.
5. J. van den Berg. Subject retrieval in pictorial information systems. In Proceedings of the 18th international congress of historical sciences, Montreal, Canada, pages 21{29, 1995.http://www.iconclass.nl/texts/history05.html.
6. T. Peterson. Introduction to the Art and Architecture thesaurus, 1994. http://shiva.pub.getty.edu.
7. A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photoannotation. *IEEE Intelligent Systems*, 16:66–74, May/June 2001.
8. G. Schreiber, I. Blok, D. Carlier,W. van Gent, J. Hokstam, and U. Roos. A mini-experimentin semantic annotation. In I. Horrocks and J. Hendler, editors, *The Semantic Web – ISWC2002. First international semantic web conference*, number LNCS 2342, pages 404–408.Springer–Verlag, Berlin, 2002.

9. Peter F. Patel-Schneider, Patrick Hayes, Ian Horrocks, “OWL Web Ontology Language Semantics and Abstract Syntax, W3C Working Draft 31 March 2003”, http://www.w3.org/TR/2003/WD-owl-semantics-20030331.
10. D. Brickley, R. Guha (eds.), “Resource Description Framework (RDF) Schema Specification, W3C Candidate Recommendation 27 March 2000, http://www. w3.org/TR/2000/CR-rdf-schema-20000327.
11. T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web”, Scientific Am., vol.284, no.5, May 2001, pp.34-43.
12. F.Wolter and M.Zakharyaschev. “Modal description Logics: Modalizing roles”, *Fundamenta Informaticae*, 39:411-438, 1999.
13. V. Haarslev, C.Lutz, and R. Moller, “A description logic with concrete domains and a role-forming predicate operator,” Journal of Logic and Computation 9(S), pp. 351-384, 1999
14. T.Y. Jen and P. Boursier, "A Model for Handling Topological Relationships in a 2D Environment", Sixth International Symposium on Spatial Data Handling, Edinburg, Scotland, Uk.
15. V. Haarslev, C. Lutz, and R. Moller, “Foundations of spatioterminological reasoning with description logics,” Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning(KR ’98), pp. 112-123, June 1998.
16. M. Erwig and M. Schneider, "Query-By-Trace: Visual Predicate Specification in Spatio-Temporal Databases", 5th IFIP Conf. on Visual databases, 2000
17. A. Cohn, Z. Cui, and D. Randell, “A spatial logic based on regions and connection,” Proc. Third International Conference on Principles of Knowledge Representation and Reasoning(KR ’92), 1992.
18. B. Nebel and J. Renz, “On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the region connection calculus,” Artificial Intelligence, 1992.
19. Guarino, N, and Giaretta, P., “Ontologies and Knowledge bases: towards a terminological clarification”, In N. Mars, Ed. Toward Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995, PP. 25-32.
20. A. R. Shariff, M. J. Egenhofer and D. Mark "Natual-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English Language Terms", International Journal of Geographical Informaion Science, 12(3): 215-246, 1998.
21. W. Kim, H. Kong, K. Oh, Y. Moon and P. Kim, “Concept Based Image Retrieval Using the Domain Ontology”, Computational Science and Its Applicatons(ICCSA 2003), PP 401-410, 2003.
22. B. Chandrasekaran, J. Josephson, and R. Benjamins, “What Are Ontologies, and Why do we Need Them?”, IEEE Intelligent Systems, 14,1:20-26, 1999.
23. M. Andrea Rodriguez, Max J. Egenhofer, Andress D. Blaser, “Query Pre-processing of topological Constaints:Comparing a Composition-based with Neighborhood-Based Approach”, SSTD 2003, LNCS 2750, pp. 362-379, 2003

Ridgelets Frame

Tan Shan[1], Licheng Jiao[1], and Xiangchu Feng[2]

[1] National Key Lab for Radar Signal Processing and Institute of Intelligent Information Processing, Xidian University, 710071 Xi'an, China
tanshan5989@yahoo.com.cn, lchjiao@mail.xidian.edu.cn
[2] College of Science, Xidian University, 710071 Xi'an, China

Abstract. In this paper, a new system called ridgelets frame in $L^2(R^2)$ is constructed. To construct the new system, we use other orthonormal wavelet rather than Meyer wavelet, which was used in the construction of orthonormal ridgelets by Donoho. Due to the losing of two special closure properties of Meyer wavelet, the new system is a tight frame with frame bound 1 instead of orthonormal basis for $L^2(R^2)$. As an example, we demonstrate the potential power of the new constructed system by showing its ability of recovering the line structure in images in the presence of noise.

1 Introduction

In paper [1], Donoho constructed a new system, namely, orthonormal ridgelets, which can effectively represent the two-dimension function smooth away from straight singularity. To obtain orthogonality of orthonormal ridgelets, Donoho made use of two special properties of Meyer wavelet, i.e., closure property under reflection about the origin in the ridge direction: $\psi_{j,k}(-t) = \psi_{j,1-k}(t)$, and closure property under translation by half a cycle in the angular direction: $w_{i,l}(\theta+\pi) = w_{i,l+2^{i-1}}(\theta)$. Note that the later closure property would not hold for other prominent wavelet families, for example, Daubechies' compactly supported wavelet families. It is the closure properties that make it possible to construct orthonormal basis by removing the duplications. In paper [2], from the viewpoint of frequency domain and radon domain respectively, the author constructed systems called ridgelets packets that provide a large family of orthonormal basis in $L^2(R^2)$.

In this paper, as an extension of the ridgelets packets and an implementation of the principle proposed in paper [2], we construct a new system called ridgelets frame using orthonormal wavelet not only restricted to Meyer wavelet. Due to the losing of special closure properties of Meyer wavelet, the new system is a tight frame with frame bound 1 instead of orthonormal basis in $L^2(R^2)$, and we call the new frame ridgelets frame. As its forerunner, namely, orthonormal ridgelets and ridgelets packets, the ridgelets frame retains the key idea that deals with straight singularities by transporting it to point singularities. Therefore, the ridgelets frame is good at recov-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 479–486, 2004.
© Springer-Verlag Berlin Heidelberg 2004

ering the line structures in images in the presence of noise, just the same as the ridgelets and their derivations, i.e., monoscale ridgelets and curvelets. We show the powerful ability of ridgelets frame to recover line structure, through comparing it with wavelets based method on synthesis test image.

This paper is organized as follows. In Section 2, the construction of ridgelets frame using orthonormal wavelet is proposed and the associated proofs are given. Then, in section 3, based the ridgelets frame, a method for image denoising are introduced, and the denoising results are compared with that of wavelets based method both visually and in terms of the PNSR. Finally, concluding remarks are given in Section 4.

2 Construction of Ridgelets Frame

We represent the construction of ridgelets frame in this section. In the field of computed tomography [1], [3], it is well known that there exists an isometric map from Radon domain to spatial domain $L^2(R^2)$. So, for the purpose to construct a tight frame in spatial space $L^2(R^2)$, one can construct a tight frame in Radon domain. In this paper, a tight frame with frame bound 1 is constructed first in Radon domain using orthonormal wavelet basis. Then, it is obvious that the image of the tight frame under the isometric map constitutes a tight frame also for $L^2(R^2)$.

To construct tight fame in Radon domain, we start from an orthonormal basis in $L^2(R \otimes [0,2\pi))$. And the orthonormal basis is obtained from tensor product of one dimension wavelet basis respectively for $L^2(R^2)$ and for $L^2([0,2\pi))$.

For convenience below, we denote the orthonormal basis in $L^2(R \otimes [0,2\pi))$ by w''_λ ($\lambda \in \Lambda$), where Λ is the collection of index λ. Now, let $w'_\lambda := 2\sqrt{\pi} w''_\lambda$.

Define orthoprojector $P_{\Re}$ from $L^2(R \otimes [0,2\pi))$ to Radon domain by

$$(P_{\Re} F)(t,\theta) = (F(t,\theta) + F(-t,\theta+\pi))/2, \tag{1}$$

where $F \in L^2(R \otimes [0,2\pi))$.

Then, applying $P_{\Re}$ on w'_λ, we obtain

$$w_\lambda := P_{\Re}(w'_\lambda) = (\frac{I + T \otimes S}{2}) w'_\lambda = 2\sqrt{\pi} P_{\Re}(w''_\lambda), \tag{2}$$

where operator T is defined by $(Tf)(t) = f(-t)$ and operator S is defined by $(Sg)(\theta) = g(\theta+\pi)$.

We will show that w_λ is a tight frame with frame bound 1 in Radon domain. First, we prove several Lemmas.

Lemma 1. w_λ is complete in Radon domain $\Re$.

Proof. For $\forall F \in \Re$, it is obvious that $F \in L^2(R \otimes [0,2\pi))$ and we have $F = P_\Re F$ due to the definition of $P_\Re$. We obtain

$$F = \sum_{\lambda\in\Lambda} <\mathrm{F}, w_\lambda''> w_\lambda'' = \sum_{\lambda\in\Lambda} <\mathrm{F}, w_\lambda''> P_\Re\left(\frac{1}{2\sqrt{\pi}} w_\lambda'\right) = \frac{1}{2\sqrt{\pi}} \sum_{\lambda\in\Lambda} <\mathrm{F}, w_\lambda''> w_\lambda . \tag{3}$$

So, w_λ is complete in $\Re$. ■

Lemma 2. For $\forall F \in \Re$,

$$< w_\lambda'', F > = <(\mathrm{T}\otimes\mathrm{S}) w_\lambda'', F > . \tag{4}$$

Proof. It is easy to obtain the relationship by computing both $< w_\lambda'', F >$ and $<(\mathrm{T}\otimes\mathrm{S}) w_\lambda'', F >$ respectively. In fact,

$$\begin{aligned}
<w_\lambda'', F> &= \int_0^{2\pi}\int_{-\infty}^{\infty} w_\lambda''(t,\theta)\overline{F}(t,\theta)dtd\theta \\
&= \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta)\overline{F}(t,\theta)dtd\theta + \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta+\pi)\overline{F}(t,\theta+\pi)dtd\theta \\
<(\mathrm{T}\otimes\mathrm{S}) w_\lambda'', F> &= \int_0^{2\pi}\int_{-\infty}^{\infty} \psi(-t)\omega(\theta+\pi)\overline{F}(t,\theta)dtd\theta \\
&= \int_0^{\pi}\int_{-\infty}^{\infty} \psi(-t)\omega(\theta+\pi)\overline{F}(t,\theta)dtd\theta + \int_0^{\pi}\int_{-\infty}^{\infty} \psi(-t)\omega(\theta+2\pi)\overline{F}(t,\theta+\pi)dtd\theta \\
&= \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta+\pi)\overline{F}(-t,\theta)dtd\theta + \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta+2\pi)\overline{F}(-t,\theta+\pi)dtd\theta \\
&= \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta+\pi)\overline{F}(t,\theta+\pi)dtd\theta + \int_0^{\pi}\int_{-\infty}^{\infty} \psi(t)\omega(\theta)\overline{F}(t,\theta)dtd\theta
\end{aligned}$$

So, we have $< w_\lambda'', F > = <(\mathrm{T}\otimes\mathrm{S}) w_\lambda'', F >$. ■

In Radon domain $\Re$, define inner product for $\forall F, G \in \Re$ as

$$[F,G] = \frac{1}{4\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} F(t,\theta)\overline{G}(t,\theta)dtd\theta . \tag{5}$$

Then, we have

Lemma 3. For $\forall F \in \Re$,

$$< w_\lambda'', F > = 2\sqrt{\pi}[w_\lambda, F] . \tag{6}$$

Proof.

$$\begin{aligned}
[w_\lambda, F] &= \frac{1}{4\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} w_\lambda(t,\theta)\overline{F}(t,\theta)dtd\theta = \frac{1}{4\pi}\times 2\sqrt{\pi}\int_0^{2\pi}\int_{-\infty}^{\infty} \frac{w_\lambda'' + (\mathrm{T}\otimes\mathrm{S}) w_\lambda''}{2}\overline{F}(t,\theta)dtd\theta \\
&= \frac{1}{2\sqrt{\pi}}\times\frac{1}{2}\{< w_\lambda'', F > + <(\mathrm{T}\otimes\mathrm{S}) w_\lambda'', F >\} = \frac{1}{2\sqrt{\pi}} < w_\lambda'', F > . \blacksquare
\end{aligned}$$

Theorem 1. The collection w_λ $(\lambda \in \Lambda)$ is a tight frame with frame bound 1 in Radon domain $\Re$.

Proof. The theorem is equivalent to that for $\forall F \in \Re$,

$$\|F\|^2 = [F,F] = \sum_\lambda |[F, w_\lambda]|^2 \quad . \tag{7}$$

We have

$$\begin{aligned}[F,F] &= [\frac{1}{2\sqrt{\pi}} \sum_{\lambda\in\Lambda} <\mathrm{F}\ , w''_\lambda > w_\lambda, F] \\ &= \sum_\lambda [F, w_\lambda][w_\lambda, F] = \sum_\lambda |[F, w_\lambda]|^2 \ . \blacksquare\end{aligned}$$

Then, from the properties of classical frame theory, we have

$$F = \sum_\lambda [F, w_\lambda]\ w_\lambda \quad . \tag{8}$$

$$\sum_\lambda |[F, w_\lambda]|^2 = \|F\|^2 \quad . \tag{9}$$

By now, we have constructed a tight frame in Radon domain $\Re$ using orthonormal wavelet basis. As mentioned above, we can exactly obtain a tight frame by mapping the one in Radon domain $\Re$ to spatial $L^2(R^2)$. And the resulting tight frame in $L^2(R^2)$ is with the same frame bound 1 as its counterpart in Radon domain $\Re$. We call the tight frame in $L^2(R^2)$ ridgelets frame.

It is worth emphasizing that one can obtain the orthonormal ridgelets if Meyer wavelet is used in the above construction and the redundancy of resulting tight frame is removed also. Generally, the ridgelets frame can be considered as an extension of orthonormal ridgelets. An element of orthonormal ridgelets that is constructed using Meyer wavelet and an element of ridgelets frame that is constructed using Danbechies-8 wavelet are displayed in Fig. 1.

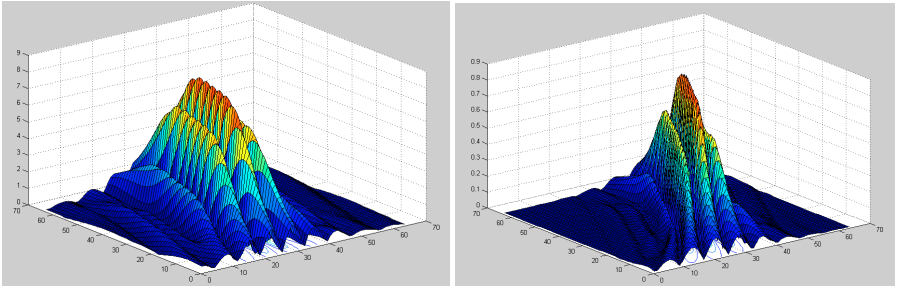


Fig. 1. An element of orthonormal ridgelets *(left)* and an element of ridgelets frame *(right)*

The orthonormal ridgelets can effectively represent two dimensions function smooth away from straight singularity. And the key reason is that it transports the straight singularity to point singularity in Radon domain then deal with the resulting point singularity using Meyer wavelet. The effectiveness of orthonormal ridgelets to represent the straight singularity, therefore, is due to the effectiveness of Meyer wavelet to represent point singularity. Note that the ridgelets frame is constructed in Radon Domain. As a result, the ridgelets frame retains the ability to effectively represent straight singularity also.

3 Image Denoising Using Ridgelets Frame

In section 2, we have constructed a tight frame with frame bound 1 in $L^2(R^2)$. Different from the use of Meyer wavelet in orthonormal ridgelets, the ridgelets frame can be constructed using broader orthonormal wavelet families. Undoubtedly, the ridgelets frame provides a powerful tool in various applications, especially in image processing task.

In this section, we shall investigate the ability of ridgelets frame to recover the edges in image in the presence of noise.

Based on the localization principle and subband decomposition, the monoscale ridgelets and curvelets were proposed [4], [5], both of which were derived from the ridgelets system and can efficiently deal with smooth images with smooth edges including straight and curve singularity. It is easy to check that we can extend the ridgelets frame to monoscale ridgelets, and the resulting monoscale ridgelets constitutes a tight frame too.

Because our main aim is to investigate the ability of ridgelets frame to recover line structure in image, we only use a simple hard threshold algorithm instead of sophisticated ones. And note that the hard threshold algorithm is commonly used in wavelet domain in the literature of image denoising.

We carried out the experiments on a synthesis image, which is shown in Fig. 2. And the image is contaminated with additive Gaussian white noise with different variance levels. We compared the quality of the denoising algorithm based on ridgelets frame with those based on the decimated wavelets (DWT) and undecimated wavelet (UDWT). In Table 1, the PSNR of different algorithms are listed for noise level with different standard variance, where the PSNR is expressed using dB.

From Table 1, it is obvious that the method based on ridgelets frame outperforms substantially those based on wavelets for all noise levels.

In addition to the comparison in terms of PSNR, we display also the denoised images and their crops for visual comparison in Fig. 2 and Fig. 3.

Table 2. Comparison of performance of different transforms for image doising

Noise level	DWT	UDWT	Ridgelets frame
15	30.9639	34.318	35.3746
20	29.2861	32.2668	34.0756
25	27.8041	30.7049	33.0962
30	26.6904	29.4834	32.0935
40	24.7231	27.8679	30.676
60	22.3687	25.7513	28.2237
80	20.6152	24.3358	26.2391
100	19.2911	23.3936	24.8497

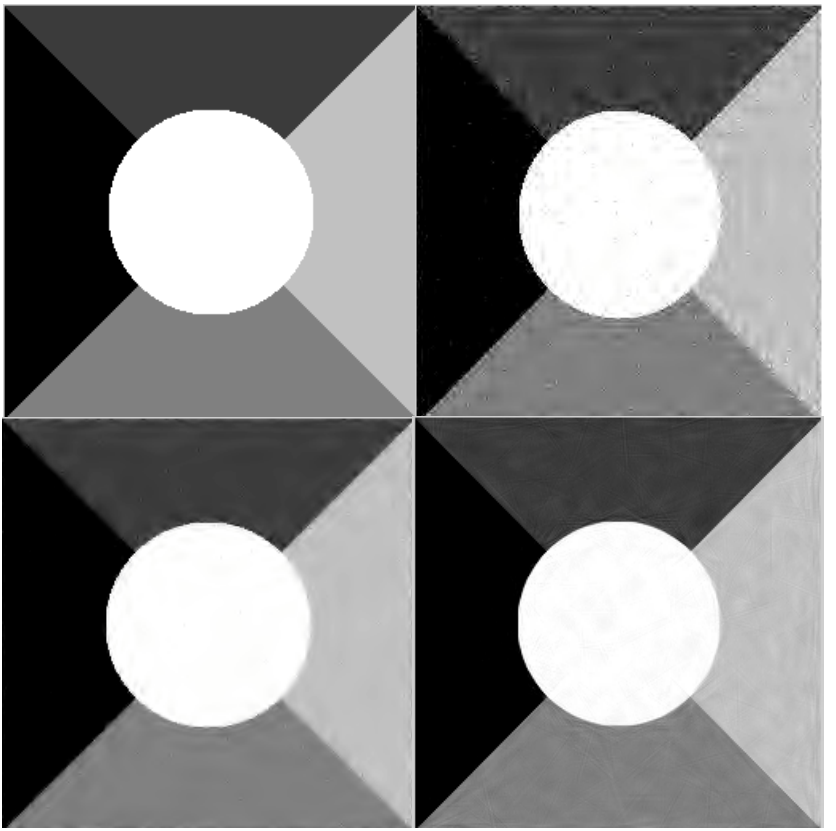


Fig. 2. Visual comparison of denoised results by different methods in the presence of noise with standard variance $\sigma = 20$. Top-left: original image; Top-right: denoising result using DWT, PSNR=29.2861; Bottom-left: denoising result using UDWT, PSNR=32.2668; Bottom-right: denoising result using ridgelets frame, PSNR=34.0756

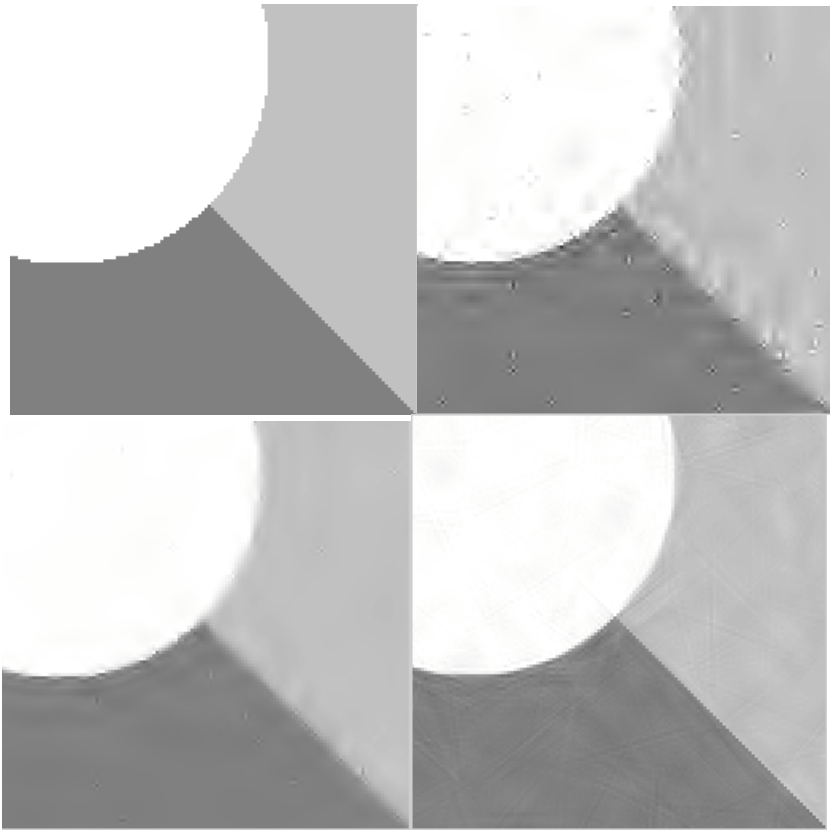


Fig. 3. Visual comparison of crop of denoised results by different methods in the presence of noise with standard variance $\sigma = 20$. Top-left: original image; Top-right: denoising result using DWT, PSNR=29.2861; Bottom-left: denoising result using UDWT, PSNR=32.2668; Bottom-right: denoising result using ridgelets frame, PSNR=34.0756

The ability of ridgelets frame to recover the line structure is well revealed by the comparison of visual effect in Fig. 2 and Fig. 3. For decimated wavelet, the resulting image is blemished by artifacts seriously, exactly the same as the usual case in the literature of image denoising. In the case of undecimated wavelet, there are few artifacts, however the line structures blur obviously. On the contrary, the line structure in image is well recovered when using the ridgelets frame.

4 Conclusion

In this paper, we have constructed a new system called ridgelets frame, which is a tight frame with frame bound 1. The ridgelets frame is characteristic of representing

effectively the line structure as its forerunner, orthonormal ridgelets. And the effectiveness of ridgelets frame to recover line structure in noisy image is revealed by experiments.

The ridgelets frame provides a powerful tool for various applications, especially for image processing tasks. However, there is much wok necessary to do, for example, to find new application of the ridgelets frame and to establish statistic model in ridgelets domain, as have been done in wavelet domain by many researchers.

References

1. Donoho, D.L.: Orthonormal Ridgelets and Linear Singularities. SIAM J. Math Anal. 5 (2000) 1062–1099
2. Flesia, A.G., Helor, H.A., Averbuch, E.J., Candès, E.J., Coifman, R.R., Donoho, D.L.: Digital Implementation of Ridgelet Packets. Stanford Univ., Stanford, CA, Tech. Rep. (2002)
3. Deans, S. R.: The Radon Transform and Some of Its Applications. Wiley, New York (1983)
4. Candès, E.J.: Monoscale Ridgelets for the Representation of Images with Edges. Dept. Statist., Stanford Univ., Stanford, CA, Tech. Rep. (1999)
5. Candès, E.J., Donoho, D.L.: Curvelets—a Surprisingly Effective Nonadaptive Representation for Objects with Edges. In: Cohen, A., Rabut, C., Schumaker, L.L. (eds.): Curve and Surface Fitting. Van-derbilt Univ. Press, Nashville (1999)

Adaptive Curved Feature Detection Based on Ridgelet

Kang Liu and Licheng Jiao

National Key Lab for Radar Signal Processing
and Institute of Intelligent Information Processing,
Xidian University, 710071 Xi'an, China
liukang1981@hotmail.com

Abstract. Feature detection always is an important problem in image processing. Ridgelet performs very well for objects with linear singularities. Based on the idea of ridgelet, this paper presents an adaptive algorithm for detecting curved feature in anisotropic images. The curve is adaptively partitioned into fragments with different length, and these fragments are nearly straight at fine scales, then it can be detected by using ridgelet transform. Experimental results prove the efficiency of this algorithm.

1 Introduction

Edge detection is always an important problem in image processing. Recently, several methods based on wavelets had been proposed for edge detection. Wavelets perform very well for objects with point singularities and are shown to be optimal basis for representing discontinuous functions in one dimension and functions with point-like phenomena in higher dimensions. However, edges always represent 1-dimensional singularities and wavelets are not the optimal basis for representing them. To resolve this problem, Candès introduce a new analysis tool named ridgelets in his Ph.D. Thesis [1]. The bivariate ridgelt function is defined as follow:

$$\psi_{a,b,\theta}(x) = a^{-1/2}\psi\left(\left(x_1\cos\theta + x_2\sin\theta - b\right)/a\right). \tag{1}$$

$a>0$, $b\in R$, $\theta\in[0,2\pi)$. Given an integrable bivariate function $f(x)$, its ridgelet coefficients are defined by [1], [4]:

$$R_f(a,b,\theta) = \int \psi_{a,b,\theta}(x) f(x) dx. \tag{2}$$

Ridgelets can effectively deal with linelike phenomena in dimension 2. But to objects with curved singularities, the approach performance of ridgelet is equal to wavelet and not the optimal basis. Candès present a method named monoscale ridgelets analysis that we can smoothly partition the image into many blocks with same size and each fragment of the curve in the block are nearly straight at fine scales [2]. This is a non-adaptive method for representing the image. It is difficult for us to decide the size of the partitioned block. The size of the block being too large would produce errors after detection and too small would increase the cost of the computation.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 487–494, 2004.
© Springer-Verlag Berlin Heidelberg 2004

This paper advances a novel adaptive algorithm based on ridgelet transform for detecting curved feature in an image in the frame of ridgelet analysis. We apply this method to sar images and results prove the efficiency of this algorithm.

We firstly outline an implementation strategy of discrete ridgelet transform and next introduce the basic ideas and the implementation process of our curved feature detection algorithm in detail. We present results of several experiments in section 3 and make an analysis. Finally we propose the conclusion and possibilities for future work.

2 Adaptive Curved Feature Detection Based on Ridgelet

2.1 Discrete Ridgelet Transform

Ridgelet analysis can be construed as wavelet analysis in the Radon domain and the ridgelet transform is precisely the application of a 1-D wavelet transform to the slices of the Radon transform [6]. We have

$$R_f(a,b,\theta) = \int Rf(\theta,t)\, a^{-1/2} \psi\left((t-b)/a\right) dt, \tag{3}$$

where $Rf(\theta,t)$ is the Radon transform of the function which is given by $Rf(\theta,t) = \int f(x_1,x_2)\delta(x_1\cos\theta + x_2\sin\theta - t)\, dx_1 dx_2$ where $(\theta,t) \in [0,2\pi)\times R$ and δ is the Dirac distribution. $\psi((t-b)/a)$ is a 1-D wavelet function. So linear singularities in the image can be send to point singularities by Radon transform, and then wavelets are fully efficient at dealing with point-like singularities, which is equal to that ridgelet perform well to linear singularities.

So the key step of ridgelet transform is accurate Radon transform. To a digital image, a widely used approach of the Radon transform is applying the 1-D inverse Fourier transform to the 2-D Fourier transform restricted to radial lines going through the origin [3]. It can obtained by the following steps (let $(f(i_1,i_2))$, $1 \le i_1, i_2 \le n$ be a digital image):

①. 2-D FFT. Compute the 2-D FFT of f giving the array $(\hat{f}(k_1,k_2))$, $-n \le k_1, k_2 \le n-1$ after padding the array f $n\times n$ to be $2n\times 2n$ by adding extra rows and columns of zeros in every tow rows and columns respectively.

②. Using the interpolation scheme and regarding the center of the image $\hat{f}(k_1,k_2)$ as the coordinate origin, we obtain the Cartesian-to-polar conversion and the radial slice of the Fourier transform. There are in total $2n$ direction angles, each direction corresponding to a radial array composed of $2n$ points. We used trigonometric interpolation rather than nearest-neighbor interpolation used by Starck in [5]. Fig. 1 shows the geometry of the polar grid, and each line crossing the origin denote a direction. (where n=8, so there are in total 16 directions).

③. 1-D IFFT. Compute the 1-DIFFT along each line. We denote it by $R(u_1, u_2)$, $1 \le u_1, u_2 \le 2n$, where u_1 denote the distance between the center of the block and the line, u_2 denote the angle of the line in the block.

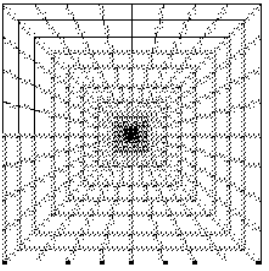


Fig. 1. The geometry of the polar grid *(n=8)*

Because of our using 1-D IFFT of length $2n$ on $2n$ lines, the total work takes $O(N \log N)$, where $N = n^2$ .To complete the ridgelet transform, we must take a 1-D wavelet transform along the radial variable in Radon domain. We choose the dyadic wavelets transform [7], defined by:

$$Wf(u, 2^j) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{2^j}} \phi\left(\frac{t-u}{2^j}\right) dt = f * \overline{\phi}_{2^j}(u), \tag{4}$$

where $\overline{\phi}_{2^j}(t) = \phi_{2^j}(-t) = \frac{1}{\sqrt{2^j}} \phi\left(-\frac{t}{2^j}\right)$. Because of its undecimated property it can capture as many as possible characteristics of a signal or an image and make us can measure the position and the magnitude of the point-like singularities. We use 3-order B-spline wavelet which is wildly used in edge detections.

2.2 The Basic Idea of the Proposed Algorithm

Ridgelet only performs very well to detect linear features in an image. However, edges are typically curved rather than straight. The object with curved singularities is still curvilineal one and not a point after Radon transform. So in the Radon space, its wavelet coefficients are not sparse and the ridgelet alone can not yield efficient representations. Candès introduce monoscale ridgelets transform that the image is partitioned into several congruent blocks with fixed side-length, and at sufficiently fine scales, a curved edge is almost straight then it can be detected by using ridgelet transform [2].

We promise only one line exist in each block. However, because of the limitation of the pels, we can not partition the image infinitely. To a n by n image, we have $2n$ directions while we make randon transform to it. With the size of the block

smaller, the number of the directions would decrease more and more. Then it would produce errors when we detected the direction of the line and the computation cost would increase. If the size of the block were too large, we can not detect the position and the length of the curved features accurately. Figure 2 give out four cases which maybe produce errors after detection because of too large block.

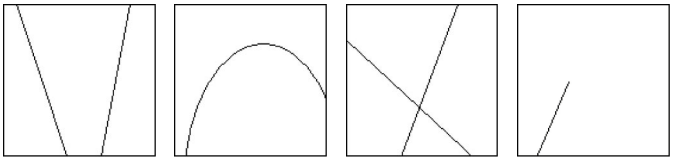


Fig. 2. The cases which the size of the block is too large and maybe produce errors after detection

2.3 Adaptive Algorithm Based on Ridgelet for Detecting Curved Features

We present an adaptive algorithm based on ridgelet transform for detecting curved features here. The size of the block can be changed adaptively. An image is partitioned into several congruent blocks with initial side-length. When each of four cases shown above came forth, that block should be partitioned into four parts with the same size. Then in each part, we use ridgelet to detect it again.

Firstly, we make ridgelet transform to each block partitioned using proposed method. We get ridgelets coefficients array denoted as $W_R f(u_1, u_2)$.

Because of Radon transform, linear singularities are be sent into point singularities, wavelets coefficients of these points are local maximum values. So we search for the maximum value whose absolute value is the biggest in $W_R f(u_1, u_2)$, and write it as Mmax. Then search for the maximum absolute value denoted by Mmax2 in this block except the small region with Mmax being the center. The size of this small region is set to be 5×5 in experiments. Let T be the threshold.

• While Mmax>T and Mmax2>T, it is corresponding to that two lines or one curve which radian value is large exist in one block. Then we partition the block into four parts with the same size, and each part will be deal with again.

• While Mmax $\leq$ T, it shows that no lines or curves exist in this block. Then it wouldn't be detected.

• While Mmax>T and Mmax2 $\leq$ T, it represent that only one line exists in the block, and there are two cases:

①. While Mmax>kT, it is corresponding to that the line cross the whole block. Then this block should be detected immediately. We define k=1.5 in experiments.

②. While T<Mmax $\leq$ kT, it shows that the line don't cross the whole block. Then we should partition it into four parts and each part should be judged or detected again.

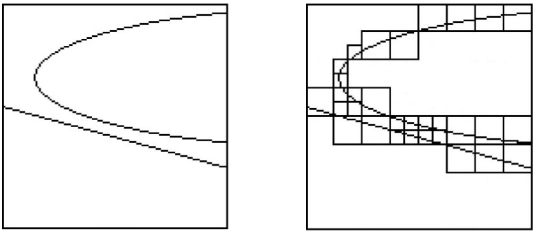


Fig. 3. (*Left*) original image with curved features, (*Right*) partitioned image using adaptive method

Fig. 3 shows the sketch map after partitioned using our method. Now we present the detailed process of the adaptive algorithm based on ridgelet:

Step1. Apply the method introduced by Hou Biao to a digital image $f(i_1, i_2)$, $1 \le i_1, i_2 \le n$ to form ridgelet subbands [8], and denote it by $f_i(i_1, i_2), (i = 1, 2)$. And partition $f_2(i_1, i_2)$ into non-overlapping blocks with size $L \times L$, denoting it by $G_{i,j}(k, l)$ $(1 \le k, l \le L \quad 1 \le i, j \le n / L)$. Initialize $L = L_{\max}$, the smallest side-length of the block is $L_{\min}$.

Step2. Make Radon transform mentioned in sec.2.1 to each block. Results are given by $R_{i,j}(u_1, u_2)$, $(1 \le u_1, u_2 \le 2L \quad 1 \le i, j \le n / L)$.

Step3. Take dyadic wavelet transform to each column of $R_{i,j}(u_1, u_2)$ and results are written as $W^0_{i,j} f(u_1, u_2)$, $W^m_{i,j} f(u_1, u_2)$ $(m = 1, 2, 3)$. Then find out the value Mmax and Mmax2 in $W^1_{i,j} f(u_1, u_2)$ using the method mentioned above. Set T be the threshold.

Step4. While $L = L_{\min}$, this block can not be partitioned into four parts and while Mmax>T, go to step 5, otherwise don't detect it.

While $L > L_{\min}$, do:

If Mmax $\le T$, this block would not be detected and move to the next block. Go back to step2.

If T <Mmax $\le kT$ and Mmax2 $\le T$, then partition this block to four parts and repeat processes mentioned above from step 2.

If Mmax2> T, then partition this block to four parts and repeat processes mentioned above from step2.

If Mmax2 $\le T$ and Mmax> kT, then detect this block immediately. Go to step 5.

Step5. Search for the location of the maximum absolute value of ridgelet coefficients array $W^1_{i,j} f(u_1, u_2)$ and record the corresponding coordinate,

$(u_{1\max}, u_{2\max})$. From $(u_{1\max}, u_{2\max})$ we can obtain the distance form the line to the center of the block and angel of the line, writing them as (t, θ).

Step6. Define an array which is a zero matrix with the same size of the block. Regard the center of this block as the coordinate origin. On the line across the origin and of direction angle $\theta + \pi/2$, we find out the point the distance from which to the origin is equal to t. Then the line across this point and of the slope θ is the desired one. We find out the coordinates of the two points on the line that intersect the borderlines of the block, and then computer all the coordinates of the points on the line between two points using linear interpolation.

Step7. Go to step2 and detect the next block. Finally synthesize a binary edge image composed of several blocks obtained above.

Because the block after partitioning are non-overlapping, while one curve is across the corner of this block, the part of the curve in the block is weak and it is difficult to detect it accurately on the effect of the noises. In experiments, we can see that the curve in the result have several broken parts (shown in Fig. 4). To resolve it, we search for broken parts and use linear interpolation to connect the adjacent line segments. Because these broken parts are always small, the result after interpolation is not likely to change the original result too much.

Applying the above steps to an image, we can detect curved singularities efficiently, each of which is composed of many linear segments with different length. It also can detect the length of curves or lines accurately. It is hard to do for classical Radon transform and Hough transform.

3 Experiments

Based on the algorithm mentioned above, three images: a basic curve, two noisy circularities (standard deviation Sigma=40, PSNR=16.0973) and a sar image are worked out. And experiment results is shown in fig. 4, fig. 5 and fig. 6 respectively. In the experiment we initialize the size of blocks to be 16×16, i.e. $L_{\max}$ =16. And the size of the smallest partitioned block is 8×8. The smallest distance between two circles (see fig.5) is no more than 16. To the sar image (see fig.6), we detect the river after filtering because this can decrease the effect of the speckle noises, where we use median filter. Then we apply wavelet and our method to detect river edges respectively. We choose 3-order B-spline wavelet basis and use uniform threshold method [7] in experiments.

In fig. 4, we can see that our method not only performs well for detecting general curved singularities in an image, but also the part whose curvature is large. Though the result has many broken parts which are small, the direction of the curve has been detected accurately. After we search for broken parts and fill them using linear interpolation method, we can obtain the exactly full detected results (see fig.5 and fig.6). From results of SAR image after detection in fig. 6, we can see that our method is better than the method based on wavelets for restraining the effect of speckle noises on edges of this image. The whole contour of the river has been detected accurately, and we can locate the positions of the curves and compute the length of the curves.

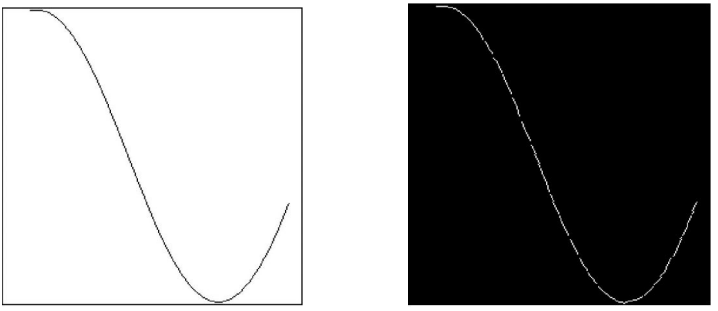


Fig. 4. (*Left*) a basic curve image, (*Right*) the result after detection

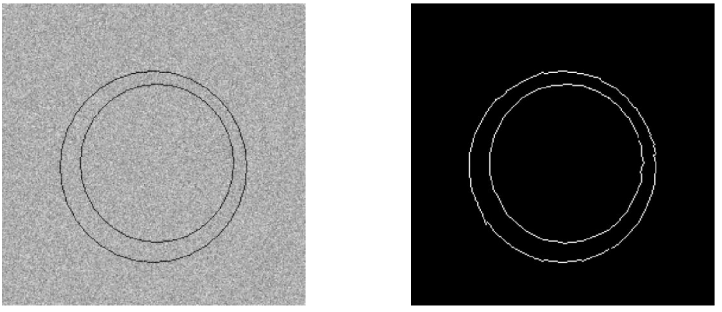


Fig. 5. (*Left*) an original image with two noisy circularities (*standard deviation Sigma=40, PSNR=16.0973*), (*Right*) the result after detection

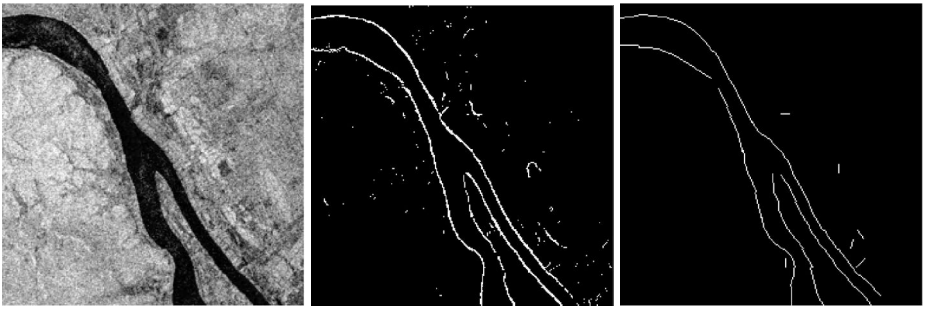


Fig. 6. (*Left*) a SAR image with speckle noises, (*Middle*) the result after detection using wavelets, (*Right*) the result after detection using our method

4 Conclusion

Ridgelets send linear singularities into point singularities by using its capability of reducing dimension. It can capture the linear singularities in the image rapidly and this is hard for wavelet to do. Based on ridgelet transform we change problem of curved singularities detection to the problem of linear singularities detection by using the idea of adaptively partitioning the image. We can locate the position of each linear segment and its length. The results of the experiments prove the efficiency and advantage of our algorithm. However, because the blocks partitioned are non-overlapping, we can see several broken parts in results. How to avoid these? Can we partition the image into many blocks which are overlapping? But the overlapping blocks must lead the cost of the computation to increase. These give us a next question to discuss.

References

1. Candès, E.J.: Ridgelets: Theory and Application. PhD Thesis, Department of Statistics, Stanford University (1998)
2. Candès, E.J.: Monoscale Ridgelets for the Representation of Images with Edges[R]. Department of Statistics, Stanford University (1999)
3. Averbuch, A., Coifman, R.R., Donoho, D.L., Israeli, M. and Walden, J.: Fast Slant Stack : A Notion of Radon Transform for Data in a Cartesian Grid which is Rapidly Computible, Algebraically Exact, Geometrically Faithful an Invertible. Department of Statistics, Stanford University (2001)
4. Candès, E.J. and Donoho, D.L.: Recovering Edges in Ill-Posed Inverse Problems: Optimality of Curvelet Frames. Department of Statistics, Stanford University (2000)
5. Starck, J.L., Candès, E.L. and Donoho, D.L.: The Curvelet Transform for Image Denoising. IEEE Transactions on Image Processing, 11 (2002) 670–684
6. Candès, E.J. and Donoho, D.L.: Ridgelts: A Key to Higher-Dimensional Intermittency? Department of Statistics, Stanford University (1999)
7. Mallat, S.: A Wavelet Tour of Signal Processing, Second Edition. CA: Academic Press (1999)
8. Hou, B., Liu, F. and Jiao, L.C.: Linear Feature Detection Based on Ridgelet. Science in China, Ser.E., 46 (2003) 141–152

Globally Stabilized 3L Curve Fitting

Turker Sahin and Mustafa Unel

Department of Computer Engineering, Gebze Institute of Technology
Cayirova Campus
41400 Gebze/Kocaeli Turkey
{htsahin,munel}@bilmuh.gyte.edu.tr

Abstract. Although some of the linear curve fitting techniques provide improvements over the classical least squares fit algorithm, most of them cannot globally stabilize majority of data sets, and are not robust enough to handle moderate levels of noise or missing data. In this paper, we apply "ridge regression regularization" to strengthen the stability and robustness of a linear fitting method, 3L fitting algorithm, while maintaining its Euclidean invariance.

1 Introduction

Implicit polynomial (IP) models have proven to be more suitable than parametric representations for fitting algebraic curves to data with their advantages like global shape representation, smoothing noisy data and robustness against occlusion [1,2,3,4,5,6,7,8]. Nonlinear optimization methods have been commonly applied for IP curve modelling; however, they suffer from high computational complexity and cost [2,3,4]. Recently linear approaches to curve fitting have started to emerge, which address such problems [9,10,11]. However, these techniques usually cannot provide globally stabilized fits for many cases and are not robust versus perturbational effects like noise.

In this paper as a way to overcome these problems, we apply ridge regression regularization to the 3L fitting method. We have observed that the ridge regression regularization of Gradient1 method expressed in [11] does not provide satisfactory results for reasons like oversensitivity to changes in parameters and normalization. We have obtained better results with regularization of 3L, which we present for verifying the noticeable improvements in global stability and robustness, as well as insensitivity to parameter changes.

2 Implicit Curve Models and Data Set Normalization

2.1 Algebraic Curves

Algebraic curves are represented by implicit polynomial models of arbitrary degree, n, as:

$$f_n(x,y) = a_{00} + a_{10}x + a_{01}y + \ldots + a_{n0}x^n + a_{n-1,1}x^{n-1}y + \ldots + a_{0n}y^n$$

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 495–502, 2004.
© Springer-Verlag Berlin Heidelberg 2004

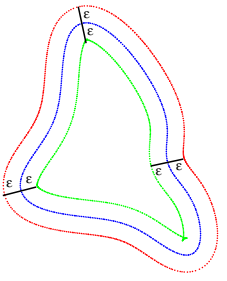


Fig. 1. The 3 Levels of data for a free from boundary

$$= \underbrace{\begin{bmatrix}1 & x & y & \dots & y^n\end{bmatrix}}_{Y^T} \underbrace{\begin{bmatrix}a_{00} & a_{10} & a_{01} & \dots & a_{0n}\end{bmatrix}^T}_{A} = Y^T A = 0 \tag{1}$$

where Y^T is the vector of monomials and A is the vector of IP coefficients.

2.2 Normalization

An important concept integrated into many fitting methods is data set normalization to reduce the pathological effects in the resulting IP's, which usually arise because of their high degree terms taking large values. In our experiments radial distance normalization is used. This is a linear process, based on dividing every data point, $(x_i, y_i)_{i=1}^N$, with the average radial distance of the set $\frac{1}{N}\sum_i (x_i^2 + y_i^2)^{1/2}$ after the center of data has been shifted to origin. It has been observed to give better results than other proposed normalizations [11] in our experiments.

3 The 3L Linear Fitting Method

The objective of all linear IP curve fitting techniques is to approximate a given data set with a polynomial as closely as possible by minimization of their algebraic distance. The adopted 3L algorithm [9] uses the following principle for this minimization procedure: closed-bounded IP's should have zero values at the data points, negative values for inside points and positive values for outside points, or vice versa. Thus any data set to be curve fitted is first integrated with two more data sets with points at a distance, ε inside and outside the original data as in figure 1. Accordingly the IP function is forced to take $+1$ value at the outer layer, -1 at the inner level, and 0 at the intermediate layer. Thus a b vector and the matrix of 3 layers of data as M is prepared such that:

$$b = \begin{bmatrix}+1 \dots +1\ 0 \dots 0\ -1 \dots -1\end{bmatrix}^T_{(3N\times 1)}, \quad M = \begin{bmatrix}M_{+\varepsilon} \\ M_0 \\ M_{-\varepsilon}\end{bmatrix} = \begin{bmatrix}Y_1^T \\ Y_2^T \\ \dots \\ Y_{3N}^T\end{bmatrix}_{3N\times c}$$

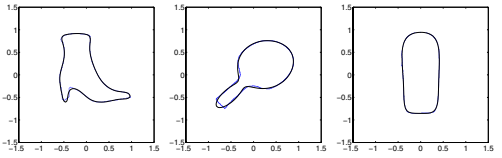


Fig. 2. Some stable 4th and 6th degree 3L fits

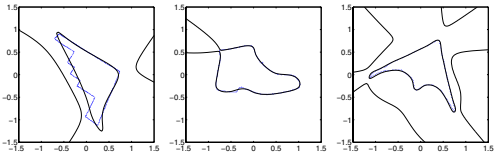


Fig. 3. Examples of unstable 3L fits (a) a 4th degree fit for a B21 plane (b) 6th degree fit of a Shoe; and (c) 8th degree for a Glider

where Y_i are the vectors of monomials for the 3 layers of data, N is the number of data points, n is the degree of polynomial and $c = (n+1)(n+2)/2$ is the number of the coefficients of the IP curve. The resulting curve coefficient vector is obtained by:

$$A = M^{\dagger} b \tag{2}$$

where $M^{\dagger} = (M^T M)^{-1} M^T$ is the pseudo-inverse matrix for M. This method is Invariant under Euclidean transformations, as the two synthetic layers is formed by using the distance measure ε.

4 Global Stability by the Ridge Regression Regularization

Linear curve fitting techniques achieve local stability around the data points; however, are weak in providing global stability. A reason for this is the near collinearity in the data, which cause the $M^T M$ matrix of products of the monomials to be almost singular with some eigenvalues much smaller than the others. Such eigenvalues do not contribute to the fit around the data set and cause extra open unstable branches.

Ridge regression is a computationally efficient method for reducing data collinearity and the resulting instability [11]. By this technique, the condition number of $M^T M$ is improved, and the extra curves are moved to infinity, where they disappear giving a stable closed bounded fit. To achieve this, a κD term is applied to equation (2) as:

$$A_{\kappa} = (M^T M + \kappa D)^{-1} M^T b \tag{3}$$

Here κ is the ridge regression parameter, which is to be increased from 0 to higher values until a stable closed bounded curve is obtained.

The other part of the ridge regression term is the diagonal D matrix, which has the same number of terms as the the coefficient vector, A_κ. The entries of D can be obtained by:

$$D_{ii} = \beta_{j+k} \frac{j!k!}{(j+k)!} \tag{4}$$

where the index for each diagonal element is calculated according to variation of the degrees of the x and y components in equation (1) by $i = k + \frac{(j+k)(j+k+1)}{2} + 1$. Also β_{j+k} is chosen to be :

$$\beta_{j+k} = \sum_{r,s \geq 0; r+s=j+k} \frac{(r+s)!}{r!s!} \sum_{l=1}^{N} x_l^{2r} y_l^{2s} \tag{5}$$

or when expanded:

$$\beta_0 = \sum_{l=1}^{N} x_l^0 y_l^0 = N$$

$$\beta_1 = \sum_{l=1}^{N} x_l^2 + y_l^2 = \sum_{l} (x_k^2 + y_k^2)^1$$

$$\beta_2 = \frac{2!}{2!0!} \sum_{l} x_l^4 + \frac{2!}{1!1!} \sum_{l} x_l^2 y_l^2 + \frac{2!}{0!2!} \sum_{l} y_l^4 = \sum_{l} (x_l^2 + y_l^2)^2$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\beta_n = \sum_{l} (x_l^2 + y_l^2)^n$$

with "n" the degree of the resulting IP and $(x_l, y_l)_{l=1}^{N}$ are the elements of the normalized object data. As a result the entries of D are set to the invariantly weighted sum of the diagonal elements of $M^T M$, which is an Euclidean invariant measure. Therefore inherent Euclidean invariance properties of fitting methods are preserved by this approach.

5 Experimental Results and Discussion

In this section many fits obtained from ridge regression technique are compared to those of non-regularized 3L, to depict the resulting stability and robustness improvements. The 3L fitting technique occasionally gives reasonable results. Some examples are a boot, a Racket and a Siemens mobile phone in Figure 2. Here the boot is a $6th$ degree fit example, while the others are $4th$ degree IP's. However, generally this fitting method has weak global stability properties, which cause significant problems for applications. One related issue is that usually a

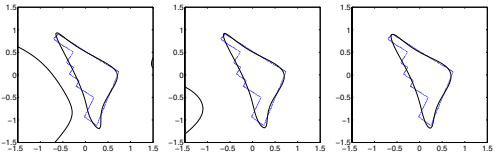


Fig. 4. Stabilization of the B21 plane fit by ridge regression method: (a) $\kappa = 10^{-4}$ (b) $\kappa = 5 \times 10^{-4}$ (c) $\kappa = 10^{-3}$

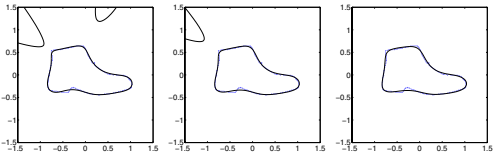


Fig. 5. Stabilization of the Shoe fit by ridge regression method: (a) $\kappa = 10^{-4}$ (b) $\kappa = 5 \times 10^{-4}$ (c) $\kappa = 1.5 \times 10^{-3}$

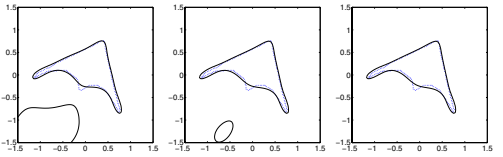


Fig. 6. Stabilized 8th degree Glider fit by ridge regression method: (a) $\kappa = 10^{-4}$ (b) $\kappa = 4 \times 10^{-4}$ (c) $\kappa = 5 \times 10^{-4}$

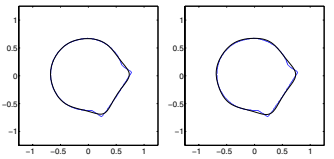


Fig. 7. The regularized fit for Cd Box of (a) $\varepsilon = 0.02$ (b) $\varepsilon = 0.07$. Stabilizing $\kappa = 4 \times 10^{-3}$ for both cases

data to be modelled can be fit stably by IP's of one or two different degrees, but not by others. For example, among the objects in Figure 2, the boot does not have stable $4th$ or $8th$ degree fits, while the racket cannot be stabilized above $6th$ degree IP's. A further problem is this method cannot cannot give stable fits for many important data at all. Some examples are depicted in Figure 3. These curves are with a $4th$ degree IP for the B21 plane, $6th$ degree for the shoe and $8th$ degree for the glider. Each of these data cannot be stabilized for fits of either

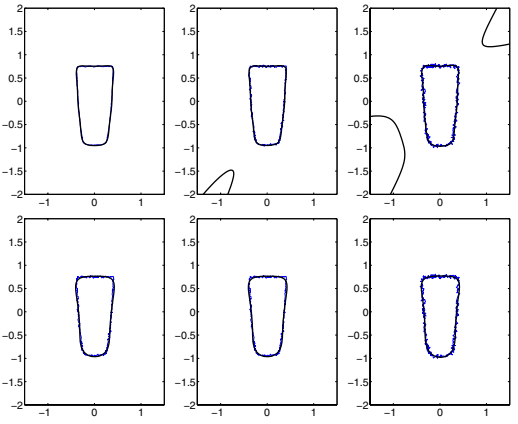


Fig. 8. Robustness improvement of ridge regression method versus noise. First row depict the degradation of 3L Vase fit, and the second row is the robustness of ridge regression based fit, both subjected to no noise in left subplots; moderate noise of $\sigma = 0.0075$ in the middle; and much higher noise of $\sigma = 0.02$ in the right.

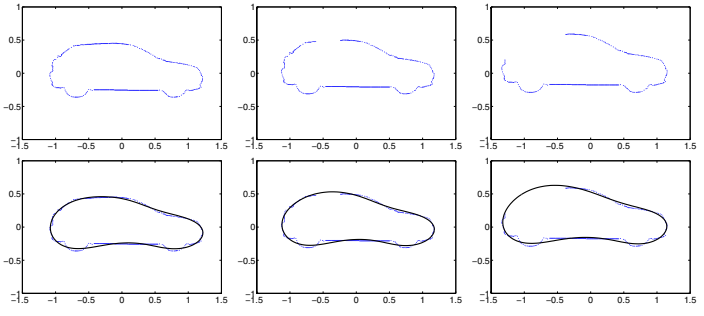


Fig. 9. The robustness of regularized curve fitting approach against Occlusion: In the first row plots, the left one is data is with no occlusion, middle is with 10% occlusion, and right is with 20% occlusion; In the second row are the corresponding fits.

$4th$ or $6th$ or $8th$ degrees, but they were modelled by IP's of one of these three degrees for more compact examplification.

All data data in Figure 3 can be globally stabilized by ridge regression regularization, which is depicted in figures 4 - 6. Again these are $4th$, $6th$ and $8th$ degree fits for the B21, the shoe and the glider which indicate the ability of this technique to model data with various degree curves. As presented in their subplots, when the κ parameter is increased from zero to the range $10^{-4} - 10^{-2}$, the extra unstable curves tend to move away from the actual data set and disappear. The insensitivity of ridge regression based 3L method to variations in the ε parameter is examplified in Figure 7 for two CD Box fits of $\varepsilon = 0.02$ and

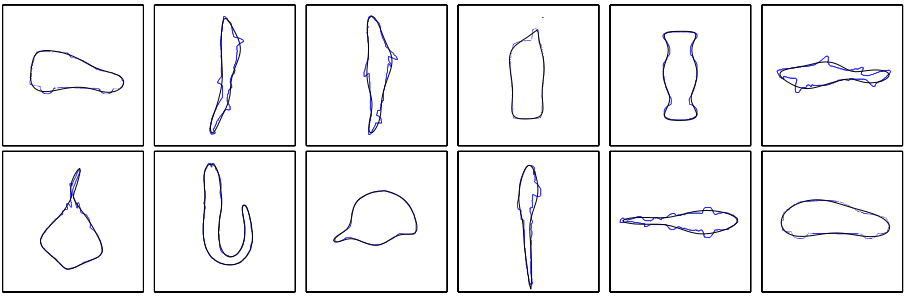


Fig. 10. Some globally stabilized curves of marine creatures and man-made objects. The first four objects are modelled by 8th degree, the next six of them are modelled by 6th degree and the last two are modelled by 4th degree algebraic curves

0.07. As observable these fits cannot be distinguished from each other; moreover the stabilizing κ's are the same for both cases. Thus a single ε can be used for modelling all data sets. In deed, $\varepsilon = 0.05$ has been used in all other example figures.

Robustness of the this approach has also been verified by fits to noisy and occluded cases. Figure 8 shows the improvement in robustness against noise by ridge regression application over the 3L only case. Here the top row is the non-regularized fits, while the second row is the ridge regression based fits. Also the left subplots depict noise free cases; the middle plots are with moderate noise level of $\sigma = 0.0075$; and the right ones are for the much higher noise of $\sigma = 0.02$. It can be observed the vase data can be stably fit in presence of much higher noise levels with ridge regression technique, which verifies the remarkable robustness of this technique to Gaussian perturbations.

In figure 9 robustness of this regularization for occlusion or data loss has been examplified. The first row depict the employed car data for cases of no occlusion, 10% of data chopped, and 20% occlusion from left to right. The lower row shows the corresponding curve fits to each case, which are all stable and very near to each other in shape. Thus this method can also be applied to cases of missing data with reasonable accuracy.

Finally we present $4{-}8th$ degree fits of various objects in Figure 10. Nearly all these data could be globally stabilized using κ values of no more than $5{\times}10^{-3}$ and none require κ's more than 10^{-2}, which indicate the strong global stabilizability properties of this method.

6 Conclusions

The ridge regression Regularization dramatically improves the poor global stability properties of the 3L technique it has been applied. Thus by application of this global stabilization method, a much wider range of data can be accurately fit by IP's of all degrees. Ridge regression also improves the robustness of curves

to occlusion and especially noise significantly. The parameter tuning process of this approach is also much simpler, as it is much less sensitive to parameter and normalization changes. Moreover, this method preserves the Euclidean invariance, thus should be very suitable for many important applications in motion identification, pose estimation and object recognition.

Acknowledgments. This research was supported from GYTE research grant BAP #2003*A*23. The data of employed marine creatures are courtesy of University of Surrey, UK.

References

1. M. Pilu, A. Fitzgibbon and R. Fisher, "Ellipse Specific Direct Least Squares Fitting," Proc. IEEE, International Conference on Image Processing, Lausanne, Switzerland, September 1996.
2. G. Taubin, "Estimation of Planar Curves, Surfaces and Nonplanar Space Curves Defined by Implicit Equations with Applications to Edge and Range Segmentation," IEEE TPAMI, Vol. 13, pp. 1115-1138, 1991.
3. G. Taubin, et al. "Parametrized Families of Polynomials fo Bounded Algebraic Curve and Surface Fitting," IEEE Transactions on Pattern Analysis and Machine Vision, 16(3):287-303, March 1994.
4. D. Keren, D. Cooper and J. Subrahmonia, "Describing Complicated Objects by Implicit Polynomials," IEEE Transactions on Pattern Analysis and Machine Vision, Vol. 16, pp. 38-53, 1994.
5. W. A. Wolovich and M. Unel, "The Determination of Implicit Polynomail Canonical Curves," IEEE TPAMI, Vol. 20(8), 1998.
6. M Unel and W. A. Wolovich, "On the Construction of Complete Sets of Geometric Invariants for Algebraic Curves," Advances In Applied Mathematics 24, 65-87, 2000.
7. M Unel and W. A. Wolovich, "A New Representation for Quartic Curves and Complete Sets of Geometric Invariants," International Journal of Pattern Recognition and Artificail Intelligence, Vol. 13(8), 1999.
8. J. Subrahmonia, D. B. Cooper and D. Keren, "Practical Reliable Bayesian Recognition of 2D and 3D Objects using Implicit Polynomials and Algebraic Invariants", IEEE TPAMI, 18(5):505-519, 1996.
9. M.Blane, Z.Lei et al., The 3L algorithm for Fitting Implicit Polynomial Curves and Surfaces to Data, IEEE Transaction on Pattern Analysis and Machine Intelligence, Bol.22, No.3, March 2000.
10. Z. Lei and D. B. Cooper, "New, Faster, More Controlled Fitting of Implicit Polynomial 2D Curves and 3D Surfaces to Data," IEEE Conference on Computer Vision and Pattern Recognition, June 1996.
11. T. Tasdizen, J-P Tarel and D. B. Cooper, " Improving the Stability of Algebraic Curves for Applications," IEEE Transactions on Image Processing, Vol. 9, No: 3, pp. 405-416, March 2000.

Learning an Information Theoretic Transform for Object Detection

Jianzhong Fang and Guoping Qiu

School of Computer Science, The University of Nottingham
{jzf, qiu}@cs.nott.ac.uk

Abstract. We present an information theoretic approach for learning a linear dimension reduction transform for object classification. The theoretic guidance of the approach is that the transform should minimize the classification error, which, according to Fano's optimal classification bound, amounts to maximizing the mutual information between the object class and the transformed feature. We propose a three-stage learning process. First, we use a support vector machine to select a subset of the training samples that are near the class boundaries. Second, we search this subset for the most informative samples to be used as the initial transform bases. Third, we use hill-climbing to refine these initial bases one at a time to maximize the mutual information between the transform coefficients and the object class distribution. We have applied the technique to face detection and we present encouraging results.

1 Introduction

Representation plays a key role in the success of computer vision and pattern recognition algorithms. An effective representation method should be compact and discriminative. It is desired that the representation should have low dimensionality to combat the "curse of dimensionality" problem and to improve computational efficiency. The representation should also ideally be in a space where different classes of objects are well separated.

Classical techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) [7] are well studied in the literature. Although PCA can produce compact representation, it cannot enhance the discriminative power. Since LDA only makes use of covariance, it is only optimal for classes having unimodal Gaussian density with well-separated means. In many applications, it may be beneficial to exploit higher than second order statistical information.

Theoretically, information theoretic approaches [8] have a number of advantages. For example, mutual information measures general statistical dependence between variables rather than the linear correlation. The mutual information is also invariant to monotonic transformations performed on the variables.

In this paper, we present a learning procedure for developing a dimension reduction linear transform based on the mutual information criterion, and apply it to object

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 503–510, 2004.
© Springer-Verlag Berlin Heidelberg 2004

detection. The organization of the paper is as follows. Section 2 gives a brief background overview on the Shannon information theory and Fano's inequality on the relationship between mutual information and a lower bound of misclassification error [2]. Section 3 describes a 3-step learning procedure for deriving a mutual information maximizing linear dimension reduction transform. Section 4 presents experiments and results of applying the method to human face detection. Section 5 concludes the paper.

2 Information Theory Background

Let ensemble X be a random variable x with a set of possible outcomes, $A_X = \{a_1, a_2, \ldots. a_n\}$, having probabilities $\{P(x = a_i)\}$, and ensemble Y be a random variable y with a set of possible outcomes, $A_Y = \{b_1, b_2, \ldots. b_m\}$, having probabilities $\{P(y = b_j)\}$. Let $p(x, y)$, $x \in A_X$, $y \in A_Y$ be the joint probability. We can define the following Shannon information theory functions

The entropy of X is defined as

$$H(X) = -\sum_{x \in A_x} P(x) \log(P(x)) \tag{1}$$

The joint entropy of X and Y is defined as

$$H(X,Y) = -\sum_{x \in A_x} \sum_{y \in A_y} P(x,y) \log(P(x,y)) \tag{2}$$

The mutual information between X and Y can be defined as (other forms of definition also exist)

$$I(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{3}$$

The entropy measures the information content or uncertainty of the random variable. The mutual information measures the average reduction in uncertainty of x as a result of learning the value of y, or vice versa. Another interpretation of the mutual information measure is that it measures the amount of information x conveys about y.

2.1 Fano's Mutual Information Bound

In the context of object classification, Fano's inequality [2] gives a lower bound for the probability of error (an upper bound for the probability of correct classification). Our present application uses Fano's inequality in much the same way as it is used by other authors [3, 4]. The classification process can be interpreted as a Markov chain as illustrated in Fig. 1.

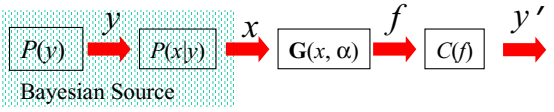


Fig. 1. Interpreting the classification process as a Markov chain [3, 4], y is the object class random variable, x are the observations generated by the conditional probability density function $P(x \mid y)$. The observations are subjected to a transform G, which produces a new feature f from input x. The classifier C then estimates the class identity of input x as y' based on the transformed feature f.

The probability of misclassification error in the setting of Fig. 1, $P_e = \mathrm{P}(\mathrm{y} \neq \mathrm{y}')$, has the following bound [2]

$$P(y \neq y') \geq \frac{H(Y) - I(Y,F) - 1}{\log(m)} \tag{4}$$

where F is the ensemble of random variable f, and m is the number of outputs of y (number of object classes). The form of the classifier, C, has not been specified. Eq. (4) quantifies at best how well we can classify the objects using the features f. However, an upper bound of the probability of misclassification error cannot be expressed in terms of Shannon's entropy. The best one can do is to minimizes the lower bound to ensure an appropriately designed classification algorithm does well. Since both m and $H(Y)$ are constants in (4), we can maximize the mutual information $I(Y, F)$ to minimize the lower bound of the probability of misclassification error. The task now becomes that of finding the transform function G that minimizes this lower bound. In the next section, we propose a three-stage solution.

3 Learning a Linear Informative Transform

Our objective is to find a dimension reduction linear transform G that minimizes the lower bound in (4). Because the observations x and the transformed feature f and class variable y are all normally multidimensional vectors, directly estimating an optimal G that maximizes $I(Y, F)$ is computationally extremely difficult. Assume x is an l-d column vector and f is a k-d column vector, ($k << l$), then $f = Gx$, G is a k (rows) by l (columns) transform matrix. In this section, we present an engineering solution to estimate G, which we will demonstrate in the next section it works satisfactorily.

The developed 3-stage process is as follows:

Stage 1: Use a support vector machine (SVM) [7] to select the training samples that are near the class boundaries

Stage 2: Use a maximum mutual information criterion to select an initial set of transform base vectors.

Stage 3: Use a hill-climbing algorithm to refine the transform base vectors one at a time.

Starting from labeled training samples, we use the raw data to train a support vector machine to classify the samples directly. After training, the "support vectors" of this SVM are those samples near the decision surfaces. Conceptually, this set of training samples (support vectors) are the most difficult to classify. We reason this set of samples is potentially the most "informative" because there is more uncertainty in relation to the class identity. We will therefore only use this set of samples to find the transform G. The rationale is that by reducing the number of samples, we can reduce the training time; also, if we can separate those samples near the decision boundaries, the rest of the samples can be separated easily. It is to be noted that the only use of the SVM at this stage is to select samples near the class boundaries.

Once the training samples have been selected, we use a constructive procedure to build an initial set of transform base vectors based on the maximum mutual information criterion. Let $X = \{x_1, x_2, \ldots x_N\}$ be the N labeled training samples that constitute the support vectors of the SVM, $Y = \{y_1, y_2, \ldots y_N\}$ their corresponding class labels, $G = [g_1, g_2, \ldots, g_k]^T$, and g_i be the ith transform base. We use the following procedure to find the initial value of G

Proc. Initialize$G(X,Y)$
  for $i = 1$ **to** k **do**
    for $m = 1$ **to** N **do**
      for $n = 1$ **to** N **do**
        $F(m, n) = \langle x_m, x_n \rangle$ //inner product
      End for
      Proc. Estimate joint probability $P(Y, F(m, \bullet))$
      Proc. Compute $I(m) = I(Y, F(m, \bullet))$
                //Compute mutual information (3)
    End for
    If $I(j) > I(m), \forall m$ **Then** $g_i = x_j / \|x_j\|$
    Remove x_j and y_j from X and Y respectively
    $N = N - 1$
    for $m = 1$ **to** N **do**
      $x_m = x_m - \langle x_m, g_i \rangle g_i$
    End for
  End for
End Proc.

To find the first transform base, we select one sample at a time, and project all other training samples onto that selected sample. The projection (a scalar) and the sample identity can be used to estimate the joint probability, which in turn can be used to estimate the mutual information of the projection and the class distribution. The sample with projection output that maximizes the mutual information is selected as the first transform base. This base is then removed from the training sample set. All remaining samples are then made orthogonal to the first base and used as training samples to find the second transform base. The process continues until all required k bases are found. From the procedure it is not difficult to see that all k initial bases are orthonormal.

It is clear that the bases are selected individually based on a maximum mutual information criterion. Ideally, these bases should be optimized jointly. However, estimating the joint probability of high dimensional vectors is computationally prohibitive. The mutual information function with respect to the transform G is non-differentiable. This makes a closed form optimization algorithm difficult to derive (if not impossible). Therefore some form of heuristic techniques have to be employed to refine the initial transform bases. We decided to use hill-climbing [10] to accomplish the task.

This is an iterative process. We refine the bases, g_i, one at a time. For each hill-climbing step, the criterion is the maximization of the mutual information between the projection of the training samples onto that base and the samples class distribution. Starting from the 1st base, hill climbing is used to refine the base such that the mutual information between the projections of the original data sample onto this base and the sample's class distribution reaches the highest possible value. Once a local maximum is reached, this base is normalized and fixed. All training samples are made orthogonal to the new base to form the new training samples to be used to refine the next base. The final transform bases all will have been made to have a unit length but are not necessarily orthogonal to each other.

The process first finds a single base, onto which the projection of the original signal will produce a scalar, whose distribution and the object class has maximum mutual information. We then make the signal orthogonal to the first base to form the residue signal. From this residue signal, we attempt to find another transform base that will maximize the mutual information between the projection of this residue signal onto the second base and the class distribution. This process repeats until a fixed number of bases are created.

An intuitive understanding of the method can be thought of as follows: from the original sample, we find a direction that conveys the maximum information about the object class distribution. We then remove what is already known by making the signal orthogonal to this base to form the residue signal. We then find a direction in the residue signal space that conveys the maximum information about the object classes. What is already known about this base is again removed by making the signal orthogonal to this base and the process continues.

Because each base (except the 1st one) is trained on the residue signals from the transform of the previous base, the transform should reflect this and the new maximum mutual information linear dimension reduction transform is illustrated in Fig. 2 as a neural network style diagram.

4 Experiments

In this section, we use human face detection [5] as an application example of the approach developed in section 3. We first collected 9390 face and nonface samples from various sources (the numbers of face/nonface samples are roughly 1 : 2). The original samples are of various sizes and we normalize them to a uniform size of 32 x 32 pixels. We first use these 1024d vectors to train a support vector machine [1], when it is

converged, there are 2217 support vectors, of which 986 are face and 1231 are non face samples. Using these support vector samples, we then follow the procedure in section 3 to develop the transform bases. Fig. 3 shows examples of 16 such bases.

In the following experiments, 64 transform bases are used, that is the input vector of 1024d is reduced to 64d for detection (16 : 1 compression). To search for faces in images, we use detection windows of 30 different sizes ranging from the smallest of 20 x 24 pixels to the largest of 426 x 520 pixels. For each of these windows, it is first resized to 32 x 32, then the 32 x 32 window is passed to the transform to reduce its dimension to 64. The 64 dimensional vector is then passed to a support vector machine, which has been trained to determine whether the current window is a face [1].

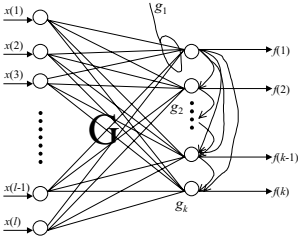


Fig. 2. Schematic of the maximum mutual information dimension reduction linear transform. See eq. (5)**.**

Fig. 3. Examples of maximum mutual information linear dimension reduction transform bases for face/nonface objects

The transform output is defined as (5). In the next section, we use human face detection as an application example to evaluate the validity of the approach.

$$\begin{aligned} f(1) &= g_1 x \\ f(2) &= g_2\left(x - f(1)g_1\right) \\ &\vdots \\ f(i) &= g_i\left(x - f(i-1)g_{i-1} - f(i-2)g_{i-2} - \ldots - f(1)g_1\right) \\ &\vdots \\ f(k) &= g_k\left(x - f(k-1)g_{k-1} - f(k-2)g_{k-2} - \ldots - f(1)g_1\right) \end{aligned} \tag{5}$$

In the first experiment, we tested all 500 upright frontal view face images under the "/inp" directory in the FERET data set [9]. The detector correctly detected 495 faces from the data set achieving a detect rate of 99%. There were 22 false positives with a universal threshold. The testing result is comparable to recent work using this data, e.g. [11]. Fig. 4 shows some examples of detection results.

In a second experiment, we use 130 photographs from the CMU website [6]. This is the set of images used extensively by researchers in face detection. The 130 images contain 507 faces. For the 130 images, our experiment evaluated a total of 52,129,308 patterns. The receiver operating characteristic (ROC) curve of the detection is shown in Fig. 5. Detection examples are shown in Fig. 6. These results are quite good and are comparable to state of the art. This demonstrates that the new method is effective and its potential is very encouraging.

Fig. 4. Examples of experimental results on the FERET database.

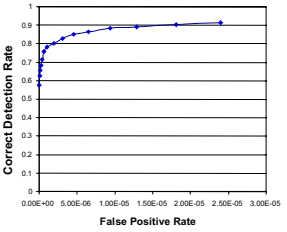


Fig. 5. Receiver operating characteristics of face detection using maximum mutual information dimension reduction linear transform (data representation) and support vector machine (decision mak ing). 130 testing images with 507 faces and 52,129,308 evaluated patterns.

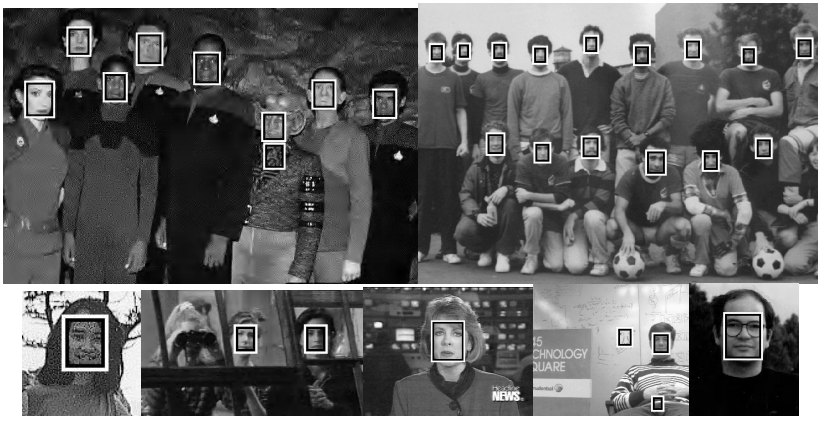


Fig. 6. Examples of face detection result performed on the CMU database

We have also performed some initial comparisons to other transform techniques, in particular principal component analysis (PCA). We found that at lower dimension (high compression), the new informative transform clearly outperforms PCA, the advantage of the new method is less pronounced when very high dimensions are used. Fig. 7 shows 3D plots of 1500 face samples and 1500 nonface samples in the first 3 dimensions of PCA and the new transform space. It is seen that the face/nonface patterns are better separated in the new maximum mutual information transform space. It is also seen that samples belonging to the same class are closer to each other in the new transform space.

5 Concluding Remarks

In this paper, we have presented a learning procedure to create a linear dimension reduction transform based on an information theoretic criterion. We have successfully

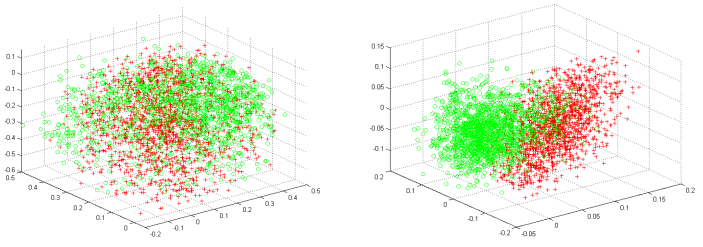


Fig. 7. 3D plots of 1500 face (green o) and 1500 nonface (red +) patterns in the first 3 dimensions of the PCA space (left) and the new maximum mutual information transform space (right).

applied the transform to face detection. Our initial results indicate that the new technique is very effective. Information theoretic approaches have many advantages compared to other conventional methods. Our work here and recent work by others, e.g. [12], have clearly demonstrated the potential of information theoretic approaches to computer vision and pattern recognition problems.

References

1. E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: an application to face detection”. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Pages 130-136, 1997.
2. R. M. Fano, Transmission of Information: A Statistical Theory of Communications, MIT Press, Cambridge, MA, 1961.
3. J. W. Fisher III and J. C. Principe, “A methodology for information theoretic feature extraction”, World Cogress on Computational Intelligence, March 1998
4. T. Butz, J. P. Thiran, "Multi-modal signal processing: an information theoretical framework". Tech. Rep. 02.01, Signal Processing Institute (ITS), Swiss Federal Institute of Tech- nology (EPFL), 2002.
5. M-H Yang, D. Kriegman and N. Ahuja. "Detecting face in images: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, January, 2002.
6. CMU website: http://www.cs.cmu.edu/~har/faces.html
7. S. Haykin, Neural Networks: A Comprehensive Foundation (2nd Edition) , Englewood Cliffs, NJ: Prentice-Hall
8. T. M. Cover and J. A Thomas, Elements of Information Theory, Wiley, 1991
9. P.J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The feret database and evaluation procedure for face-recognition algorithms", Image and Vision Computing, 16(5):295-306, 1998. 27
10. D. H. Ackley. A Connectionist Machine for Genetic Hillclimbing. Boston: Kluwer Academic Publishers, 1987.
11. C. Liu, "A Bayesian discriminating features method for face detection". IEEE Transaction on PAMI, Vol. 25, No. 6, June 2003
12. M. Vidal-Naquet, S. Ullman, "Object recognition with informative features and linear classification", ICCV 2003, Nice, France

Image Object Localization by AdaBoost Classifier

Władysław Skarbek and Krzysztof Kucharski

Warsaw University of Technology
W.Skarbek@ire.pw.edu.pl

Abstract. AdaBoost as a methodology of aggregation of many weak classifiers into one strong classifier is used now in object detection in images. In particular it appears very efficient in face detection and eye localization. In order to improve the speed of the classifier we show a new scheme for the decision cost evaluation. The aggregation scheme reduces the number of weak classifiers and provides better performance in terms of false acceptance and false rejection ratios.

1 Introduction

Face detection using AdaBoost classifier was introduced by Viola and Jones in their seminal paper [4]. Their really novel approach has shown how local contrast features found in specific positions of the face can be combined to create a strong face detector.

AdaBoost is known from the late eighties as a multi-classifier and a training procedure for a collection of weak classifiers, e.g. having the success rate about 0.5, to boost them by suitable voting process to very high level of performance. Viola and Jones applied an early and suboptimal heuristics of Shapire and Freunde for AdaBoost training algorithm [2]. However, their face recognition system described in [1] (April 2003) which is also based on AdaBoost concept, already exploits an optimal training scheme which is methodologically sound. The proposed scheme is a generalization of the scheme described by Schapire in [3].

The AdaBoost approach for face detection has several advantages:

- high speed (real time is reported for 15 fps on Pentium III);
- no special requirements on image quality; (even face images with variance less than four were detected);
- no use of color (what can be of merit in embedded systems);
- very mild restrictions on pose (which are even milder when we consider the proposal of Xiao, Li, and Zhang [5]);
- no limits on number of faces in the image;
- it can be generalized to detection of other facial features (however enough image resolution must be provided for them);
- it can be also generalized to localization of image objects of any other class, e.g. cars, buildings, doors, and windows.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 511–518, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The main disadvantage of the algorithm is the very long training time (it is the matter of weeks on contemporary general purpose machines) which prohibits advanced optimizations of the scheme.

In this paper after reviewing the basic principles of Viola and Jones approach, we present new elements in training of the AdaBoost classifier and show experimental results of face and eye detection algorithm.

2 Survey of AdaBoost Image Object Detector

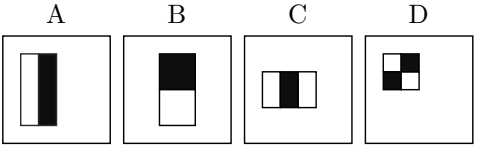
2.1 Region Contrasts

Definition 1. *Region contrast*
Let f_R be a subimage of the image window $f : D \to [0, 255]$ defined on a region R included in the pixel domain D, $R \subset D$. Let R^+ be the positive subregion of R and R^- – its negative subregion. Then the region contrast $c_R(f)$ *is defined as the difference of the luminance sums in the subregions R^+ i R^-:*

$$c_R(f) \triangleq \sum_{(x,y)\in R^+} f(x,y) - \sum_{(x,y)\in R^-} f(x,y)$$

Viola and Jones [4] in their frontal face detection system use the analysis window of size 24×24, scaled with $s_x = s_y = 1.25$, a certain number of times. The four region types designed by them are shown in the figure below. Negative regions are drawn in black while positive ones are white. Each region is identified by four parameters x, y, a, b, where (x, y) is the location of left upper region's corner, a multiple of a is the region's width, and a multiple of b is its height.

region	type A	type B	type C	type D
width	2a	a	3a	2a
height	b	2b	b	2b



However, other class of objects may require other design of contrasting regions. For instance, as it was shown in [5], faces in general pose require more types of image masks.

2.2 The AdaBoost Algorithm

In general the AdaBoost algorithm selects a set of classifiers from a family of weak classifiers $\{C_\omega\}$, indexed by a compound parameter ω. A concerted action of the classifier produces a strong classifier. For each object o, the classifier C_ω elaborates a hypothesis $\delta_\omega(o) \in \{-1, +1\}$ on the membership of the object o to one of two classes labelled by -1 and $+1$. If $\delta_\omega(o) = 1$, then the cost of such decision $\gamma_\omega(o) = \alpha_\omega$, otherwise $\gamma_\omega(o) = \beta_\omega$. The cost of the decision is a real number and it can be negative.

The AdaBoost selects the best classifier ω_{best} which achieves the minimum of the average classification error for the training sequence (o_1, w_1, y_1), ..., (o_L, w_L, y_L) by calling a procedure $getBestWeakClassifier$ (in short $getBWC$):

$$\omega_{best} \leftarrow \arg\min \epsilon(\omega) \triangleq \arg\min \sum_{i=1}^{L} w_i |\delta_\omega(o_i) - y_i| \tag{1}$$

If AdaBoost selects classifiers $C_{\omega_1}, \ldots, C_{\omega_T}$, then the strong classifier $\mathcal{C}_T$ elaborates a hypothesis $\Delta(o)$ for the object o by summing costs of individual decisions and comparing the result with zero:

$$\Delta_T(o) \triangleq \begin{cases} +1, \text{ if } \sum_{t=1}^{T} \gamma_{\omega_t}(o) > 0 \\ -1, \text{ otherwise} \end{cases} \tag{2}$$

In order to get costs of weak decisions AdaBoost calls $getGoodDecisionCosts$ (in short $getGDC$) which returns the cost α for the positive hypothesis (class) and the cost β for the negative hypothesis (class).

AdaBoost algorithm:

Input:
- Training data: $(O_1, y_1), \ldots, (O_L, y_L)$, where $y_i \in \{-1, +1\}$ assigns the example O_i to the class -1 or $+1$
- Number of weak classifiers to be found: T
- Family of weak classifiers: $\{C_\omega\}$
- Procedure $getBWC : [\omega, \epsilon] \leftarrow getBWC(\ldots)$
- Procedure $getGDC : [\alpha, \beta] \leftarrow getGDC(\ldots)$

Output:
- Parameters of selected classifiers: $\omega_1, \ldots, \omega_T$
- Costs of positive hypothesis: $\alpha_1, \ldots, \alpha_T$
- Costs of negative hypothesis: $\beta_1, \ldots, \beta_T$

Method: – Execute steps $1, 2, 2a, 2b, 2c, 3$

1. Initialize weights: for $i = 1, \ldots, L : w_{i,1} \leftarrow 1/L$;
2. For $t = 1, \ldots, T$:

   2a. Select the optimal classifier, its hypotheses and its error:
   $[\omega_t, y'_1, \ldots, y'_L, e_t] \leftarrow getBWC(o_1, \ldots, o_L; w_{1,t}, \ldots, w_{L,t}; y_1, \ldots, y_L)$;

   2b. Select costs of positive and negative hypotheses:
   $[\alpha_t, \beta_t] \leftarrow getGDC(w_{1,t}, \ldots, w_{L,t}; y_1, \ldots, y_L; y'_1, \ldots, y'_L)$;

   2.c Update object weights

   For $i = 1, \ldots, L : w_{i,t+1} \leftarrow w_{i,t} e^{-\gamma_t(o_i) y_i}$;

   Normalize weights: $Z_t \leftarrow 0$;

   For $i = 1, \ldots, L : Z_t \leftarrow Z_t + w_{i,t+1}$;

   For $i = 1, \ldots, L : w_{i,t+1} \leftarrow w_{i,t+1}/Z_t$;
3. Return $[\omega_1, \ldots, \omega_T, \alpha_1, \ldots, \alpha_T, \beta_1, \ldots, \beta_T]$.

◁

3 Bounds on Detection Error Rate

If the object o_i, $i \in I \triangleq [1, L]$, belongs to the class y_i, and the classifier $\mathcal{C}_T$ issues the hypothesis $y_i' = \Delta_T(o_i)$ then the following mutually excluding events can occur:

1. $y_i' = y_i$ – *true hypothesis.* Then the set of indices i such that $\mathcal{C}_T$ for o_i issues the true hypothesis is denoted by I_+.
   a) $y_i' = y_i = 1$ – *true acceptance* (the corresponding index set is I_+^+).
   b) $y_i' = y_i = -1$ – *true rejection* (I_+^-).
2. $y_i' \neq y_i$ – *false hypothesis* (I_-).
   a) $y_i' = 0 \neq y_i = 1$ – *false rejection* (I_-^-)
   b) $y_i' = 1 \neq y_i = -1$ – *false acceptance* (I_-^+).

In order to distinguish the results of the classifier after T iterations we use the notation $I_a^b(T)$, $a, b = +/-$. If a decision refers to a weak classifier selected in the iteration t the relevant set of indices is denoted by $J_a^b(t)$. For instance $J_-^+(t)$ denotes the set of indices for objects falsely accepted by the classifier C_t.

The basic advantage of AdaBoost over other voting schemes in multi-classifier approach is very effective heuristics for costs of negative and positive hypothesis of the selected weak classifier.

The heuristics uses the functions Z_t, $t = 1, \ldots, T$, for bounding of detection error rate. The normalization factor Z_t has been defined in step 2c of AdaBoost 2.2. It can be written compactly:

$$Z_t \triangleq \sum_{i=1}^{L} w_{i,t} e^{-\gamma_t(o_i) y_i}, \tag{3}$$

Let I_-^{-1} and I_{-1} be equivalent notation to I_-^-, I_-. Then:

$$\begin{aligned} I_+^b(T) &= \{i \in I : \Delta_T(o_i) = b = y_i\} \\ I_-^b(T) &= \{i \in I : \Delta_T(o_i) = b \neq y_i\} \\ I_a &= \{i \in I : |\Delta_T(o_i) - y_i| = 1 - a\} \end{aligned} \tag{4}$$

Theorem 1. *Upper bounds for the number of errors*

$$|I_-(T)| \leq B_-(T) \triangleq \sum_{i \in I_-(T)} e^{-y_i \Gamma_T(o_i)} \tag{5}$$

$$|I_-^+(T)| \leq B_-^+(T) \triangleq \sum_{i \in I_-^+(T)} e^{-y_i \Gamma_T(o_i)} = \sum_{i \in I_-^+(T)} e^{\Gamma_T(o_i)} \tag{6}$$

$$|I_-^-(T)| \leq B_-^-(T) \triangleq \sum_{i \in I_-^-(T)} e^{-y_i \Gamma_T(o_i)} = \sum_{i \in I_-^-(T)} e^{-\Gamma_T(o_i)} \tag{7}$$

$$|I_-| \leq B_-(T) = B_-^+(T) + B_-^-(T) < \sum_{i=1}^{L} e^{-y_i \Gamma_T(o_i)} = L \prod_{t=1}^{T} Z_t \tag{8}$$

3.1 Hypothesis Costs by Schapire and Freund

Using as the upper bound for I_-, the expression $L\prod_t Z_t$ from (8) and assuming $\beta_t = -\alpha_t$, Shapire and Freund [2] found the minimum of Z_t in the following way.

$$Z_t(\alpha_t) \triangleq \sum_{i=1}^{L} w_{i,t} e^{-\gamma_t(o_i)y_i} = \sum_{i\in J_+(t)} w_{i,t} e^{-\alpha_t} + \sum_{i\in J_-(t)} w_{i,t} e^{\alpha_t}$$

Hence $Z_t = W_+(t)e^{-\alpha_t} + W_-(t)e^{\alpha_t}$, where
$W_+(t) \triangleq \sum_{i\in J_+(t)} w_{i,t}, \quad W_-(t) \triangleq \sum_{i\in J_-(t)} w_{i,t}$.

The stationary point of the function $Z_t(\alpha_t)$ gives us the optimal solution:

$$\alpha_t = \frac{1}{2}\ln\frac{W_+(t)}{W_-(t)} = \frac{1}{2}\ln\frac{1-W_-(t)}{W_-(t)} = \frac{1}{2}\ln\frac{2-e_t}{e_t} \tag{9}$$

where the last equality follows from $W_-(t) = e_t/2$, and $e_t = \epsilon(\omega_t)$ is the error of the classifier C_t found in the iteration t.

3.2 Our Estimation of Hypothesis Costs

Similarly to Shapire and Freund in this approach as the upper bound for I_-, the expression $L\prod_t Z_t$ from (8) is used but it assumed that $\beta_t = -k_t\alpha_t$, $k_t > 1$. In this case the weights of false rejected examples are powered k_t times more than weights of false accepted examples. Then

$$\begin{aligned} Z_t(\alpha_t) = {} & \textstyle\sum_{i\in J_+^+(t)} w_{i,t} e^{-\alpha_t} + \sum i \in J_-^+(t) w_{i,t} e^{\alpha_t} + \\ & + \textstyle\sum_{i\in J_-^-(t)} w_{i,t} e^{k_t\alpha_t} + \sum i \in J_+^-(t) w_{i,t} e^{-k_t\alpha_t} \end{aligned}$$

Hence $Z_t(\alpha_t) = W_+^+(t)e^{-\alpha_t} + W_-^+(t)e^{\alpha_t} + W_-^-(t)e^{k_t\alpha_t} + W_+^-(t)e^{-k_t\alpha_t}$, where

$$\begin{aligned} W_+^+(t) &\triangleq \textstyle\sum_{i\in J_+^+(t)} w_{i,t}, \quad W_-^+(t) \triangleq \sum_{i\in J_-^+(t)} w_{i,t} \\ W_-^-(t) &\triangleq \textstyle\sum_{i\in J_-^-(t)} w_{i,t}, \quad W_+^-(t) \triangleq \sum_{i\in J_+^-(t)} w_{i,t} \end{aligned} \tag{10}$$

Looking for stationary point of the function $Z_t(\alpha_t)$ leads us to the problem of searching for zero points of the following functions:

$$\begin{aligned} \tfrac{dZ(\alpha)}{d\alpha} &= -W_+^+e^{-\alpha} + W_-^+e^{\alpha} + kW_-^-e^{k\alpha} - kW_+^-e^{-k\alpha} = 0 \\ f(z) &= z^{-k}\left(kW_-^-z^{2k} + W_-^+z^{k+1} - W_+^+z^{k-1} - kW_+^-\right) = 0 \\ g(z) &= kW_-^-z^{2k} + W_-^+z^{k+1} - W_+^+z^{k-1} - kW_+^- = 0 \end{aligned}$$

where $z = e^{\alpha}$.

Property 1. Crossing zero conditions
Let $g(z) = kW_-^-z^{2k} + W_-^+z^{k+1} - W_+^+z^{k-1} - kW_+^- = 0$, $z \geq 0, k > 1$. Then the following facts are true:

- $g(0) \leq 0$.
- If $g(z_0) = 0, z_0 > 0$ then for $z > z_0 : g(z) > 0$
- $g(z) \to \infty$ for $z \to \infty$
- There exists exactly one zero crossing point of $g(z)$ for $z \geq 0$.

The above properties lead to a simple algorithm for finding the unique crossing zero point which uses marching steps concept with step halving and march reversing when $g(z)$ changes its sign.

Property 2. Choice of k

- The unique crossing zero point is greater than one if and only if

$$k(W_-^- - W_+^-) + W_-^+ - W_+^+ < 0$$

- The AdaBoost algorithm can make a progress, i.e. it can find a new weak classifier if and only if

$$k(W_-^- - W_+^-) + W_-^+ - W_+^+ < 0$$

- If $W_+^- > W_-^-$ then AdaBoost algorithm makes the progress if and only if

$$k > \max\left(1, \frac{W_+^+ - W_-^+}{W_+^- - W_-^-}\right)$$

- If $W_+^- < W_-^-$ then AdaBoost algorithm makes the progress if and only if

$$1 < k < \frac{W_-^+ - W_+^+}{W_-^- - W_+^-}$$

3.3 Hypothesis Costs by Viola and Jones

Viola and Jones [1] considered the same upper bound for I_- as Shapire and Freund. However, they assumed that costs α_t and β_t are independent. Then

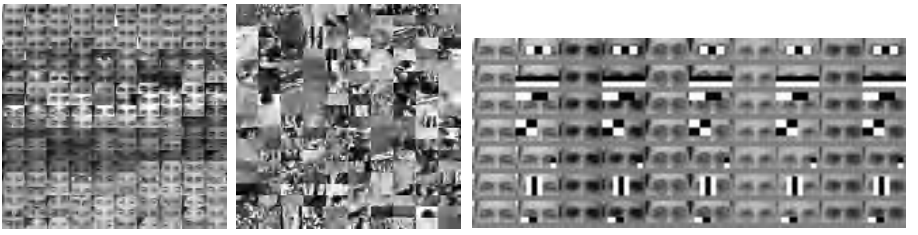
$$\begin{aligned} Z_t(\alpha_t, \beta_t) = \textstyle\sum_{i \in J_+^+(t)} w_{i,t} e^{-\alpha_t} + \sum i \in J_-^+(t) w_{i,t} e^{\alpha_t} + \\ + \textstyle\sum_{i \in J_-^-(t)} w_{i,t} e^{-\beta_t} + \sum i \in J_+^-(t) w_{i,t} e^{\beta_t} \end{aligned}$$

Hence $Z_t = W_+^+(t)e^{-\alpha_t} + W_-^+(t)e^{\alpha_t} + W_-^-(t)e^{-\beta_t} + W_+^-(t)e^{\beta_t}$, where

$$\begin{aligned} W_+^+(t) \triangleq \textstyle\sum_{i \in J_+^+(t)} w_{i,t}, \quad W_-^+(t) \triangleq \textstyle\sum_{i \in J_-^+(t)} w_{i,t} \\ W_-^-(t) \triangleq \textstyle\sum_{i \in J_-^-(t)} w_{i,t}, \quad W_+^-(t) \triangleq \textstyle\sum_{i \in J_+^-(t)} w_{i,t} \end{aligned} \tag{11}$$

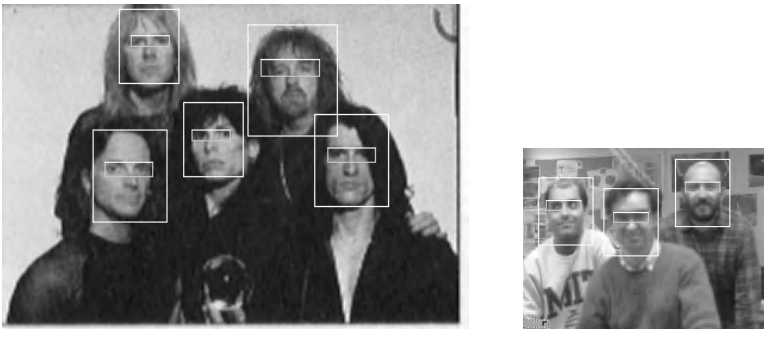
Again the stationary point of $Z_t(\alpha_t, \beta_t)$ gives us the optimal solution:

$$\begin{aligned} \alpha_t &= \tfrac{1}{2} \ln \frac{W_+^+(t)}{W_-^+(t)} \\ \beta_t &= \tfrac{1}{2} \ln \frac{W_-^-(t)}{W_+^-(t)} \end{aligned} \tag{12}$$



(a) (b) (c)

Fig. 1. Training data for eyes detector: (a) positive examples, (b) negative examples. (c) Contrast regions found in first iterations of AdaBoost



(a) (b)

Fig. 2. Detection results: (a) higher resolution image, (b) lower resolution image.

3.4 Properties of Minimal Upper Bounds

Property 3. Minimal value of $Z_t(\alpha)$ and $Z_t(\alpha_t, \beta_t)$
The minimal values for error upper bound are as follows:

$$\min_{\alpha_t}[Z_t(\alpha_t)] = \sqrt{\epsilon_t(2-\epsilon_t)} \tag{13}$$

$$\min_{\alpha_t,\beta_t}[Z_t(\alpha_t,\beta_t)] = 2\left[\sqrt{W_+^+(t)W_-^+(t)} + \sqrt{W_-^-(t)W_+^-(t)}\right] \tag{14}$$

$$\min_{\alpha_t}[Z_t(\alpha_t)] = \min_{\alpha_t}[Z_t(\alpha_t,-\alpha_t)] \geq \min_{\alpha_t,\beta_t}[Z_t(\alpha_t,\beta_t)] \tag{15}$$

From the inequality $\ln(1+x) \leq x$ for $x \in (-1,\infty)$ follows the relation $\sqrt{y(2-y)} \leq e^{-(1-y)^2/2}$ for $y \in [0,2]$ which implies the following important property of AdaBoost algorithm.

Property 4. Freund and Shapire exponential bound
In Shapire and Freund scheme we have the exponential upper bound for $|I_-|$:

$$\begin{aligned} &|I_-(T)| \leq L\prod_{t=1}^T Z_t = L\prod_{t=1}^T \sqrt{\epsilon_t(2-\epsilon_t)} \leq Le^{-\sum_{t=1}^T (1-\epsilon_t)^2/2} \\ &\text{If } \epsilon_t < 1-\tau,\ 0<\tau<1,\ t=1,\ldots,T \text{ then } |I_-(T)| < L\left(e^{-\tau^2/2}\right)^T \end{aligned} \tag{16}$$

Since $\min[Z_t(\alpha_t, \beta_t)] \leq \min[Z_t(\alpha_t)]$ then in Viola and Jones scheme the above exponential upper bound holds, too. This upper bound implies the exponential convergence of the number of false hypotheses in AdaBoost as the function of T.

Property 5. Exponential convergence of error to zero
The upper bound (16) implies the exponential convergence of the number of false hypotheses in AdaBoost algorithm as the function of T.

4 Experimental Results and Conclusions

In figure 3.3 (a) and (b) we illustrate our experimental positive and negative training examples. In part (c) there are shown contrasting regions found within eyes area by the beginning weak classifiers.

We have applied the new scheme for hypothesis costs evaluation in many training and testing experiments for face and eyes detection on real data in both, one stage and many stages (cascade) circumstances. In all experiments consistently our approach gives better performance than Shapiro and Freund approach in terms of detection speed and false acceptance and false rejection rates.

In 30 stage cascade we reach on test data the false acceptance rate below 10^{-5} and false rejection rate below $5 \cdot 10^{-2}$. The detection speed is appropriate for real time video applications using general purpose PCs.

Acknowledgment. The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

References

1. Jones M., Viola P.: *Face Recognition Using Boosted Local Features*, Mitsubishi Electric Research Laboratories, TR20003-25, April 2003.
2. Freund Y, Schapire R.E.: *A decision theoretic generalization of on-line learning and an application to boosting,* Journal of Computer and Systems Sciences, 55(1):119-139, August 1997.
3. Shapire R.E.: *The Boosting Approach to Machine Learning – An Overview*, MSRI Workshop on Nonlinear Estimation and Classification, 2002.
4. Viola P., Jones M.: *Rapid Object Detection using a Boosted Cascade of Simple Features*, Computer Vision and Pattern Recognition, 2001.
5. Xiao R., Li M.-J., Zhang H.-J.: *Robust Multipose Face Detection in Images,* IEEE Trans. on Circuits and Systems for Video Technology, January 2004.

Cost and Information-Driven Algorithm Selection for Vision Systems*

Mauricio Marengoni [1], Allen Hanson[2], Shlomo Zilberstein[2], and Edward Riseman [2]

[1] Marist College, Department of Computer Science
Poughkeepsie, NY 12590
mauricio.marengoni@marist.edu
[2] University of Massachusetts, Computer Science Department
Amherst, MA 01003
{hanson, shlomo, riseman}@cs.umass.edu

Abstract. A computer vision application can be defined as a sequence of image processing, feature extraction, and interpretation operations that are used to solve a specific task. In most of the computer vision systems this sequence is pre-defined and static. The work presented here shows a dynamic technique for algorithm selection based on both the value of information of each operation, and the computational complexity of the operations, which is modeled in terms of cost. This technique is used in the Ascender II system to select visual operators to perform aerial image interpretation of urban regions. The results show that the cost and information driven method presented here leads to performance gains. Moreover the hierarchical structure of the system simplifies the addition of new visual operations.

Topic Area: Image Recognition

Keywords: vision system, algorithm selection, cost of operators, decision making in vision systems

1 Introduction

The problem of algorithm selection has been studied not only in the computer vision community but also in other areas of computer science. Fink [1] uses a technique based on statistical analysis of the past performance of a problem solving method in order to select the method to be used in a new problem. Lobjois and Lemaitre [2] used a method called Selection by Performance Prediction to select a Branch and Bound algorithm among several in order to solve a particular problem.

In computer vision a simple task such as finding a rectangular area in an image can be decomposed into a sequence of operations: filter the image (what filter should be used?), find lines in the image (what "line finder" should be used?), find corners in

* Funded by the National Council for Scientific Research-CNPq, Brazil, grant number 260185/92.2, by DARPA under contract number DACA76-97-k-0005, and by the Army Research Office, under contract numbers DAAG55-97-1-0188, and DAAG55-97-1-0026..

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 519–529, 2004.
© Springer-Verlag Berlin Heidelberg 2004

the image (what "corner finder" should be used?), build hypothesis for rectangles, and test each hypothesis to check the ones that matches the "rectangle model".

Most of the vision systems developed so far use a static sequence of operations to solve a problem [3,4,5,6]. However, no static sequence will be able to deal with illumination changes, weather changes, different resolutions, and other types of commonly changes occurring in a scene or to deal with uncertainties that are common when performing scene analysis.

Some work has been done on designing dynamic methods for algorithm selection. Shaaban and Schalkoff [7] present a dynamic and adaptive system for the image classification problem based on a decision graph. This approach can be learned from human experts and an optimization process is used to select algorithm parameters. Potkonjak and Rabaey [8] use an algorithm throughput optimization and a cost optimization (in terms of running time) and they treat algorithm selection as finding a critical path in a graph. Lagoudakis and Littman [9] model algorithm selection as a Markov Decision process and use Q Learning combining Monte Carlo and Temporal Difference methods to learn how to perform algorithm selection.

The approach used here for algorithm selection is dynamic and takes into account the contribution of the algorithm to the classification process (the value of information) as well as the algorithm "cost" (modeled in terms of computational complexity). The experiment results show that there are performance gains when using this approach to algorithm selection.

The next section in this paper presents the Ascender II system. Section 3 discusses how the value of information is used for algorithm selection and section 4 introduces the idea of cost in terms of computational complexity and shows how cost was empirically modeled. Section 5 shows the results of experiments performed using the Ascender II system, and section 6 presents conclusions and outlines future work.

2 The Ascender II System

The Ascender II system was designed for aerial image interpretation, particularly for the 3D reconstruction of urban areas [10]. The system is divided into two independent parts. The first part is the reasoning subsystem which has a knowledge base composed of a set of Bayesian networks organized hierarchically. The knowledge base was designed to allow variations in the imagery, such as a varying number of views, different image resolutions and viewpoints, and different number of objects in each class. This subsystem is responsible for selecting visual operators and for making decisions throughout the inference process. The second part is the visual subsystem that has a database of images, models, visual operators and other image features extracted during the inference process [10].

The Ascender II has been designed as a general purpose vision system. It has a set of focus-of-attention regions as input, and the system's goal is to select visual operators, recognize objects in the scene, and reconstruct these objects automatically in 3D. The Bayesian networks in the Ascender II system are used for the following tasks: select a feature to be measured in the image, combine evidence from different fea-

tures once they have been collected, and decide when to accept the current hypotheses and stop the reasoning process [11].

The Bayesian networks were designed using the HUGIN system [12]. Each feature is quantized into ranges such that a feature value can be used to discriminate between the classes in the root node; these ranges were defined empirically, and they are part of the knowledge engineering process. The general structure of the networks is shown in Figure 1.

3 Information Driven Algorithm Selection

Utility theory is a probabilistic technique for decision making and it fits well in a Bayesian network system. Utility theory provides mechanisms for selecting the decision that has the highest expected utility [13]. In the discussion that follows, we use the following notation:

- $R_j \overset{def}{=}$ the event that region *R* belongs to Class *j*.
- $DR_j \overset{def}{=}$ the decision that region *R* is identified as Class *j* (the decision based on the outcome).
- $E \overset{def}{=}$ all the evidence collected so far.
- $F_m \overset{def}{=}$ feature *F* is discretized into *m* states.

The expected utility *(EU)* of each decision is computed using the probability that a region belongs to a class *j* given all knowledge available, $P(R_j|E)$, and the utility of deciding that a region is in class *i* given that the region belongs to class *j*, $U(DR_i|R_j)$:

$$EU(DR_i \mid E) = \sum_{j=1}^{N} U(DR_i \mid R_j) * \Pr(R_j \mid E) \ . \qquad (1)$$

The system's utility is defined as the maximum value among the expected utilities of each decision. In our problem domain the system has to decide the most likely identity (e.g. label) of a region. Features are selected based on their value of information, which gives the expected gain, in terms of utility, that is achieved by knowing the feature value. This value is computed as follows: for each feature currently available compute the expected utility of the system given that information about the feature is known, *EU(Fm)*. The feature is divided into M discrete intervals and each of these intervals might lead to a different object class with a corresponding utility.

The utilities found for each interval have to be weighted by the probability that the feature value is in the corresponding interval:

$$EU(DR_i \mid E, F_m) = \sum_{m=1}^{M} P(F_m) * \max(EU(DR_i \mid E, F_m) \ . \ (2)$$

Now, compute the value of information of each feature by subtracting the current utility of the system from the expected new utility of the system when information about the feature is known:

$$VI(F_m) = EU(DR_{\alpha'} \mid E, F_m) - EU(DR_{\alpha} \mid E) . \ (3)$$

and select the feature with the highest value of information. Intuitively, the value of information measures the expected improvement in the utility once the result of an operator becomes available. Notice that the best decision can change once the most valuable feature is physically measured and its evidence is entered into the system. This reasoning process is presented in details in [14].

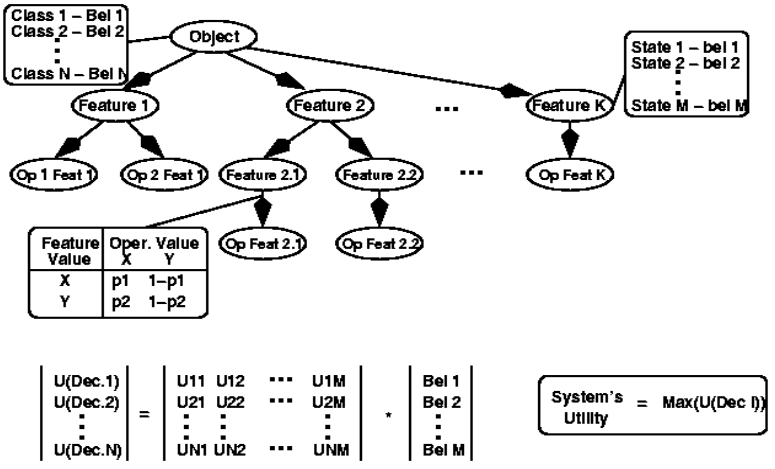


Fig. 1. The general structure of a Bayesian networks used in the Ascender II system [14]. The root node gives the possible classes used to classify a region. The feature nodes have the features that can be used to discriminate among the classes in the root node. The operator's nodes have the operator's available to measure the feature attached to it.

4 Modeling Computational Complexity as Cost

In the previous section the cost of applying an operator was ignored (all operators were assumed to have zero cost). This is not generally true and a vision system should be able to incorporate varying costs into its algorithm selection process. These costs may be based on processing time, algorithm complexity, resources required, etc.

A cost based on processing time requires a mapping function that outputs the running time of an operator given the number of pixels (or other primitive feature) to be processed. In this scenario, ideally, all algorithms should have their implementation optimized. A badly implemented operator, in this case, could have a high cost and therefore never be selected.

A cost based on resources required, such as memory, processor speed, or number of processors, would not apply in our case. It is true that each operator requires some memory to compute a feature over a region, but it is bounded by the image size, which is kept in memory all the time. The operators run on the same machine, so they all use the same processor speed. None of them is a parallel operator and they all run on a single processor machine. Thus, this type of cost function is not appropriate for most centralized vision systems.

The cost function presented here is based on the algorithmic complexity of each operator. The operators were divided into three complexity classes and an associated cost was attributed to each of these classes. A short description of some operators is presented in Table 1. The first class contains all operators that compute their output based only on the information from a small set of pixels. These operators run in constant time and they belong to the complexity class O(1).

The second class has the operators that need to compute features based on the lines found in the image. The running time of these operators is linear with the number of lines in the image and they form the complexity class O(L). The third class has the operators that need information from all pixels inside the region being processed to compute a feature. The running time of these operators is linear with the number of pixels being processed. These operators form the O(N) complexity class.

Table 1. List of some visual operators, their corresponding complexity classes, and a short description of them

Operator	Class	Description
2	O(L)	Finds # of T junctions in a region.
6	O(N)	Computes planar fit in a region.
8	O(L)	Computes line density in a region.
10	O(N)	Computes the average height.
15	O(L)	Finds a center line in a region.
19	O(1)	Computes height difference.
25	O(1)	Computes the area of a region.

The cost associated with the complexity class of an operator has to be expressed in terms of utility so it can be combined with the "information-based" utility presented in the previous section.

Without cost the value of information considers only the increment in the system's utility given by the knowledge of a certain feature. In this case the value of information can be defined as the expected gain, in terms of utility, that we can expect if we know the value of the feature:

$$Gain(F_m) = EU(DR_{\alpha'} \mid E, F_m) - EU(DR_{\alpha} \mid E). \quad (4)$$

The value of information when using cost is computed as:

$$VI(F_m) = Gain(F_m) - Cost(F_m) \quad . \quad (5)$$

Examining equation 5 we can see that the cost for an operator should not be too high or else the operator could have negative value of information and would never be called, making it useless and forcing the system to use only cheap and perhaps unreliable operators, leading to poor classifications. On the other hand, the cost of an operator should not be too low, or else the system would behave as if all operators were zero cost operators, leading to a system that could select very expensive operators, even when they would not be required.

The operator's cost was defined in terms of utility values based on the operator's complexity class. We define the maximum cost (MC) of an operator as a linear function of the mean of the expected gain given by an operator over the sequence of op-

erators (called average gain or AG), that is, *MC=F*AG*, where F is called the cost factor and it is a weight value to balance the average gain.

The cost function for each complexity class was defined as a linear function of the maximum cost, as presented in Table 2. The average gain was computed over all regions in the data set for all cases where evidence over the sequence of operators were converging, that is, $\alpha' = \alpha$ in equation 4. The value of the average gain obtained is 1.083 out of 10, which was approximated to 1.1. Further experiments were performed to tune the values of F, A, B, and C.

A set of experiments were performed using the Avenches dataset (Figure 2) [15] to tune the cost functions in the Ascender II system and to study the system behavior when using cost. In the first experiment the goal was to determine a value for *F* in the expression *MC=F*AG*. A first guess was made for the constants A, B and C in the cost functions presented in Table 2.

Table 2. The generic function of an operator's cost based on the operator's complexity class

Complexity	O(1)	O(L)	O(N)
Cost	A*MC	B*MC	C*MC

Fig. 2. The Avenches dataset and the regions of discourse used in the experiments. The dataset has 4 views with 0.1 m for ground sample distance (gsd), and a DEM with 0.25 m for gsd.

The value of *AG* was set to 1.1 and the value of *F* was changed from 0 to 1; for each value of *F* the operators selected were applied in the Avenches data set and the regions were classified properly. The resulting curve is shown in Figure 3 left. The number of operators applied in the image was reduced when the cost factor (*F*) increased, as expected. The average of the final beliefs in each level did not change significantly (less than 3%), and the final classification of the regions for the Avenches data set was the same as when *AG = 0*. The use of cost proved to be very efficient: the system maintained the same accuracy and used fewer resources to perform the same task. Figure 3 (left) shows a clear turning point when *F=0.5*; at this point the number of operators used still decreases but the rate of decrease is smaller

than before (less than 5%). The value of *F* could be selected as any value between 0.5 and 1.0 without any significant change. We set the value of *F* as 0.5.

For any region we know that *O(1)<O(N)*. We can also say that for all lines inside a region *O(1)<O(L)<O(N)*. The problem, in our case, is that the *O(L)* operators consider all lines in the image and, as a first step, clips the lines in the image against the region boundaries. Because of this step it is hard to compare an *O(L)* operator with an *O(N)* operator. If the region is small and the image has a large number of lines we can find that *O(L)>O(N)*.

Consequently, the second set of experiments was performed to empirically estimate the value of B. Its value was changed from 0.02 (it has to be greater than 0) to 1.5 (50% more expensive than the full cost). The result is shown in Figure 3 (right). The average final beliefs in each level in the hierarchy have not changed significantly (less than 1%) and the final classification of the regions was the same. When the cost factor of an *O(L)* operator is between 0.2 and 1, the variation in the total number of operators used by the system is small (67 to 66). Considering that the overall variation in the total number of operators used is not too large (around 15%) and considering the time required to the *O(L)* operators was, on average, smaller than the time required for the *O(N)* operators in the Avenches data set we decided, empirically, to set the value of the constant *B* for the *O(L)* operators to 0.4.

Based on these two experiments and that an *O(1)* operator should cost much less than an *O(N)* operator we empirically fixed the values of the constants *A* and *C* in the cost function for the complexity classes *O(1)* and *O(N)* as 0.02 and 1 respectively.

In order to define the cost functions for each complexity class in the Ascender II system only the average gain (*AG*) remains to be investigated. In this third experiment we used *F=0.5* and the cost functions as defined in Table 3; the value of *AG* was changed from 1.1 to 10 and the results are presented in Figures 4 (left and right).

Figure 4 (right) shows that as the average gain increases, the number of operators used decreases, but when the average gain increases above 5 the number of regions correctly classified decreases (Figure 4 left). In particular, there was a dramatic decrease in the beliefs at level 0 when *AG* increased beyond 5.

From the experiments just described the maximum cost of an operator was defined as *MC=0.5*AG*. The cost function for each complexity class was kept as presented in Table 3. From the experiments performed in the Avenches data set the value of the average gain can be set anywhere between 1.1 and 5, because the average final beliefs in each level are about the same and the classification of the regions was the same. However, higher values of *AG* in this range dramatically reduced the number of operators used.

Table 3. This table shows the function of an operator's cost based on the operator's complexity class

Complexity	O(1)	O(L)	O(N)
Cost	0.02*MC	0.4*MC	1.0*MC

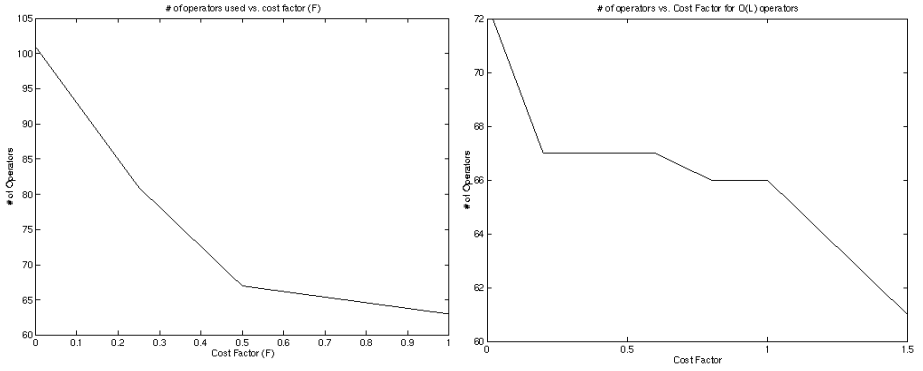


Fig. 3. Left: the graph shows the number of operators used by the system in the Avenches data set when the value of F goes from 0 to 1. For the Avenches data set the number of regions correctly classified was the same. Right: the graph shows the number of operators used by the system in the Avenches data set when the constant B in the cost function for the O(L) operators goes from 0.02 to 1.5. The overall classification did not change.

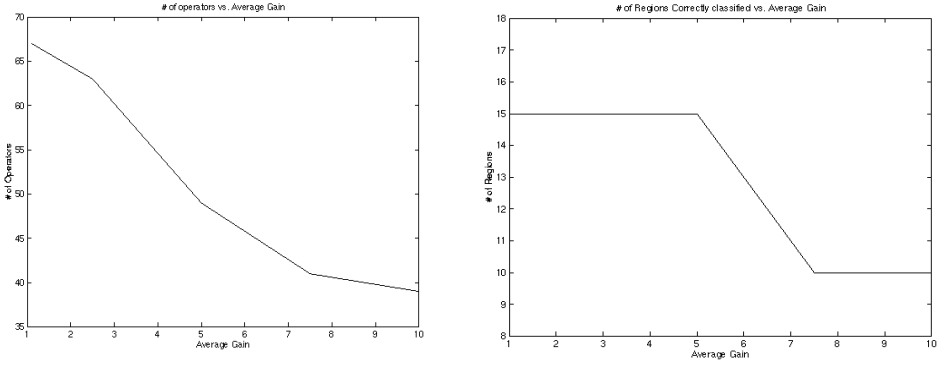


Fig. 4. Left: the graph shows the number of operators used by the system in the Avenches data set when the average gain goes from 1.1 to 10. Right: the graph shows the number of regions correctly classified in the Avenches data set when the average gain goes from 1.1 to 10.

5 Experiments and Results

Experiments were performed using the Ascender II system on other two datasets: Fort Hood (7 views, varying image resolutions, and a DEM with 0.35 m of gsd – Figure 5 left) and Fort Benning (2 views, 0.14 m of gsd, and a DEM with the same resolution – Figure 5 right). Although we could have used a value of 5 for *AG*, this value proved to be too high for the Fort Benning data set, decreasing the number of regions correctly classified considerably (results not presented here).

Three sets of experiments were performed. In the first experiment the average gain (AG) value was set to 0, thus checking the efficiency of the information-driven algo-

rithm selection technique without cost. In the second experiment the average gain value was set to 1.1, and in the third experiment an average gain of 2.5 was used.

Overall, 79 regions were considered in the three data sets (Fort Hood, Avenches, and Fort Benning). A total of 906 visual operators could be applied to all 79 regions in order to classify them. The first column presented in Table 4 shows the performance of the Ascender II system with no cost (number of operators used in each dataset and the number of regions correctly classified in parentheses). The second column in Table 4 shows the results in terms of number of operators and regions correctly classified when the average gain was set to 1.1.

The third column shows the results for an average gain of 2.5. The information-driven algorithm selection method reduced the number of algorithms applied in the data sets in about 57%. The percentage of regions correctly classified using the information-driven method is about 84%.

The introduction of cost into the Ascender II system has only a small change in the system's performance with respect to the number of regions correctly identified. The number of regions correctly classified was reduced by 1.3% for an AG = 1.1 and by 2.6% for AG=2.5. The results also show a significant decrease in the number of operators used by the system (27% for AG=1.1 and 30% for AG=2.5). The changes in the average final beliefs, which gives the reliability (or quality) of the final classification in each level, were too small to be significant (less than 3%).

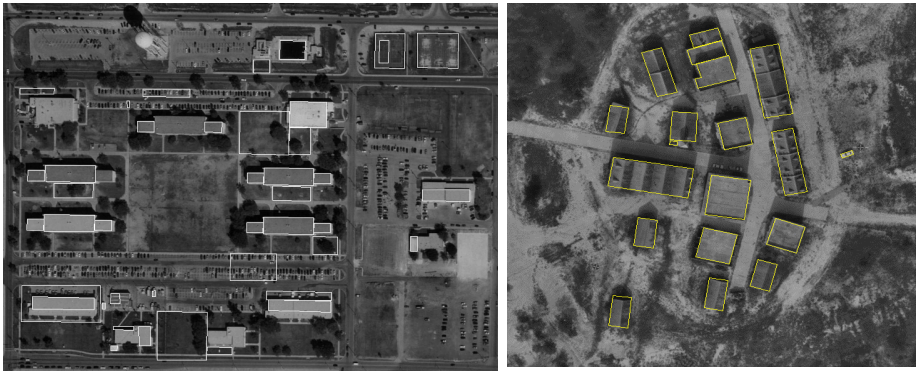


Fig. 5. Left: a view of the Fort Hood dataset. Right: a view of the Fort Benning dataset.

Table 4. Number of operators used set and the number of regions correctly classified (in parentheses) in each data when the average gain changes

Data Set	AG = 0	AG=1.1	AG=2.5
Fort Hood	178 (36)	119 (36)	114 (35)
Avenches	101 (15)	67 (15)	63 (15)
Fort Benning	113 (15)	100 (14)	97 (13)
Total	392 (66)	286 (65)	274 (63)

6 Conclusions and Future Work

In this paper it was shown how utility theory can be used in an information-driven method for algorithm selection. The information-driven technique presented here proved to be very efficient in terms of resources used to correctly classify regions in aerial images.

It was also shown how computational complexity can be modeled in terms of utility and how it can be used as cost for an algorithm. This model of cost was included in the information-driven method and the results in the experiments performed here proved to be even more efficient in terms of resources, with a small reduction in the number of regions correctly classified.

The algorithm selection technique is dynamic and the selection is done "on the fly" depending on the system's current state of knowledge. The technique also allows variations in the imagery, such as number of views, and image resolution.

The current system has only one visual operator associated with features. It is possible to add *more operators per feature*, and to decide which operator to use once a feature is selected. The idea would be to extend the feature selection process currently used to select the operator that will measure the feature.

References

1. E. Fink, How to solve it automatically, selection among problem-solving methods, *Proc. of the 4th Int. Conf. on Artificial Intelligence Planning Systems*, pp 128-136, AAAI Press 1998.
2. L. Lobjois and M Lemaitre, Branch and Bound algorithm selection by performance prediction, *Proc. of the 15th National Conference on Artificial Intelligence*, pp 353-358, CA 1998.
3. R.A. Brooks, Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence* 17 (1981), 285-348.
4. R. Rimey, and C. Brown, Task-oriented vision with multiple bayes nets. In *Active Vision*, Blake A. and Yuille A., Eds. The MIT Press, 1992.
5. W.B. Mann, and T.O. Binford, An example of 3D interpretation of images using bayesian networks. *DARPA Image Understanding Workshop* (1992), 793-801.
6. V.P. Kumar, and U.B. Desai, Image interpretation using bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(1) (1996), 74-77.
7. K.M. Shaaban and R.J. Schalkoff, Image processing and computer vision algorithm selection and refinement using an operator assited meta algorithm, *SPIE Proc. Vol 2488 Visual Information Processing IV*, Orlando, FL 1995 pp. 77-87
8. M. Potkonjak and J. Rabaey, Algorithm selection: a quantitative computation intensive optimization approach, *IEEE/ACM Int. Conf. on Computer Aided Design*, CA, 1994.
9. M.G. Lagoudakis and M.L. Littman, Algorithm selection using reinforcement learning, *Proc. of the 17th National Conf. on Artificial Intelligence AAAI 00*, TX 2000, pp. 108.
10. M. Marengoni, C. Jaynes, A. Hanson, and E. Riseman, Ascender II, a visual framework for 3D reconstruction. In *Proc. of the Int. Conf. on Vision Systems*, Las Palmas, Spain (1999).

11. M. Marengoni, A. Hanson, S. Zilberstein, and E. Riseman, Control in a 3D reconstruction system using selective perception. In *Proceedings of the International Conference in Computer Vision*, Greece, 1999.
12. S.K. Andersen, K.G. Olesen, F.V. Jensen, and F. Jensen, HUGIN – a shell for building bayesian belief universes for expert systems. In *Proceedings of the 11th International Congress on Uncertain Artificial Intelligence* (1989), pp 1080-1085.
13. D.V. Lindley, *Making Decisions: Second Edition*. John Wiley and Sons, 1985.
14. M. Marengoni, A. Hanson, S. Zilberstein, and E. Riseman, Decision making and Uncertainty Management in a 3D reconstruction system, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2003, *(Vol. 25 No. 7) pages 852-858*.
15. International Society for Photogrammetry and Remote Sensing (ISPRS): ftp.vislist.com/IMAGERY/PTERRAIN.

Gesture Recognition for Human-Robot Interaction Through a Knowledge Based Software Platform

M. Hasanuzzaman[1], Tao Zhang[1], V. Ampornaramveth[1], M.A. Bhuiyan[2], Yoshiaki Shirai[3], and H. Ueno[1]

[1] Intelligent System Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.
hzamancsdu@yahoo.com
[2] Jahangirnagor University, Dhaka-1342, Bangladesh.
[3] Department of Computer Controlled Mechanical Systems, Osaka University, Suita, 565-0871 Japan.

Abstract. The task of real-time hand gesture recognition is extremely challenging due to a number of DOFs of hand pose and motion. However, for human-robot interaction in natural ways, gesture can provide a powerful interface tool for commanding a robot to perform a specific task. This paper presents a vision-based real-time gesture recognition system by segmenting the three largest skin color components and template-matching techniques with multiple features. Gesture commands are generated whenever the combinations of three skin-like regions at a particular image match with the predefined gestures. These gesture commands are sent to robots through a knowledge based software platform for human-robot interaction. The effectiveness of our method has been demonstrated by interacting with an entertainment robot named AIBO.

1 Introduction

Research on humanoid robots has been increasing rapidly for the last decade. Recent advances in computer science and robotics make robots more applicable to our daily life. Robots and human will co-exit with sharing and co-operating tasks according to Symbiotic Information System (SIS) concept [1]. However, most robots so far still lack of the ability to interact with user in natural ways and work independently. Recent technologies for multi-modal communication can provide various communication channels like voice and gestures. Using gesture to interact with a robot provides a natural and powerful interface that can be used to command a robot to perform tasks or to control its operations. Considering the facts it is necessary to develop efficient and real time gesture recognition system to perform more human like interaction between human and robot. Two approaches are commonly used to interpret gestures for human machine interaction. One is gloved based approach [2] that requires wearing of cumbersome contact devices and generally carrying a load of cables that connect the device to a computer. Another approach is vision based technique that does not require wearing any of contact devices with human body part, but uses a set of video cameras and computer vision techniques to interpret gestures. Gesture recogni-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 530–537, 2004.
© Springer-Verlag Berlin Heidelberg 2004

tion based on vision technology has been emerging with the rapid development of computer hardware of vision system in recent years and in future it will dominate in both Human-Computer and Human-Robot interactions. For gesture interpretation system gestures modeling is the first step that mainly depends on the intended application of those gestures. Gesture modeling can follow appearance based or model based approach. Model based approach is very hard to implement in real time because they usually use very complicated algorithms to extract accurate joint angles. An appearance-based algorithm is a strong tool for object recognition. Here, a variety of object appearances are stored as a statistical model and used in the recognition task. The gestures are modeled by relating the appearance of any gesture to the appearance of the set of predefined template gestures.

Watanabe et. al. [3] used maskable template based on minimum distance between template and partial block of an input image for gesture recognition. Hongo et. al. [4] has developed a system that can track multiple faces and hands by using multiple cameras to focus on face and gesture recognition. Utsumi et. al. [5] detected hand using hand shape model and tracked using extracted color and motion. They also propose multiple cameras for data acquisition to reduce the occlusion problem. But in this process there incurs complexity in computations. However all this paper [3-17] did not consider face and two hands gestures at the same time that we have considered here.

In this paper we present a simple and faster method for recognizing gestures with skin-color segmentation and multiple features-based template matching techniques. In this method three larger skin like regions are segmented from the input images by skin color segmentation technique considering that face and two hands may be present in the input images at the same time. Segmented blocks are filtered and normalized and then compared with template images for finding best match. For template matching we have used the combination of two features: correlation coefficient and minimum (Manhattan distance) distance qualifier. If the combination of three skin-like regions at a particular frame matches with predefined gesture then corresponding gesture command is generated. In this experiment we have recognized eight gestures as listed in Table 1. Gesture commands are being sent to robots through a knowledge-based software platform and their actions are being accomplished according to users predefined action for that gesture. A method has also been developed to detect left hand and right hand relative to face position, as well as, to detect the face and locate its position. We have prepared template images with different illuminations to adapt our system with illumination variation. In this method we have also considered slightly rotated face for both left and right side to make face detection algorithm a more pose invariant. As an application of our method, we have implemented a real-time human robot interaction system using a robot named AIBO.

The remainder of this paper is organized as follows. In section 2, we briefly describe skin-like regions segmentation, filtering as well as normalization. Section 3 focuses on multiple features based template-matching techniques for face and hand poses detection and gestures recognition method. Section 4 presents our experimental results and discussion. Section 5 concludes this paper.

2 Skin Color Segmentation, Filtering, and Normalization

This section introduces a color segmentation based approach for determining skin parts (mainly face and hands) from color images. This system uses video camera for data acquisition and from color input image isolate three larger skin-like regions by using skin-color information. HSV or YIQ color model is used for skin color segmentation, since color footprint is more distinguishable and less sensitive to illumination changes in the hue saturation space than the standard RGB color space. We use YIQ (Y is luminance of the color, I, Q are chrominance of the color) color representation system since it is typically used in video coding and provides an effective use of chrominance information for modeling the human skin color.

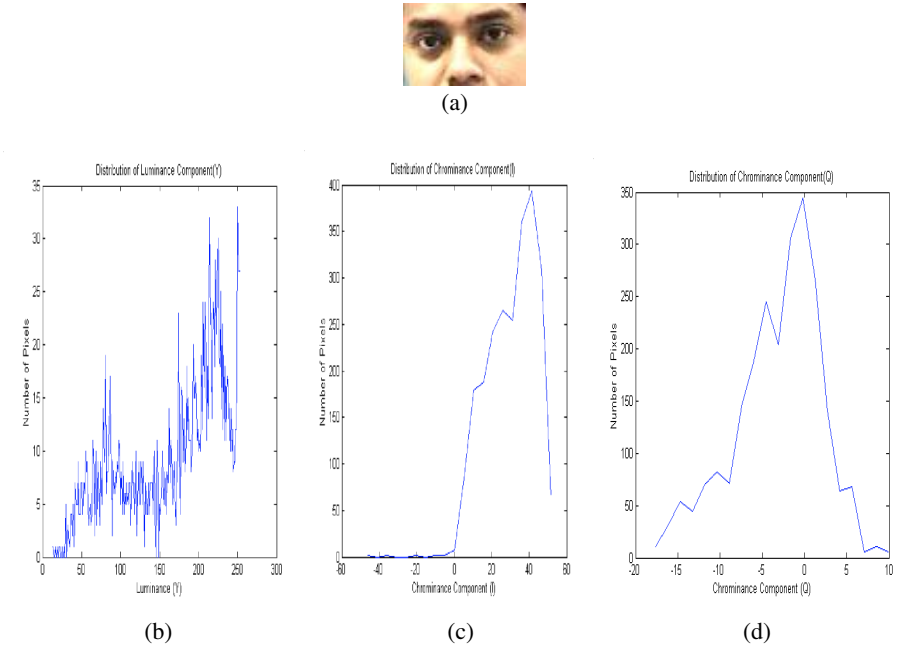


Fig. 1. a) Sample skin region, b), c) and d). Y, I, Q distributions respectively

The RGB image taken by video camera are converted to YIQ color space and threshold it by the skin color range [6]. Fig. 1., shows sample skin region and corresponding Y, I, Q distributions for every pixels. Chrominance component I, play an important role to distinguish skin like regions from non-skin regions, because it is always positive for skin regions and negative for non-skin region. Locations of the probable hands and face are determined from the image with three larger connected regions of skin-colored pixels. In this experiment, 8-pixels neighborhood connectivity is employed. In order to remove the false regions from the isolated blocks, smaller connected regions are assigned by the values of black-color. Noise and holes are filtered by morphological dilation and erosion operations and normalization is done to con-

vert the segmented images to gray images and resized the image as template images size. Sample output of the skin-color segmentation is shown in Fig. 2.

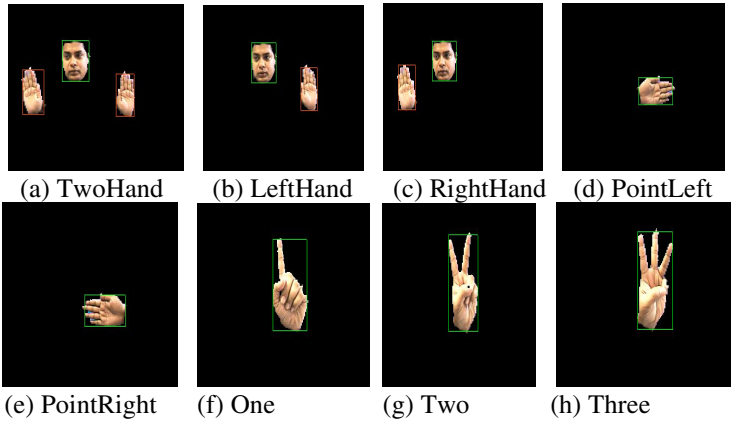


Fig. 2. Sample output of Skin color Segmentation and filtering

3 Face and Hand Poses Detection for Gesture Recognition

3.1 Hand Poses and Face Detection Using Multiple Features Based Template Matching

As template-matching approach is a more natural approach of pattern recognition, we use it to recognize gesture from an unknown input image. First we have prepared noise free version of faces and hands as templates as shown in Fig. 3. To support small rotation we included some slightly rotated images within our template images. To adapt illumination variation we have formed templates in different illumination. For template matching we have considered two features: one is maximum correlation coefficient and another is minimum distance classifier (Manhattan distance) between two same size images. Correlation coefficient is calculated using following equation,

$$\alpha_t = M_t / P_t \quad (0<\alpha_t \quad 1) \tag{1}$$

where M_t is total number of match pixels (white pixels with white pixels and black pixels with black pixels) with t-th template, P_t is number of total pixels in t-th template and t is a positive number. For exact matching α_t is 1, but for practical environment we have chosen threshold value for α_t through experiment for optimal matching. Minimum distance can be calculated by using following equation,

$$\delta_t = \sum_{1}^{x \times y} | I - G | \tag{2}$$

where, I(x, y) is the input image and $G_1(x,y)$, $G_2(x,y)$,----------,$G_t(x,y)$ are template images. There is more than one way to define $_t$ corresponding to different ways of measuring distance. Two of the most common are: Euclidean metric and Manhattan metric. In our experiment we have used Manhattan metric. For exact matching $_t$ is 0 (zero) but for practical purpose we have used a threshold value through experiment for finding optimal matching.

Fig. 3. Example template images

We have combined output of these two matching methods to make our system more accurate to recognize gestures. In our method we have grouped template images into eight classes (each for different hand poses), for example face class (C_1) and left hand class (C_2) and right hand class (C_3), etc. If max $\{\alpha_t\}>th_1$ is true then corresponding class (C_α) is identified and if min $\{ _t\}<th_2$ is true then corresponding class (C) is identified, where th_1 and th_2 are thresholds for correlation coefficient and minimum distance qualifier respectively. If both methods identified the same class then corresponding class is detected, otherwise ignored.

3.2 Gesture Recognition

Similar calculation is done for all three segments simultaneously for a particular image frame and with the combinations of three segments at a particular time form a gesture. Gestures are recognized using rule-based system according to Table 1. For examples: if two hands and one face are present in the input image then recognized it as "TwoHand" gesture. If one face and one hand are present in the input image frame then recognize it as either left hand or right hand respective to face position using following equation,

$$\varepsilon = f_x - h_x \tag{3}$$

where, f_x is the x-coordinate of the centre position of face segment and h_x is the x-coordinate of the centre position of hand segment (when one hand is present). If the distance is negative then it is detected as "RightHand" gesture and if it is positive then it is detected as "LeftHand" gesture. If pointing left hand present in the input images then it is recognized as "PointLeft" gesture and similarly others gesture are recognized. According to gestures recognition corresponding gesture command is generated and sent to interact with robot through a knowledge base software platform (SPAK) [7]. SPAK (Software Platform for Agent and Knowledge Management) consists of a frame based knowledge management system and a set of extensible autonomous software agents representing object inside the environment and support human

robot interaction and collaborative operation with distributed working environment. Our approach has been implemented on a robot named AIBO.

Table 1. Input segment combinations and corresponding gestures.

Three input segment combinations			Gesture
Face	Left hand palm	Right hand palm	Twohand
Face	Right hand palm	X	Righthand
Face	Left hand palm	X	LeftHand
Face/(no face)	Point left hand	X	PointLeft
Face/(no face)	Point right hand	X	PointRight
Face/(no face)	Index finger raise	X	One
Face/(no face)	Form V sign with index and middle fingers	X	Two
Face/(no face)	Index, middle and ring fingers raise	X	Three

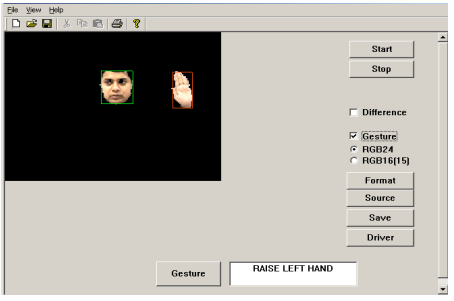
[X=Absence of predefined hand or face poses]

4 Experimental Results and Discussion

This section describes experimental procedures, as well as experimental results of the gestures recognition system and human-robot interaction system by gesture. This system uses a standard video camera for image data acquisition. Each capture image is digitized into a matrix of 320×240 pixels with 24-bit color. First we prepare pure templates for face and different hand poses. All the templates are 60×60 pixels gray images. Template images are consisted a total of 480 frontal images of faces, different hands poses of different persons related to gestures. Fig. 3 shows the example template images. We have tested our system for real time input images. The sample output of our gesture recognition system is shown in Fig. 4(a). This shows gesture command at the bottom text box corresponding to matched gesture, in case of no match it shows “no matching found”.

In this part we have explained a real time human-robot interaction system using recognization gestures. We have implemented this application by off-board configuration that’s means gesture recognition program run in client PC, not in robot. We have considered robot as a server and our PC as a client. Communication link has been established through SPAK. Initially, we connected the client PC with SPAK server and then gestures recognition program was run in the client PC. As a result of gestures recognition program client PC has sent gesture commands to SPAK. In SPAK we have designed frames corresponding to each gesture. After getting gesture command related frame is activated and robot acted according to users predefined actions. Fig. 4(b) shows our experiment setup for human-robot interaction using gestures. The actions of AIBO are: STAND, WALK FORWARD, WALK BACKWARD, TURN LEFT, TURN RIGHT, KICK (by right leg), SIT and LIE in accordance with gesture

"One", "Two", "Three", "PointLeft", "PointRight" "RightHand" "TwoHand", and "LeftHand" respectively. We have considered for human-robot interaction that gesture command will be effective until robot finished corresponding action for that gesture.



(a) Output for "LeftHand" gesture.

(b) AIBO STAND-UP for "One" gesture.

Fig. 4. Sample output of gesture based human-robot interaction system.

5 Conclusions

This paper describes a real-time hand gesture recognition system using skin color segmentation and multiple features based template-matching techniques. For the matching algorithm we have used combinations of minimum distance qualifier (Manhattan distance) and correlation coefficient based matching approaches, that's why it is more robust than any single feature based template-matching techniques. We have compared only precisely segmented images with templates, and used a group of template for a single pose, that's why it is more reliable and it has less computational cost. One of the major constrain of this system is that the background should be non-skin color substrate. If we used infrared camera then it is possible to overcome this problem just by a minor modification of our segmentation technique, other modules will remain the same.

We have also successfully implemented gestures based human-robot interactive system using a robot named AIBO. A particular user may assign distinct command to specific hand gesture and thus control various intelligent robots using hand gestures. The significant issues in gesture recognition for our method are the simplification of the algorithm and reduction of processing time in issuing commands for the robot. Our next approach is to make the detection system more robust and to recognize more facial and hand gestures for interaction with different robots such as AIBO, ROBOVIE, SCOUT, MELFA, etc. Our ultimate goal is to establish a symbiotic society for all of the distributed autonomous intelligent components so that, they share their resources and work cooperatively with human beings.

References

1. Vuthichai Ampornaramveth, Haruki Ueno, "Concepts of Symbiotic Information System and Its Application to Robotics", proceedings of 11th European-Japanese Conference on Information Modeling and Knowledge Bases, 2001, pp. 394-407.
2. Vladimir I. Pavlovic, Rajeev Sharma and Thomas S. Huang, " Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review" IEEE PAMI, 1997, Vol. 19, No. 7, pp. 677-695.
3. Takahiro Watanabe, Chil-Woo Lee, Akitoshi Tsukamoto, and Masahiko Yachida, "Real-Time Gesture Recognition Using Maskable Template Model", Proc. of the International Conference on Multimedia Computing and Systems (ICMCS'96), 1996, pp. 341-348.
4. Hitoshi Hongo, Mitsunori Ohya, Mamoru Yasumoto, Yoshinori Niwa, and Kazuhiko Yamamoto, "Focus of Attention for Face and Hand Gesture Recognition Using Multiple Cameras", AFGR00, IEEE, pp. 156-161.
5. Akira Utsumi, Nobuji Tetsutani and Seiji Igi, "Hand Detection and Tracking using Pixel Value Distribution Model for Multiple-Camera-Based Gesture Interactions", Proc. of the IEEE workshop on knowledge Media Networking (KMN'02), 2002, pp. 31-36.
6. Md. Al-Amin BHUIYAN, Vuthichai AMPORNARAMVETH, Shin-yo MUTO, Haruki UENO "Face Detection and Facial Feature Localization for Human-machine Interface", NII Journal. March-2003, No. 5, pp. 25-39.
7. Vuthichai Ampornaramveth, Haruki Ueno, "Software Platform for Symbiotic Operations of Human and Networked Robots", NII Journal, Vol.3,pp 73-81,2001
8. Matthew Turk and Alex Pentland " Eigenface for Recognition" Journal of Cognitive Neuroscience, 1991, Vol. 3, No.1, pp. 71-86.
9. Yu Huang, Thomas S. Huang, Heinrich Niemann, "Two-Hand Gesture Tracking Incorporating Template Warping With Static Segmentation", AFGR'02, IEEE, 2002. pp. 260-265.
10. M.A. Bhuiyan, V. Ampornaramveth, S. Muto, and H. Ueno, " ON TRACKING OF EYE FOR HUMAN-ROBOT INTERFACE", International Journal of Robotics and Automation, 2004, Vol. 19, No. 1, pp. 42-54.
11. Hyeon-Kyu Lee and Jin H. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition" 1999, Vol. 21, pp. 961-973.
12. Nobutaka Shimada, Yoshiaki Shirai, "3-D Hand Pose Estimation and Shape Model Refinement from a Monocular Image Sequence", Proc. of VSMM'96 in GIFU, pp.23-428
13. Radek Grzeszczuk Gray Bradski Michael H chu Jean-Yves Bouguet, "Stereo Based Gesture Recognition Invariant to 3D pose and lighting", CVPR'00, IEEE, 2000, pp. 1826-1833.
14. Yunato Cui and John J. Weng, " Hand Segmentation Using Learning-Based prediction and verification for hand Sign Recognition, Proc. of the Conference on Computer Vision and pattern Recognition (CVPR'96), IEEE, 1996, pp. 88-93.
15. Yoichi Sato, Yoshinori Kobayashi, Hideki Koike "Fast Tracking of hands and Fingertips in Infrared Images for Augmented Desk Interface", AFGR'00, IEEE, 2000, pp. 462-467.
16. Charles J. Cohen, Glenn Beach, Gene Foulk. "A Basic Hand Gesture Control System for PC Applications" Proc. of the 30th Applied Imagery Pattern Recognition Workshop (AIPR'01), IEEE, 2001, pp. 74-79
17. Dong, Guo, Yonghua Yan and M. Xie, "Vision-Based Hand Gesture Recognition for Human-Vehicle Interaction", Proc. of the International conference on Control, Automation and Computer Vision, 1998, Vol. 1, pp. 151-155.

Appearance-Based Object Detection in Space-Variant Images: A Multi-model Approach*

V. Javier Traver[1], Alexandre Bernardino[2], Plinio Moreno[2], and José Santos-Victor[2]

[1] Dep. de Llenguatges i Sistemes Informàtics, Universitat Jaume I, Castelló (Spain), vtraver@uji.es

[2] Instituto Superior Técnico, Instituto de Sistemas e Robótica, Lisboa (Portugal) {alex,plinio,jasv}@isr.ist.utl.pt

Abstract. Recently, log-polar images have been successfully used in active-vision tasks such as vergence control or target tracking. However, while the role of foveal data has been exploited and is well known, that of periphery seems underestimated and not well understood. Nevertheless, peripheral information becomes crucial in detecting non-foveated objects or events. In this paper, a multiple-model approach (MMA) for top-down, model-based attention processes is proposed. The advantages offered by this proposal for space-variant image representations are discussed. A simple but representative frontal-face detection task is given as an example of application of the MMA. The combination of appearance-based features and a linear regression-based classifier proved very effective. Results show the ability of the system to detect faces at very low resolutions, which has implications in fields such as visual surveillance.

1 Introduction

The combination of space-variant images and active-vision systems represent a biologically plausible approach to reduce the complexity of visual tasks. Attentional mechanisms [5] allow potential-interest objects be detected in periphery, so that the fovea can be directed to the selected object for its fine-detail inspection. As a remarkable space-variant image model, log-polar images have a central fovea with a very high resolution, which decreases with the eccentricity. The size of these images is given by $R \times S$ (the number of rings R and sectors S in which the original cartesian space is sampled). The particular log-polar model used in this work leaves a central circle unmapped (the *blind spot*). In an example of log-polar transformation (Fig. 1), it is worth noticing the important data reduction achieved, by comparing the sizes of images in Fig. 1(b) and Fig. 1(c).

* Research partly funded by grants E-2003-03 from *Fundació Caixa-Castelló Bancaixa*, European IST 2001 37540 (CAVIAR project), and CTIDIB/2002/333, from *Conselleria de Educació, Cultura i Ciència, Generalitat Valenciana*.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 538–546, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Fig. 1. Log-polar mapping: (a) grid layout example (10×16); (b) original cartesian image (256×256); (c) cortical image (32×64); (d) retinal image (256×256) reconstructed from (c) by the inverse mapping.

Most of past work on visual attention has focused on salience computation [6] in static images, and scarce work has been done in active-vision and foveal systems. While the benefits of the high-acuity fovea and the implicit focus of attention of log-polar images have been exploited in some active-vision problems in the past [1], the role of coarse-resolution peripheral information has been mostly neglected. However, for active-vision systems to get a real benefit of space-variant vision, such as a significant data reduction, *both* foveal and peripheral information needs to be considered appropriately.

In this paper, we focus on the problem of model-based detection of objects across the field of view (top-down attention [9]). Since the appearance of imaged objects changes significantly with their position, we introduce a multi-model approach, and illustrate it with the important problem of human-face detection. While computers and robots interacting with people require this kind of social ability, past research only considered uniformly-sampled images. Interesting results have been achieved [11], but the problem remains very difficult due to the complexity and wide variability in human faces and environmental conditions. In addition, further challenges arise when faces are to be detected within space-variant images. Therefore, the aim of the paper is *not* to propose a face detector outperforming existing systems, but to suggest a framework for object detection in space-variant images.

It is worth stressing that the problem of face detection *within* log-polar images should not be confused with the use of log-polar mapping as a *tool* for face detection/recognition. The former is the problem considered here, and, to the best of our knowledge, it has only been studied before in [7]. A few more works exist on the latter.

Section 2 describes our proposal: the multiple-model approach. For face representation/detection, a PCA-based technique has been studied (Sect. 3), and some classifiers have been tried (Sect. 4). Results are later presented (Sect. 5), before the final discussion (Sect. 6).

2 Multiple Spatial Models

In practice, top-down attention consists of using some *a priori* model of a target being searched. While just a single model is usually considered, when it comes to space-variant sensing, it makes sense to have multiple models. In the case of log-polar images, due to its lack of translation invariance and its varying resolution, the appearance of an object is non-linearly distorted in different parts of the visual field (Fig. 2). Therefore, we propose to have M models of the target (faces here). Each model $i \in \{1, \ldots, M\}$ is defined by a set of image positions $\mathcal{L}_i$, and a set of features $\mathcal{F}_i$ describing the target as viewed at $\mathcal{L}_i$. Thus, instead of having a single model representing the target and distort it *on-line* while searching for instances of it, the target is mapped *off-line* to a set of locations and a multiple-model representation of that target is built. We argue that, for space-variant images, a multiple-model approach (MMA) have significant advantages over a single-model approach (SMA):

- It is ***intuitive***. Because resolution is different at different sensor locations, it makes sense to have different models at different spatial positions.
- It offers a ***natural*** solution to a number of issues. Particular conditions such as the central blind spot, varying resolution between and within models, targets partially visible, etc., represent no problem at all, because targets imaged at different locations are never compared one to each other.
- Target detection can be as ***efficient*** in MMA as in SMA. Model acquisition usually requires a learning stage. Under MMA, each model requires its own learning process, but this is only at *off-line* time, while the *on-line* detection stage can proceed as fast as, or faster than, in SMA.
- It exploits ***data reduction***. Peripheral models can benefit from the fact of targets being imaged with fewer log-polar pixels. This is in clear contrast with the approach in [7], where a SMA is adopted. In SMA, for the feature set of all models to be directly comparable, they *all* must have the same length, which, in turn, requires image data oversampling. As a result, most feature sets are bigger than strictly necessary, and contain redundant information.

Examples of targets at different model locations are shown in Fig. 2. Notice that faces further away from the center take up less log-polar pixels. The MMA makes full sense in active vision scenarios. For example, because only a discrete number M of models can be considered, the detection of faces at positions different to $\mathcal{L}_i$ requires the collaboration of purposeful movements of a robotic head. With such motions, the same set of models $\{\mathcal{L}_i, \mathcal{F}_i\}$ can be reused, but each time observing different parts of the scene, where potential faces might be. Thus, a larger number of *virtual models* are possible under an active vision setup.

It is worth stressing that the application to face detection is just an example, and the multi-model framework is perfectly suited to detecting general visual classes of objects. Similarly, while PCA is used here for illustrating an actual application of MMA, a variety of techniques are possible under the MMA. Finally, even though we are particularly interested in log-polar images, the idea of MMA is easily applicable to other space-variant models and, more generally, whenever image distortions may happen (omnidirectional images, fish-eye lenses, etc.).

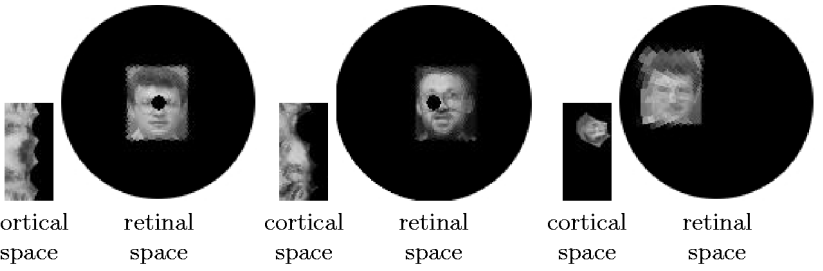


Fig. 2. Multiple models: example of human faces as viewed at different spatial locations

3 Principal Component Analysis (PCA) Under MMA

PCA is a well-known technique allowing a high-dimensional space (e.g., images) be represented in a low-dimensional one (the eigenspace)[1]. The *eigenface* technique (PCA applied to faces) has been used since the early 90s [10], mostly for face recognition and, much less, for face detection. One of the things that should be decided when using PCA is how many eigen vectors k to use: the smaller k, the more compact the representation, but the worse the samples in the training set (TS) can be approximated from the eigen vectors. Several heuristics have been suggested to choose k. Consider the eigen values $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$, with n the size of the TS, and let $V(k) = \sum_{i=1}^{k} \lambda_i$. Here, we choose the smallest k such that $\frac{V(k)}{V(n)} > \theta_\lambda$, $\theta_\lambda \in [0, 1]$. Thus, the higher θ_λ, the more variability in the TS is accounted for [8].

Under the MMA, for each model i, a rectangle of the size of the face subimage is placed over a log-polar grid at an eccentricity ρ_i and an angularity θ_i, so that the set of log-polar image positions $\mathcal{L}_i = \{l_j = (u_j, v_j),\ j = 1, \ldots, m_i\}$ underlying this rectangular region is found. Then, for each model i, a feature vector $\mathbf{I}_s^i$ is computed from the gray-level values of each face image in the TS, I_s, $s \in \{1, \ldots, n\}$, being mapped to a log-polar image A, and a function f being applied: $\mathbf{I}_s^i = \{f(A(l_j)),\ l_j \in \mathcal{L}_i\}$. The function f weights the gray-level values taking into account the size of receptive fields and to attenuate background information present in the (non-segmented) face database. Finally, each feature set $\mathcal{F}_i$ encapsulates the first k_i eigenfaces computed from all $\mathbf{I}_s^i$.

If we consider $V(k)$ for models at different eccentricities ρ, we find that more peripheral models require less eigen vectors to account for the same variability θ_λ (Fig. 3). This makes sense because a set of faces observed at coarser resolution look like more similar between them (there is less variability in the set). Therefore, peripheral models are computationally efficient, not only because the reduced number of log-polar pixels they occupy (i.e., $m_i < m_j$ for $\rho_i > \rho_j$), but also because fewer eigenfaces are needed (i.e., $|\mathcal{F}_i| < |\mathcal{F}_j|$ for $\rho_i > \rho_j$).

[1] A discussion on PCA *vs.* ICA, which is beyond the purpose of this paper, can be found, e.g., in [2].

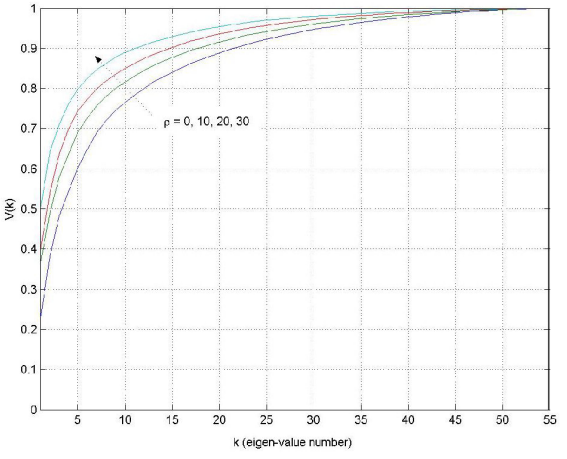


Fig. 3. $V(k)$ for different eccentricities ρ

Notice that the proposed PCA-based approach implies a learning mechanism to discover a face model for each location. The suitability of a learning-based scheme to represent and detect faces in space-variant images is further confirmed by recent evidence showing the superiority of learning over mathematical models heuristically defined [3].

To measure the distance of an input image to face space, two distances (or a combination of them) have typically been used: the distance *from* feature space and the distance *in* feature space. In theory, we could use this *faceness* for face/non-face classification, but there is one important problem: the need to set distance thresholds, a generally tricky and undesirable task. Next section proposes an effective alternative strategy overcoming these difficulties.

4 Exemplar-Based Face/Non-face Classification

Because of the drawbacks of PCA-based distances, and our negative experience using them, we looked into different ways to approach the problem. One interesting idea was having two sets, each representing the face and non-face classes, so that the system is better at discriminating between them. The problems now are that examples of both classes are needed, and that the non-face class has a huge variability and can be difficult to represent compactly. Even with these disadvantages (which, otherwise, are common in many techniques [11]), we believe this approach is preferable to having to choose thresholds.

As in any classification task, two things have to be defined: the features representing the samples and a classifier. As for the features, we have tried several possibilities, but in this paper we report on the one using the projections to face space as the feature vectors. Regarding the classifier, several alternatives have also been explored, but here we use the *Linear Regression of an Indicator Matrix* (LRIM) [4], a technique we found simple yet effective.

Fig. 4. Examples of images

Table 1. Number of images used

PCA	LRIM clusters		Test set	
Training set	Faces	Non-faces	Faces	Non-faces
130	64	64	120	190

In LRIM, the p features of N training instances belonging to one of K classes (here, $K = 2$) are given in the *model matrix*, $\mathbf{X}_{(N\times(p+1))}$, with one leading column of 1's. The *indicator response matrix*, $\mathbf{Y}_{(N\times K)}$, has N rows of K indicators $y_k,\ k = 1, \ldots, K$, with $y_c = 1, y_j = 0, j \neq c$ to indicate class c. Thus, $\mathbf{Y}$ is a matrix of 0's and 1's, with a single 1 per row, representing the class of feature vector in the corresponding row in $\mathbf{X}$. With these two matrices as input, a *coefficient matrix* is estimated as $\hat{\mathbf{B}}_{((p+1)\times K)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

To classify a new observation (a p-feature vector) $\mathbf{x}$, the fitted output $\hat{\mathbf{f}}$ is computed as $\hat{\mathbf{f}} = ([1\ \mathbf{x}]\cdot\hat{\mathbf{B}})^T$, and its largest component, identifies the predicted class c to which $\mathbf{x}$ belongs: $c = \arg\max_k \hat{\mathbf{f}}_k(\mathbf{x})$.

We note that, for classification efficiency, it is valuable the fact that the size of $\hat{\mathbf{B}}$ depends on the number of features, but *not* on the number of samples in $\mathbf{X}$. And because the feature vector in our case is small (projections in eigenspace), the choice of LRIM as a classifier turns out to be particularly convenient when used in conjunction with PCA.

5 Experimental Results

In the experiments, we used the AT&T face database (`www.uk.research.att.com/facedatabase.html`), consisting of 400 images (40 individuals, each at 10 different approximately frontal views). We collected 190 non-face images for the LRIM stage and for testing purposes. Some of the non-faces images are upside-down faces (which are not in the *face* class, as meant here). This helps the system classify test images that could otherwise be misclassified. Examples of face and non-face images are shown in Fig. 4.

The distribution of images in different stages of the algorithm is summarized in Table 1. The face cluster for LRIM was a subset of the TS used in PCA. The test set included faces of subjects not included in the TS, as well as different views of faces of subjects in the TS. Regarding the log-polar images, the considered size was as small as $R \times S = 32 \times 64$.

With all the elements comprising this experimental setup, there are many and interesting issues that deserve exploration. The influence on the performance of some factors is plotted in Fig. 5, where, besides recognition error, recognition rates for faces and non-faces are shown, so that the number of false positives and

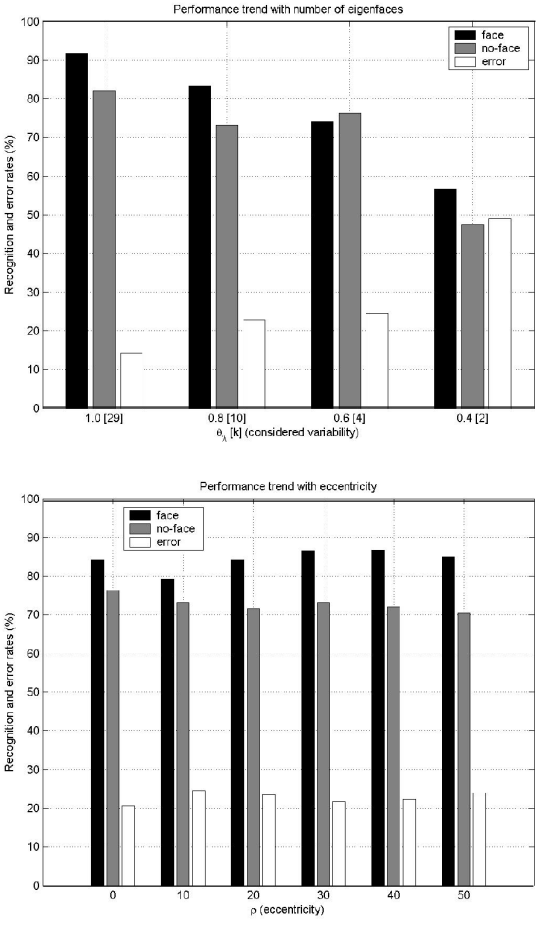


Fig. 5. Performance trends (see text for details)

false negatives can readily be perceived. On the other hand, more important than the actual rates, it is the performance *trend* when varying these factors.

In the first experiment, we used the central model ($\rho = 0$), and the number of eigenfaces was decreased by changing θ_λ, which accounts for the variability in the training set. For these θ_λ, the number of eigen vectors were 29, 10, 4 and 2, respectively. The less eigen vectors, the faster the classification can be, but as the results reveal (Fig. 5 (top)), the representativeness of a face diminishes, and so the recognition rates. It is worth noticing how, with as few as 10 eigen vectors, recognition is reasonable good, which points to the appropriateness of PCA as a technique to represent face patterns economically.

Next, the effect of eccentricity is analyzed. We kept θ_λ fixed to 0.7, while the training and test images were placed at increasing distance from the center,

ρ, thus loosing visual acuity with eccentricity. The classification results, shown in Fig. 5 (bottom), are somehow surprising and unexpected: there is no significant decay in performance. The possible explanation of this seemingly counter-intuitive result is that the important, discriminative features in the face pattern can be observed at sufficient detail even at very coarse resolution. This somehow reminds us the amazing human ability of spotting human faces even in adverse conditions, and points to the key role that the periphery in log-polar images may play in attentional tasks.

6 Conclusions

A multiple-model approach (MMA) has been proposed as a general framework for modeling classes of objects in the context of space-variant images and active vision setups. The motivation behind this is the use of log-polar images for top-down visual attention tasks, and the exploration of the role that peripheral information can play. While MMA is a general approach, it has been illustrated with an application to the problem of face detection in log-polar images.

PCA has been used within MMA (i.e., multiple PCAs) as a technique for an efficient representation of a face class (one face class per spatial model), and has been shown to be effective, although quite sensitive to out-of-class variations. To overcome the encountered limitations of existing PCA-based distances, a simple but effective exemplar-based strategy to classify input patterns into face/non-face classes has been provided. Importantly, the use of LRIM in tandem with PCA turned out to be both efficient and effective.

Although traditionally underestimated, peripheral data, even at coarser resolution, plays an important role for fast attentional tasks, as experimental evidence reveals.

Acknowledgments. We want to acknowledge comments from Matthew Turk, clarifications from DoJoon Jung, discussion with Roger Freitas and Lorenz Gerstmayr, comments and Matlab code for PCA from Matthew Dailey, and proofreading from Adolfo Martínez and Raul Montoliu.

References

1. C. Capurro, F. Panerai, and G. Sandini. Dynamic vergence using log-polar images. *Intl. Journal of Computer Vision*, 24(1):79–94, 1997.
2. J. Draper, K. Baek, M. Bartlett, and J. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding (CVIU)*, 91:115–137, 2003.
3. H. M. Gomes and R. Fisher. Learning-based versus model-based log-polar feature extraction operators: a comparative study. In *XVI Brazilian Symp. on Computer Graphics & Image Processing (SIBGRAPI)*, San Carlos, Brazil, Oct. 2003.
4. T. Hastle, R. Tibshirani, and J. Friedman. *The elements of statistical learning; data mining, inference, and prediction.* Springer, 2001.

5. L. Itti. Modelling primate visual attention. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*, pages 635–655. CRC Press, 2003.
6. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, July–Aug. 2000.
7. F. Jurie. A new log-polar mapping for space variant imaging. Application to face detection and tracking. *Pattern Recognition*, 32:865–875, 1999.
8. H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *Intl. Journal of Computer Vision*, (14):5–24, 1995.
9. Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146:77–123, 2003.
10. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, Maui, Hawaii, 1991.
11. M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):34–58, Jan. 2002.

3D Object Recognition from Appearance: PCA Versus ICA Approaches

M. Asunción Vicente, Cesar Fernández, Oscar Reinoso, and Luis Payá

Industrial Systems Engineering Department,
Miguel Hernandez University, Avda. de la Universidad s/n,
03202 Elche, Spain
{suni, c.fernandez, o.reinoso, lpaya}@umh.es
http://lorca.umh.es

Abstract. Two feature extraction techniques (PCA/ICA) for recognition of 3D objects from appearance are compared with respect to different recognition approaches (universal/object subspaces). A class separation ratio is defined, and several recognition experiments are performed using the COIL-100 database. The results show that both techniques produce similar recognition rates when universal subspaces are used; but, when object subspaces are used, ICA representation greatly outperforms the earlier PCA technique due to its ability to separate classes.

1 Introduction and Motivation

Object recognition is a fundamental ability of any visual system. Recognizing real three-dimensional objects in controlled backgrounds or scenes is quite easy provided an adequate 3D visual model of the object is available. However, non-controlled backgrounds make it almost impossible to apply such model-based techniques because in these situations it is very difficult to segment the target object from the scene. Appearance-based recognition approaches are a powerful alternative to model-based techniques when it is difficult to obtain geometrical models of the objects [1] and when the images have non-controlled backgrounds [12] [9]. First appearance-based systems found in the literature used principal component analysis (PCA) as a feature extraction technique to reduce the dimensionality of the object classes or models [7][10], while recently the use of the independent component analysis (ICA) for feature extraction is preferred by some authors [2]. The performances of both techniques have not been compared in terms of their applicability to object recognition and sometimes contradictory conclusions have been drawn. The aim of this paper is to clarify the advantages and drawbacks of both techniques as feature extractors in object recognition systems.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 547–555, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In [7] Murase addresses the problem of automatically learning object models for recognition and pose estimation. PCA is used and the objects are represented in two different *eigenspaces*: the *universal eigenspace*, computed using an image set of all objects of interest, and the *object eigenspaces*, computed using only images of an object. The universal eigenspace is best suited for discriminating between objects, whereas the object eigenspace is better for pose estimation. In this paper the structure of the PCA-based Murase recognition system is used to perform the comparison of ICA and PCA approaches with the COIL100 library [8]. The results presented throughout the paper show how ICA clearly outperforms PCA when using object subspaces. The reason can be found in its class separation capability, as stated by Bressan [3].

2 Recognition of 3-D Objects from Appearance

The appearance of a 3D object in a 2D image depends on its shape, its colour, its pose in the global scene, its reflectance properties and the sensor and illumination characteristics. An object image may be considered as a vector of pixels where the value of each entry in the vector is the greyscale (or colour) value of the corresponding pixel. For example, a NxN image may be unwrapped and treated as a vector of length N^2. The image is said to sit in N-dimensional space, which is considered to be the original space of the image. In appearance-based systems the whole object image is projected to a lower dimensional space using different techniques or subspaces, the most frequently used is the subspace created by the eigenvectors of the covariance matrix of the training images (or PCA) [7] [10], another common subspace is the one created by the basis vectors obtained using Linear Discriminant Analysis or the subspaces computed by Independent Component Analysis [6].

2.1 Constructing the *Object Subspaces*

Two different representations are used: namely the *universal subspace* and the *object subspaces* following Murase nomenclature [7]. Each object of interest can be represented or modelled using an *object subspace*. This subspace is constructed from a set of training images belonging to different views of the same object. The set of M training images, $\vec{x}_m$, is defined by G_j:

$$G_j = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_M\}. \tag{1}$$

and there are as many sets as different objects of interest. With $\vec{x}_m$, we denote a vector image that groups the three color image matrices. Each image, $\vec{x}_m$, from the training set is filtered by the feature channel selected to model the object:

$$\vec{f}_m = F_{channel}(\vec{x}_m). \quad (2)$$

Typical feature channels are: the image in a modified colour space, the image histogram, the border image, or special areas from the image grouped in feature windows (local image features) [2][9]. Once the feature channel is selected, the new set of M training feature vectors, G_j' is defined:

$$G_j' = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_M\}. \quad (3)$$

The object subspace, W, is computed from G_j' for each object *j*, extracting the principal components (PC) or the independent components (IC). After that, all the training image features are projected in that subspace creating an object representation, named as *manifold.* That projection can be represented as:

$$\vec{p}_m = W \cdot \vec{f}_m. \quad (4)$$

where W is the matrix containing the PC's or the IC's from each object.

2.2 Constructing the *Universal Subspace*

The universal subspace is constructed from a set of training image features belonging to different objects. The set of J training images is defined by U :

$$U = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_J\}. \quad (5)$$

The universal subspace, W , is also computed by PC or IC extraction. But in this case there is only one subspace for all the objects. So, the model of each object is composed of its views projected in the universal subspace.

3 Selection of Features Using Linear Transformations

3.1 PCA Approach

PCA subspace or eigenspace is computed by finding the eigenvectors of the covariance matrix created from the set of training vectors [4]. The eigenvectors corresponding to non-zero eigenvalues of the covariance matrix represent an orthonormal basis that projects the original vectors (of length N^2) in the M-dimensional space ($M<<N^2$). So in PCA approach, W is the matrix containing the eigenvectors. In this work the *snapshot* method [10] has been used to compute the eigenspace in order to avoid the high dimensionality of the original covariance matrix. PCA enables us to create and use a reduced set of variables. A reduced set (the classes obtained from the training images) is much easier to analyze and interpret than the original variables (the training images themselves).

3.2 ICA Representation

The ICA of an N^2 dimensional random vector is a linear transform that minimizes the statistical dependence between its components. This analysis has a great number of applications such as data analysis and compression, blind source separation, blind deconvolution, denoising, etc.

If the random vector we wish to represent through ICA has no noise and is zero-centered, the ICA model can be expressed as:

$$\vec{x} = A \cdot \vec{s}. \tag{6}$$

where $\vec{x}$ is the random vector representing our data, $\vec{s}$ is the random vector of independent components with dimension $M \leq N^2$, and A is the mixture matrix. The pseudoinverse of A, represented by W, is called the projection matrix and it provides an alternative representation of the ICA model:

$$W \cdot \vec{x} = \vec{s}. \tag{7}$$

Various objective functions have been proposed for the estimation of the projection matrix such as nongaussianity, likelihood, mutual information, and tensorial methods [6]. In this paper the FastICA [5] method is used. FastICA estimates the whole decomposition by minimizing mutual information, and estimates the individual independent components as projection pursuit directions. FastICA uses PCA as a preprocessing step for data whitening.

4 Experimental Results with the COIL-100

The experiments are performed with object images from the COIL-100 database [8]. The selected feature channel to carry out the comparison has been the *I* component (*HIS* colour space).

4.1 Comparing *Universal Subspaces*

In order to compare the performance of PCA and ICA universal subspaces were computed from the main image of each object in the database (10 to 100 object sets were used in different experiments). The rest of the images were used as test examples and the classifier selected was the nearest neighbour (k-NN with k=1 and L2-norm as distance measure). In Fig. 1 the experimental results from both feature selectors are shown: both techniques performed almost equally. Fig. 1a shows the recognition rates obtained using both ICA and PCA when only 10 components are used to represent each subspace. As the number of different objects increases, there is a reduction in classification accuracy with both systems. In Fig. 1b the influence of the dimensionality of the subspace is shown: all experiments are carried out with 20 different objects, and the number of components varies from 2 to 20. As expected, the recogni-

tion rate grows as the dimensionality of the subspace increases. Experiments performed with bigger object sets gave similar results.

The universal subspaces perform well with symmetrical objects, as all the views of an object are very similar; but it produces disastrous results with quite non-symmetrical objects. This is shown in Fig. 2: the symmetrical objects, (2, 4, 5, 18, 24, 25, 26, 30, 32, 33, 34, 35, 47, 50, 56, 58, 61, 64, 70, 72, 83, 86, 87, 88, 94, 95), give a recognition rate of 100% in both types of subspaces whereas non-symmetrical objects are hardly recognized. We can also notice than ICA works only slightly worse than PCA as there are 26 symmetrical objects in PCA's case and just 21 in ICA's (5, 18, 24, 25, 26, 30, 32, 34, 35, 47, 50, 56, 58, 61, 70, 72, 83, 86, 87, 94, 95). These similar results were expected as ICA representation is based on a dimensionality reduction obtained by using PCA in the pre-processing step of the FastICA algorithm and the further statistical independence forced by ICA does not necessarily improve the results.

As a conclusion, universal subspaces with a distance measure can be used to recognize objects when they are quite symmetrical or they always show the same view to the camera. Universal subspaces are mostly used in face recognition systems with the processed frontal face (*mugshots*) as the training images.

If a more complex classifier is used instead of a distance measure, the recognition rates can be improved even with non-symmetrical objects; on the other hand, this requires the use of more training views in the learning step [12].

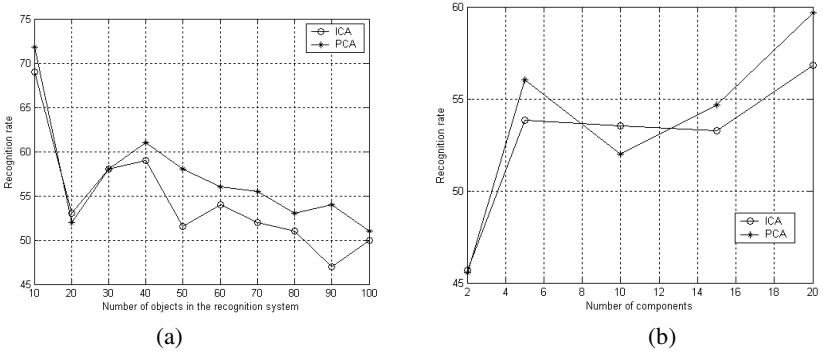
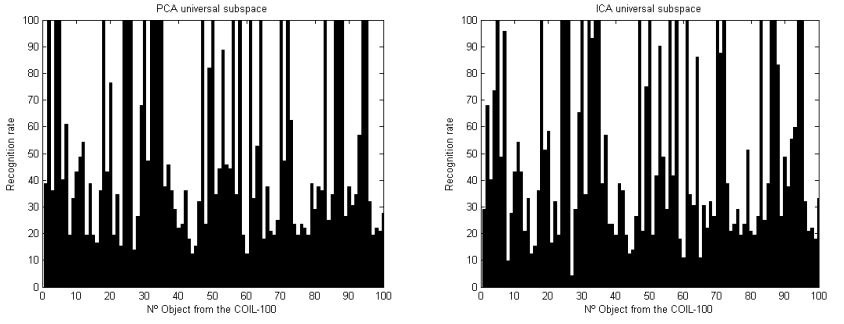


Fig. 1. Comparing universal subspaces: Fig 1.a shows ICA and PCA recognition rates using just 10 dimensions, and Fig 1.b is referred to the influence of the subspace dimensionality.

4.2 Comparing *Object Subspaces*

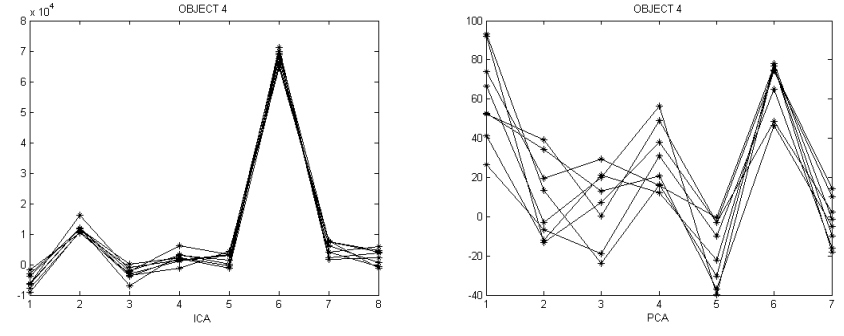
Object subspaces generated with PCA were used by Murase in [7] in order to find out the pose of the object and not to identify it, but they are also useful for recognition as described in [11]. Even more, if ICA is used instead of PCA, this technique greatly outperforms the universal subspace method as the ICA manifolds of each class are more apart from each other. The reason can be found in the close relationship existing



(a) (b)

Fig. 2a. Symmetrical objects in PCA subspace. **Fig. 2b.** Symmetrical objects in ICA subspace.

between sparse coding and ICA: sparse coding is a coding of the data such that only a few components of the code will be significantly active (nonzero) In Fig. 3a it is possible to observe the high sparsity in the ICA manifolds (each object fires a certain component, while the remaining ones are kept close to zero). PCA manifolds do not represent a sparse coding as each component is not associated to a certain object; the components are instead ordered in decreasing values of their variance in the training data (Fig. 3b).

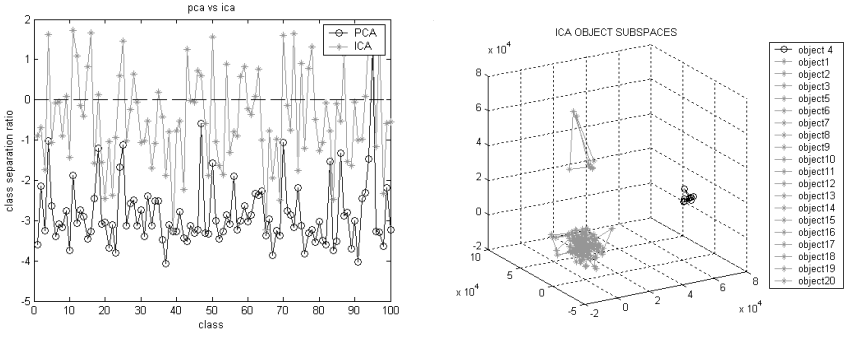


(a) (b)

Fig. 3a. Values of the ICA projections over an 8 component subspace (object 4). The sparsity of the code is evident: all but one of the components are close to zero. **Fig. 3b.** Values of the PCA projections over a 7 component subspace (same object). It can be seen that the components are ordered in decreasing values of variance.

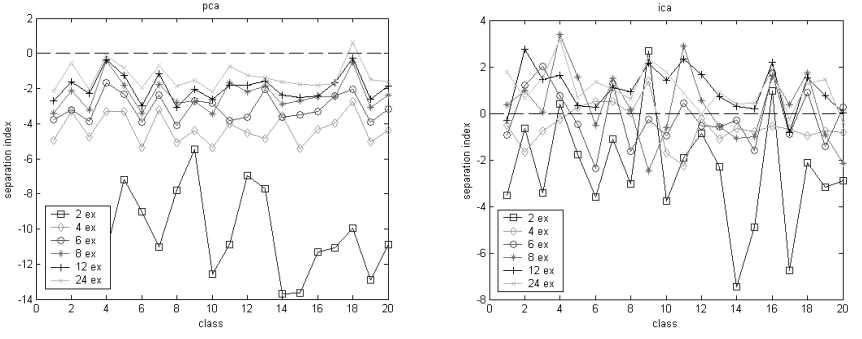
In order to compare the two approaches a class separation ratio has been defined. This ratio represents the degree of separation of a certain class, and it can be expressed as follows:

$$\alpha_j = \log\left[\frac{\max(d_j)}{\min(D_j)}\right]. \tag{8}$$



(a) (b)

Fig. 4a. Class separation ratio for ICA and PCA object subspaces for the whole library. The subspaces are computed using just 8 training images. **Fig. 4b.** In this figure the reduced manifolds of the 20 first objects from COIL-100 are displayed, the three mostly fired components have been selected for a 3D representation .



(a) (b)

Fig. 5. Class separation ratio for PCA(a) and ICA(b) object subspaces for the 20 first objects from COIL-100 using different dimensionalities of the manifolds (from 2 to 24 components). Class separation increases with the dimensionality (larger training sets) in both cases, but ICA clearly outperforms PCA even with a lower number of components.

where α_j is the class separation ratio for class j, d_j is the L2 norm between two projection vectors belonging to manifold j (distances between elements of the same class) and D_j is the L2 norm between two projection vectors, only one of them belonging to manifold j (distances between elements of different classes).

High α_j values represent easily separable classes, and particularly, when α_j is greater that zero, the classes can be trivially separated by any simple classifier. The class separation ratio for ICA and PCA object subspaces for the whole library is shown in Fig. 4a : most of the ica-classes from the COIL-100 are above or close to the threshold $\alpha_j = 0$. However the pca-classes are in almost all objects below this threshold, meaning that they are not so easily separable.

This fact is also shown in Fig.5a and 5b., where each approach is tested with a different number of training examples (from 2 to 24). It can be seen that PCA class separation ratio falls below the $\alpha_j = 0$ threshold even with 24 training examples; whereas the ICA class separation ratio is always higher than the PCA one and even more, only 4 components are required to trivially separate some of the objects. Several classification experiments have been performed using a linear perceptron, a k-NN rule and a naive Bayes classifier, and the recognition rates obtained confirm the results: with just a few components (8 o more) ICA obtains recognition rates close to 100%, whereas PCA does not reach a 100% classification rate even with 24 components.

5 Conclusions and Future Work

Two main conclusions are drawn from this work:

- The universal subspace technique with a distance based classifier is only applicable to symmetrical objects or objects that always present the same view to the camera. In this scenario ICA does not improve PCA results.
- The object subspace technique benefits from using ICA instead of PCA. ICA produces a reduced model of the object due to the sparse coding of the CI's; even more, class separation is much more relevant even with only a few components.

An ICA-based object subspace technique allows extremely compact representations of the objects. This fact makes this technique suitable for the combination of different feature channels in order to improve the recognition capability of the visual system. Future works will consider simultaneously color, shape and texture information combined to get an almost invariant representation of the object, useful for object recognition in non-controlled backgrounds.

References

1. Biederman, I: Visual object recognition. In: S. F. Kosslyn and D. N. Osherson (Eds.). An Invitation to Cognitive Science, 2nd edition, Volume 2., Visual Cognition. MIT Press. Chapter 4, pp. 121-165. 1995.
2. Bressan M., Guillamet D. and Vitrià J.: Using an ICA representation of local color histograms for object recognition. Pattern Recognition 36(3): 691-701 (2003)
3. Bressan M. and Vitrià J.: On the Selection and Classification of Independent Features, IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 25 n.10, pp. 1312-1317, 2003.
4. Duda Richard.O., Hart Peter E. and Stork David G., Pattern Classification, Wiley-Interscience, 2001
5. FastICA software, http://www.cis.hut.fi/projects/ica/fastica/
6. Hyvärinen A., Karhunen J., and Oja E.: "Independent Component Analysis", Wiley-Interscience, 2002.

7. Murase H. and Nayar S.: "Visual learning and recognition of 3D objects from appearance", International Journal of Computer Vision, no.14, 1995, pages 5-24.
8. Nene S. A., Nayar S. K. and Murase H.: “Columbia Object Image Library (COIL-100)”, Technical Report CUCS-006-96, 1996, http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html.
9. Ohba K. and Ikeuchi K., Detectability, Uniqueness, and Reliability of EigenWindows for Stable Verification of Partially Occluded Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19 n. 19, pp. 1043-1048, 1997.
10. Turk M. and Pentland A., Eigenfaces for Recognition, Journal of Cognitive Neuroscience, vol. 3, no. 1, pp.71-86,1991.
11. Vicente M.A., Reinoso O., García N., Sabater J.M., Fernández C., Jiménez L.: Experiments with COIL database using PCA-based Object Recognition Techniques, ISRA’2002 3rd International Symposium on Robotics and Automation, 2002.
12. Vicente M.A. , Reinoso O. , Fernández C.: "Recognition and location of real objects using eigenimages and a neural network classifier", Proceedings of SPIE: Visual Communications and Image Processing, vol. 5150, pp. 385-392, 2003.

A Stochastic Search Algorithm to Optimize an N-tuple Classifier by Selecting Its Inputs

Hannan Bin Azhar[1] and Keith Dimond[2]

[1] Department of Electronics, University of Kent,
Canterbury, Kent, England
maha2@kent.ac.uk
[2] Master, Keynes College, University of Kent,
Canterbury, Kent, England
k.r.dimond@kent.ac.uk

Abstract. The N-tuple method [4] is a statistical pattern recognition method, which decomposes a given pattern into several sets of *n* points, termed "*N* tuples". The input connection mapping of the N-tuple classifier determines the sampling and defines the locations of the pattern matrix. Realizing the fact that the classification performance of the N-tuple classifier is highly dependant on the actual subset of the input bits probed [3][7], we have introduced an approach based on a **R**eward a**n**d **P**unishment (RnP) scheme to select input mappings of the classifier. We termed the classes with high error rates as critical classes. Different groups of tuples have been formed for different classes. The strategy was to employ more number of tuples to a critical class-group than an easily distinguishable class. In order to illustrate the capabilities of the RnP based measure the task of recognizing hand-written digits from NIST [10] database has been chosen.

Keywords: N-tuple classifier, Pattern recognition, Stochastic search.

1 Introduction

The N-tuple classifier [4] is one of the oldest pattern recognition methods and has been successfully used in many application domains [8][2]. The Random Access Memory (RAM) is the basic component of the classifier. A RAM node (tuple) can be viewed as implementing a look-up table, the entries of which are determined by the RAM inputs. Sampling of *n* specific data locations of the input constitutes a 'feature' of the pattern. A pattern is classified as belonging to the class for which it has the most features in common with at least one training pattern of that class.

The classification performance of the N-tuple classifier is highly dependant on the input bits probed [7][3]. The number of possible connections for a 32x32-bit pattern matrix is enormous. Let us consider an N-tuple classifier of 140 tuples with the tuple-size 8 bits. For an input binary image of 32 bits high and 32 bits wide the total number of pixels are 1024 bits. Now if we consider that the same pixel doesn't repeat in a single tuple then the possible number of tuples that can be formed is found by the formulae of combination, $M = {}^{1024}C_8 \approx 2.91 \times 10^{19}$. Now M tuples when divided into

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 556–563, 2004.
© Springer-Verlag Berlin Heidelberg 2004

groups of 140 then total number of combinations will be $B = {}^{M}C_{140}$, which is a very big number. Therefore an exhaustive search for B mappings is impossible.

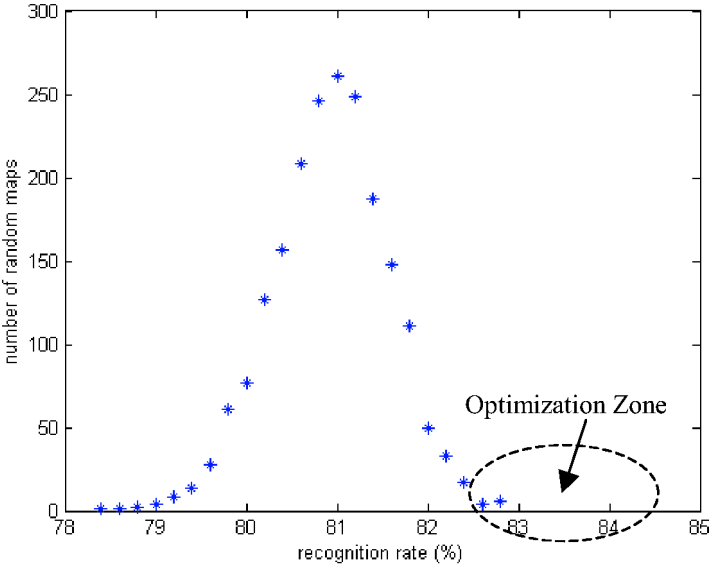


Fig. 1. Classification histogram with 2000 different maps to recognize handwritten characters from NIST Special Database 19 [10] with 140 tuples of tuple-size 8 bits.

The classification performance is a function of input mappings and it approximates to a normal distribution [1], where the majority of the mappings give average performance. Figure 1 demonstrates the classification histogram for a character database where two thousands maps were generated randomly and the frequency of the maps was plotted. If we generate all B mappings as we explained earlier, the tail on the right side of the histogram would go much farther as shown by the imaginary dotted area in Figure 1. To find the mappings in this dotted optimization zone is extremely challenging. The detection of mappings in this area by random search is governed by chance. Jung et al. [5] presented a method for selecting an optimal set of n-tuples in order to recognize classes of characters. The relative confidence [9] of a class discriminator [9] has been measured in [3] to find different sampling sequences for characters those are too similar. Jorgensen et al. [7] proposed a measure, which combines the concept of cross-validation and Shannon information theory to select the optimum connections needed to achieve a given performance.

Unlike others our inspiration was to develop an algorithm that would find maps in the optimization zone of Figure 1 and would improve the classification performance considerably. Our proposed method will measure the performance of an individual map based on a reward and punishment scheme and the best scored maps over a number of iterations will be selected for classification task to recognize handwritten characters from the NIST [10] database. The next section will describe our proposed stochastic search method and Section 3 will describe the reward and punishment

based performance measure to evaluate input connections. Section 4 will show the experimental outcomes and finally Section 5 will conclude the paper.

2 Tuple Search Algorithm

Our objective was to find an optimal set of input connections for N-tuple classifier that will give higher recognition rate than the random case. The performance of each connection map during search will be evaluated by a reward and punishment technique that will be explained in the next section. We termed the classes with high error rates as critical classes. As critical classes are giving low recognition rate, one strategy would be to optimize or tune a number of tuples so that they will give high scores to only one critical class. We distributed the total available tuples among classes proportionately to the error rates. This way classes with the higher error rates will get more tuples. This scheme ensures that more care has been taken to the critical classes by engaging tuples specific to a class. It has been found that class-specific tuples best describe a certain class by giving high scores evaluated by the objective function (Section 3) and might also give moderately high scores for few other classes. Thus tuples engaged to a specific class best describe the features of that class and also have the features for other classes to some extent.

The pseudo code of the search algorithm is given below:

```
LET i = 0;//class index
    Tf = 0;//number of successful tuples in any
           //iteration
    p'i = number of tuples responsible to best describe
         class Ci;
REPEAT
GENERATE ([P-Tf] set of tuples randomly);
FIND SCORE (P set of tuples based on class Ci);
RANK (P set of tuples based on their scores);
CARRY (Tf set of successful tuples to next iteration);
IF (Tf = p'i) THEN {
              SAVE(p'i set of tuples as mature tuples);
              SET (i = i + 1, Tf = 0);}
UNTIL (i < number of classes);
```

To understand the search algorithm to find class specific tuples, consider a class index *i* to identify a class '*Ci*'. Let us consider an objective function (Section 3) that gives scores for class-specific tuples. Let's say *P* is the total number of tuples in any iteration. Our target is to find *N* number of class-specific tuples in total. The distribution of *N* tuples among the classes is proportionate to the error rates. The more the class is critical, the greater number of tuples it gets. Let us assume *p'i* is the number of tuples that will be matured for class *Ci*. Thus the summation of all *p'i* ($\Sigma p'i$) will be equal to *N*. In every iteration, scores for *P* sets of tuples are evaluated according to the objective function based on *Ci.* So the RnP based objective function will give the scores only for the class *Ci.* Then *P* sets of tuples are ranked based on their scores for the class *Ci.* The number of tuples, say *Tf*, whose scores are higher than a predefined threshold are treated as the successful tuples for a certain iteration and they are being carried to the next iteration by virtue of their good scores. So in the next iteration only *P-Tf* tuples will be created randomly. Before moving to the next iteration a check has to be made if the number of successful tuples (*Tf*) have met the number of class-specific tuples (*p'i*) in the class *Ci*. If *Tf* = *p'i,* then the mappings for these successful tuples will be saved and these tuples will be treated as the matured tuples those are specific for the class *Ci.* The whole process repeats until all matured class-specific tuples ($\Sigma p'i = N$) for all classes have been sought and later these matured tuples will be used for the final recognition task.

3 RnP Based Performance Measure

A trained classifier can either recognize or misclassify or reject a test pattern. In the reward and punishment (RnP) scheme a reward is associated with the correct recognition of the pattern and the penalties for misclassification and rejection. The whole pattern data are divided in three parts: training set, evaluation set and test set. Let us consider S_l is the total number of samples for training the classifier, S_e is the number of samples available for evaluation purpose and S_t denotes the number of samples in the test data set. If *S* is the total number of available samples then $S = S_l + S_e + S_t$. Now for optimization purpose the network is trained with S_l and evaluated with S_e dataset. For the final recognition task both the S_l and S_e are used for training the network and S_t is used for testing. Finally, the dataset S_e can be considered to have three parts: Sc_i, S_r and S_m. Sc_i is the number of samples recognized only as the class *Ci*, S_r is the number of samples rejected and S_m denotes the number of samples misclassified. Considering all these definitions the objective function for class *Ci,* denoted as *O(i)*, will be evaluated by the following equation:

$$O(i) = Sc_i \times R + S_r \times P_r + S_m \times P_m \tag{1}$$

Where, R = Positive Points associated for reward for recognition,
P_m = Negative Points associated for misclassification,
P_r = Negative Points associated for rejection.

The point scheme for the objective function (Equation 1) will be explained in the next section. As explained in the previous section, in every iteration the search algorithm selects the number of tuples as the successful tuples according to a threshold. Consequently a threshold function was developed for the system, which is shown in the Equation 2:

$$O(i) \geq O(i)max \times (1 - k(t)) \tag{2}$$

$$k(t) = \exp(-\tau / t) \tag{3}$$

In the Equation 2, *O(i)max* is the score of the tuple which is the maximum among all scores in the current iteration and the value of *k* defines the percentage of the *O(i)max* that will be set as a threshold to be required to become a successful tuple. To accelerate searching, *k* can be varied over time based on the Equation 3. In this equation τ is the time constant which should be chosen suitably to set the exponential decay of the threshold over the number of iterations. Like any other stochastic search, the RnP based search will give better results if more time is given for the search. So the value of τ should be carefully chosen and varied throughout the search as a tradeoff between the performance and the speed.

3.1 Point Scheme for RnP

The point scheme determines what values we should set for *R, Pm* and *Pr* in the Equation 1. In general a rejection is thought to be more favorable than a misclassification. It is equivalent to the system getting confused rather than making the wrong decision. To find out what should be the point to set for the reward consider a point for misclassification as *Pm* = –1. If *J* is the criterion deciding minimum number of samples of the class *Ci* that must be recognized by one individual tuple to maintain a predefined minimum score, say *Omin*, then the value of *R* can be found by the following formulae:

$$J \times S_v \times R + P_m \times S_v(1 - J) = Omin \tag{4}$$

In the above formulae if *Omin* is set to 500 and *J*=5%, S_v=500; then *R* comes out as 39. So rewarding points should be 39 for a recognition of a pattern when *Pm*=-1. As rejection is more favorable than misclassification *Pr* can be set empirically as –0.5. *J*=5% indicates that the pattern is so complex that even when the least 5% of the samples of the evaluation data set is recognized by a single tuple, the evaluation function (Equation 1) will give a minimum score of 500. When the tuple-size is larger, a single tuple can recognize more patterns of the evaluation set. So for the larger tuple-size higher percentage of *J* can be chosen. It is more convenient to choose lower *J* value as it works for both higher and lower tuple-size, on the contrary algorithm with higher *J* value works only for the larger tuple-size.

4 Application of RnP Based Optimization to NIST Data Set

We applied our developed stochastic search method to recognize handwritten characters from the NIST database. We used the partition *hsf_{0}* [10] of the Special Database 3 for training and the partition *hsf_{4}* [10] from the Special Database 7 for testing. NIST recommends using images of *hsf_{4}* for testing as they are more difficult from other partitions and this ensures the heterogeneity between the training and testing set (see analysis in [10]), a fact which is reflected in our results. The numeric data set consisting only the digits (0,1…9) was used for the experiments. Each character is a binary image with the dimension 32 by 32. All digits are scaled into same dimension and centered. In the partition *hsf_{0}* there are 1000 images per class for training and in the partition *hsf_{4}* there are 1000 images (S_t) per class for testing. For the experiments the total train samples were again divided into two halves by the holdout method [6] and one part (S_l) was used to train the network in the evaluation phase and other part (S_e) to evaluate the RnP based objective function. After finding the mature tuples, all training images ($S_l + S_e$) of *hsf_{0}* were used to train the classifier and the images from *hsf_{4}* were used to test.

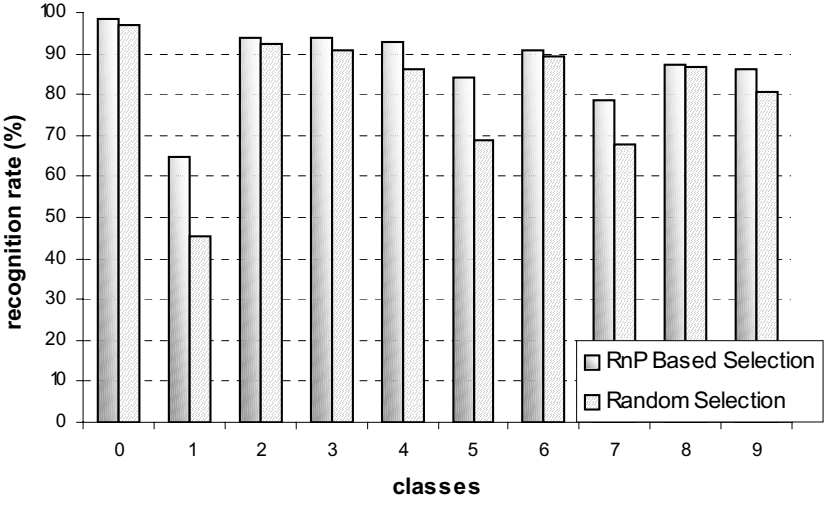


Fig. 2. Class-wise comparison of recognition rates between random selection and RnP based selection. For a class, in RnP based selection all tuples were tuned only for that class.

Two experiments were performed. The first one was to demonstrate how the recognition can be improved for a class when all the tuples in the network are tuned only for that particular class. In total 140 tuples were used for the network with the tuple-size 8 bits. The results are shown in the Figure 2. It can be seen that character 1 is the most critical class to be recognized in the NIST database. For the random case the recognition of class 1 was 45.26, which was the mean of 2000 runs. The RnP based optimization improves the recognition of class 1 by 19.66% (Fig. 2). It improves the recognition rate for class 5 by 15.56%, class 7 by 11.02% and all other classes by

several percents. The recognition rates found by RnP based search (Fig. 2) correspond to the mean of ten runs.

Table 1. Experimental Settings for RnP Optimization

Parameters	Values
Time constant (Eqn. 3), τ	100
Constant (Eqn. 4), J	5%
Point for misclassification, Pm	-0.5
Point for rejection, Pr	-1
Point for reward, R	39
Population size, P	200
For experiment 2: {*p`0,p`1, p`2,p`3, p`4, p`5,p`6,p`7,p`8,p`9* }	{2,42,6,7,10,24,8,25,11,15}

Table 2. Improved overall recognition rate by RnP based optimization

Methods	Average Recognition Rate (%)	Best Recognition Rate (%)
Conventional randomly selected N-tuple	80.93	82.83 (in 2000 runs)
RnP Based Stochastic approach	83.67	84.5 (in 10 runs)

The second experiment was required to demonstrate the improvement of the overall recognition rate by the stochastic search method. The overall rate was the average of all recognition rates of all classes. The total number of tuples for the experiment was 150 with the tuple-size 8 bits. Tuples were distributed among classes proportionately to the error rates of the classes. Thus being the most critical class, character 1 gets 42 tuples out of 150. The number of tuples for other class-groups (*p`i* in Section 2) can be found in Table 1. The randomly selected network was run for long enough (2000 iterations) to give it a chance to find better input maps that could be comparable with the maps found by our stochastic search. After long simulations we found the best overall recognition rate by the random network to be 82.83% and the average rate of 80.93%. In case of stochastic optimization (Table 2) the average overall recognition was 83.67%, which was 2.74% superior to the random case. We used a T-test with 99.9% confidence to determine if results between algorithms were significantly different. Our calculated t-value exceeded the tabulated value at the level of significance of 0.001. This shows that the increase in recognition rate by RnP based approach over traditional N-tuple method is statistically “very highly significant”. The results were obtained with the code running under Windows XP on a 2 GHz Pentium 4 machine.

5 Conclusions

In this article we focused on a stochastic search technique to select an optimal set of n-tuples. We achieved only 2.74% of improvement in recognition rate by RnP based approach over conventional randomly selected N-tuple method [4]. This small improvement in recognition was due to the fact that in this paper we didn't consider any type of "diversity" while forming tuples through the stochastic process. Without diversity a pixel may copy itself several times while forming a tuple. Repetition of same pixels in a tuple can obstruct the exploration of new features in the image. Thus some sort of diversity control could benefit the system to achieve even better recognition rate than the one we achieved in this paper. We will try to explore this diversity issue in our follow-up work in future. This article will be a good reference to realize the underlying methodology of the RnP based stochastic process for N-tuple classifier. From the application point of view we expect our RnP based approach will help to reduce the requirement of hardware resources considerably as it can achieve a given level of recognition rate with less number of tuples.

References

1. Aleksander, I., and Stonham, T.J., "Guide to Pattern Recognition using random –access Memories," Computers and Digital Techniques,2,29-40 (1979)
2. Aleksander, I., W. V. Thomas, and P. A. Bowden, WISARD: A radical step forward in image recognition. Sensor Review 4 (3), 120—124. (1984)
3. Bishop, J.M., Crowe, A.A., Minchinton, P.R. and Mitchell, R.J., "Evolutionary Learning to Optimise Mapping in n-Tuple Networks", IEE Colloquium on "Machine Learning", Digest 1990/117. (1990)
4. Bledsoe, W. and Browning, I. Pattern recognition and reading by machine, Proceedings of Eastern Joint Computer Conference, pp.225-232, Birmingham (1959)
5. Dz-Mou Jung, M. S. Krishnamoorthy, George Nagy, Andrew Shapira: N-Tuple Features for OCR Revisited. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(7): 734-745. (1996)
6. Hand, D.J., "Recent Advances in Error Rate Estimation," Pattern Recognition Letters, 4, 335 -346. (1986)
7. Jorgensen, T. M., S. S. Christensen, and C. Liisberg, Crossvalidation and information measures for RAM based neural networks. In D. Bisset (Ed.), Proc. of the Weightless Neural Networks Workshop (WNNW95), University of Kent at Canterbury, UK, pp. 87-92. (1995)
8. R. Rohwer and D. Cressy. Phoneme classification by boolean networks, Proceedings of the European Conference on Speech Communication and Technology, pages 557--560, (1989)
9. R J Mitchell, P R Minchinton : "Optimising memory usuage in n-tuple networks", Mathematics & Computers in Simulation, 40, pp: 549-563, (1996)
10. Wilkinson, R., Geist, J., Janet, S., Grother, P., Burges, C., Creecy, R., Hammond, B., Hull, J., Larsen, N., Vogl, T., and Wilson, C., The first census optical character recognition systems conference. Technical Report NISTIR 4912, National Institute of Standards and Technology, Gaithersburg, USA.(1992)

A Multi-expert Approach for Shot Classification in News Videos

M. De Santo[1], G. Percannella[1], C. Sansone[2], and M. Vento[1]

[1] Dipartimento di Ingegneria dell'Informazione e di Ingegneria Elettrica
Università di Salerno - Via P.te Don Melillo, 1 I-84084, Fisciano (SA), Italy
{desanto,pergen,mvento}@unisa.it
[2] Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II"
Via Claudio, 21 I-80125 Napoli (Italy)
carlosan@unina.it

Abstract. In this paper we propose a multi-expert approach for anchor shot detection in news videos. The proposed Multi-Expert System (MES) combines three algorithms selected among those presented in the literature that are model-free and do not require a specific training phase. In order to assess the performance of the proposed MES, we built up a large database, significantly wider than those ones typically used in the field. Reported experimental results show not only that proposed system performances are better than the ones of original experts but also the smart use made by the MES of strengths and weaknesses of each expert.

1 Introduction

Among all the different sources of video material nowadays available, news videos received great attention by the scientific community. This is mainly due to the fact that broadcasters are interested in building large digital databases of their resources, so to allow reuse, after a suitable indexing procedure, of the archived material for other TV programs. A first step towards an effective indexing is the segmentation of a news video into *stories*. At a first stage, it implies the partition of the video into sequences of frames, called *shots*, obtained by detecting transitions that are typically associated to camera changes. Once the shots have been individuated, they are classified on the basis of their content. Two different classes are typically considered, an *anchor shot* and a *news report shot* class. Successively, a news video can be segmented into stories; each story is obtained by linking each anchor shot together with all successive news report shots, until another anchor shot, or the end of the news video, occurs.

In the following, we mainly address the shot classification problem. Most of the authors in the current literature exploit the video source information by using a model matching strategy [1,2]. For each shot, a distinctive frame, called key-frame, is extracted. Then, the key-frame is matched against a set of pre-defined models in order to classify it. Obviously enough, these approaches are strongly dependent on the accuracy of the model of the specific video program. This is a severe limitation, since it is difficult to construct all the possible models for the different news videos and since the style of a particular news video program can change over the time.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 564–571, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Other authors use a face detection approach to identify anchor shots [3]. However, face detection in video is generally too time-consuming for practical applications. A different approach based on the frame statistics is presented in [4], where the authors use a Hidden Markov model (HMM) to classify frames. The HMM parameters are evaluated during a training phase by using the ground truth of a news video.

Finally, some authors [5,6,7] propose methods that are substantially unsupervised and do not require the explicit definition of a specific model of the anchor shot. In particular, in [5] a graph-theoretical cluster analysis method is employed to classify video shots. This approach fails in presence of news shots that are identical to each other in different stories of the same TV news, or when a particular anchor shot model appears only once in a whole TV news. In [6] a template-based method is proposed. However, the authors assume that different anchorperson models have the same background. This is not true for most news stations: because of different camera angles, different models can have different backgrounds. In [7] shot classification is also performed on the basis of the motion features of the anchor shot. In some TV news, however, zooming effects can be used also in an anchor shot; if motion is contained also in the background, the anchor shot can be missed.

All summarized, a definitive solution for the shot classification problem does not exist up till now. On the other side, efforts for increasing the performance of a given shot classification technique seem to be unjustified, since it is always possible to find a case in which the assumptions of that technique are not valid.

Starting from these considerations, our proposal is to exploit a Multi-Expert approach for the shot classification problem. The underlying idea of a Multi-Expert System (hereinafter MES) is to combine a set of rather simple classifiers (also called experts), in a system taking the classification decision on the basis of the classification results provided by each of the experts involved [8]. The rationale of this approach lies on the assumption that the performance obtained by suitably combining the results of a set of experts is better than that of any single expert.

Another important issue addressed in the paper regards the testing of the MES. We used a database that is twice the biggest database reported until now in the scientific literature [5]. In order to obtain more robust results from a statistical point of view, a k-fold cross validation was also performed. All the tests show that the proposed MES performs better than each of the considered experts.

The organization of the paper is as follows: in section 2 the proposed system architecture is presented. In section 3 the database used is described, while in section 4 the tests carried out in order to assess the performance of the proposed system are reported. Finally, some conclusions are drawn in section 5.

2 System Architecture

According to the rationale inspiring a multi-expert approach, a common strategy that is followed when building a MES is to consider experts whose features are as complementary as possible. However, among all the techniques described in the introduction, we choose to select only those not using a model-based approach neither requiring a specific training phase. This choice can be justified by considering that there are too many news videos in the world so that is practically infeasible to build a general model which is good for each of them. It could be reasonable, in a real

application scenario, to limit our attention on a single TV network. Even in that situation, it is frequent to see the model of the news video repeatedly change over the time. In this case a model-based approach could provide good results only for a small period of time, while a continuous re-modeling of the considered news video is required to guarantee acceptable performance for a longer period. For the same reason, also methods requiring a specific training phase are not very suitable for the application at hand, since it would be necessary to re-train the system each time the news video changes its style.

So, we considered only shot classification algorithms that are substantially unsupervised. Since the definition of a specific news video model is not required, the obtained results can be also considered quite independent of the specific style of the news program. In particular, on the basis of the best results available in the literature and of the complementariness of the features used, the experts used in our system implement the shot classification algorithms proposed by Bertini *et al*. in [7], Gao and Tang in [5] and Hanjalic *et al.* in [6]. Hereinafter, for the sake of simplicity, we will refer to these three experts with the terms BER, GAO and HAN, according to the first three letters of the first author's name. A brief description of their main features follows:

BER Expert - the shot classification is performed on the basis of a statistical approach and of the motion features of the anchor shots. The statistical approach is based on the consideration that anchor shots are repeated at variable length throughout the video. The obtained classification is successively refined by considering motion features: for the authors, in fact, it is reasonable to assume that in an anchor shot both the camera and the anchorperson are almost motionless.

GAO Expert - in this case, video shots are classified by using an algorithm based on graph-theoretical cluster analysis. It automatically groups similar key-frames into clusters, on the basis of their color histograms. The key-frames composing a cluster are classified as potential anchorperson frames if the cluster size is greater or equal to 2. Then, a spatial difference metric (SDM) is used to refine the shot classification. In some situations, in fact, the key-frames in a cluster may have similar histograms but different content. If a cluster has an average SDM value higher than a suitable threshold, it is removed from the anchorperson frame list.

HAN Expert - a template-based method is proposed in this case. It is based on the assumption that an anchorperson shot is the only type of video shot that has multiple matches of most of its visual content along the whole video. The template is found in a robust and unsupervised way and it does not depend on a particular threshold.

As regards the combining rule, the Behavior Knowledge Space (BKS) rule [9] was chosen, taking into account its performance, as reported in the literature [10], and the fact that it does not explicitly require the classifier independence assumption. This characteristic is important since even if the selected shot classification techniques adopt complementary features, they all work on the same input source, i.e. a video shot. Therefore, the complete independence of their results cannot be completely guaranteed.

The BKS rule is based on the definition of the so-called Behavior-Knowledge Space, a K-dimensional space where each dimension corresponds to the decision of an expert. Given a sample to be assigned to one of M possible classes, the ensemble of the experts can in theory provide M^K different decisions. Each one of these decisions constitutes one *unit* of the BKS. The BKS rule operates in two phases: a training one

and an operating one. In the training phase each BKS unit can record M different values e_i, one for each class. Given a suitably chosen training set, each sample x of this set is classified by all the experts and the unit that corresponds to the particular experts' decision (called *focal* unit) is activated. It records the actual class of x, say C_j, by adding one to the value of e_j. At the end of this phase, each unit can calculate the best representative class associated to it, defined as the class that exhibits the highest value of e_i. This class corresponds to the most likely class, given an experts' decision that activates that unit. In the operating mode, for each sample x to be classified, the K decisions of the experts are collected and the corresponding focal unit is selected. Then x is assigned to the best representative class associated to its focal unit.

The proposed system, a parallel MES, is sketched in Fig. 1: each one of the three experts receives in input the list of all the shot boundaries of a news video calculated by a shot change detection module together with the video itself and provides its own classification for each news video shot. Then, the combination module of the MES furnishes the final classification for each shot on the basis of the outputs of the three experts and of the BKS combining rule.

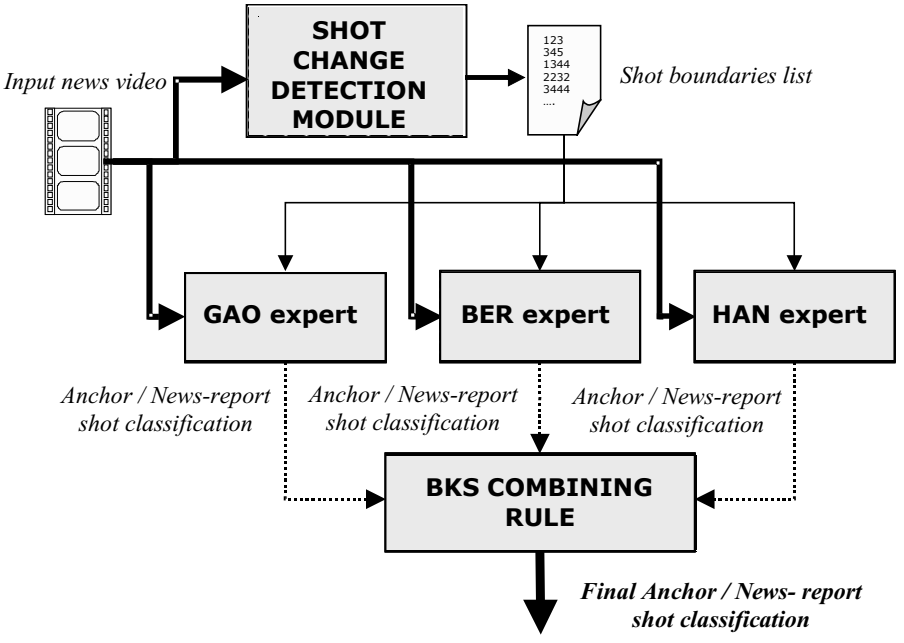


Fig. 1. A sketch of the proposed multi-expert architecture for shot classification.

3 The News Video Database

Some efforts have been spent in the recent past by other researchers in building video databases for benchmarking purposes; in particular in [11] a database was built in order to characterize the performance of shot change detection algorithms. This database, however, is not adequate for our aims, since it is made up not only of news

video but also of sport events and sitcom videos, and the duration of news videos is only 20 minutes. Therefore, we have built a new database. The acquisition was performed by means of a digital satellite decoder. Then, the videos were transferred on a PC: in this way, the broadcasting quality has been preserved. We encoded the videos in the MPEG-1 format using the TMPGEnc encoder (ver. 2.01). The video is coded at 2 Mbit/s with a 704x576 frame size. Finally, we built the ground truths of the shot boundaries and of the anchor/news report shots, respectively.

When building a database, it is important to reproduce as much as possible the variability of the phenomenon under study. In our case, the variability is due to both the different news video editions of a single TV-network and the news video editions of different TV-networks. However, while in the first case the different models are usually less than ten, in the latter the number of models is huge. As a consequence, we preferred to perform tests on a large database composed by TV-news captured from a single TV-network. Even if some archiving companies work with large quantities of videos from different sources, this approach fits the most realistic use case for the proposed system. A typical TV-network, in fact, should be interested to employ such a system for analyzing all the editions of its news videos. This is particularly true for those TV-networks not yet in possess of their own digital news video database.

The database used in this paper (about nine hours) is composed by 33 news videos from the main Italian public network (namely, RAI 1). Particular care was taken in order to include in the database all the video news editions from this TV-network. As it is evident from Table 1, the size of our database is very large, especially if compared to the databases used in the papers proposing the experts GAO, BER, and HAN.

Table 1. Composition of the databases used in the current paper and in [5] [6] and [7].

Paper	Total length (hh:mm:ss)	# of videos	# of Anchor/news report shots
This	08:52:34	33	444 / 5344
[5]	05:05:17	14	253 / 3654
[6]	00:37:00	2	22 / ---
[7]	02:41:00	12	66 / 665

4 Experimental Results

When dealing with unbalanced data sets, like in this case, where the anchor shot samples outnumber news report shot samples, the system performance is typically reported in terms of *Precision* and *Recall* [11,12].

Moreover, we will characterize the system performance by using also a unique figure of merits F [12] that combines *Precision* and *Recall* as reported in equation (1):

$$F = \frac{2 \cdot \mathrm{Pr}\,ecision \cdot \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{1}$$

In order to obtain a more realistic estimate of the performance of the proposed MES, a four-fold cross-validation was performed. Therefore, we divided the database in four subsets. Then, we performed four tests: in each one, two subsets of the

database were used as training set and one subset for testing. In each fold, the training set was used to calculate the Behavior-Knowledge Space matrix needed by the BKS rule, while the test set was used to test the three experts and the proposed MES. Hence, all the reported results were obtained as the average of the performance on the four test sets. A fair comparison of the performance of the proposed MES and of the three experts should be independent of the possible errors made by the employed shot change detection system. To this aim, the tests were carried out by using the shot boundaries ground truth. Nevertheless, the hypothesis of having no errors at the shot boundary detection stage is pretty realistic. In fact, in the recent years different solutions that assure very high recognition rates have been proposed in the literature. As an example, the MES system proposed in [10] is able to detect almost 100% of shot boundaries using the MAX rule, even if it is paid with a higher false detections percentage. However, in a system devoted to shot classification, false shot boundary detections cost less than missed detections. The effect of false alarms is the oversegmentation of the video that may be eliminated after the anchor shot classification with simple heuristics.

As a first step of our tests, we compared the performance obtained by the three experts. Table 2 reports the global performance of each expert: *%cd* and *%f* represents the percentages of the anchor shot correct detections and false alarms respectively.

Table 2. The performance of the three considered experts averaged on the four test sets.

Expert	*%cd*	*%f*	*Recall*	*Precision*	*F*
GAO	92.86	1.59	0.929	0.842	0.881
BER	81.61	0.10	0.816	0.987	0.892
HAN	62.29	2.40	0.623	0.692	0.655

A first and unexpected result is the discrepancy among the performance reported in Tab. 2 and the results reported in the original papers. This is particularly true for the HAN expert. Moreover, the GAO expert performs better in terms of *Recall*, while in [5] its behavior was the opposite one. Analogously, the BER expert exhibits on our database a *Precision* value higher than the *Recall* one, differently from the behavior described in [7]. Since the anchor shots/news-report shots ratio of our database is quite similar to those of the databases used in the original papers (excluding HAN, where the ratio is not specified), such discordance is mainly due to the different size of the database used here for the tests. This clearly confirms the importance of testing algorithms on a large and significant database.

The results reported in Tab. 2 point out that, even if the BER expert outperforms the remaining two in terms of *Precision* and *F*, GAO exhibits the best value of the *Recall*. On the other hand, the HAN expert performs sensibly worse than BER and GAO, but the difference between its values of *Recall* and *Precision* is less than those obtained by the other two experts. These considerations confirm the good level of complementariness exhibited by the three experts' behavior. This was one of the main assumptions we made and it is mandatory for a successful implementation of a multi-expert system.

When dealing with the evaluation of a MES, it is useful to consider the performance of the so-called "*oracle*". The oracle is the theoretic MES that correctly classifies a shot as an anchor shot or as a news-report shot if at least one of the

employed experts is able to provide the *correct* classification. It is evident that for a defined set of experts, the performance of the oracle is the upper bound of all the MES's obtainable from the same set of experts by using any combining rule.

In table 3, the performance of the *oracle*, of the proposed MES and of the best expert in terms of *F* (BER) is reported. From the table it is possible to notice how the proposed MES outperforms the best expert in terms of the both *Precision* and *Recall*, so giving rise to a better value of *F*.

Table 3. The performance of the oracle, of the proposed MES and of the best expert (BER).

	%cd	*%f*	*Recall*	*Precision*	*F*
Oracle	95.28	0.00	0.953	1.000	0.976
Proposed MES	86.94	0.00	0.869	1.000	0.930
BER	81.61	0.10	0.816	0.987	0.892

Another very interesting way to examine the obtained results is showed by considering the data reported in table 4. Here, the performance of the proposed architecture is expressed in terms of the relative improvement with respect to the performance of the *oracle*. Such improvement has been calculated as:

$$RI_{par} = \frac{par_{proposedSystem} - par_{Expert}}{par_{oracle} - par_{Expert}} \tag{2}$$

being $par \in \{\%cd, \%f, Recall, Precision, F\}$ one of the defined parameters for evaluating the performance.

The results reported in table 4 show that the proposed architecture is able to improve of about the 45% of *F* with respect to the BER expert. This represents a valuable result if we consider that this expert is already characterized by a good performance.

Table 4. The relative improvement introduced by the proposed MES with respect to the performance that can be obtained by using the *oracle*.

Expert	$RI_{\%cd}$	$RI_{\%f}$	RI_{Recall}	$RI_{Precision}$	RI_{F}
GAO	-245.3%	100%	-245.3%	100%	51.7%
BER	39.0%	100%	39.0%	100%	45.4%
HAN	74.7%	100%	74.7%	100%	85.7%

5 Conclusions

Automatic shot classification in news videos is still an open problem. In spite of the efforts of many researchers involved in this field, yet there are not techniques that provide fully satisfactory performance.

In this context, we employed a Multi-Expert approach to combine some of the best algorithms available in the scientific literature. Therefore, we implemented three

unsupervised algorithms for shot classification and combined them in a parallel Multi-Expert System (MES) using the Behavior-Knowledge Space rule. Experimental results on a large database demonstrated that the proposed MES performs better than each single classifier. In particular, it gave rise to no false alarms on the anchorperson shot detection on a set of over 4,000 shots; being able in the same time to keep the number of missed anchorperson shots low.

References

1. S.W. Smoliar, H.J. Zhang, S.Y. Tao, Y. Gong, "Automatic parsing and indexing of news video", Multimedia Systems, vol 2, no. 6, pp. 256-265, 1995.
2. B. Furht, S.W. Smoliar, H. Zhang, Video and Image Processing in Multimedia Systems, Kluwer Publishers, Boston (MA), 1996.
3. Y. Avrithis, N. Tsapatsoulis, and S. Kollias, "Broadcast news parsing using visual cues: A robust face detection approach", Proc. IEEE Int. Conf. on Multimedia and Expo, vol. 3, pp. 1469–1472, 2000.
4. S. Eickeler, S. Muller, "Content-based video indexing of TV broadcast news using Hidden Markov Models", Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing, pp. 2997-3000, 1999.
5. X. Gao, X. Tang, "Unsupervised Video-Shot Segmentation and Model-Free Anchorperson Detection for News Video Story Parsing", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 9, pp. 765-776, 2002.
6. A. Hanjalic, R. L. Lagendijk, J. Biemond, "Semi-Automatic News Analysis, Indexing, and Classification System Based on Topics Preselection", Proc. of SPIE: Electronic Imaging: Storage and Retrieval of Image and Video Databases, San Jose (CA), 1999.
7. M. Bertini, A. Del Bimbo, P. Pala, "Content-based indexing and retrieval of TV News", Pattern Recognition Letters, vol. 22, pp. 503-516, 2001.
8. J. Kittler, M. Hatef, R.P. W. Duin and J. Matas, "On Combining Classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), (1998), 226-239.
9. Y. S. Huang and C. Y. Suen, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, no. 1, pp. 90-94, 1995.
10. M. De Santo, G. Percannella, C. Sansone, M. Vento, "A Multi-Expert System for Shot Change Detection in MPEG Movies", International Journal of Pattern Recognition and Artificial Intelligence, 2004 (in press).
11. U. Gargi, R. Kasturi, S.H. Strayer, "Performance Characterization of Video-Shot-Change Detection Methods", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, pp. 1-13, 2000.
12. L. Chaisorn, T.-S. Chua, C.-H. Lee, "A Multi-Modal Approach to Story Segmentation for News Video", World Wide Web, vol. 6, pp. 187–208, 2003.

Motion-Compensated Wavelet Video Denoising

Fu Jin, Paul Fieguth, and Lowell Winger

Dept. of Systems Design Engineering, Univ. of Waterloo
Waterloo, Ontario, Canada N2L 3G1
{fjin, pfieguth,lwinger}@uwaterloo.ca

Abstract. We study a low complexity, motion-compensated recursive video denoising scheme in the wavelet domain. To preserve *weak* spatial textures, both spatial and temporal filtering are carried out in the wavelet domain. We justify the proposed approach by evaluating the relative increase of mean squared error (MSE) with respect to the optimal estimator. Various multiresolution motion estimation schemes are studied to allow us to exploit the high temporal correlation present in most video. Kalman filtering is then applied to the wavelet coefficients along motion trajectories to efficiently suppress noise.

We show experimentally that our wavelet-based recursive denoising compares favorably with other wavelet-based denoising approaches. Specifically, we can preserve both strong and weak spatial details while removing noise.

Keywords: Video denoising, motion compensation, wavelet transform, Kalman filtering.

1 Introduction

With the maturity of digital video capturing devices and broadband transmission networks, numerous applications have been emerging. These include teleconferencing, remote surveillance, multimedia services and digital television, to name a few. However, the video signal is almost always corrupted by noise from the capturing devices or during transmission due to random thermal or other electronic noises. Usually, noise reduction can considerably improve visual quality and benefit the subsequent processing tasks, such as video compression.

There are many existing video denoising approaches in the spatial domain [1,2,4,9], which can roughly be divided into two classes:

Temporal-only: A temporal-only approach utilizes only the temporal correlations, neglecting spatial information. Since video signals are strongly correlated along motion trajectories, motion estimation/compensation is normally employed. In those cases where motion estimation is not accurate, motion detection may be used to avoid blurring. These techniques can preserve spatial details well, but the resulting images usually still contain removable noise since spatial correlations are neglected.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 572–579, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Spatio-temporal: More sophisticated methods exploit both spatial and temporal correlations, such as simple adaptive weighted local averaging [3], *3-D* order-statistic algorithms [2], *3-D* Kalman filtering [4] and *3-D* Markov random models [5].

Although there have been many papers addressing the application of wavelet transforms to image denoising, comparatively few have addressed wavelet-based video denoising. Roosmalen *et. al.* [6] proposed video denoising by thresholding the coefficients of a specific *3-D* wavelet representation and Selesnick *et. al.* [7] found an *efficient 3-D* orientation-selective wavelet transform, *3-D complex* wavelet transforms, which avoided the time-consuming motion estimation process. The main drawbacks of the *3-D* wavelet transforms include a long time latency and the inability to adapt to fast motions.

In most video processing applications a long latency is unacceptable, so recursive approaches are widely employed. Pizurica *et. al.* [8] proposed sequential *2-D* spatial and *1-D* temporal denoising, in which they first do sophisticated wavelet-based image denoising for each frame and then recursive *temporal* averaging. However, *2-D* spatial filtering tends to introduce artifacts and to remove *weak* details along with the noise.

In this paper, we propose wavelet-domain recursive video denoising. We use a *2-D* spatial wavelet, and filter recursively to preserve low latency. But unlike [8], our temporal filtering works on the wavelet coefficients themselves (instead of on the spatial pixels). In such a way we minimize spatial blurring to preserve the *weak* spatial details in still areas. The key to processing directly in the wavelet domain is an efficient shift-invariant transform, allowing spatial motion to be meaningfully reflected in the wavelet coefficients. We study several schemes for robust wavelet-domain motion estimation and then proceed to motion-compensated temporal filtering in the *wavelet domain* using an adaptive Kalman filter.

2 Wavelet-Based Video Denoising

In standard wavelet-based image denoising [9] the *2-D* wavelet transform is used to get a compact representation. Thus it would seem natural to select *3-D* wavelets for video denoising [6,7]. However, there are compelling reasons to choose a *2-D* spatial wavelet transform with recursive temporal filtering for video denoising:

1. There is a clear asymmetry between the spatial and temporal axes, in terms of correlation and resolution. A recursive approach is naturally suited to this asymmetry, whereas a *3-D* wavelet is not.
2. Recursive filtering can significantly reduce time delay and memory requirements.
3. For autoregressive models the optimal estimator can be achieved recursively.
4. Motion information can be efficiently exploited with recursive filtering.

A. Problem Formulation

The video denoising problem can be modeled as follows: given video measurements y with spatial indices i, j and temporal index k

$$y(i,j,k) = x(i,j,k) + v(i,j,k), \;\; i,j = 1,2,...,N, \;\; k = 1,2,...,M \tag{1}$$

corrupted by i.i.d Gaussian noise v, we need to estimate the true image sequence x. Define $\mathbf{x}(k)$, $\mathbf{y}(k)$ and $\mathbf{v}(k)$ to be the column-stacked images at time k, then (1) becomes

$$\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k), \;\; k = 1,2,...,M \tag{2}$$

We propose to do denoising in the wavelet domain; let H be the *2-D* wavelet transform matrix, then (2) is transformed as

$$H\mathbf{y}(k) = H\mathbf{x}(k) + H\mathbf{v}(k) \tag{3}$$

Denoting wavelet coefficient vector as $\mathbf{y}_H(k)$, (3) is rewritten as

$$\mathbf{y}_H(k) = \mathbf{x}_H(k) + \mathbf{v}_H(k) \tag{4}$$

Since we seek a recursive temporal filter we assert an autoregressive form for the signal model

$$\mathbf{x}(k+1) = A\mathbf{x}(k) + B\mathbf{w}(k+1) \tag{5}$$

thus

$$\mathbf{x}_H(k+1) = A_H \cdot \mathbf{x}_H(k) + B_H \cdot \mathbf{w}_H(k+1) \tag{6}$$

where $A_H \equiv HAH^{-1}$, $B_H \equiv HBH^{-1}$. It should be noted the wavelet domain state model still has an autoregressive form. Therefore, *optimal* filtering can be achieved in a recursive way in the wavelet domain.

B. An Example: Recursive Image Filtering in the Spatial and Wavelet Domains

As a quick proof of principle, we can denoise *2-D* images using a recursive *1-D* wavelet procedure, analogous to denoising *3-D* video using *2-D* wavelets. We do not propose this as a superior approach to image denoising, rather as a measure of promise in the video case. We use an autoregressive image model and do *1-D* wavelet transform on each column, followed by recursive filtering column by column. We assess estimator performance in the sense of relative increase of MSE:

$$\delta_{MSE} = \frac{MSE - MSE_{optimal}}{MSE_{optimal}} \tag{7}$$

where $MSE_{optimal}$ is the MSE of the optimal Kalman filter.

Table 1. Percentage increase δ_{MSE} in estimation error relative to the optimal estimator. Estimating wavelet coefficients independently introduces only slight error.

SNR(dB)	10	0	−10
δ_{MSE} (spatial)	1.88%	14.10%	18.65%
δ_{MSE} (wavelet)	0.1%	1.09%	6.51%

Table 2. Comparison of PSNR (in dB) of the proposed method and *3-D* complex DWT for sequence *Paris*.

PSNR (Original)	28.2	22.1	18.6	16.1
PSNR (Proposed method)	37	30	27	26
PSNR (*3-D* complex DWT [7])	35	28	24	23

We use a common image model

$$x(i,j) = \rho_v x(i-1,j) + \rho_h x(i,j-1) - \rho_v \rho_h x(i-1,j-1) + w(i,j), \quad \rho_h = \rho_v = 0.95 \tag{8}$$

which is a causal MRF model and can be converted to a vector autoregressive model [10].

The optimal recursive filtering requires the joint processing of entire columns. As this would be completely impractical in the video case, for reasons of computational complexity we recursively filter the wavelet coefficients *independently*, ignoring inter-coefficient relationships. As shown in Table 1, scalar processing in the wavelet domain leads to only very moderate increases in MSE relative to the optimum, whereas this is not at all the case in the spatial domain. It should be noted that the wavelet-based scalar processor is comparable to the optimal filter when $SNR > 0dB$, a condition satisfied in many practical applications.

3 The Denoising System

The success of *1-D* wavelet image denoising motivates the extension to the video case. There are three crucial aspects: (1) the choice of *2-D* wavelet transform, (2) wavelet-domain motion estimation, and (3) the recursive filtering applied to the motion-compensated wavelet coefficients. These steps are detailed below.

2-D **wavelet transform**: We apply a *2-D* wavelet transform to each frame, rather than a *3-D* transform for the whole image sequence, and the coefficients are then filtered recursively.

A huge number of wavelet transforms have been developed (e.g., orthogonal / non-orthogonal, real-valued/complex-valued, decimated / redundant). For image denoising problems, three criteria are desired:

1. Shift invariance: to suppress frequency aliasing and related artifacts;
2. Direction selectivity: of importance when image has dominant oriented features;
3. Low complexity.

The *2-D* dual-tree complex wavelet proposed by Kingbury [7] satisfies these requirements very well. Unfortunately it is not convenient for motion estimation since the motion information is related to the coefficient phase, which is a non-linear function of translation. Alternatively, specially designed *2-D* wavelet transforms (e.g., curvelet, contourlet) are sensitive to feature directions but are too complex for computation. In this paper, we choose to use an over-complete wavelet representation proposed by Mallat *et. al.* [11]. Although this wavelet representation does not have very good directional selectivity (e.g., it mixes 45-degree and -45-degree features), several researchers have used it for natural image denoising with impressive results (e.g.[8]).

Multiresolution Motion Estimation (MRME): We utilize the well-used block matching (BM) technique for motion estimation. Compared with other motion estimation approaches, such as optical flow and pixel-recursive methods, block matching is straightforward to compute and less sensitive to noise.

Single-resolution block matching approaches have been studied extensively [1] and are successfully used in modern video compression standards. Multi-resolution block matching (MRBM) was first proposed by Zhang *et. al.* [1] for wavelet-based video coding, and recent developments can be found in [12]. The basic idea of standard MRBM is to start block matching at the coarsest level, using this estimate as a prediction for the next finer scale. Oddly, a critically decimated wavelet was used [12], which implies that the motion is not a constant function of scale. A much more sensible choice of wavelet, used in this paper, is the overcomplete transform which is shift-invariant, leading to consistent motion as a function of scale, except in the vicinity of motion boundaries. Clearly, this high inter-scale relationship of motion should be exploited to improve accuracy. We developed a sequence of four approaches of increasing accuracy:

1. The standard MRME scheme [12].

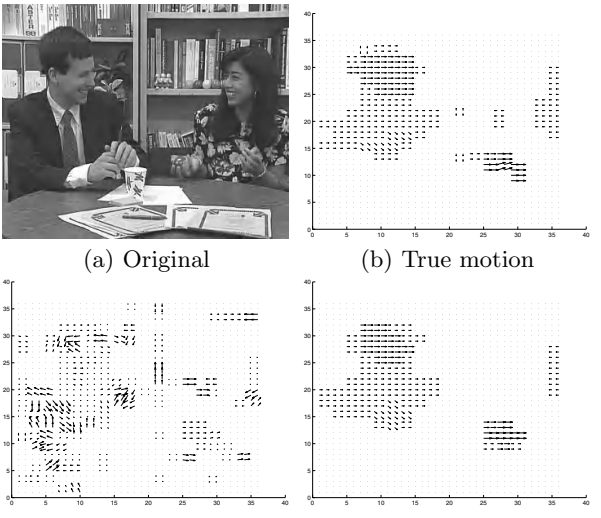
2. Block matching separately on each level, combined by median filtering.

3. Joint block matching simultaneously at all levels:
Let $\epsilon^l(\mathbf{i}, \mathbf{v})$ denote the displaced frame difference (DFD) at position $\mathbf{i}$ of level l with displacement $\mathbf{v}$). Then the total DFD over all levels is defined as

$$\epsilon(\mathbf{i}, \mathbf{v}) = \sum_{l=1}^{J} \epsilon^l(\mathbf{i}, \mathbf{v}) \tag{9}$$

and the displacement field $\mathbf{v}(\mathbf{i}) = [v_x(\mathbf{i}), v_y(\mathbf{i})]$ is found by minimizing $\epsilon(\mathbf{i}, \mathbf{v})$.

4. Block matching with smoothness constraint:
The above schemes did not assert any spatial smoothness or correlation in the motion vectors, which we expect in real-world sequences. This is of considerable importance when the additive noise levels are large, leading to irregular estimated motion vectors. Therefore, we introduce an additional



(a) Original (b) True motion

(c) Standard (d) Smoothness-constrained

Fig. 1. Compare MRBM results of the standard approach (c) and the proposed one (d). The standard does not exploit the smoothness property (*a prior* knowledge) of motion fields. Therefore, it performs poorly in the presence of noise. For sharp comparison, our proposed approach give much better results.

smoothness constraint and perform BM by solving the optimization problem

$$arg\min_{\mathbf{v}} \left\{ \sum_{\mathbf{i}} \left[\gamma \cdot \epsilon(\mathbf{i}, \mathbf{v}) + (|v_x(\mathbf{i}) - \frac{1}{M} \sum_{m \in N_b(\mathbf{i})} v_x(\mathbf{i}+\mathbf{m})| \right. \right.$$
$$\left. \left. + |v_y(\mathbf{i}) - \frac{1}{M} \sum_{n \in N_b(\mathbf{i})} v_y(\mathbf{i}+\mathbf{n})|) \right] \right\} \tag{10}$$

where N_b is the neighborhood set of the element $\mathbf{i}$ and M is the number of elements in N_b. $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N]^T$ is the motion field to be estimated. γ controls the tradeoff between $\epsilon(\mathbf{i}, \mathbf{v})$ (see (9)) and smoothness.

Experimentally, we have found Approach 4 to be the most robust to noise and yield reasonable motion estimates. In the simulation results given below we will use this latter approach for motion estimation.

Wavelet coefficient filtering: The key to our approach is to support both spatial and temporal filtering, as appropriate. Specifically, when the motion information is unambiguous, i.e.,

$$|x_H(m,n,k) - \rho(x_H(m,n,k-1))| < 2\sqrt{2}\sigma_{v_H}, \tag{11}$$

where $\rho(\cdot)$ is the motion compensation function, we restrict the filtering to be purely temporal to avoid any spatial blurring. However, when motion estimates are poor, i.e.,

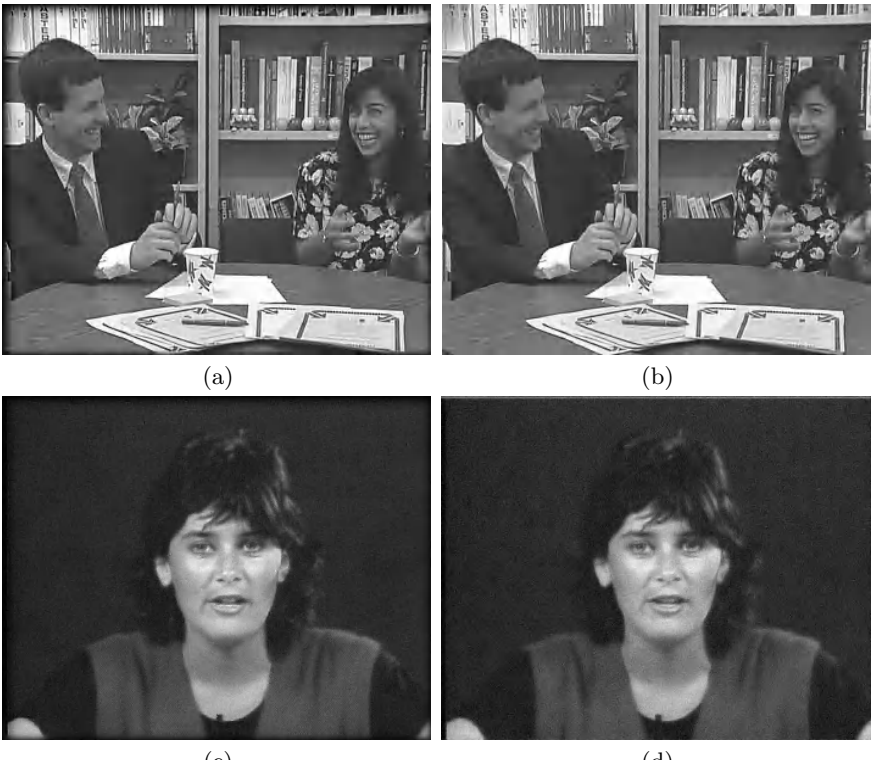


Fig. 2. Denoised images : (a) & (c) the proposed method (b) & (d) *3-D* complex DWT. $\sigma_{v_H} = 10$ for *Paris*, $\sigma_{v_H} = 20$ for *MissAmerica*.

$$|x_H(m,n,k) - \rho(x_H(m,n,k-1))| \geq 2\sqrt{2}\sigma_{v_H}, \tag{12}$$

we apply wavelet shrinkage to exploit spatial correlation. Although this introduces some blurring, wavelet shrinkage is state-of-the-art in static image denoising.

4 Experimental Results

We have tested all four of the MRBM approaches listed in the previous section. Due to space limit we show only the results of standard MRBM (hierarchical prediction-updating approach) and our proposed smoothness-constrained motion estimation in Fig. 1. It is clear that MRBM becomes much less sensitive to noise if the smoothness constraint is applied, so we propose to use this method for the subsequent video denoising tests.

To show denoising performance of the proposed method we have experimented on several standard video sequences (e.g., *Paris, Salesman, MissAmer-*

ica). We compare the denoising results of the proposed method with those from the *3-D* complex DWT [7] (http://taco.poly.edu/WaveletSoftware/), in terms of Peak SNR (PSNR) in Table 2, and our method gives a consistent improvement in performance. Several denoised images are shown in Fig. 2, where a careful examination of spatial textures (in the face and table top) shows the supperiority of our proposed approach. In particular, there are fewer aliasing artifacts, and stationary objects (such as the bookshelf) are denoised much more crispy.

References

1. Brailain J., Kleihorst R., Efstratiadis S., Katsaggelos A. and Lagendijk A.: Noise reduction filters for dynamic image sequences: A review. Proc. of IEEE **83**(9) (1995) 1272–1292
2. Arce G: Multistage order statistic filters for image sequence processing. IEEE Trans. Acoust., Speech, Signal Proc. **39**(5) (1991) 1147–1163
3. Ozkan M. and Sezan I. and Tekalp A.: Adaptive motion-compensated filtering. IEEE Trans. on CSVT.,**3**(8) (1993) 277–290
4. Kim J. and Woods J.: Spatio-temporal adaptive *3-D* Kalman filter for video. IEEE Trans. Imag Proc.,**6**(3) (1997) 414–424
5. Brailean J. and Katsaggelos A.: Simultaneous recursive displacement estimation and restoration of noisy-blurred image sequences. IEEE Trans. Imag Proc.,**4**(9) (1995) 1236–1251
6. Roosmalen P., Westen S., Lagendijk R. and Biemond J.: Noise reduction for image sequences using an oriented pyramid thresholding technique. Proc. Intern. Conf. Image Proc., (1996) 375–378
7. Selesnick I and Li K.: Video denoising using *2-D* and *3-D* dual-tree complex wavelet transforms. Proc. SPIE Wavelet Applications in Signal and Image Processing X,2003,700–707
8. Pizurica A., Zlokolica V. and Philips W.: Combined wavelet domain and temporal video denoising. Proc. Advanced Video and Signal Based Surveillance,2003,334–341
9. Chang S., Bin Y. and Vetterli M.: Spatially adaptive wavelet thresholding with context modeling for image denoising. IEEE Trans. Image Proc., **9**(9), (2000) 1522–1531
10. Rosenfeld A. and Kak A.: Digital picture processing. Acadamic Press (1982)
11. Mallat S. and Zhong S.: Characterization of signals from multiscale edges. IEEE Trans. PAMI, **14**(9),(1992) 710–732
12. Zan J, Ahmad M.O. and Swamy S.: New techniques for multi-resolution motion estimation. IEEE Trans. CSVT, **12**(9),(2002) 793–802
13. Lee J.:Digital image enhancement and noise filtering by use of local statistics. IEEE Trans. PAMI, **2**(6),(1980) 165–168

Alpha-Stable Noise Reduction in Video Sequences

Mohammed El Hassouni and Hocine Cherifi

University of Bourgogne
LIRSA, Faculty of science Mirande
BP 47870, 21078 Dijon cedex, France
{Mohamed.Elhassouni,Hocine.Cherifi}@u-bourgogne.fr

Abstract. In this paper, a nonlinear motion-compensated filter is described for removing α-stable noise from video sequences. To address this problem, we propose a spatio-temporal adaptive weighted myriad and l_p-norm filters. Prior to filtering, motion compensation is performed by using a block-matching algorithm with p-norm matching error function. Then, we apply the proposed filter to the reconstructed frames. The effective performance of the proposed scheme is illustrated through computer simulations involving the filtering of video sequences.

1 Introduction

Video sequences occur naturally in digital video broadcasting, teleconference, video phone, medical imaging, surveillance and defense. The demand for enhancement of video sequences has increased not only to improve visual quality but also to aid subsequent processing [1]. Second-order Statistics based filtering techniques have been used in many image processing applications. These classical techniques assume that images are second order processes and the noise is additive white Gaussian but the observation do not support this assumption [2]. There are applications, for which the noise is characterized by other distributions, such as the uniform, Laplacian, or a combination of them. In this case, higher-order statistics (HOS) exhibit improved performance compared to the second-order based estimation methods [3,4]. However, in many real-world problems the noise encountered is more impulsive in nature than that presented by a Gaussian distribution. Important non-Gaussian impulsive noise, found in fixed and mobile communications for example, can be efficiently modelled by infinite variance processes for which the theory of SOS and HOS is not useful [5].

There exists a class of distributions, called α-stable distributions that can be used to model these types of noise. A variety of fractional lower order statistics FLOS-based algorithms have been proposed to filter α-stable random processes [6]. This property has been used extensively in one-dimensional (1-D) signal processing, e.g. in removing some undesired effects in communication channels or synthesizing a deconvolution filter [7,8]. Developing this techniques for two-dimensional (2-D) signal processing (i.e. image processing) has been proposed [9, 10].

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 580–587, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper, the emphasis is on the development of a spatio-temporal filters for video sequences corrupted with α-stable noise. In general, spatio-filter performance is will be improved with accurate motion estimation and compensation [11,12]. Most motion estimators used in these noise filtering schemes use block matching, dense optic flow or parametric motion models. Block-matching gives good approximations of apparent motion for a considerably lower computation time. In [13], the displaced frame difference (DFD) of images sequences has been modelled as an α-stable distribution. In this paper, motion compensation along the temporal direction is performed by the block matching method. Motion field of the sequences is determined by a p-norm cost function in order to be robust to α-stable noise effects. After motion compensation, the 3-D myriad filter is applied to a motion compensated frames. Weighted myriad filters (WMyF) have been used with considerable success in robust communications and image processing [14]. They have been derived based on the Cauchy distribution, which is a special case of α-stable distributions. Weighted myriad filters (WMyF) are inherently more powerful than weighted median filters (WMF), which along with other order statistics filters, have been used widely in image processing, due to their edge preservation and outlier rejection properties. Then, 3-D l_p-norm filter is applied to a motion compensated frames. The adaptive algorithm is based on fractional lower order statistics (FLOS). The performance of the algorithm is compared to those of the normalized least mean square-type algorithm which is developed under the Gaussian assumption.

The paper is organized as follows. In Section 2, we present the problem formulation with some properties of the α-stable processes. In Section 3, we present the used motion estimation technique. Section 4 is then devoted to the presentation of spatio-temporal filters. Finally, experimental results and some concluding remarks are given in Section 5 and Section 6, respectively.

2 Problem Formulation

Let us consider that an observed n^{th} frame $g_n(i,j)$ can be expressed as a sum of noise-free n^{th} frame $f_n(i,j)$ plus additive noise, i.e.,

$$g_n(i,j) = f_n(i,j) + \eta_n(i,j) \tag{1}$$

where (i,j) denotes the pixel coordinates. The random noise $\eta_n(i,j)$ is modelled as a symmetric α-stable process ($S\alpha S$). The stable distribution law is a direct generalization of the Gaussian distribution and in fact includes the Gaussian as a limiting case. The main difference between the non-Gaussian stable distribution and the Gaussian distribution is that the tails of the stable density are heavier than those of the Gaussian density. The characteristic of the stable distribution is one of the main reasons why the stable distribution is suitable for modelling signals and noise of impulsive nature (see figure 1). The α-stable distributions do not have close form probability density function except the cases $\alpha = 1$ (Cauchy distribution) and $\alpha = 2$ (Gaussian distribution).

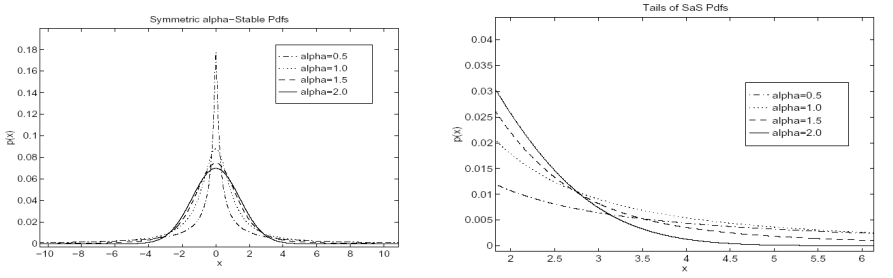


Fig. 1. Various α-stable pdfs with varying characteristic exponent (alpha) values. (a) Whole pdfs, (b) detail from the tails. Distribution gets more impulsive (heavy-tailed) α as a decreases.

The $S\alpha S$ probability density function (pdf) is defined by means of its characteristic function

$$\phi(\omega) = \exp(\delta i\omega - \gamma|\omega|^{\alpha}) \tag{2}$$

where

- *i*) $\alpha(0 < \alpha \leq 2)$ is the characteristic exponent, controlling the heaviness of the pdf tails,
- *ii*) $\gamma(\gamma > 0)$ is the dispersion, which plays an analogous role to the variance, and
- *iii*) δ is the location parameter, the symmetry axis of the pdf.

Due to the heavy tails, stable distributions do not have finite second- or higher-order moments, except the limiting case of $\alpha = 2$. More precisely, for X, an α-stable random variable with $(0 < \alpha < 2)$

$$E[|X|^p] = \infty \qquad \text{if} \qquad p \geq \alpha \tag{3}$$

However, for $0 < p \leq \alpha$, the fractional lower order moment (FLOM) is finite, i.e.,

$$E[|X|^p] < \infty \qquad \text{if} \qquad 0 \leq p < \alpha \tag{4}$$

If $\alpha = 2$, then

$$E[|X|^p] < \infty \qquad \text{for} \qquad p \geq \alpha \tag{5}$$

The fractional pth-order moment of an $S\alpha S$ random variable with zero location parameter, $\delta = 0$, is given by

$$E[|X|^p] = C(p,\alpha)\gamma^{p/\alpha}, \qquad \text{for} \quad 0 < p < \alpha \tag{6}$$

where

$$C(p,\alpha) = \frac{2^{p+1}\Gamma(\frac{p+1}{2})\Gamma(-p/\alpha)}{\alpha\pi\Gamma(-p/2)} \tag{7}$$

$\Gamma(.)$ denotes the gamma function.

3 Motion Estimation

In a typical block-matching algorithm (BMA), the current frame of a video sequence is divided into square blocks of pixels of size $N \times N$. For each reference block in the current frame, BMA searches for the best matched block within a search window of size $(2W + N) \times (2W + N)$ in the previous frame, where W stands for the maximum allowed displacement. A matching error function $D_p(i, j)$ is defined over all the positions to be searched, i.e.,

$$D_p = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} |g_n(i + x, j + y) - g_{n-1}(i + x + k, j + y + l)|^p \tag{8}$$

where $-W \leq k, l \leq W$, $g_n(i, j)$ is the reference block of its upper left pixel at the coordinate (i, j) in the current frame $g_{n-1}(i + k, j + l)$ is a coordinate block of its upper left pixel at the coordinate $(i + k, j + l)$ in the previous frame. This computation can be carried out efficiently for finite-order L_p norms. The traditional matching criteria used in the L_1 (sum of absolute difference (SAD)), and L_2 (sum of squared differences (SSD)). Unfortunately, in the presence of α-stable noise, L_1 and L_2 are not finite. However, for $0 < p \leq \alpha$, the fractional lower order moment (FLOM) is finite. The common way for determining p value of the algorithm is $p = \alpha - \epsilon$, where $\epsilon << \alpha$ as proposed in [15].

4 Spatio-temporal FLOS-Based Filters

Consider a set of observations $\{g_r\}_{r=1}^{M}$ and the set of filter weights $\{w_r\}_{r=1}^{M}$. Where M is the size of the spatio-temporal filter support and r is a scalar running index. We define the observation vector as

$$\mathbf{g} = [g_1, g_2, ..., g_M]^T \tag{9}$$

and the weight vector as

$$\mathbf{w} = [w_1, w_2, ..., w_M]^T \tag{10}$$

Let's define s as a scalar running index which can be used instead of the pixel coordinates. Henceforth, a 1-D notation will be adopted for simplicity.

4.1 Adaptive Weighted Myriad Filter

Given $K > 0$, the weighted myriad filter (WMyF) output is given by

$$\begin{aligned} \hat{f}_n(s) &= myriad(K; w_1 \circ g_1, w_2 \circ g_2, ..., w_M \circ g_M) \\ &= \arg\min G_K(\beta, \mathbf{w}, \mathbf{g}) \end{aligned} \tag{11}$$

where

$$G_K(\beta, \mathbf{w}, \mathbf{g}) = \prod_{r=1}^{M} [K^2 + w_r(g_r - \beta)^2] \tag{12}$$

will be called the myriad objective function since this function is minimized by the weighted myriad. $w_r \circ g_r$ denotes the weighting operation.

Adopting the Mean Absolute Error (MAE) criterion, we obtain the cost function

$$J(s) = E\{|e_n(s)|\} = E\{|\hat{f}_n(s) - f_n(s)|\} \tag{13}$$

where The filtering error is defined as

$$e_n(s) = \hat{f}_n(s) - f_n(s) \tag{14}$$

Necessary conditions for optimality by equating the gradient of the cost function with respect to the weights to zero has been proved in [14]. To update the filter weights, we employ gradient-based methods. We derive an adaptive algorithm based on the steepest descent method to design the filter weights. Incorporating the constraint of non-negative weights into the steepest descent method, we obtain the following gradient descent algorithm:

$$w_r(s+1) = P\left[w_r(s) - \mu \frac{\partial J}{\partial w_r}(s)\right] \tag{15}$$

where $w_r(s)$ denotes the r^{th} weight at the s^{th} iteration, μ denotes the step-size of the algorithm and $P[.]$ projects the updated weight onto the constraint space of the weights. In this case we have

$$P[u] = \begin{cases} u, u \geq 0 \\ 0, u < 0 \end{cases} \tag{16}$$

In practice, we set $P[u]$ to a small positive value ϵ if $u < 0$. It is difficult to estimate the gradient since this requires knowledge of the statistics of the desired image f and the observation vector $\mathbf{g}$. In order to simplify computation and also to be able to operate in non-stationary environments, we use instantaneous estimates for the gradient [14].

4.2 Adaptive Least l_p-Norm Filter

The intensity estimate at image location s is given by:

$$\hat{f}(s) = w^T(s)g(s) \tag{17}$$

where $g(s)$ is the observations vector and $a(s)$ is the filter coefficient vector that minimizes criterion between the estimated and the desired images.

Let $\varepsilon(s)$ be the estimation error at pixel s, i.e., $\varepsilon(s) = \hat{f}(s) - f(s)$. The LMS algorithm has sever convergence problems for signals with more probability mass in the tails, than the Gaussian distribution. Recently filter theory for $S\alpha S$ signals has been developed, and the least Mean p-norm (LMP) algorithm has been proposed. The objective is to minimize the dispersion of the error (p-norm cost function) which is defined by :

$$J(s) = E\left\{|\varepsilon(s)|^p\right\} \tag{18}$$

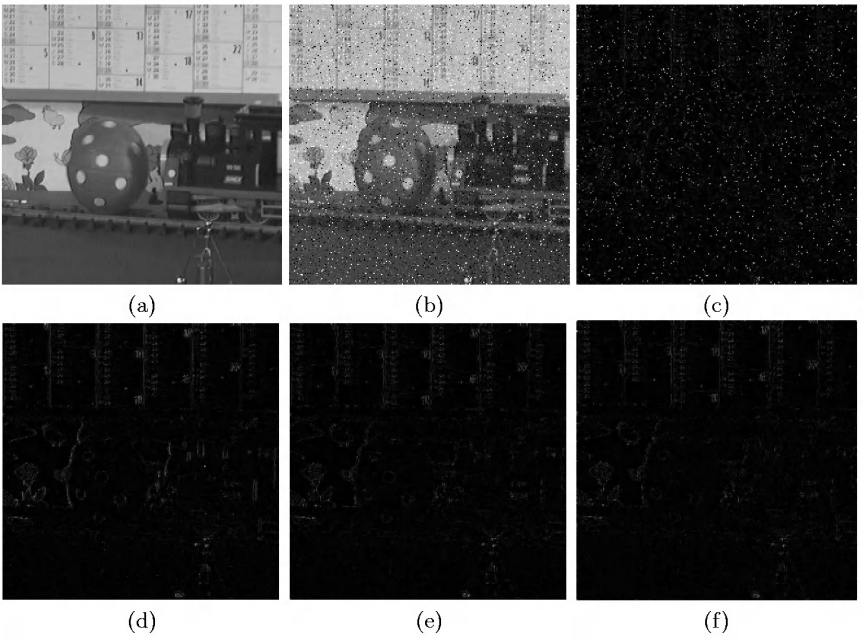


Fig. 2. (a) Original frame, (b) Corrupted frame with ($\alpha = 0.8$)(c) error image using NLMS filer (d) error image using weighted median filter (e) error image using weighted myriad filter (f) error image using NLMP filter

The filter update equation is defined by:

$$\hat{w}(s+1) = \hat{w}(s) + \mu p |e(s)|^{p-1} sign(\varepsilon(s)) g(s) \tag{19}$$

Although LMP is much more robust to impulsive noise than LMS, in some cases with the appearance of extremely impulsive noise it can become unstable. Motivated by the stability and increased speed of the normalized version of LMS, namely NLMS, an adaptive version for LMP, namely normalized LMP (NLMP), which can be expressed by:

$$\hat{w}(s+1) = \hat{w}(s) + \mu p \frac{|\varepsilon(s)|^{p-1} sign(\varepsilon(s))}{\|g\|_p^p(s) + \lambda} g(s) \tag{20}$$

where μ and λ are algorithm parameters. In all of these algorithms, the coefficient vector can be conveniently initialized to a good guess or to a zero vector if no prior information is available.

5 Experimental Results

In this section, we present a set of experiments in order to assess the performance of the proposed spatio-temporal Myriad and l_p-norm filter that we have

Table 1. FSNR (dB) achieved by the spatio-temporal filters

α values	0.5	0.8	1.0	1.5
3-D Normalized LMS Filter	11.24	12.13	12.31	11.72
3-D WMF (Weighted Median Filter)	14.84	14.03	13.32	12.61
3-D WMyF (Weighted Myriad Filter)	16.92	15.95	15.64	14.81
3-D LMP (Least P-norm filter)	14.22	14.93	15.02	14.86
3-D NLMP (Normalized Least P-norm filter)	15.02	16.25	16.44	16.32

discussed so far. A monochrome video sequence which is part of the *Calendar train* sequence, was used as original video sequence. This sequence is characterized by temporally varying objects along the temporal direction in a slowly moving background. The sequence is corrupted by symmetrical α-stable noise with different values of $(\alpha = 0.5, ..., 1.5)$, $(\gamma = 1)$ and $(\delta = 0)$. The spatio-temporal weighted median and NLMS filers have been used as a comparison reference. Figure 2(a) shows the original third frame of mobile sequence. Figure 2(b) shows the image corrupted by symmetric α-stable noise with $\alpha = 0.8$. The error image for a spatio-temporal NLMS filer is shown in figure 2(c). The error image of a spatio-temporal weighted median filter is shown in Figure 2(d). The error image of the spatio-temporal weighted myriad filter with $(K = 0.2)$ is shown in Figure 2(e) and figure 2(f) shows the output error of the spatio-temporal l_p-norm filter. The last two filters do a good job at removing the noise and retaining the detail of the image.

For objective quality evaluation, a criterion has been employed, namely, the fractional order SNR (FSNR[15]) defined as the logarithm of the ratio of the pth-order moments of the noise and the signal where $0 < p < \alpha$ instead of their powers, that is,

$$FSNR(dB) = 10 \log_{10} \frac{\sum_{i=1}^{K} \sum_{j=1}^{L} |f_n(i,j)|^p}{\sum_{i=1}^{K} \sum_{j=1}^{L} |\hat{f}_n(i,j) - f_n(i,j)|^p} \tag{21}$$

Table 1 summarizes the FSNR achieved by the proposed filters with different values of α. The motion estimation uses L_p-norm BMA. The values of p depend on α and they have been tuned experimentally is order to give the best matched blocks. According to these results, the filtered images using spatio-temporal myriad and l_p-norm filter have higher FSNR and give good improvement in term of visual quality. In some cases, we observe that the adaptive l_p-norm filter outperforms the myriad. This is because the α-stable noise characterization in both spatial and temporal direction is not a realistic assumption even if the myriad filter is optimal in such case.

6 Conclusion

In this paper, we presented a 3-D FLOS-based filters for α-stable noise suppression in video sequences. Experimental results on calendar train image sequence

showed marked improvement in visual and numerical qualities when using the p-norm block matching based algorithm and 3-D weighted myriad and l_p-norm filers. These filters are very suitable for α-stable and impulsive noise removal.

References

1. S. D. Kim, J.B. Ra. Efficient block-based video encoder embedding a wiener filter for noisy video sequences. *Journal of Visual Communications and Image Representation*. vol. 14, pp. 22-40, 2003.
2. A. Srivastava, A.B. Lee, E.P. Simoncelli, S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, Vol. 18, pp. 17-33, 2003.
3. O. Tanrikulu and A.G. Constantindes, "Least-mean Kurtosis : A novel higher-order statistics based adaptive filtering algorithm", *Electronics Letters*, vol. 30, no. 3, pp. 189-190, February 1994.
4. M. El Hassouni, H. Cherifi. Noise Reduction in Color Video Sequences Using Multichannel Motion-Compensated *L*-Filter. *Proc. of ICIP 2003*. September 14-17, 2003. Barcelona, Spain.
5. J. G. Gonzales, G.R. Arce. Zero-order statistics : a signal processing framework for very impulsive processes. *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, Banff, Canada, July 1997, pp. 254-258.
6. M. Shao, C. L. Nikias. Signal processing with fractionnal lower order moments : Stable processes and their applications. *Proc. IEEE*, vol. 81, pp. 986-1009, 1993.
7. G. Aydin, O. Arikan, and E. Cetin. Robust adaptive filtering algorithms for α-stable random processes. *IEEE Trans. On Circuits and Systems II*, vol. 46, N°. 2, pp. 198-202, February 1999.
8. P. Kidmose. Adaptive filtering for non-Gaussian noise processes. *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, ICASSP2000, pp. 424-427.
9. A. Ben Hamza and H. Krim. Image denoising : A nonlinear robust statistical approach. *IEEE Trans. on Signal Processing*, Vol. 49, N0. 12, Decembre 2001.
10. M. El Hassouni, H. Cherifi. A 2-D adaptive least l_p-norm filter for impulsive noise Cancellation in still images. *Proc. of ISSPA 2003*. 1-4 July 2003, Paris, FRANCE.
11. J. C. Brailean, R. P. Kleihorst, S. Efstratiadis, A. K. Katsaggelos, ad R. L. Lagendijk. Noise reduction filters for dynamic image sequences : A review. *Proc. IEEE*, vol. 83, pp. 1270-1291, Sept. 1995.
12. K.J. Boo and N.K. Bose. A motion-compensated spatio-temporal filter for image sequences with signal dependent noise. *IEEE Trans. Cir, and Syst. for Vid. Tech.*, vol. 8, no. 3, pp. 287-298, June. 1998.
13. E. Sayrol. Modelling the displaced frame difference as an alpha-stable distribution. *IEEE International Conference on Image Processing*, Volume 2, pages 195-198, Santa Barbara, CA, October 1997.
14. J. G. Gonzalez, G. R. Arce. Weighted myriad filters: A robust filtering framework derived from alpha-stable distributions. *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, May 1996, Vol. 5, pp. 2833-2836.
15. Ercan E. Kuruoglu, Peter J. W. Rayner, William J. Fitzgerald. Least l_p-norm impulsive noise cancellation with polynomial filters. *Signal Processing*. v. 69, pp. 1-14. 1998.

Automatic Text Extraction in Digital Video Based on Motion Analysis

Duarte Palma, João Ascenso, and Fernando Pereira

Instituto Superior Técnico – Instituto de Telecomunicações, 1049-001 Lisboa, Portugal
{Duarte.Palma, Joao.Ascenso, Fernando.Pereira}@lx.it.pt

Abstract. It is well known that the text that appears in a video scene or is graphically added to it is an important source of semantic information for indexing and retrieval, notably in the context of video databases. This paper proposes an improved algorithm for the automatic extraction of text in digital video; its major strengths are its robustness in terms of text skew and its improved performance in dealing with scene text. The system is based on a segmentation approach, using geometrical and spatial analyses for text detection. After, temporal redundancy is exploited to improve the detection performance by means of motion analysis. The output of the text detection step is then directly passed to a standard OCR software package in order to obtain the detected text as ASCII characters.

1 Introduction

The technological advances seen in recent years in the area of audiovisual representation technology have led to a boom in the usage of audiovisual information, namely accessed through the Internet, by a growing number of users. The increasing amount of audiovisual information being deployed has led many relevant content players such as audiovisual content producers and television operators to show interest in creating digital libraries which should allow the efficient storage and indexing of audiovisual information for future management and retrieval. Nowadays, the task of annotation is typically performed manually, by a human operator; this process is very expensive, time consuming and many times suffers from the subjectivity associated to the human operator. To address the need and overcome the problem, it is necessary to develop systems capable of automatically processing audiovisual information, targeting its efficient indexing, storage, searching, transmission, and viewing. Thus the development of automatic and efficient systems able to analyze, describe, filter and retrieve audiovisual information is necessary. With this objective in mind, several methods and even international standards have been developed in the last few years. Much of this technology targets the textual information that exists in images and video sequences since this is a source of highly semantic information and thus, if available, would allow the filtering and searching of audiovisual data by users in a more intuitive and natural way. The objective of this paper is to propose an automatic text extraction solution for digital video based on motion analysis. The most significant novelty of this proposal regarding the solutions already available in the literature is its capacity to deal both with graphical and scene text, oriented in any direction, with a high performance, notably a small number of

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 588–596, 2004.
© Springer-Verlag Berlin Heidelberg 2004

false detections. Graphical text is the text separately produced from the video shooting which is overlaid on the scene in a post-processing stage; scene text appears as part of the scene and is recorded together with the scene. Both types of text may be composed by characters with various sizes, fonts, colors and can appear in different directions; of course, scene text is typically more varied.

There have been several approaches proposed in the last few years for the automatic extraction of text in digital videos. Existing text extraction methods can be classified into three main categories: methods based on segmentation, on edge detection and on supervised learning. The methods based on the segmentation of the visual data treat the text as regions with special characteristics, e.g. in terms of texture, size, shape and alignment constraints. For example in [1], color is used to form homogeneous regions; after, heuristic methods are used to detect characters in these homogeneous regions. The methods based on edge detection exploit the typical high concentration of edges in characters. For example in [2], the problem of video indexing is addressed assuming that text consists in strokes with high contrast; thus the method searches for vertical edges which are grouped into rectangles. The methods based on supervised learning typically use neural networks for text detection. For example in [3], a hybrid wavelet/neural network is used to locate text in videos; the high-frequency coefficients are the input to the network. In [4], text lines are identified by using a complex-valued multi-layer feed-forward network trained to detect text with a fixed scale. The output of the neural network for all scales and positions is integrated into a single text-map, serving as a starting point for the detection of candidate text lines. Many of these methods show a lower detection performance when dealing with scene text and non-horizontal text.

This paper is structured as follows: Section 2 presents the automatic text detection algorithm here proposed. Section 3 shows the experimental results and thus the performance of the proposed text detection solution for various types of text. Finally, Section 4 concludes the paper and refers to future work.

2 A Proposal for the Automatic Extraction of Text in Video

This paper intends to propose an improved algorithm for the extraction of text in digital video considering both graphical and scene text. The overall architecture for the text extraction solution proposed is shown in Fig. 1 and considers, as usual, two major phases: text detection and text recognition. Due to space limitations, it is not possible to describe all the modules in the architecture in full detail; for further information see [6].

2.1 Text Detection

The principal objective of the text detection phase is to identify a set of regions classified as text candidate regions. The text detection is carried out in three steps [2, 3, 4]: video monitoring, formation of text sequences and motion analysis.

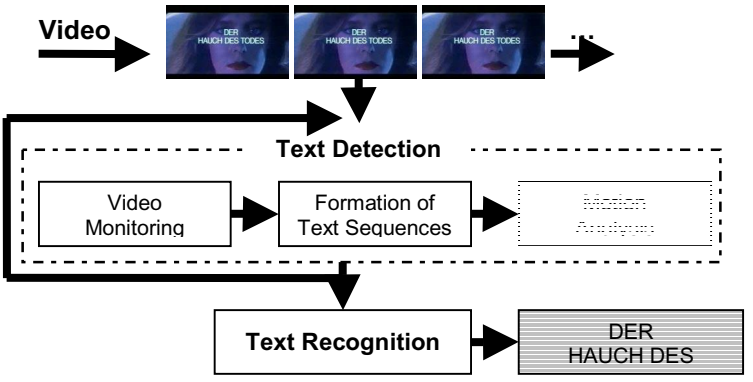


Fig. 1. Architecture of the proposed system.

2.1.1 Video Monitoring

The main purpose of the video monitoring module is to detect text in a video sequence. In order to reduce the overall computational complexity, text detection is first performed using a low temporal resolution of the video, this means only a few (periodic) frames are processed out of the full set. It is known from vision research that humans need between 2 and 3 seconds to process a complex scene [4, 5]. Therefore, it is safe to assume that in order for humans to recognize text, it needs to be visible for at least about one second; for a 25 fps video, this means that human text processing lasts for sure at least 25 frames (1 s) and thus it is enough to process one out of 25 frames in order to detect the presence of text. After the presence of text is detected at this low granularity, a finer detection phase starts targeting the frame precise detection of the text, see the formation of text sequences module.

The proposed video monitoring algorithm includes four steps: frame simplification, frame segmentation, character detection and word formation, detailed in the sequel:

1. *Frame simplification* – The purpose of this step is to decrease the influence of some unwanted effects such as noise or a too high number of colors. Frame simplification shall preserve as much as possible the original edges in each frame. So each frame is filtered with a filter combining edge detection with a median filter (3×3) [6]. The edge detector used is based on the Canny operator [7].
2. *Frame segmentation* **–** The purpose of frame segmentation is to split the image into homogeneous regions based on texture data (luminance). Each one of these regions has a certain probability of being a text character. The split and merge algorithm used in this paper was proposed by Cortez et al. [8] and is based on the hierarchical decomposition of each frame. However, when the frames have a significant amount of noise, this method results in over-segmentation, which degrades the quality of the characters edges. In order to eliminate this effect and improve the quality of the characters edges, it is necessary to eliminate small regions that share an edge with more than one region and that in some way contribute to the degradation of the characters boundaries. In order to identify these small regions for posterior merging and improve the characters edges, a technique presented by Lienhart [1] combining edge detection with dominant contrast local orientation is used: while edges are located using the Canny

detector [7], the dominant local orientation is determined by the inertia tensor [9].

3. *Character detection* – In order to detect each text character, the regions resulting from the segmentation step are filtered based on adequate geometrical restrictions related to their height, width, height-to-width ratio and compactness $(=A_i/A_{bb})$ where A_i, is the area of region i, and A_{bb}, is the area of the *bounding box* for region i. The thresholds used for each restriction depend on the range of character sizes selected for detection.
4. *Word formation* – To form words, it is considered that text consists of groups of characters aligned in a certain direction with characters sufficiently close to each other to form words. The word formation technique proposed here is based on the spatial analysis of each region previously classified as text. The following spatial criteria are used to form words:
   - *Proximity* – The regions shall be sufficiently close to each other;
   - *Alignment* – The regions shall be aligned in a certain direction;
   - *Height* – The regions shall have a minimum difference in height;
   - *Luminance* – The regions shall have similar luminance values;
   - *Dimension* – A word has to include at least three regions; smaller sets of related regions are discarded since they are meaningless.

The full description of the word formation process is described in [6]. Each analyzed frame is considered to contain text if at least one valid word exists, in any possible direction.

2.1.2 Formation of Text Sequences

The objective of the formation of text sequences module is to identify in the video sequence sets of contiguous frames where text exists, in order to apply later motion analysis. A text sequence is limited at the beginning and at the end by at least three contiguous frames without text. If the text monitoring phase detects some text in a frame, a finer detection of text starts for all the intermediate frames to determine the precise limits of the text sequence. This temporal finer detection is first performed backwards from the frame where the text has been initially detected and then forward in order to establish more precisely the first and the last frame where the text appears and the location of the text for each frame. This backward and forward detection ends when at least three contiguous frames without text are found in each direction. The detected interval represents a text sequence and the detected regions will be subject to motion analysis targeting the exploitation of temporal redundancy to increase the chance of detecting text and removing false alarms in individual frames since they are usually not stable throughout time. In order to detect text in the intermediate frames, the same algorithm used in the text monitoring phase is used.

2.1.3 Motion Analysis

In the motion analysis module, the correlation between frames is exploited to improve the results of the text detection phase. The major goals are to refine the text detection in terms of removing false alarms in individual frames, interpolating the location of accidentally missed text characters and words in individual frames, and temporally localizing the beginning and end of each word as well as its spatial location within each frame. The method proposed for motion analysis is based on the comparison of regions in successive frames and includes five steps:

1º Step – Text Tracking

This step is responsible for tracking the already detected text along the frames that constitute each text sequence targeting the formation of temporally related chains of characters. Each character chain represents the same character during its existence in the video sequence and consists in a collection of similar regions, occurring in several contiguous frames (although very likely with motion, e.g. scrolling). Every time a character region is detected for the first time, a position is stored and a signature is computed for that character region, using the following features: luminance, size and shape. Each frame contributes with one region classified as a character for the construction of a character chain.

The text tracking process creates the character chains through the following steps:

1. Each character region C_i, which belongs to frame n is compared with all the character chains CC_j, $j \in \{1,..., t\}$ that exist in frames n-8 to n-1. If the caracter C_i signature is similar enough to an already existing caracter chain signature and the caracter position is close to the estimated position for the already existing chain in the current frame, then that caracter is included in that caracter chain. The estimated position $\left(x_{CC_{jk}}, y_{CC_{jk}}\right)$ for the chain CC_j, in the frame F_k, is given by:

$$\left(x_{CC_{jk}}, y_{CC_{jk}}\right) = \left(x_{CC_{jk-1}}, y_{CC_{jk-1}}\right) + (dx, dy) \tag{1}$$

   Where (dx, dy) is the medium displacement for CC_j, among two frames, given by:

$$(dx, dy) = \left(\frac{x_{CC_{j1}} - x_{CC_{jk-1}}}{n}, \frac{y_{CC_{j1}} - y_{CC_{jk-1}}}{n} \right) \tag{2}$$

   Where n represents the duration of CC_j and $x_{CC_{j1}}$, $y_{CC_{j1}}$, $x_{CC_{jk-1}}$, $y_{CC_{jk-1}}$ represent the CC_j coordinates in the frame where it was created and in the frame before the current frame, respectively.
2. If caracter C_i is not similar to any existing character chain, a new character chain is created and initialized with the features/signature of that caracter.
3. After the processing of each frame, it is necessary to verify which character chains ended and which can still propagate to the next frame. A character chain is declared ended if it does not meet at least one of the following criteria:
   - Creation in frame n-1 but not present in frame n.
   - Not tracked in the last 0.32 seconds, i.e. 8 frames for a 25 fps.
   - Estimated position is outside the borders of the image.

   After a character chain is declared stopped, it is has to be globally validated by means of the following criteria:
   - Duration greater than 0.25 seconds, i.e. 6 frames for a 25 fps, and effective appearance in more than 3 frames;
   - Motion smaller than 250 pels per second, i.e. 10 pels/frame at 25 fps.

The thresholds for the above criteria were obtained empirically after extensive tests. Character chains classified as invalid are discarded in order to avoid false detections for individual frames. The output of the text tracking phase is a group of character chains, each one corresponding to a single character; they represent the temporal evolution of each character present in the video.

2º Step – Text Integration

The text integration step is responsible for grouping the character chains in order to form words. The word formation technique proposed here is based on the spatial and temporal analysis of each character chain. Besides the spatial criteria presented in Section 2.1.1, the following temporal criteria are used to form words:

- *Temporary coexistence* – The character chains shall exist in about the same frames, more precisely in at least 75% of the smaller character chain duration;
- *Duration* – The character chains shall have a minimum difference in duration;
- *Motion* – The character chains shall have a similar motion;

The chains not included in words at this phase are considered as noise and are discarded. The full description of the word formation process is described in [6].

3º Step – Character Recovery

The character recovery step explores the video temporal redundancy to complete the words with missing characters at least for some frames, e.g. due to noise or too textured background. In order to complete these words, they are extended to the size of their biggest chain of characters and the characters missing in the chains are recovered by means of temporal interpolation with motion compensation. Thus, by using temporal redundancy, the text detection for each frame is improved by 'completing' the words with missing characters for some frames.

4º Step – Elimination of Overlapped Words

This step is important to improve the performance for scene text; it performs the elimination of words which are overlapped with other words, i.e. words which bounding boxes overlap. Usually, overlapping of words occurs when false words are detected, e.g. due to shadows or three-dimensional text. Every time two or more words overlap, more precisely their bounding boxes overlap at least for one frame, the words with the smaller areas are discarded.

5º Step – Text Rotation

This step performs the rotation of the text to the horizontal position in order to be (better) recognized by the OCR system. Text rotation is performed in two steps:

1. *Calculation of the rotation angle* – The rotation angle is defined as the angle between the *xx* axis and the line that passes through the center of the regions where the word in question begins and ends. Considering the word $P = \{C_1,...,C_n\}$ with bounding box of coordinates $\{(x_{C_1}, y_{C_1}),...,(x_{Cn}, y_{Cn})\}$, its rotation angle is given by:

$$\theta_T = \tan^{-1}\left(\frac{\beta}{\alpha}\right) \tag{3}$$

   Where $\alpha = \max\{x_{C_1},...,x_{Cn}\} - \min\{x_{C_1},...,x_{Cn}\}$, $\beta = y_{\max\{x_{C_1},...,x_{Cn}\}} - y_{\min\{x_{C_1},...,x_{Cn}\}}$;

2. *Text rotation* – To rotate the text, the rotation angle θ computed is used along with the algorithm proposed by Alan Paeth [10].

Fig. 2 illustrates the effect of exploiting temporal redundancy for the extraction of text in video sequences. The result of extracting text from a single frame is shown in Fig. 2 (b) while Fig. 2 (c) shows the result of extracting text for the same frame but exploiting the temporal redundancy in the context of a video sequence.

(a) (b) (c)

Fig. 2. Example of the benefits of motion analysis for text extraction: (a) original image; (b) text detection for a single frame; (c) (better) text detection for the same frame but using temporal redundancy for recovering lost characters.

2.2 Text Recognition

The text bitmap produced using the words formed in the previous phase is recognized by standard OCR engines. In this paper, the OCR system used was OmniPage Pro 12.0.

3 Experimental Results

In order to evaluate the proposed automatic text extraction solution, a group of 13 video sequences with a total of 14.2 minutes and with plenty of text was used. These sequences where selected from several TV channels and were acquired using a Pinnacle Linx Video Input Cable, video capture card, with resolutions ranging from 352×208 pels to 384×288 (luminance) pels. The video sequences contain both scene and graphical text, aligned in any direction, with multiple fonts and sizes, in a total of 5852 characters. For example, the test video sequences were obtained from movie opening sequences and titles where graphical text is common, as well as TV commercials and news programs where scene text is common. The sequences with graphical text contain still as well as moving text (with scrolling motion).

3.1 Text Detection

Before processing each video sequence with the proposed text detection algorithm, a ground truth in terms of characters to be detected and recognized was determined. After, the correct detection or not of the ground truth characters has been determined by visual inspection of the images created by the text detection algorithm. For evaluating the text detection performance, the precision and recall metrics have been used: *Recall=CCD/CGT; Precision=CCD/TCD where CCD* represents the number of characters correctly detected by the algorithm, *CGT* represents the number of ground truth characters and *TCD* represents the total number of characters detected by the algorithm. The results in terms of detection for all the text defined as ground truth, i.e. horizontal, vertical and skewed text, for each type of text, can be observed in Table 1.

The best detection results, both in terms of recall and precision, were obtained for the videos with graphical text Recall results of approximately 94% indicate that only

Table 1. Detection results for all the video sequences.

Type of text	Frames in test set	Characters in test set	Recall	Precision
Scene text	8319	907	0.879	0.836
Graphic text	12979	4926	0.953	0.958
Total text in the ground truth	**21298**	**5833**	**0.941**	**0.938**

Fig. 3. Text extraction result for a video sequence: (a) video sequence; (b) text detection for individual frames (c) image resulting from the integration of the text in the full video sequence.

6% of the text that was considered relevant according to the defined conditions was not detected. Precision results of approximately 93% indicate that only 7% of the characters detected were falsely detected.

Fig. 3 illustrates the result of text extraction for a video sequence. The result of extracting text from the individual frames is shown in Fig. 3 (b) while Fig. 3 (c) shows the final integrated result for the full video sequence.

3.2 Text Recognition

For evaluating the text recognition performance, also precision and recall metrics have been used: *Recall=CCR/CGT; Precision=CCR/CR where CCR* represents the number of characters correctly recognized by the OCR system, *CGT* represents the number of ground truth characters and *CR* represents the number of characters output by the OCR system. The results in terms of recognition are presented in Table 2.

Table 2. Detection and recognition results for all the video sequences.

Type of text	Recall	Precision
Scene text	0.803	0.811
Graphic text	0.920	0.922
Total text in the ground truth	**0.902**	**0.905**

Similar to the detection results, the best recognition results were obtained for the videos with graphical text, both in terms of recall and precision. The characteristics of graphical text are the main justification for the different performance: normally, graphical text has a higher contrast and is better defined than scene text; on the other hand, scene text shows a larger number of fonts and sizes, making it more difficult to discriminate from other structures present in the scene.

4 Conclusions and Future Work

In this paper an improved automatic text extraction solution for digital video has been proposed. The major novelty of the solution is its improved performance in dealing with both types of text, oriented in any direction, with a small number of false detections. The performance of the proposed algorithm has been evaluated using various types of videos and text, notably title sequences of feature films, commercials and newscasts that contain both scene and graphical text, aligned in any direction with multiple fonts and sizes. Although it is difficult to compare the performance of the proposed algorithm with the few results available in the literature, notably because different sets of test material are used, it is possible to state that the present solution performs at least as well as the available state-of-the-art solutions and this for more varied text, notably in terms of scene text and text orientation. Our future research will concentrate on text with fewer constraints, e.g. text with random movements, characters may touch each other, and the font size may be smaller.

References

1. R. Lienhart e W. Effelsberg, “Automatic Text Segmentation and Text Recognition for Video Indexing”, Multimedia Systems, Vol. 8, Nº 1, (2000) 69 – 81.
2. T. Sato, T. Kanade, E. K. Hughes, M. A. Smith e S. Satoh, “Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions”, Multimedia Systems, Vol. 7, Nº5, (1999) 385 – 395.
3. H. Li, D. Doermann e O. Kia, “Automatic Text Detection and Tracking in Digital Video”, IEEE Transactions on Image Processing, Vol. 1, Nº 1, (2000) 147 – 156.
4. R. Lienhart e A. Wernicke, “Localizing and Segmenting Text in Images and Videos”, IEEE Transactions on Circuits and Systems for Video Technology, Vol.12, Nº4, (2002) 256 – 268.
5. P. H. Lindsay e D. A. Norman, ”Introduction into Psychology – Human Information Reception and Processing”, Springer – Verlag (1991).
6. D. Palma, “Automatic Text Extraction in Digital Video Sequences”, Instituto Superior Técnico, Lisboa, Master Thesis, (2004).
7. J. Canny, “A Computational Approach to Edge Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, Nº 6, (1986) 679 – 698.
8. D. Cortez, P. Nunes, M. Sequeira, F. Pereira, “Image Segmentation Towards New Image Representation Methods”, Signal Processing: Image Communication, Vol. 6, Nº 6, (1995) 485 – 498.
9. B. Jähne: Digital Image Processing, “Concepts, Algorithms, and Scientific Applications”, Springer (1997).
10. A. Paeth: A Fast Algorithm for General Raster Rotation. Graphics Interface, (1986) 77–81

Fast Video Registration Method for Video Quality Assessment*

Jihwan Choe and Chulhee Lee

Dept. Electrical and Electronic Eng., Yonsei University
134 Shinchon-Dong, Seodaumoon-Ku, 120-749 Seoul, KOREA
chulhee@yonsei.ac.kr

Abstract. In this paper, we propose a block-based video registration method that can be used for various applications such as objective video quality assessment. Instead of using the entire frames of a reference video sequence and a processed video sequence, we use a limited number of frames which are selected under a certain criterion and the selected frames are divided into a number of sub-blocks. Then, we select a small number of sub-blocks which have large spatial gradients. In order to find such sub-blocks, a spatial filtering is applied to the sub-blocks of the selected frames and we use energy of filtered sub-blocks to measure spatial gradient. The proposed registration method is fast and experimental results show that the method provides accurate registrations for a wide range of videos.

1 Introduction

The subjective video quality assessment [1] has been widely used for analog television services. The subjective test is considered to be the most accurate method that reflects human perception. However the subjective video quality assessment requires a number of human observers and it is impossible to perform quality assessment in real time. Thus, there has been a great interest in developing objective methods for video quality measurement and ITU (International Telecommunication Union) have been taking standardization activities [2-7]. In general, the objective video quality measurement can be divided into three categories: full reference, reduced reference and no reference methods [8]. The full reference method requires the reference video and processed video sequences to evaluate an objective quality score. In the reduced reference method, a set of parameters are extracted from reference and processed video sequences and theses features are used to evaluate video quality. In the no reference method, video quality is evaluated only using the processed video sequence.
Most objective video quality assessment methods require a certain calibration procedure. Quite often, spatial-temporal registration is the most important preprocessing.

* This work was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center (ITRC) support Program.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 597–604, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Spatial and temporal shifts frequently occur in practical video applications. Many video encoders introduce cropping, scaling and spatial shifts. Spatial mismatch also occurs when video resizing or format translation is performed. Transmission of encoded data also introduces temporal shifts. If the reference video sequence and the corresponding processed video sequence do not match exactly in the spatial and temporal directions, the performances of most objective quality assessment methods are adversely affected. In order to address these problems, we need video registration that matches the processed video sequence with the corresponding reference video sequence.

The registration procedure can be done by minimizing the error between a reference video sequence and a processed video sequence [9]. However, video registration can be time-consuming when the whole video sequences are used. In addition, if we have to deal with sub-pixel registration, we also require interpolation operations which are also time-consuming. In this paper, we propose a block-based registration method for objective video quality assessment. The proposed method is fast enough to be implemented in real-time application and provides robust performances over a wide range of videos.

2 Image Registration for Video Quality Assessment

2.1 Quality Assessment System

Fig. 1 shows a full reference quality assessment system for measuring the end-to-end performance of video sequences. In the full reference system, we use the reference video and the processed video sequences to assess the objective video quality [8]. The HRC (Hypothetical Reference Circuit) can be found in television broadcasting networks. Although the reference video and the decoder output may be analog (NTSC/PAL) or digital (CCIR 601), the digital video is commonly used for objective quality assessment. In a reduced reference system, a set of parameters extracted from the reference video and the processed video are used for quality assessment as illustrated in Fig. 2 [8]. The reduced reference system needs an auxiliary channel to transmit the features. A full reference system produces more accurate quality score than a reduced reference system. However a reduced reference system is more suited for practical video quality monitoring systems.

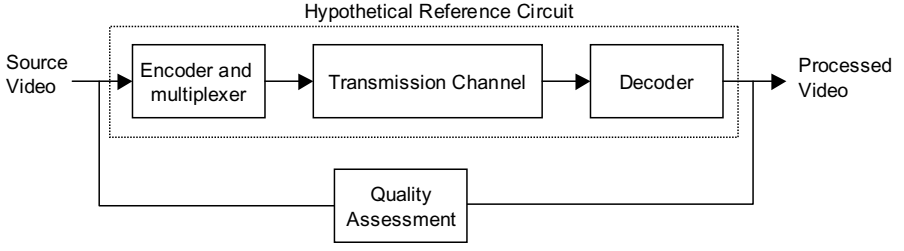


Fig. 1. Full reference objective video quality assessment system

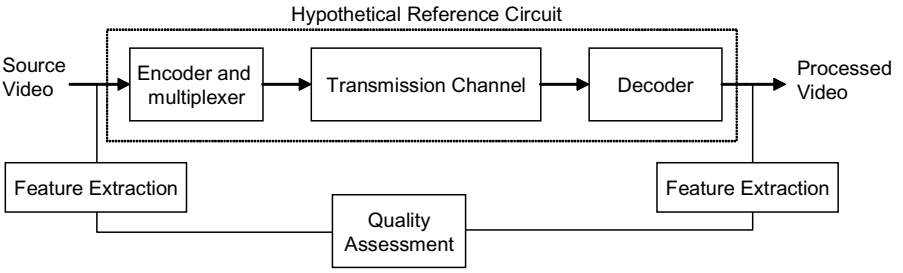


Fig. 2. Reduced reference objective video quality assessment system

2.2 Video Registration

A video registration is to find the best match between two video sequences. In video quality assessment, we need to find how much the processed video is shifted in the spatial or temporal directions. If translation is represented as displacement vector $\mathbf{d} = [d_1, d_2, d_3]^{\mathrm{T}}$, the mean square error between a reference video sequence and a processed video sequence that is translated by $\mathbf{d}$ is computed by

$$MSE(d_1, d_2, d_3) = \frac{1}{LMN} \sum_{l}^{L} \sum_{m}^{M} \sum_{n}^{N} (U(m,n,l) - V(m+d_1, n+d_2, l+d_3))^2 . \quad (1)$$

The best estimate for the displacement vector is obtained by minimizing the *MSE* :

$$\hat{\mathbf{d}} = \arg \min_{(d_1,d_2,d_3)} MSE(d_1, d_2, d_3) . \quad (2)$$

The resolution of spatial displacement vector may be integer pixel or sub-pixel. If it is sub-pixel, we need an interpolation technique such as the bilinear interpolation or the cubic spline interpolation. Generally, the precision of the temporal component of the displacement vector is one video frame for a progressively scanned video sequence as shown in Fig. 3. For an interlaced video sequence, it is one field: about 1/60 sec for 60 Hz interlaced video and about 1/50 sec for 50 Hz interlaced video as shown in Fig. 4. In the interlaced format, we need to construct a complete frame from each field to find the spatial displacement of interlaced videos. In other words, we need de-interlacing techniques.

2.3 Block Based Video Registration

Typically, it requires a long processing time to find the best match using equations (1) and (2). If we use all the frames of video sequences to find *MSE*, it cannot be done in real time. In order to reduce the processing time, we propose to use smaller number of sub-blocks which are selected from the video sequence. In other words, we find the best match between two video sequences by computing *MSE* in the sub-blocks *B*:

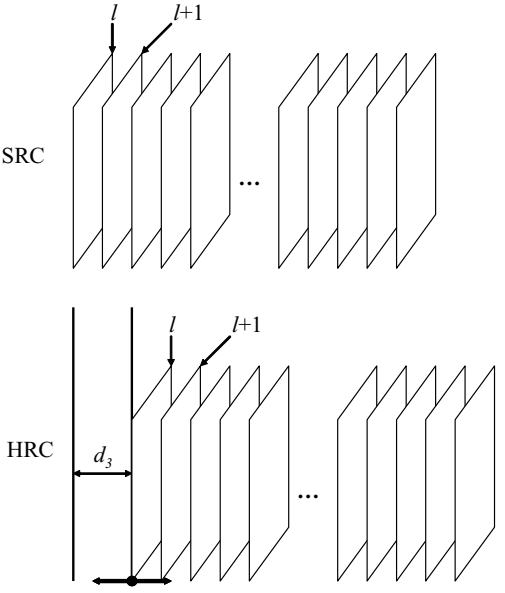


Fig. 3. Temporal registration of the progressive video sequence. The unit of d_3 is one video frame

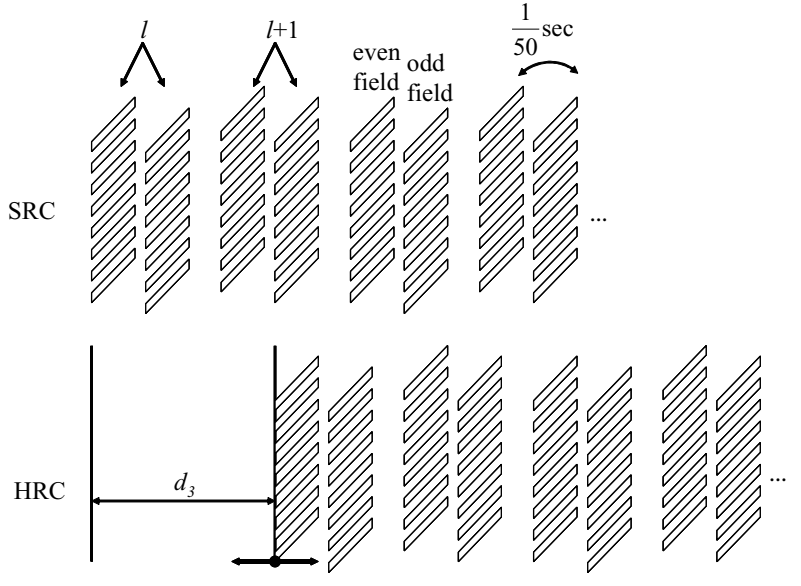


Fig. 4. Temporal registration of interlaced video sequences. The unit of d_3 is one field (1/50 sec for 50 Hz interlace video sequence)

$$MSE_B(d_1,d_2,d_3)=\frac{1}{K}\sum_{(m,n,l)\in B}(U(m,n,l)-V(m+d_1,n+d_2,l+d_3))^2 . \quad (3)$$

It is observed that rapidly changing areas of video sequences provide valuable information in video registration. Therefore, we find such areas and use them for video registration. In order to find rapidly changing scenes, the sum of absolute difference of two adjacent frames is computed as follows:

$$\begin{aligned} SAD(l)=&\frac{1}{MN}\sum_{m}^{M}\sum_{n}^{N}\left|U(m,n,l)-U(m,n,l-1)\right| \\ &+\frac{1}{MN}\sum_{m}^{M}\sum_{n}^{N}\left|U(m,n,l)-U(m,n,l+1)\right|. \end{aligned} \quad (4)$$

After computing *SAD* for the entire frames of the reference video sequence, we select a number of frames that have the largest *SAD*. These frames can be considered as the most rapidly changing scenes in the temporal direction. Furthermore, we have an additional restriction that intervals of the selected frames should exceed a certain amount of time period.

Then, we apply a spatial filtering to the selected frames to obtain the areas which have high spatial gradient. In particular, we use a 2-D wavelet transformation using the Haar filters. The corresponding 1-D output $y[n]$ is computed by

$$y[n]=\sum_{k}x[k]h[2n-k] \quad (5)$$

where $x[n]$ is an 1-D input signal and

$$h[n]=\begin{cases}1 & n=0\\ -1 & n=1\\ 0 & otherwise.\end{cases}$$

Then the HH sub-band of the 2-D Haar wavelet transform is divided into a number of 8×4 blocks and the energy of each block is calculated. Sub-blocks that have largest energy are selected as the areas that have high spatial gradient and the corresponding 16×8 blocks in the original frames are used for spatial-temporal registration. Thus, we select a small number of sub-blocks from the frames that have the largest *SAD*. Finally, using equation (3), we find the best displacement vector which provides the minimum *MSE* in the sub-blocks.

3 Experimental Results

The proposed registration method was applied to the video data provided by VQEG Phase I [4]. There are two video formats (525 and 626 lines) and each format has 160 video sequences (10 test sequences × 16 HRCs). Each sequence is 8 seconds in dura-

tion. It is reported that the VQEG Phase I video sequences are already registered in the spatial or temporal directions. The cubic spline interpolation was used to find missing lines of each field and to compute sub-pixel values. We divide an 8-seconds video sequence into three parts. Each part consists of 80 frames for 525/60 format sequences that have total 260 frames. The first ten frames and the last ten frames are ignored. We choose three frames (one frame from each part) which have the maximum *SAD* in each part. Once a frame is selected from each part, we apply the spatial filtering to the three frames. We selected ten sub-blocks from each frame and then found the best displacement vector which provided the minimum *MSE*. The search areas are ±20 pixels in the vertical and horizontal directions and ±4 fields in the temporal direction. Table 1 shows the registration results of 625/50 format sequences. The proposed registration method provides accurate results for most videos. The range of displacement vectors does not exceed ±0.5 except for one source video. Since the source video is almost repetition of a still picture and has no motion in temporal direction, the errors in the temporal displacement vector components of the videos are not critical. The experiments with the 525/60 format sequences have similar results that are not presented due to page limit.

In order to evaluate the consistency of the proposed method, we selected a processed video sequence encoded at a low bit rate. The processed video sequence is severely degraded and we tried to register it with the corresponding source video sequence. First, we randomly selected 3 frames and then randomly choose various numbers of blocks. The number of blocks varies from one to sixty for each frame. The registration results are shown in Fig. 5(a). As can be seen, displacement vectors show inconsistent performances. On the other hand, the results of the proposed method are consistent no matter how many blocks are used as can be seen in Fig. 5(b).

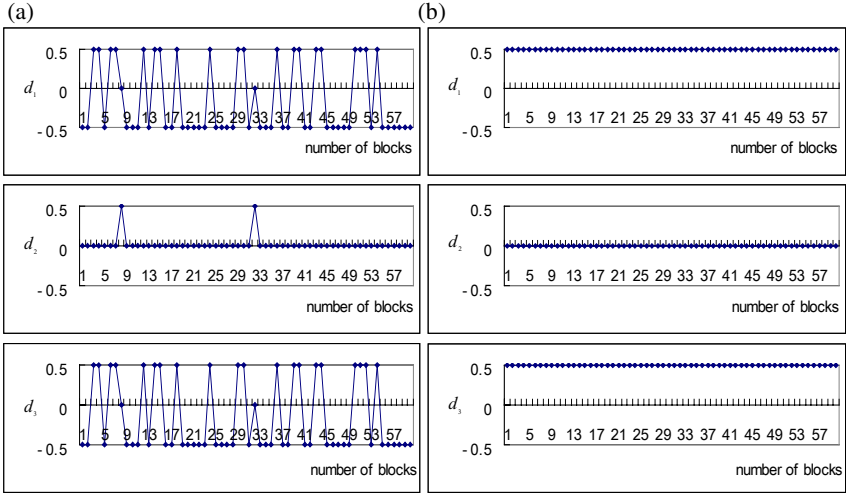


Fig. 5. Registration results of SRC10-HRC16 video sequence; (a) random selection of sub-blocks; (b) the proposed method

Table 1. Registration results. 625/50 format sequences of VQEG FRTV Phase I

no	d_1	d_2	d_3
1	-0.5	0	-1
2	0	0	-1
3	0	0	-1
4	0	0	-1
5	0	0	-1
6	0	0	-1
7	0	0	-1
8	0	0	-1
9	0	0	-1
10	0	0	-1
11	0	0	-1
12	0	0	-1
13	0	0	-1
14	0	0	-1
15	0	0	-1
16	0	0	-1
17	-0.5	0	0
18	0	0	0
19	0	0	0
20	0	0	0
21	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0
26	0	0	0
27	0	0	0
28	0	0	0
29	-0.5	-0.5	0
30	0	-0.5	0
31	0.5	0.5	0
32	0.5	0	-0.5
33	-0.5	0	0
34	0	0	0
35	0	0	0
36	0	0	0
37	0	0	0
38	0	0	0
39	0	0	0
40	0	0	0
41	0	0	0
42	0	0	0
43	0	0	0
44	0	0	0
45	0	0	0
46	0	0	0
47	0	-0.5	0
48	0	-0.5	0
49	-0.5	0	0
50	0	0	0
51	0	0	0
52	0	0	0
53	0	0	0
54	0	0	0
55	0	0	0
56	0	0	0
57	0	0	0
58	0	0	0
59	0	0	0
60	0	0	0
61	0	0	0
62	0	0	0
63	-0.5	0.5	0
64	-0.5	0.5	0
65	0	0	0
66	0	0	0
67	0	0	0
68	0	0	0
69	0	0	0
70	0	0	0
71	0	0	0
72	0	0	0
73	0	0	0
74	0	0	0
75	0	0	0
76	0	0	0
77	0	0	0
78	0	0	0
79	0	-0.5	0
80	0	-0.5	0
81	-0.5	0	0
82	0	0	0
83	0	0	0
84	0	0	0
85	0	0	0
86	0	0	0
87	0	0	0
88	0	0	0
89	0	0	0
90	0	0	0
91	0	0	0
92	0	0	0
93	0	0	0
94	0	0	0
95	0	-0.5	0
96	0	-0.5	0
97	-0.5	0	0
98	0	0	0
99	0	0	0
100	0	0	0
101	0	0	0
102	0	0	0
103	0	0	0
104	0	0	0
105	0	0	0
106	0	0	0
107	0	0	0
108	0	0	0
109	0	0	0
110	0	0	0
111	0	0	0.5
112	0	0	0.5
113	-0.5	0	0
114	0	0	0
115	0	0	0
116	0	0	0
117	0	0	0
118	0	0	0
119	0	0	0
120	0	0	0
121	0	0	0
122	0	0	0
123	0	0	0
124	0	0	0
125	0	0	0
126	0	0	0
127	0	-0.5	0
128	0	-0.5	0
129	0	0	0
130	0	0	0
131	0	0	0
132	0	0	0
133	0	0	0
134	0	0	0
135	0	0	0
136	0	0	0
137	0	0	0
138	0	0	0
139	0	0	0
140	0	0	0
141	0	0	0
142	0	0	0
143	0	-0.5	0
144	0	-0.5	0
145	-0.5	0	0
146	0	0	0
147	0	0	0
148	0	0	0
149	0	0	0
150	0	0	0
151	0	0	0
152	0	0	0
153	0	0	0
154	0	0	0
155	0	0	0
156	0	0	0
157	0.5	0	0.5
158	0	0	0
159	-0.5	0	-0.5
160	0.5	0	0.5

4 Conclusion

In this paper, we proposed an efficient registration method that can be used for video quality objective assessment and other video registration problems. Instead of using the entire frames to find best match between the reference video and processed video sequences, we use a small number of sub-blocks that are obtained by using spatial filters and frame differences. Using a small number of sub-blocks, the registration can be fast and efficient. Experiments show that the proposed method provides satisfactory performances.

References

1. ITU-R, Recommendation BT.500-8, "Methodology for the subjective assessment of quality of television pictures," 1998.
2. Video Quality Experts Group, "Video Quality Experts Group: Current Results and Future Directions," 2000.
3. S. Olsson, M. Stroppiana and J. Baina, "Objective methods for assessment of video quality: state of the art," IEEE Trans. Broadcasting, vol.43, no.4, pp.487-495, 1997.
4. Video Quality Experts Group, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," 2000.
5. Video Quality Experts Group, "Draft VQEG final report of FR-TV Phase II validation test," 2003.
6. C. Lee, O. Kwon, T. Jung, S. Cho, H. Kim, "Weighted PSNR for objective measurement of video quality," IASTED, pp.130-133, 2002.
7. C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," Optical Engineering, vol. 42, pp. 265-272, 2003.
8. ITU-T Recommendation J.143, "User requirements for objective perceptual video quality measurements in digital cable television," 2000.
9. ITU-T Recommendation P.931, "Multimedia communications delay, synchronization, and frame rate measurement," 1998.

Hidden Markov Model Based Events Detection in Soccer Video

Guoying Jin, Linmi Tao, and Guangyou Xu

Key Laboratory of Pervasive Computing, Ministry of Education
Department of Computer Science and Technology
Tsinghua University, Beijing 100084 P.R.China
jingy97@mails.tsinghua.edu.cn
{linmi, xgy-dcs}@tsinghua.edu.cn

Abstract. This paper presents an algorithm developed based on hidden Markov model for cues fusion and event inference in soccer video. Four events, shoot, foul, offside and normal playing, are defined to be detected. The states of the events are employed to model the observations of the five cues, which are extracted from the shot sequences directly. The experimental results show the algorithm is effective and robust in inferring events from roughly extracted cues.

1 Introduction

One of the most challenging issues in content-based retrieval is to tackle the semantic gap between the low-level features and the high-level semantics. Event analysis has been considered as a promising approach to bridge this gap. In this approach, events with semantic meanings are inferred from fusing relevant cues, which are extracted from low level features under the guidance of context and domain knowledge.

Video events detection techniques have attracted more and more researchers in recent years. However, there is no uniform conception of "event" up to now, as the definition of "event" is related to the application domain. The events that users concern with are different under the different backgrounds even in the same video. Thus, the definitions of event and the methods of event detection and recognition are distinct from each other.

However, events hold relevantly clear semantics in sports video analysis. Rui[1] and Zhang[2] presented methods for parsing baseball videos. Examples for analyzing basketball video are listed in [3]. Assfalg[4] decomposed each shot into its visual and graphic content elements and capture semantic content at a higher level of significance. Using a unique domain-specific feature grass-area-ratio, Xu[5] classified each soccer video frame into three kinds of view, namely global, zoom-in, and close-up. Then the play/break status of soccer game is obtained by heuristic rules. The dominant color region detection algorism proposed by Ekin[6] can robustly handle the temporal variations in the dominant color due to field, weather and lighting conditions throughout a sports video. It can

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 605–612, 2004.
© Springer-Verlag Berlin Heidelberg 2004

also be applied in shot boundary detection and shot type classification in sports video, referee and player of interest detection, penalty-box detection[7], and a fast algorithm for play-break event detection.

Hidden Markov Model (HMM) is a powerful tool that has been shown to be of great use in speech and signal processing[8]. The HMM theory is increasingly being applied to the areas of video analysis, such as detection of slow-motion replay segments in sports video for highlights generation[9], adaptive background estimation in surveillance applications[10], and classification of different types of video[11].

In this paper, we propose a HMM-based algorithm for video events detection based on cues fusion and inference. The video shot is regarded as the unit of video analysis. Under the guidance of context and domain knowledge, relevant cues are extracted and fused into the observations of HMM to study the relationship between the observations and the states (shots' contents) that could not be seen directly. Therefore we can model and detect specific events in video. Our primary experiments on soccer videos show that HMM-based method is effective in modeling and detecting of video events based on the well-selected cues.

The rest of the paper is organized as follows. Section 2 describes the HMMs for inferring soccer video events. In Section 3, the cues extraction methods are presented, while in Section 4 we propose the HMM-based soccer events detection method. Experiment results on soccer video are shown in Section 5. Section 6 concludes this paper.

2 The Hidden Markov Models for Soccer Video Events

The keys to successfully apply HMM method on cues fusion and inference for video event modeling and detection are the cues selection and the events definition. In other words, the selected cues should be appropriate for events to be detected. Good cues reflect the statistical relation to the states of HMM effectively, so that the relative models with suitable states and observations can be efficiently used in event detection.

First of all, the events to be detected should be defined on extractable cues. For example, the appearance of playback shots is selected as an extractable cue of specific events, since playback shots usually appear for providing more details of the events to audience when the significant events (such as shoot and foul) happened in soccer game video. In this paper, we define **four events**: (1) the shoot event, (2) the foul event and (3) the offside event that contain playback shots, (4) the general game process without playback shots.

The general game process mainly includes the global shots of the soccer field, in which the actions of players could not be seen very clearly. Some zoom-in shots such as kick-off, pass and defense are intermixed, and there are some close-up shots of player's bust or even head portrait in addition. The shoot event usually starts with a global shot near the penalty box area, followed by playback of shoot that are mostly zoom-in shots. The shoot scene may be replayed multiple times from various viewpoints. There could also be several close-up shots of shoot

player or goalkeeper. The foul event may start with global/zoom-in shot, and then succeed to the playback of foul. The close-up shots of relevant players or referee may appear before or after the playback shots. The occurrence of offside event is not as frequent as other events. It may consist of a global shot, playback shots that are commonly global shots, and possibly close-up shots of relevant players or referee as well.

Secondly, the states of HMM are defined in terms of the events for generating initial estimates of HMM parameters, since good initial estimates are helpful in parameters reestimation of discrete HMM. According to the above description of events, the time-dependent changes of event contents obey some particular rules when these events happen. In our hidden Markov models, these rules are described by means of the finite states of the models,

$$S = \{s_0, \ldots, s_{N-1}\} \tag{1}$$

where $N = 6$ represents the number of states of the events (**Six states**, Fig.1). Each state represents one of the contents as below: (0) playback; (1) kick-off; (2) pass near the goal area; (3) pass on midfield; (4) ball handling/scrambling/defense; (5) close-up.

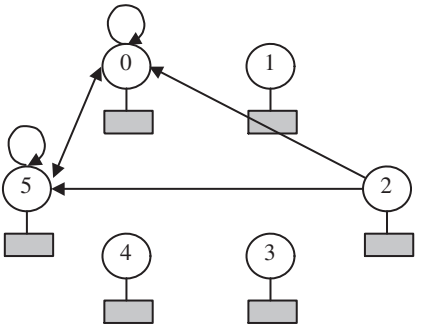


Fig. 1. An example of soccer events: the hidden Markov model of the shoot events and the relative states

The selection of cues is the balance between the correspondence to the event content and the feasibility of extracting cues in practice. **Five cues** are chosen based on our previews work:

(1) Shooting scale: global/zoom-in/close-up. It's an important attribute and is also the base for producing other cues.

(2) Playback: yes/no. This cue is essential to infer the three special events.

(3) Goalmouth appearance: yes/no. It's helpful to identify if the events happen near the penalty box.

(4) The occurrence of football trajectory: yes/no. It is usually extracted only in zoom-in shots to judge whether relate to shoot or kick-off.

(5) Blur: yes/no. The blurred frame may imply that there are fast motions in shots. It is extracted in zoom-in shots as well.

Finally, the observation of HMM is defined as the combination of these five cues. There is a finite set of possible observations relative to the cues,

$$\Theta = \{\theta_1, \ldots, \theta_M\} \tag{2}$$

In our case, the cardinality is

$$M = 3 * 2 * 2 * 2 * 2 = 48 \tag{3}$$

where five digitals represent the conditions of the cues in the order of the above list.

3 Cues Extraction

In general, model design and cues extraction should be compromised, so as to make each in its proper place. A major advantage of our method is there is no need of high accurate cues extraction algorithms, which are difficult problems to be handled. Our purpose is to validate the feasibility of video event detection by using HMM to fuse cues for inferring events.

(a) Shooting Scale

(b) Playback

(c) Goalmouth

(d) Football

(e) Blur

Fig. 2. Five cues of soccer events: (a) Shooting Scale: global, zoom-in, and close-up; (b) A playback shot with wipe transition between shots; (c) Goalmouth in global, zoom-in, and close-up shots; (d) The football track in a zoom-in shot; (e) Frame-blurred shot, player-blurred shot, and non-blurred shot

3.1 Global/Zoom-In/Close-Up Shot Detection

As the contents of global, zoom-in and close-up shots (shown as Fig.2(a)) differ from each other, the video analysis methods on them are various. The global shots give the global view of the soccer field. The zoom-in shots focus on the occurrence of specific actions or events. The close-up shots are usually the follow-scene of relevant persons. Some other cues extraction methods (for example, the goalmouth detection described in Section 3.3) are based on this cue.

The color and texture information are the most important low-level features in our method. The first step is to pick up the global-view frames according to the texture features of frames. The hue of the soccer field is then learned on these global-view frames. Significant color information except the field color is obtained among the frames whose backgrounds are mostly on the soccer field. The color models of soccer field, players and referee are created by means of mixture Gaussian model[12]. Thus we could determine the shooting scale by combining the field-ratio, relative size of person and the texture feature.

3.2 Playback Detection

The playback shots are vital in the shoot, foul and offside events. Precise detection of playback shots could help the video event detection and the shots analysis in higher semantic level.

Playback shots are determined by the methods of defining the midpoint of each shot as the starting point and running Viterbi algorithm forward/backward, based on the improved detection method of slow-motion replay segments presented in [9] (Fig.2(b)). In our method, edge change fraction is detected instead of the fourth feature based on the color histogram for capturing the gradual transition in editing effects. The HMM model of SLO-MOs is also updated for the case of continuous multiple playback shots.

3.3 Goalmouth Detection

Many events interesting people happen near the goalmouth. The visual features of goalmouth such as color of goalpost and net texture are used in goalmouth detection.

The goalmouths in three styles of shots (global, zoom-in and close-up) are shown in Fig.2(c). In zoom-in/close-up shots, the goalposts is detected by searching vertical white strip using Hough Transform, with the help of the net texture detection. As contrasted in global shots, both the position and the orientation of the soccer field are useful in finding the goalmouth, except the white goalpost detection. The parallel field lines near the penalty box area are also helpful for goalmouth detection in global shots.

3.4 Football Detection

Football is one of the most important semantic objects as the goalmouth in soccer video. Because of the small size of football in global shots and the few

appearance of football in close-up shots, we only detect and trace football in zoom-in shots. However, the accurate segmentation and tracking of football are very difficult and hard to be solved. We adopt a simpler approximate method: detecting white round area in certain amount of successive frames (Fig.2(d)).

3.5 Blurred Shot Detection

This cue is correlative to fast motions in shots, so it is used to distinguish play status (pass or defense) from break status (such as substitution) in zoom-in shots. Blur value is the reciprocal of the number of edge points. Two kinds of blurred shot should be detected: (1) The blurring of the frame, which represents the fast motion of camera. (2) The blurring of players, which reflects the fast motion of players (Fig.2(e)).

4 HMM-Based Video Events Detection

After the shot detection and cues extraction, we could obtain the observation sequences by fusing these cues. The resulting shot content sequences seem to comply with the Markov property, so that we apply the HMM method on event inference. The HMMs are used to model the temporal evolution of the shots features. The system framework of video event detection is shown in Fig.3.

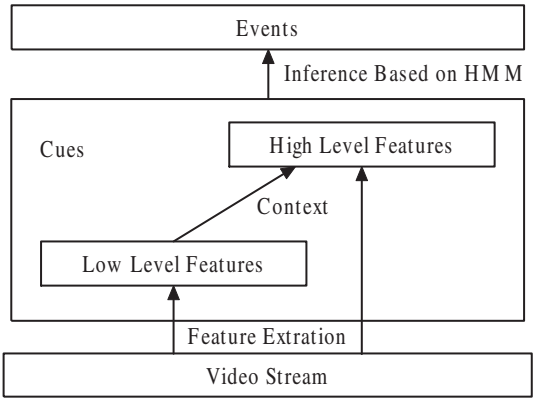


Fig. 3. The system framework of HMM-based video events detection

A hidden Markov model[8] is described by a finite set of possible states S, observations Θ, and relative parameters $\lambda = (A, B, \pi)$, where A is the transition probabilities matrix, B is the observation probabilities matrix, and π is the initial state distribution.

4.1 Hidden Markov Models of Events

Assume there are N_e events to be detected. Each $Event_i (i = 0, \ldots, N_e - 1)$ could be denoted by a HMM parameter λ_i. The hidden Markov modeling of the events is the training or learning problem: Given L_i observation sequences $O^{(l)} (l = 1, \ldots, L_i)$ produced from the training data of $Event_i$ where

$$O^{(l)} = O_1^{(l)}, O_2^{(l)}, \ldots, O_{T_l}^{(l)} \tag{4}$$

to maximize $P(O|\lambda_i)$ by adjusting the model parameters A_i, B_i and π_i. The HMM training method is based on the Baum-Welch algorism.

4.2 Events Detection Based on HMM

Given an observation sequence $O = O_1, O_2, \ldots, O_T$ produced from a shot sequence, the probability $P(O|\lambda_i)(i = 0, \ldots, N_e - 1)$, relatives to each model λ_i, is calculated by using Viterbi algorithm. Thus this shot sequence belongs to the $Enent_j$ where

$$j = \arg \max_{i=0,\ldots,N_e-1} P(Q, O|\lambda_i) \tag{5}$$

5 Experiment Results

We have experimented on 913 shots in 16 video files recorded on two soccer games, by stochastically selecting half files as the training data and the rest as the test data. The average accuracy of 100 times of experiments is 85.7%.

Table 1 shows one of the events detection results. 8 video files (include 59 events) are selected to train the HMM parameters of four events, and the other 8 video files (include 67 events) are used as test data.

There are only 2 offside events in the training video files. Hence the learned parameters cannot create the proper model of offside event. As we can see in Table 1, the accuracy of offside event detection is very low. If more video files are added in training data, the event models trained should be better.

Table 1. Events detection results

Detected Events	Amount	Results	Correct in Results	Accuracy
Normal Playing	32	33	31	93.9%
Shoot Event	21	19	18	94.7%
Foul Event	14	13	12	92.3%
Offside Event	0	2	0	0.0%
Total	67	67	61	91.0%

6 Conclusion

In soccer game video, the transition of shot contents in time domain obeys certain regular pattern. The regularity is more obvious when the special events (such as shoot events) happen.

According the above properties of video contents in soccer game, we applied the HMM theory on cues fusion and inference for the video event modeling and detection. Each video shot is viewed as a unit. Cues are extracted to form the observation of HMM. The video contents are defined as the states of HMM models to guide the generation of good initial estimates of HMM parameters. The aim of video event modeling is to learn the parameters of HMM for the events detected. The video event detection is the match problem between the observation sequence and the HMM. Experimental results show that the HMM-based method can effectively detect events in soccer video. It is also clear that this method can be applied to the video event detection in other application domains similarly.

References

1. Y. Rui, A. Gupta, and A. Acero: Automatically extracting highlights for TV baseball programs. ACM Multimedia, Los Angeles, CA (2000) 105-115
2. D. Zhang and S.F. Chang: Event detection in baseball video using superimposed caption recognition. ACM Multimedia, Juan-les-Pins, France (December 2002)
3. S. Nepal, U. Srinivasan, G. Reynolds: Automatic detection of 'Goal' segments in basketball videos. ACM Multimedia, Ottawa, Canada (September 2001)
4. J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo: Semantic annotation of sports videos. IEEE Multimedia, 9(2) (April/June 2002) 52-60
5. P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun: Algorithms and systems for segmentation and structure analysis in soccer video. ICME'01, Tokyo, Japan (August 2001) 928-931
6. A. Ekin and A. M. Tekalp: Robust dominant color region detection with applications to sports video. ICIP'2003, Barcelona, Spain (Oct. 2003)
7. A. Ekin, A. M. Tekalp, and R. Mehrotra: Automatic soccer video analysis and summarization. IEEE Trans. on Image Processing, vol. 12, no. 7 (July 2003) 796-807
8. L.R. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition. Proc.of the IEEE, 77(2) (1989) 257-285
9. H. Pan, P. Van Beek, and M.I. Sezan: Detection of slow-motion replay segments in sports video for highlights generation. ICASSP'2001
10. B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. M. Buhmann: Topology Free Hidden Markov Models: Application to Background Modeling. ICCV'2001, Vancouver, Canada, vol. I (July 2001) 294-301
11. Y. Haoran, D. Rajan, and C.L. Tien: An Efficient Video Classification System Based On HMM In Compressed Domain. IEEE Pacific-Rim Conference on Multimedia (2003), Singapore
12. LeiWang, Boyi Zeng, Steve Lin, Guangyou Xu, Heung-Yeung Shum: Automatic Extraction of Semantic Colors in Sports Video, ICASSP'2004, Montreal, Canada, May, 2004, III: 617-620

Improving Height Recovery from a Single Image of a Face Using Local Shape Indicators

Mario Castelán* and Edwin R. Hancock

Dept. of Computer Science, University of York, York YO10 5DD, United Kingdom
{mario,erh}@cs.york.ac.uk

Abstract. In this paper we propose a local-shape based method for correcting shape-from- shading information for surface height recovery of faces. The underpinning idea comes from the observation that subtle changes in the elements of a gradient field can cause notable changes in its integrated surface. One of the problems with reliable face surface reconstruction using shape-from-shading is that local errors in the needle map can cause implosion of facial features. To overcome this problem we develop a method for ensuring surface convexity. This is done by modifying the gradient orientations in accordance with critical points on the surface. We utilize a local shape indicator as a criteria to decide which surface normals are to be modified. Experiments show that altering the directions of a surface normal field of a face leads to a considerable improvement in its integrated height map.

1 Introduction

The problem of acquiring surface models of faces is an important one with potentially important applications in biometrics, computer games and production graphics. There are many ways in which surface models can be acquired, and these include the use of range-scanners, stereoscopic cameras and structured light sensors. However, one of the most appealing methods is to use shape-from-shading (SFS), since this is a non-invasive process which mimics the capabilities of the human vision system. Unfortunately, the use of SFS for face analysis has proved to be an elusive task, since the concave-convex ambiguity can result in the inversion of prominent features such as the nose. To overcome this problem, domain specific constraints must be used. For instance Zhao and Chellappa [10] have exploited facial symmetry to overcome the problem.

In general, though, SFS is an under-constrained problem since the two degrees of freedom for surface orientation (slant and tilt), must be recovered from a single measured intensity value. Hence, it is frequently posed as that of minimizing cost functionals that capture constraints on the gradient field. This is usually carried out through iterative schemes in the discrete domain. Despite sustained research activity in the field for some three decades[7,11], no SFS scheme

* Supported by National Council of Science and Technology (CONACYT), Mexico, under grant No. 141485.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 613–620, 2004.
© Springer-Verlag Berlin Heidelberg 2004

has been demonstrated to work as accurately as the *specially* constrained SFS cases of photometric stereo[3,5] and statistical SFS[1]. The first of these methods requires at least three images of the same object illuminated from different view-points, while the second uses a database of accurate surface information (height or gradient) belonging to objects of the same class, i.e. faces. When it comes to the original single-image SFS problem, the resulting data is seriously affected by the noise caused by factors such as inaccurately calculated illumination direction and unknown reflectance properties, and both of these are difficult to obtain from a single image. The restrictions imposed by most SFS schemes on the gradient field (smoothness, irradiance, integrability, unit length) are insufficient to overcome with these problems. Moreover, if misused, they may result in either oversmoothing of the gradient field, loss of fine surface detail, or overdependence on the intensity information. As a result, the recovered height surface may represent an oversmoothed reconstruction of the object under study.

However, although a precise height map is difficult to obtain by integration of the field of surface normals delivered by SFS, due to local errors in direction, this does not imply that the entire field of surface normals is in error. In fact, some regions on the image do provide surface normal information that is sufficiently faithful for good surface reconstruction. The problem originates from those locations where the combined effects of image noise, and overreliance on the consistency constraints, results in effects such as inversion of the surface convexity or concavity. When applied to the problem of face reconstruction, for instance, the effect can be to cause high curvature features such as the nose to become imploded with respect to the remainder of the surface. If such regions can be identified and the surface normal directions corrected, then the result is improved overall surface topography.

The outline of the this paper is as follows. We commence with a brief overview of the Frankot and Chellappa's global surface integration method and local shape indicators, then we develop a method that can be used to reassign the surface gradient orientations. We provide experiments to evaluate the method on a human face. Finally, we present some conclusions and identify directions for future work.

2 Integrating a Gradient Field

The integrability condition in SFS ensures that the recovered surface satisfies the following condition on the partial derivatives of the height function: $Z_{xy} = Z_{yx}$. In [4] Frankot and Chellappa proposed a method to project a gradient field to the nearest integrable solution. They suggested to use a set of integrable basis functions to represent the surface slopes so as to minimize the distance between an ideally integrable gradient field and a non integrable one.

Following [4], if the surface Z is given by

$$\tilde{Z}(x,y) = \sum_{\omega \in \Omega} \tilde{C}(\omega)\phi(x,y,\omega)$$

where ω is a two dimensional index belonging to a domain Ω, and $\phi(x, y, \omega)$ is a set of basis functions which are not necessarily mutually orthogonal, the partial derivatives of $\tilde{Z}$ can also be expressed in terms of this set of basis functions using the formulae

$$\tilde{Z}_x(x, y) = \sum_{\omega \in \Omega} \tilde{C}(\omega)\phi_x(x, y, \omega) \quad \text{and} \quad \tilde{Z}_y(x, y) = \sum_{\omega \in \Omega} \tilde{C}(\omega)\phi_y(x, y, \omega)$$

Given that $\phi_x(x, y, \omega)$ and $\phi_y(x, y, \omega)$ are integrable, then so are the mixed partial derivatives of $\widetilde{Z}(x, y)$.

In the same way, the possibly non integrable gradient field (which, indeed, is the only information we have) can be represented as

$$\hat{Z}_x(x, y) = \sum_{\omega \in \Omega} \hat{C}_1(\omega)\phi_x(x, y, \omega) \quad \text{and} \quad \hat{Z}_y(x, y) = \sum_{\omega \in \Omega} \hat{C}_2(\omega)\phi_y(x, y, \omega)$$

Note that, as $\hat{C}_1 \neq \hat{C}_2$, then $\hat{Z}_{xy} \neq \hat{Z}_{yx}$.

The goal then is to find the set of coefficients that minimize the quantity

$$d\Big\{(\hat{Z}_x, \hat{Z}_y), (\tilde{Z}_x, \tilde{Z}_y)\Big\} = \int\int \left\|\tilde{Z}_x - \hat{Z}_x\right\|^2 + \left\|\tilde{Z}_y - \hat{Z}_y\right\|^2 dxdy$$

In [4], details are given about how to solve this equation globally, in the Fourier domain.

3 Local Shape Indicators

A local shape indicator is a scalar that gives information about the local topography of a surface using its principal curvatures. The principal curvatures may be estimated using the surface normal directions to compute the Hessian matrix. They may also be estimated by fitting a local quadric surface patch $P(u, v) = \frac{1}{2}(\kappa_1 u^2, \kappa_2 v^2)$ having κ_1 and κ_2 as the principal curvatures with directions u and v respectively. κ_1 and κ_2 are the calculated eigenvalues of the Hessian matrix.

Local shape indicators are usually coupled. For instance, the HK classification[2] uses the Gaussian and mean curvatures $H = \frac{(\kappa_1+\kappa_2)}{2}$ and $K = \kappa_1 * \kappa_2$ respectively. By distinguishing between the cases in which H and K are individually negative, zero or positive, it is possible on the basis of the joint behavior to assign topographic labels to points on a surface.

A different and slightly more convenient set of attributes is the curvedness/shape-index representation developed by Koenderink and Von Doorn SC[8]. Here the principal curvatures are used to compute the shape index $S = -\frac{2}{\pi} * arctan(\frac{\kappa_1+\kappa_2}{\kappa_1-\kappa_2})$ for $\kappa_1 \geq \kappa_2$; and the curvedness, $C = (\frac{\kappa_1^2+\kappa_2^2}{2})^{1/2}$. The shape index is an angular variable that relates to the local surface topography. It varies continuously from -1 to $+1$ as the surface changes through cup, rut, saddle-rut, saddle, saddle-ridge, ridge and dome, and cup again. The curvedness relates to the degree of curvature of the surface.

4 Using the Shape-Index to Redirect SFS Gradient Fields

Inevitably, any surface gradient field delivered by SFS will be inaccurate due to noise or albedo changes which cause variations in the intensities of the input image. SFS works well for objects that are uniformly concave or convex. However, if the object under study is more complex, with both concave and convex regions, then SFS can fail. In these situations although the recovered surface normal direction is consistent with the measured image brightness, the recovered surface does not reflect the structure of the object under study. In particular, there may be inversions of the sign of the surface curvature with convex regions appearing concave and vice-versa. However, in the case of faces (and many other objects) the surface under study is largely convex.

Based on this above observation, in this paper we present a method for enforcing the convexity of the integrated surface while ensuring a global maximum on a particular position of it.

Formally stated, suppose S is a smooth surface immersed in $\mathbb{R}^3$. Let p and U_p be a critical point of S and a neighborhood of p respectively. Suppose that S is locally concave over U_p. Then, the new surface $\widetilde{S}$ constructed from S by reversing the sign of all its partial derivatives, S_x and S_y, is locally convex in U_p. Besides, a local maximum on $\widetilde{S}$ will be located on that point where the function stops increasing and starts decreasing[1]; if all the partial derivatives of $\widetilde{S}$ with respect to x, $\widetilde{S}_x$, have negative sign before reaching the position of the critical point p in the x axis and have positive sign after reaching it, and the same occurs for $\widetilde{S}_y$, then the critical point p on U_p will be the position of the global maximum[2] of $\widetilde{S}$.

The main idea underlying this paper is to enforce the condition that the integrated surface has a global height maximum where it is suitable for our problem, i.e. the tip of the nose. Such a point will serve as a division for the positive and negatively signed areas of the needle map. To enforce this condition we follow the simple rule:

$$\breve{Z}_x(x,y) = \begin{cases} abs(\hat{Z}_x(x,y)) & \text{if } x \leq a \text{ and } SI(x,y) \leq \tau_x \\ -abs(\hat{Z}_x(x,y)) & \text{if } x > a \text{ and } SI(x,y) \leq \tau_x \\ \hat{Z}_x(x,y) & \text{otherwise} \end{cases}$$

$$\breve{Z}_y(x,y) = \begin{cases} abs(\hat{Z}_y(x,y)) & \text{if } y \leq b \text{ and } SI(x,y) \leq \tau_y \\ abs(\hat{Z}_y(x,y)) & \text{if } y > b \text{ and } SI(x,y) \leq \tau_y \\ \hat{Z}_y(x,y) & \text{otherwise} \end{cases}$$

[1] Of course, $\widetilde{S}$ will present many local maxima for a face-like surface.

[2] It might be a maximum or a minimum depending on the integration method.

where $\breve{Z}_x$ and $\breve{Z}_y$ are the updated gradients, $\hat{Z}_x$ and $\hat{Z}_y$ are the original gradients, a and b are the coordinates for the position of the desired global height maximum, on the x and y axis respectively. $SI(x, y)$ is the shape-index which is compared to the thresholds τ_x and τ_y for deciding whether the element of the gradient field at the location (x, y) will be altered or not.

To illustrate the global height maximum enforcement procedure, Figure 1 (left) shows the results of applying the method to the derivatives of a sphere with radius 75 units. From left to right, transversal plots of the recovered surface on the x axis are shown. The peak coordinates (a, b) are both set to 75, 50 and 20 respectively, and τ_x and τ_y are set to one: all the derivatives are taken into account to expose the extreme case. The convexity strengthening is clearer in Figure 1 (right), where the method is applied to the Mexican hat function. Cross sections of the recovered surface are shown, and from left to right they show the original surface and recovered surface after applying the method taking as a peak point the center of the surface with $\tau_x = \tau_y = 1$. Note how the concave parts of the hat become convex.

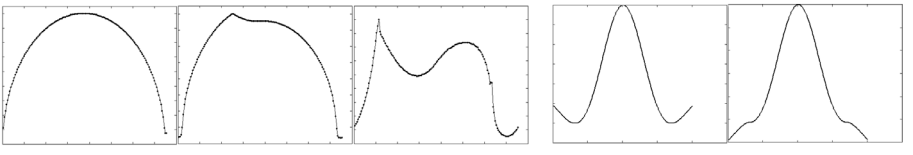


Fig. 1. Applying the method to a sphere and mexican hat.

It is evident that the peak-enforcement procedure will segment the recovered surface into four quadrants. As a result the curvedness of the recovered surface will be reduced. This is not desirable from the point of view of surface recovery from a face. The net effect will be to make the surface structure pyramidal. This problem is overcome by using the thresholding procedure to either force the normals to change direction or to allow them to remain unchanged.

In the following section some experiments will be presented in order to illustrate these points on an application involving face reconstruction using SFS.

5 Experiments

The tests were carried out on the image shown in Figure 3(top-left)[3]. To compute the surface gradients from the raw image brightness we followed the procedure described in[9]. The method involves first computing the image gradient, the surface normals are positioned on a cone whose axis is in the light source direction and whose opening angle is given by $\theta = \arccos E$. The position of the surface

[3] The face database was provided by the Max-Planck Institute for Biological Cybernetics in Tuebingen, Germany

normal on the cone is such at its projection onto the image plane is the same as the direction of the measured image gradient. This construction ensures the image irradiance equation to be satisfied. For the surface integration step we used the global method proposed in [4] which recovers surface height using the inverse Fourier transform of the field of surface normal directions.

In Figure 2 the surface gradient is shown after surface integration is performed. For each set of images we show the needle map, and the re-illuminated needle maps with light source direction $(1, 0, 0)$ and $(0, 1, 0)$ respectively. The un-altered case is shown in the top-left image. Notice how the re-illuminations suggest that the nose is imploded, since the region surrounding it is overshadowed. The remaining images present the cases of the recovered surface gradient after applying the method with $\tau_x = \tau_y = 1$ (middle-left), $\tau_x = \tau_y = 0.5$ (top-right) and $\tau_x = \tau_y = 0.4$ (middle-right). For the hardest case (middle-left), where no data is filtered by the thresholds, the four-quadrant effect is quite perceptible, but it tends to relax when the filter starts to work: the lower the threshold, the stronger the influence of the face features. It is important to highlight the difference among the top and bottom needle maps and how the change of signs in the surface normals suggests that the new derivatives arise now from the *shape* of a face (middle-right) rather than from the *image* of a face (top-left). This may be interpreted as a consequence of inverting the originally concave regions (the ones around the nose, eyes and mouth) to convex regions, making the imploded facial areas to emerge so that a new face shape can be recovered. Figure 2 makes clearer the former observation, by showing a comparison between the original and modified shape-indexes[4]: the dotted line corresponds to the original case while the solid line corresponds to the$\tau_x = \tau_y = 0.4$ case. This diagram presents only the cases when the pixels actually changed to a greater shape-index, such pixels belong to the regions surrounding the nose, mouth and eyes. Note again how the concave-like areas of the face tend to become convex-like.

The correction of the directions of the normals becomes clearer when analyzing Figure 3 (middle and right images on the top row). Each set shows the $(0, 0, 1)$ re-illumination and recovered surface as an intensity function respectively. The middle set corresponds to the unaltered case while the right set shows the $\tau_x = \tau_y = 0.4$ case. There are several features of the produced changes in the rendered surface that deserve comment. For instance, the intensities around the nose produce the appearance of a convex surface, making it appear more natural. It is also interesting to note that the difference vectors, especially the example in the right, exhibit the tendency to converge towards the nose. We take this analysis one step further in Figure 3 (middle and bottom rows) where we show profile and top-down views of the surface wire-frames. In the first column we show the original case. Here the implosion of the nose is clear. The subsequent rows are for $\tau_x = \tau_y = 4$, $\tau_x = \tau_y = 0.4$, respectively. Note in the first column the poor quality of the recovered surface in terms of its similarity to a face. For the hardest threshold (second column), the former problem is solved but the four-quadrant effect becomes obvious. By decrementing τ_x and τ_y the fine features of

[4] The shape-index scale was normalized from 0 to 1.

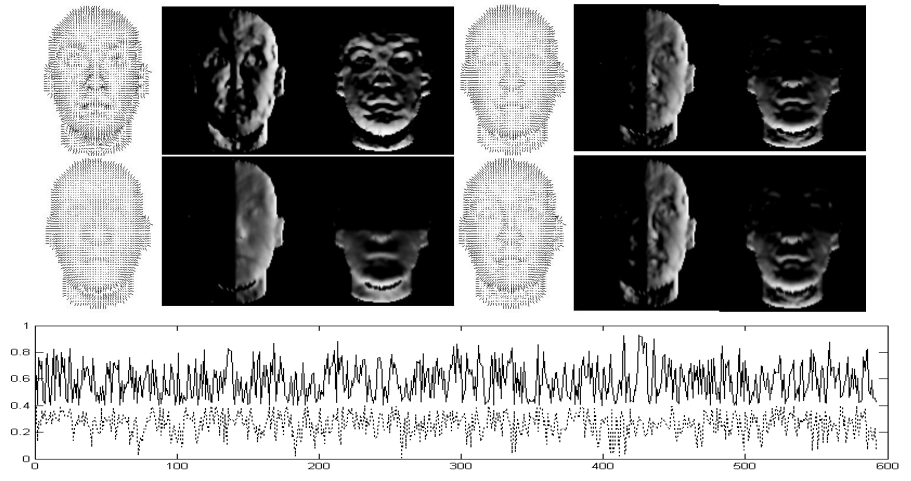


Fig. 2. Gradient analysis and convexity enforcement (see text).

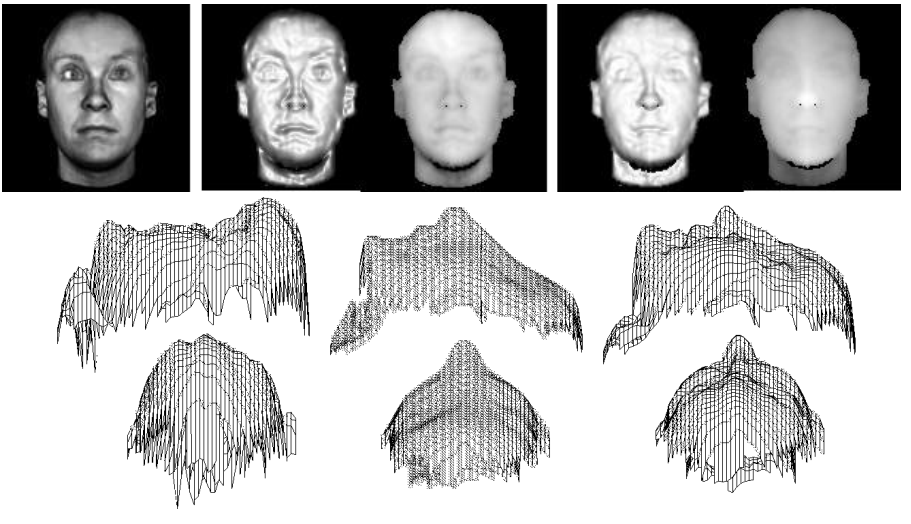


Fig. 3. Illumination, range and surface recovery analysis (see text).

the face seem to be preserved and the overall structure of the face is still sound. By choosing an appropriate threshold we are able to enhance the salient features of a particular face while maintaining the overall face composition.

6 Conclusions

We have proposed an algorithm for correcting a gradient field of a face aiming to generate a height map with a global maximum on a critical point and whose concave regions are constrained to become convex in accordance with shape-index indicators. Although the correction process reported here is effective in correcting feature implosion when the surface gradient is computed using the Worthington and Hancock method, there clearly remains scope for further improvement in computing accurate surface gradients. Of course, the gradient maximum constraint is natural for faces, since the tip of the nose is a global height maximum. However, the constrain could be used for more general surfaces in a local manner for surface height recovery where there are local regions of implosion.

References

1. Atick, J., Griffin, P. and Redlich, N. (1996), Statistical Approach to Shape from Shading: Reconstruction of Three-Dimensional Face Surfaces from Single Two-Dimensional Images, *Neural Computation*, Vol. 8, pp. 1321-1340.
2. Besl, P.J. and Jain, R.C. (1986) Invariant surface characteristics for 3-d object recognition in range images, *Comput. Vision Graphics Image Proc.*, Vol. 33,pp. 33 - 80.
3. Forsythe, D. and Ponce, J. (2001), Computer Vision: a Modern Approach, Prentice-Hall.
4. Frankot, R.T. and Chellapa, R. (1988), A Method for Enforcing Integrability in Shape from Shading Algorithms, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 10, No. 4, pp. 438 - 451.
5. Georghiades, A., Belhumeur, D. and Kriegman, D. (2001), From Few to Many: Illumination Cone Models fro Face Recognition under Variable Lighting and Pose, *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 643 - 660.
6. Horn, B.K.P. (1977), Understanding Image Intensities, *Artificial Intelligence*, Vol. 8, pp. 201-231.
7. B.K.P. Horn and M.J. Brooks (1989), Shape from Shading.*MIT Press, Cambridge, MA*.
8. Koenderink, J.J., and Van Doorn, A.J. (1992), Surface Shape and Curvature Scales, *Image en Vision Computing*, Vol. 10,pp. 557-565.
9. Worthington, P. L. and Hancock, E. R. (1999), New Constraints on Data-closeness and Needle Map Consistency for Shape-from-shading, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, pp. 1250-1267.
10. Zhao, W. and Chellapa, R. (2000), Illumination-insensitive Face Recognition Using Symmetric Shape-from-Shading. *Conference on Computer Vision and Pattern Recognition*, pp. 286-293.
11. R. Zhang, P.S. Tsai, J.E. Cryer and M. Shah (1999), Shape from Shading: A Survey, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8):690–706.

Recovery of Surface Height from Diffuse Polarisation

Gary Atkinson and Edwin Hancock

Department of Computer Science
University of York, York, YO1 5DD, UK.
atkinson@cs.york.ac.uk

Abstract. When unpolarised light is reflected from a smooth dielectric surface, it is spontaneously partially polarised. This process applies to both specular and diffuse reflection, although the effect is greatest for specular reflection. This paper is concerned with exploiting this phenomenon by processing images of smooth dielectric objects to recover surface normals and hence height. The paper presents the underlying physics of polarisation by reflection, starting with the Fresnel equations. It is explained how these equations can be used to obtain the shape of objects and some experimental results are presented to illustrate the usefulness of the theory.

1 Introduction

Polarisation has proved to be a useful source of information in the analysis of light scattering from surfaces in computer vision. There are a number of ways in which polarisation arises and can be used in surface analysis. Perhaps the most familiar example is when the incident light is polarised and the polarisation of the scattered light is analysed [7]. When the scattering process is coherent, then the polarisation of the incident light is preserved, whereas polarisation is destroyed when the scattering process is incoherent. This property can be used to remove specularities from surfaces. However, there are more subtle polarisation effects that can be exploited. For instance, under certain conditions initially unpolarised light becomes polarised as a result of the reflection [10]. This applies to both specular reflection (which we refer to as specular polarisation) and diffuse reflection (diffuse polarisation). This is due to the directionality of the molecular electron charge density interacting with the electromagnetic field of the incident light [2].

There is a considerable amount of literature on the use of polarisation for surface analysis. Most research aimed at extracting and interpreting information from polarisation data, involves placing a linear polarizer in front of a camera and taking images of an object or a scene with the polarizer oriented at different angles [10,4]. Recently however, liquid crystal technology has enabled Wolff and others to acquire sets of polarization images at a rate such that several complete sets of polarisation data can be obtained per second [11]. The polarisation

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 621–628, 2004.
© Springer-Verlag Berlin Heidelberg 2004

cameras that have been developed are particularly useful for outdoor scenes for practical reasons. For laboratory conditions however, the older method is still of use.

Most previous work on polarisation in vision relies on specular polarisation and controlled lighting conditions. Ikeuchi and co-workers [5] placed the object under investigation inside a spherical diffuser, with several light sources outside and a hole for the camera. With this set up, specular reflection occurs across the whole of the object and, since specular reflection is generally much stronger than diffuse reflection, the latter component can be ignored. Other research involved separating reflection components and includes the work of Nayer et al. who used colour [6], Hancock and Ragheb who used a probabilistic framework [7] and Wolff who used polarisation [10]. The relevant theory can then be applied to the specular or diffuse images. This existing work has demonstrated the usefulness of polarisation in surface height recovery, image segmentation and recognition and separation of reflection components.

In contrast to the bulk of previous work, the novel aspect of this paper is an investigation into whether spontaneously acquired *diffuse* polarisation can be used to recover surface shape information. In particular, we aim to explore how it can be used to estimate surface orientation near the limbs of shiny objects. To do this we make use of Fresnel Theory, which relates the reflected and incident wave amplitudes, and this provides a route to estimating the surface normals from polarisation data. The angle of maximum transmission through a polaroid analyser together with the range of measured brightnesses can be used to make estimates of the azimuth and zenith angles of the surface normal. This leads to an algorithm that uses three measured intensities to estimate the surface normal direction. We experiment with the method on real world data, and this demonstrates that measurements of diffuse polarisation can be used to make relatively accurate surface normal estimates, and that these surface normal estimates can be used for realistic surface height recovery.

2 Review of Fresnel Coefficients for Dielectrics

The Fresnel equations give the ratios of reflected wave amplitude to incident wave amplitude for incident light that is linearly polarised perpendicular to, or parallel to, the plane of specular incidence [1]. These ratios depend upon the angle of incidence and the refractive index of the reflecting medium. Since the incident light can be always be resolved into components perpendicular to, and parallel to, the plane of incidence, the Fresnel equations are applicable to all incident polarisation states. Here, it is assumed that the incident light is unpolarised.

For the geometry of Fig. 1 the amplitude reflection coefficients for perpendicularly polarised light, $r_\perp$, and parallel polarised light, $r_\parallel$, at a boundary between two media are given by

$$r_\perp = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t}, \qquad r_\parallel = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_t \cos\theta_i + n_i \cos\theta_t} \tag{1}$$

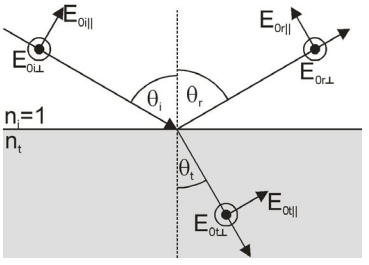


Fig. 1. Definitions with directions of electric fields indicated. N.B. $\theta_i = \theta_r$.

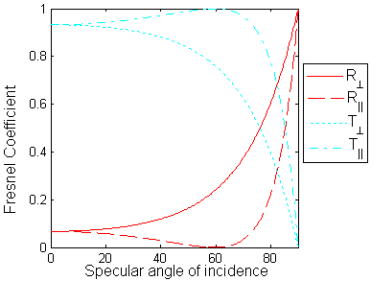


Fig. 2. Reflection and transmission coefficients for a typical dielectric ($n = 1.7$).

where n_i and n_t are the refractive indices of the first and second media and θ_i and θ_t are incident and transmitted angles respectively. θ_t can be obtained from the well-known Snell's Law:

$$n_i \sin\theta_i = n_t \sin\theta_t \tag{2}$$

Cameras measure the intensity of light, which is proportional to the square of the amplitude. With this in mind, it is possible to show [1] that the *intensity coefficients* are $R_\perp = r_\perp^2$ and $R_\parallel = r_\parallel^2$.

Figure 2 shows the Fresnel intensity coefficients for a typical dielectric. At about 60° the parallel component of reflected light is zero. This angle is known as the *Brewster* or *polarising* angle. This figure suggests that there is some information contained in the polarisation state of reflected light. The graphs show that, for specular reflection, more of the light polarised perpendicularly to the plane of incidence is reflected than that parallel to the plane. This means that the reflected light is *partially linearly polarised*, that is, the light is comprised of an unpolarised component, and a completely polarised component.

3 The Polarisation Image

As a polarizer placed in front of the camera is rotated, the measured intensity varies sinusoidally. From Fig. 2, it is clear that the maximum and minimum intensities detected for a particular surface orientation are

$$I_{\max} = \frac{R_\perp}{R_\perp + R_\parallel} I_s \;; \quad I_{\min} = \frac{R_\parallel}{R_\perp + R_\parallel} I_s \tag{3}$$

where I_s is the magnitude of the specular component of reflection (assume for now, that there is no diffuse reflection). The *degree of polarisation* or *partial polarisation*, which is frequently used in computer vision, is defined by

$$\rho = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{4}$$

Each polarisation image, i.e. the full set of polarisation data for a given object or scene, is comprised of three separate images. Firstly, there is the *intensity image*, which is simply the image that would be obtained using a normal camera. Secondly, there is the *phase image* where the intensity encodes the orientation of the polarizer corresponding to maximum transmission through it. The phase is therefore the angle of the linearly polarised component of reflected light. Note that the polarizer cannot distinguish between x and $x + 180°$, resulting in a 180° ambiguity that polarisation data alone cannot solve. The final part of the polarisation image is the degree of polarisation, as determined by (4). To obtain this, three or more images can be taken before fitting the data to a sine wave.

[11] suggests taking three images I_0, I_{45} and I_{90} corresponding to polarizer orientations of 0°, 45° and 90° respectively, and using the following to determine the phase (5) and degree of polarisation (6):

$$\phi = \frac{1}{2}\arctan\left(\frac{I_0 + I_{90} - 2I_{45}}{I_{90} - I_0}\right) + 90°$$

$$\text{if}\,(I_{90} < I_0)\,[\text{if}\,(I_{45} < I_0)\,\phi = \phi + 90°\text{else}\phi = \phi - 90°] \tag{5}$$

$$\rho = \frac{I_{90} - I_0}{(I_{90} + I_0)\cos 2\phi} \tag{6}$$

4 Shape from Polarisation

First consider shape from **specular polarisation**. Following the dichromatic reflectance model [9], the reflected light is a superposition of specular and diffuse components. For specular reflection, as Fig. 2 shows, the reflected light is attenuated more if it is polarised parallel to the specular plane. Thus, greatest transmission through the polarizer occurs when the polarizer is oriented at an angle 90° from the *azimuth* angle of the surface. Orthographic projection is assumed.

The zenith angle, the angle which the normal makes with the viewing direction, can be computed by considering the degree of polarisation. Substituting (3) into (4) gives the degree of polarisation in terms of the Fresnel coefficients. Then applying the Fresnel equations gives ρ in terms of n and the zenith angle, θ. Figure 3a shows the dependence of ρ on θ. Note that at the Brewster angle $\rho = 1$ since the light is totally polarised[1]. Also, there is a 2-to-1 mapping of ρ to θ, resulting an ambiguity that must be solved [5].

Now shape from **diffuse polarisation** is to be considered. The degree of diffuse polarisation is highest, and so most useful, near occluding edges where the zenith angle is large. Diffuse polarisation has the advantage that less controlled lighting conditions are required than for specular polarisation but disadvantages that it is not applicable to metals and is more susceptible to noise.

Diffuse polarisation is a result of the following process [10]: A portion of the incident light penetrates the surface, is partially polarised in the process and is

[1] In practice ρ will be a little less than 1 due to a small diffuse component of reflection.

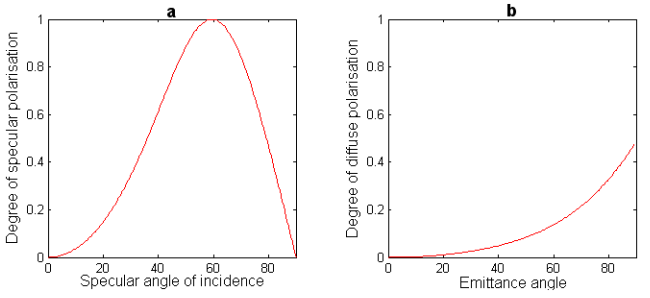


Fig. 3. Degree of (a) specular and (b) diffuse polarisation for a dielectric ($n = 1.7$).

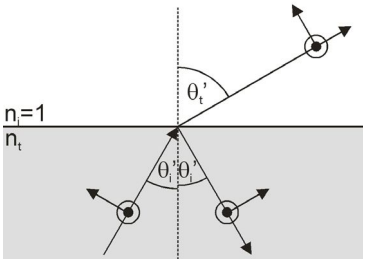


Fig. 4. Transmission of internally scattered light back into air.

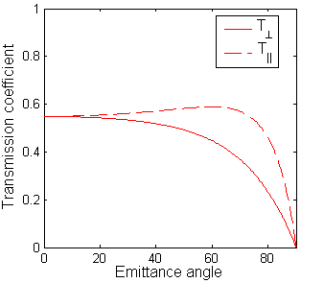


Fig. 5. Fresnel coefficients for Fig. 4 ($n = 1.7$).

internally refracted. Due to the random nature of internal scattering, the light gets depolarised. Some of the light is then refracted back into the air and, once again, is partially polarised.

When light approaches the interface from within the medium, as shown in Fig. 4, a similar process to that discussed earlier takes place but with the relative index of refraction being $1/n$ instead of n. If the internal angle of incidence is above a critical angle ($\arcsin 1/n$) total internal reflection occurs. Otherwise, Snell's Law (2) can be used to find the angle of emittance for any given angle of internal incidence. The Fresnel transmission coefficient[2] can then be calculated for a given *emittance* angle. Figure 5 shows the result of this calculation for a typical dielectric. An additional factor of $1/n$ is introduced due to a difference in wave impedance.

Using (4), the degree of diffuse polarisation is

$$\rho_d = \frac{T_{\parallel}\left(1/n, \theta_i'\right) - T_{\perp}\left(1/n, \theta_i'\right)}{T_{\parallel}\left(1/n, \theta_i'\right) + T_{\perp}\left(1/n, \theta_i'\right)} = \frac{R_{\perp}\left(1/n, \theta_i'\right) - R_{\parallel}\left(1/n, \theta_i'\right)}{2 - R_{\perp}\left(1/n, \theta_i'\right) - R_{\parallel}\left(1/n, \theta_i'\right)} \tag{7}$$

Snell's Law (2) can be used to interchange between internal angle of incidence, and the more useful angle of emittance. Figure 3b shows the degree of diffuse polarisation for different angles using the resulting equation:

[2] The transmission coefficients are simply $T_{\perp} = 1 - R_{\perp}$ and $T_{\parallel} = 1 - R_{\parallel}$

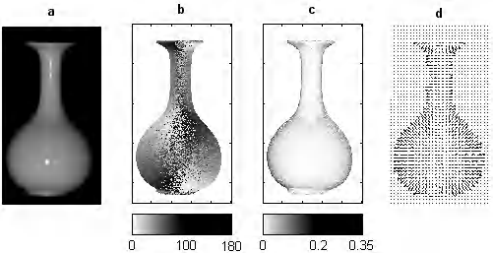


Fig. 6. (a) Greyscale image of porcelain vase. (b) Measured phase and (c) degree of polarisation. Note the noise problem at areas with low degree of polarisation. (d) Needle map (low resolution).

$$\rho_d = \frac{(n - 1/n)^2 \sin^2\theta}{2 - 2n^2 - (n + 1/n)^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}} \tag{8}$$

The zenith angle can be found from (8). The azimuth angle can be determined using exactly the same method as the one used for specular reflection, except that a phase shift of 90° is necessary. The need for a phase shift is illustrated by Fig. 5, which shows that light of parallel polarisation has a higher transmission coefficient and so greater intensity is associated with a polarizer with that orientation, this is in contrast to specular reflection (Fig. 2).

5 Experimental Results

As (5) and (6) show, three images, with the polarizer oriented at 0°, 45° and 90°, are sufficient to obtain the phase and degree of polarisation of a scene. Images were therefore taken, with a linear polarizer placed in front of a camera at these angles, of smooth porcelain objects. The walls of the room and table on which the objects lay were black so that pixels of the images having an intensity below a certain threshold could be treated as background. Figure 6 shows the result of the phase and degree of polarisation computation.

The phase image was then converted into azimuth vectors by taking the sine and cosine at each pixel. The degree of polarisation is converted to zenith angle by numerically solving (8). It is assumed throughout this work that the refractive index is 1.7. After normalising the vectors so they were unit length and applying the disambiguation routine outlined below, the needle map shown in Fig. 6d was obtained.

All of the time the camera was set up such that the diffusely reflected light intensity varies across the whole dynamic range of the camera. A noise reduction technique that we used involved processing pixels by taking median intensities over local neighbourhoods, with central pixels being counted more than once, giving them extra "weighting." This method smoothes the images, without loss of sharp edges and was improved by smoothing over larger neighbourhoods for areas of higher degree of polarisation.

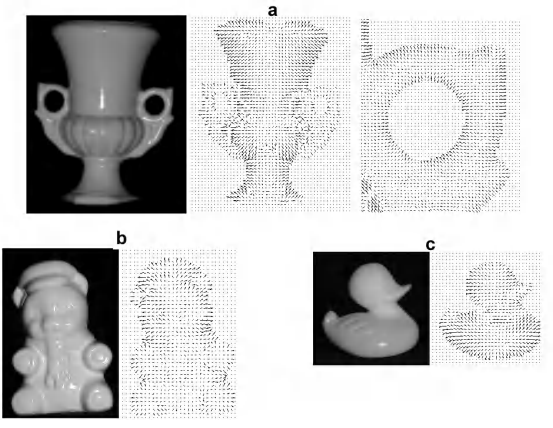


Fig. 7. (a) Needle map of a porcelain urn with close up of handle. On this occasion, four sets of images were used and the average image taken to reduce noise. Also, a few small areas have been disambiguated incorrectly. (b) Porcelain bear. (c) A plastic duck, showing that applicability is not restricted to *perfectly* smooth objects.

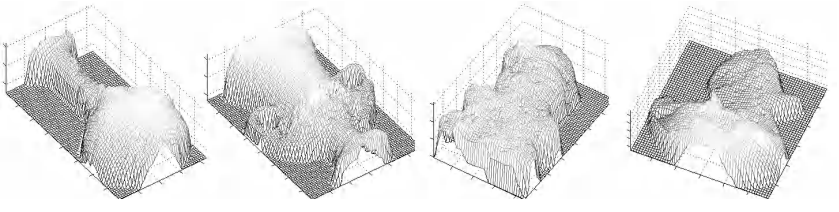


Fig. 8. Depth maps of porcelain vase, urn and bear and plastic duck computed by applying the algorithm of [8] to the needle maps and rotating the objects to different angles.

The next step in processing the images was to disambiguate the azimuth components of surface normals. This was achieved by implementing a method that orders the pixels of the image according to zenith angle with greatest angles listed first. Since we know that the vectors at the occluding contours must be pointing outward, we can propagate into the object using the order calculated previously and assuming no abrupt changes in azimuth angle (except for small zenith angles). This method gave good results for the vase, but variable results far from the limbs of more complex objects such as the urn shown in Fig. 7a.

The final step in processing is to recover the depth from surface normals. This was performed using the algorithm reported in [8], which uses a graph-spectral method. The result of the application of this algorithm is shown in Fig. 8. It is quite clear that the basic shapes have been recovered, but with noise being problematic at small zenith angles.

6 Conclusion

This study of shape from diffuse polarisation has accomplished two tasks. Firstly, it has shown that there is potential for using diffuse polarisation in geometry analysis, despite the common assumption that the difference between reflected polarisation components is negligible for diffuse reflection. Secondly however, it illustrates the fact that away from the occluding contours, diffuse polarisation is limited in applicability. The depth map of Fig. 8a is very accurate, but only due to intense smoothing techniques that would distort more complex objects, although adaptive smoothing techniques [3] improve results significantly. Diffuse polarisation is best used therefore, in conjunction with other techniques such as shape from shading or shape from specular polarisation.

Other limitations of shape from diffuse polarisation include the non-trivial disambiguation routine and the requirement of three images (although, as mentioned earlier, polarisation cameras are now removing this requirement [11]). The main advantages are that the algorithms are very reliable at obtaining azimuth and zenith angles from images of smooth dielectrics close to occluding contours and that no light source information needs to be known.

References

1. Born, M. and Wolf, E. *Principles of Optics*. New York: Pergamon, 1959.
2. Hecht, E. *Optics*, 3rd ed. Addison Wesley Longman, 1998.
3. Kervrann, C. "An Adaptive Window Approach for Image Smoothing and Structures Preserving." *Proceedings of European Conference on Computer Vision (ECCV)* (3), pp132–144, 2004.
4. Miyazaki, D., Tan, R. T., Hara, K. and Ikeuchi, K. "Polarisation-based Inverse Rendering from a Single View." *Proceedings of International Conference on Computer Vision (ICCV)*, Nice, France, vol. 2, pp.982–987, 2003.
5. Miyazaki, D., Kagesawa, M. and Ikeuchi, K. "Transparent Surface Modelling from a Pair of Polarisation Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, No. 1, 2004.
6. Nayar, S. K., Fang, X. and Boult, T. "Separation of Reflection Components Using Colour and Polarisation." *International Journal of Computer Vision (IJCV)*, 21(3), pp163–186, 1997.
7. Ragheb, H. and Hancock, E. R. "A Probabilistic Framework for Specular Shape from Shading." *Pattern Recognition*, vol. 36, no. 2, pp407–427, 2003.
8. Robles-Kelly, A. and Hancock, E. R. "A Graph-Spectral Method for Surface Height Recovery from Needle-maps." *Computer Vision and Pattern Recognition*, Kauai, Hawaii, vol. 1, pp141–148, 2001.
9. Shafer, S. "Using Colour to Separate Reflection Components," *Colour Research Applications*, vol. 10, pp. 210–218, 1985.
10. Wolff, L. B. and Boult, T. E. "Constraining Object Features Using a Polarisation Reflectance Model." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, No. 7, pp635–657, 1991.
11. Wolff, L. B. "Polarisation Vision: A New Sensory Approach to Image Understanding." *Image and Vision Computing*, 15, pp81–93, 1997.

Vectorization-Free Reconstruction of 3D CAD Models from Paper Drawings

Frank Ditrich, Herbert Suesse, and Klaus Voss

Friedrich-Schiller-University Jena, Department of Computer Science
Ernst-Abbe-Platz 1-4
D-07743 Jena, Germany
ditrich@minet.uni-jena.de, {nbs,nkv}@uni-jena.de
http://pandora.inf.uni-jena.de

Abstract. We propose a new approach for the reconstruction of 3D CAD models from paper drawings. Our method uses a combination of the well-known fleshing-out-projections method and accumulation techniques from image processing to reconstruct part models. It should provide a comfortable method to handle inaccuracies and missing elements unavoidable in scanned paper drawings while giving the user the chance to observe and interactively control the reconstruction process.

Keywords: CAD, Drawing Interpretation, 3D Reconstruction, Recognition, Applications

1 Introduction

In the past a lot of research and development work was done to find solutions for the challenging task of reconstructing 3D CAD models from 2D drawings given either on paper or in a machine-readable format like DXF.

Two main techniques can be found in the literature: The first is the fleshing-out-projections principle published by Wesley and Markovsky [8] which delivers surface models. Initially designed to handle solids with planar surfaces, it was later extended by several authors to solids with curved surfaces (e. .g [2,10]) and to handle incorrect projections with missing or additional lines (see [4]).

The second technique are algorithms which focus on the extraction of manufacturing features[1] directly from the 2D input data (e. g. [1,5,7]).

Both require high-quality complete vector data. Therefore vectorization techniques are used to convert bitmap data into vector data, but this is a very complicated process which requires a great amount of user interaction and parameter adjustment to get satisfying results. Nevertheless it is practically impossible to

[1] The term "feature" is well-known in the context of image processing and pattern recognition. However, it is also used in computer-aided design and manufacturing to denote basic construction or manufacturing elements like solids made by sweeping a profile along a line or around an axis or shapes resulting from milling or drilling.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 629–637, 2004.
© Springer-Verlag Berlin Heidelberg 2004

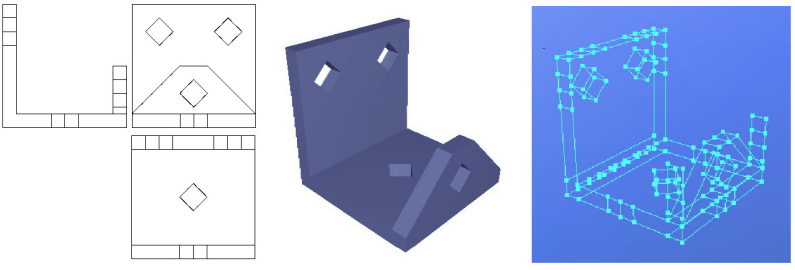


Fig. 1. A sample part with three projections and its pseudo-wireframe.

extract exact data from paper drawings since the drawings themselves are imperfect. The complete description of a part can only be obtained by interpreting the geometric sketches together with the textual information like measures or symbols.

Another group of methods wich are of interest in this context perform the extraction of manufacturing features directly from 3D boundary-representation models (for example from IGES or STEP files) [3,6].

Our solution tries to overcome the difficulties associated with the vectorization by directly combining the 2D bitmap data in 3D space and extracting manufacturing features from this generated data. The motivation behind this approach is to try to recognize objects on the drawing "as a whole", and thereby compensate for problems and errors induced by a poor drawing quality, just like a human observer would do. For example, we try to recognize an axis-parallel box directly from the scanned views and not from some possibly broken or otherwise disturbed lines produced by firstly vectorizing the drawing.

2 Basic Idea

In Fig. 1a,b a sample 3D part with three projections (left, top, front) is shown. In these projections all edges are drawn as solid lines regardless of visibility. The first stage of the fleshing-out-projections algorithm [8] is applied to these projections to construct a so-called pseudo-wireframe also shown in Fig. 1c. This pseudo-wireframe is different from the wireframe of the original part because it contains some additional vertices and edges, which are discarded in later stages of the algorithm. But to the human observer there is a great amount of similarity between the part and the pseudo-wireframe. The basic structure and shape of the part is clearly visible.

The basic idea behind the creation of the pseudo-wireframe is to combine the projections in 3D space to get candidates for vertices and edges. A somewhat similar technique can be applied also to the bitmap data instead of the vector data: Fig. 2a shows scanned pencil drawings of the three projections with some

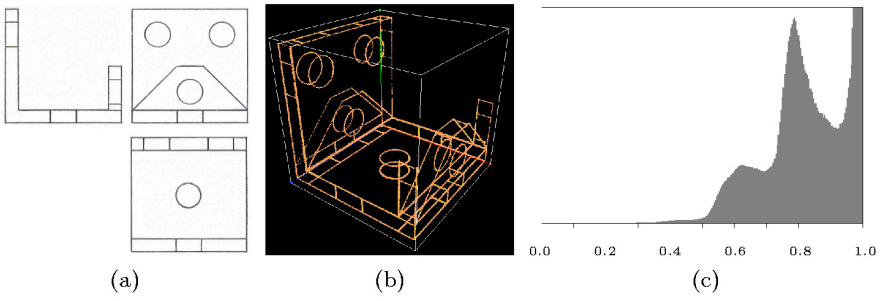


Fig. 2. Three projections of a sample part scanned from paper, the resulting voxel model (intensities from 0.0 to 0.45) and its histogram, scaled by factor 50 in y-direction.

perturbations and changes in the quality of lines as they usually appear in CAD drawings. Now the following is done: The projections are mapped to the XY, XZ and YZ planes of a coordinate system in a 3D lattice so that every projection is properly aligned with each other. That means that for example the X coordinates of the XY and XZ projections of a 3D point are the same. Now we create a 3D voxel structure from these projections by adding the intensities of corresponding projection image pixels (and normalizing the resulting intensity range to $[0.0, 1.0]$ through division by the maximum value afterwards).

Fig. 2c shows the resulting histogram of voxel intensities for our sample part (scaled in y-direction by factor 50). We distinguish four ranges: 0 to 0.5, 0.5 to 0.7, 0.7 to 0.9 and 0.9 to 1. They can be interpreted as containing different elements of the voxel model: Suppose our images would only contain black lines (intensity 0) on a white background (intensity 1). By adding the pixel intensities and performing a normalization as described above we would get four different voxel intensities: 0, $\frac{1}{3}$, $\frac{2}{3}$ and 1. Obviously we would have a small number of voxels with intensity 0 and a very large number with intensity 1 (for this reason the histogram has to be scaled to make the ranges visible). If we take the voxels from the first range, we get the structure shown in Fig. 2b. To get a better quality for later recognition tasks, here we took the range 0 to 0.45. Such a correction could easily be done manually by the user according to his visual impression.

Obviously there is a remarkable similarity between the pseudo-wireframe constructed above and the resulting voxel structure. Indeed the following can be shown: If we have projections of a part, then two voxel sets can be obtained: The first by creating the pseudo-wireframe and rastering its edges into a 3D voxel space (intensities 0 and 1, painting a voxel black if it is intersected by a line), the second by rasterizing the projections into 2D bitmaps and creating a 3D voxel model as described above. If we extract the sets of black voxels V_1 and V_2 from both models, then V_1 is a subset of V_2.

By using this method it should be possible to get 3D data to which real 3D features can be fitted without an intermediary vectorization which can introduce an additional loss of information. Besides this, the voxel model could give a per-

son whose task is to make a 3D part from a drawing a first impression of the structure of the part and could be used as a component of a comfortable user interface to a feature recognition system working on this data. The user gets the ability to interact with and guide the system supporting his work. This is useful because it seems quite sure that a fully automatic recognition/reconstruction system cannot be built due to the great variety of paper drawings and the problems concerning the quality of drawings.

One practical problem should be mentioned here: For processing larger scanned drawings a great amount of memory would be necessary if we would store the voxels in a three-dimensional array. But this can be overcome by first determining the interesting range from the histogram (which can be calculated without storing all voxels) and afterwards storing only those voxels which lie in this range in an appropriate data structure.

3 Fitting Features Using Accumulation

To fit features in the 3D space we use an accumulation technique from image processing. Its basic principle is known as *randomized Hough transform* (*RHT*, see [9]). In contrast to the (conventional) Hough transform as it is used for example to dectect lines in a 2D set of points, which calculates and accumulates several parameter tuples for every single point (in [9] this is called a *diverging mapping*) the RHT determines and accumulates one parameter tuple for chosen sets of multiple points (*converging mapping*). Using this method the peaks in the accumulator space become significantly sharper (for the above mentioned case of detecting lines the RHT accumulator contains the squared values of the Hough accumulator). Furthermore, in [9] it is shown that it is not necessary to accumulate over all possible point subsets, even the accumulation over an adequate number of randomly chosen subsets gives sharp peaks, which decreases the running time.

The RHT method can be generalized to other objects which can be described by a set of parameters. The drawback is that if we have some more parameters we also need a larger accumulation space in which only small regions are really used. One possible solution to overcome this problem is doing the accumulation not in the accumulation space itself but in lower-dimensional projections of it.

The drawback of this variant is the loss of information caused by the projection process to lower-dimensional parameter spaces. It can be very hard resp. impossible to find the right peaks in these projections since there can be pseudo-peaks which arise simply through adding all hits along a projection line instead of being produced by existing objects.

Another solution to the above mentioned problem is a clustering technique: Instead of using an accumulator array we maintain a list of clusters. Our algorithm for this accumulation process (using the random selection of points) is given in the following. It is controlled by three functions and three values (we denote with $\mathbf{p}$ a parameter tuple $(p_1, \ldots, p_k)$):

- A test function $t(\mathbf{p})$ with the return values "valid" and "invalid", which checks if a tuple $\mathbf{p}$ is a valid parameter tuple for an object we search for,
- a quality function $\mu(\mathbf{p})$ which determines a properly defined quality measure for the object determined by $\mathbf{p}$,
- a distance function $d(\mathbf{p}, \mathbf{q})$ which determines the distance of two parameter tuples,
- a number $step_{max}$ which gives the maximum number of point tuples to be chosen,
- a minimum quality measure μ_{min} which gives a minimum quality for an object to be accumulated,
- a minimum distance d_{min} to consider two parameter tuples to be different.

The algorithm gets a set of points as input and delivers a list of accumulated parameter tuples (together with a weight and quality for each). Especially the quality function μ is of great importance, since it constrains the number of clusters being created during the accumulation process. Since for every parameter tuple created for a randomly selected point set the difference to every existing cluster has to be calculated, the running time of the algorithm would make it unsuitable for practical application if there were no criterion to eliminate unusable clusters. In the above example for accumulating lines such a quality function could count the number of points near the line described by two parameters.

1. Set $step = 0$.
2. Increment $step$ by 1. If $step > step_{max}$ the algorithm is completed.
3. Otherwise choose a tuple of n points randomly and calculate the parameter tuple $\mathbf{p}$.
4. If the test function $t(\mathbf{p})$ yields "invalid" go to 2.
5. If the quality function $\mu(\mathbf{p})$ yields a value lower than a minimum quality μ_{min} then go to 2.
6. If the list of accumulation clusters is empty then initialize the list with $\mathbf{p}$ as the first element (with weight 1) and go to 2.
7. Search the accumulation list for a tuple $\mathbf{q}$ with a distance $d(\mathbf{p}, \mathbf{q}) < d_{min}$. If there are several such tuples, choose one with a minimum value $d(\mathbf{p}, \mathbf{q})$.
8. If such a tuple $\mathbf{q}$ is found (we denote its weight with $w_{\mathbf{q}}$) then replace it by $\mathbf{q}' = (w_{\mathbf{q}}\mathbf{q} + \mathbf{p})/(w_{\mathbf{q}} + 1)$ with the new weight $w_{\mathbf{q}'} = w_{\mathbf{q}} + 1$ and go to 2.
9. If no such tuple is found then add $\mathbf{p}$ with the weight 1 as a new tuple to the accumulation list and continue with 2.

3.1 Boxes

With our accumulation method not only geometric primitives but also complex shapes can be found, for example axis-parallel rectangular boxes described by six parameters $\mathbf{p} = (x_{min}, x_{max}, y_{min}, y_{max}, z_{min}, z_{max})$. We search for boxes the edges of which are contained in our voxel data. To achieve this we define the quality function in an appropriate manner: If we have a box described by the six parameters we divide every edge in a number of cells and check how many

percent of these cells contain data points (here we use 10 cells per edge). As a distance between two tuples we simply use the maximum distance between corresponding parameters:

$$d(\mathbf{p}, \mathbf{q}) = \max_{i=1,\ldots,6} |p_i - q_i| \ .$$

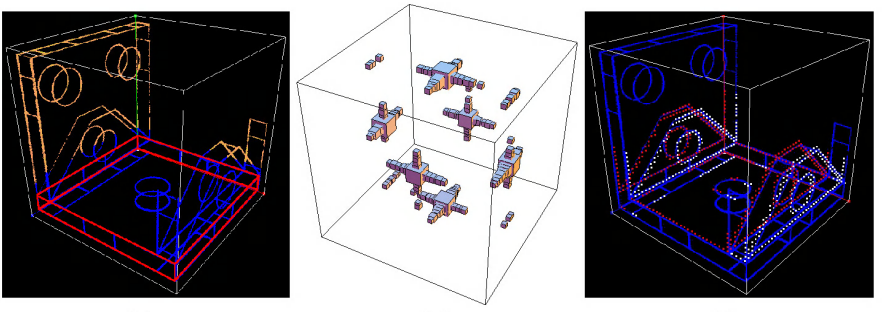
To get the six parameters we simply choose n points and determine their bounding box. The choice of the number n influences the recognition quality and should depend on the kind of data we have: If there is a point set which consists only of one box (and our task is only to determine its parameters) we should choose a higher number of points to increase the chance to hit the whole box after a few iterations. But if there is additional noise in the data (a lot of separate points not being part of the box) we should choose a smaller n since the greater n is, the greater is the chance to choose some of the wrong points, so there will be a smaller chance to get the correct box.

We carried out some experiments and found out that 4, 5, and 6 are suitable values for n for our purposes. In the following we used $n = 6$.

Besides the presence of noise, another problem is the correct recognition of several objects. If we have a scene with k boxes the probability of choosing n points belonging to the same box is $\left(\frac{1}{k}\right)^{n-1}$ (if chosing the same point multiple times is allowed), which significantly decreases with increasing k. Nevertheless it makes sense to apply our method to the voxel data: We do not try to find all boxes contained in the data automatically but give the user the chance to direct the recognition process. In our software prototype it is possible to select a region of voxels to be used for recognition. Here we use the fact that a user is able to get a principal impression of the model structure and can select interesting regions containing boxes. After the accumulation process is done the found boxes are presented to the user and he can iterate through them (they can be sorted by their qualities or the weights of their clusters) and choose the right ones. A sample is shown in Fig. 3a.

3.2 Cylinders

Another class of objects which can be found using the accumulation method are axis-parallel cylinders. They appear in the voxel structure as two circles on parallel planes (assumed the planar faces of the part are also axis-parallel). To describe them we use six parameters: The coordinate direction which is parallel to the axis of the cylinder (p_1), two coordinates to describe the position of the axis in a plane perpendicular to the axis direction (p_2, p_3), the radius (p_4) and two values for the coordinate range the cylinder covers along the axis direction (p_5, p_6). Of course we have to slightly modify our list accumulation algorithm to handle the first parameter correctly: It is a flag describing one of three directions, so the new cluster cannot be calculated as described above, instead our distance function must guarantee that tuples with different directions are never collected in the same cluster. Here the distance between two parameter tuples



(a) (b) (c)

Fig. 3. A recognized box in a selected subset of the voxels from Fig. 2b (see Section 3.1), the accumulator content after accumulating translations for an axis-parallel box and a recognized translation (see Section 3.3).

$\mathbf{p} = (p_1, \ldots, p_6)$ and $\mathbf{q} = (q_1, \ldots, q_6)$ is defined as

$$d(\mathbf{p}, \mathbf{q}) = \begin{cases} \infty & \text{if } p_1 \neq q_1 \\ \max\limits_{i=2,\ldots,6} |p_i - q_i| & \text{otherwise} \end{cases} .$$

For the calculation of parameter tuples we choose three points and calculate one possible tuple for any of the three possible axis directions. Suppose the axis direction is given it is easy to determine the position of the axis and the radius. As the covered range we use the appropriate minimum and maximum coordinate values of the three points.

To determine the quality of a parameter tuple we also divide the two circles into cells and count how many percent of them contain points (in our prototype we use 20 cells for each).

3.3 Extrusions

The accumulation cannot only be used to detect shapes but also to find transformations between point sets. This can be used for example to find features which are created by a translational sweep of a profile (usually called *extrusion* or *extrusion feature*). To find the corresponding translations we use a three-dimensional accumulator and accumulate for every pair of distinct points the translation which maps one point onto the other (we do not choose the point pairs randomly). If the voxel structure contains profiles which are "connected" through a translation we should get corresponding peaks in the accumulator.

But there is one problem: The edges contained in the voxel structure would also produce peaks since there are a lot of pairs with points lying on the same edge with the same distance. To solve this problem we should check if there are a lot of points near the line connecting the two points. An appropriate data structure can easily be built from the voxel set in a preprocessing step.

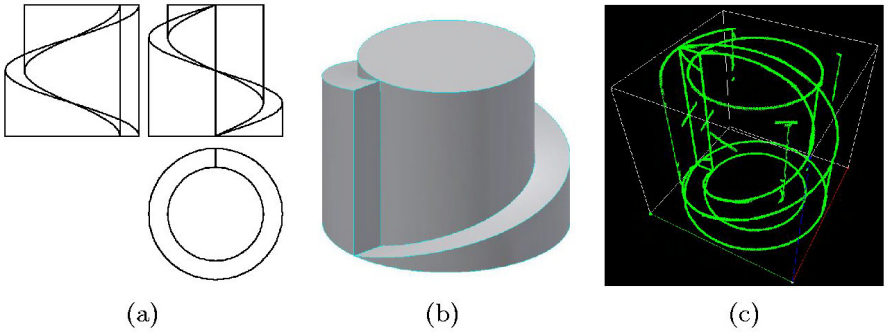


Fig. 4. Another sample part with its projections and the resulting voxel structure.

The list accumulation described above is not so well-suited for this purpose since a large number of clusters will be generated which increases the running time of the algorithm.

Fig. 3b shows the content of an accumulator for an axis-parallel box, the size of the cubes illustrates the number of hits in the cells. As expected there are prominent peaks for the six translations connecting the three pairs of rectangles forming the boundary of the box. In Fig. 3c the sample part with a detected translation is shown. Since arbitrary translations are allowed there are a lot of translations besides the six most prominent which have also a significant number of hits in the accumulator.

Of course we can limit to searching for axis-parallel translations, which could be found using three one-dimensional accumulators and would restrict the set of pairs which need to be checked.

The computational effort raised through the checking of all pairs can also be decreased by equally thinning the voxel set, if the resulting accuracy is sufficient.

If some translations are detected for every translation the two sets of points which are mapped onto the other by this translation can be built and visualized in the user interface (see Fig. 3c). The sets can be further processed for example by determining a supporting plane and extracting contours which can be used as a base for an extrusion feature.

4 Conclusion and Further Work

We presented a new idea for the reconstruction of three-dimensional CAD models from paper drawings. It could be a useful technique to support the solution of problems arising while handling real-world paper drawings (with their known quality problems which often lead to unsatisfying vectorization results).

Finally we like to mention some problems and give some prospects on further work. One very important problem which needs to be solved is the correct alignment of the three views before adding their intensities to get the voxel data. Perhaps here also a semi-automatic solution involving user interaction can be

found. Of course it is possible to extend the set of objects which can be found using the accumulation technique described above. Fig. 4 shows a part which contains "coiled" edges, which correctly appear in the voxel model. Here also fitting techniques could be applied to the 3D data to reconstruct them.

The correct placement of cross sections within the voxel data is another extension to our paradigm. A further important task is the automatic selection of interesting voxel subsets on which the accumulation procedure will be applied. Currently this has to be done by the user. A first simple solution could be the search for connected components in the voxel data, since holes appear in most cases separated from the other edges of the part. Ideally the system would be able to find those regions automatically and to extract the contained features. Then all possible combinations of the found features using boolean operations could be built to create the possible parts. Perhaps this set could be constrained using geometric properties like for example the fact that an isolated cylinder must always be a hole. From each of these generated parts the 2D views could be derived and compared to the original ones using an appropriate similarity measure.

The relationship and degree of similarity between the voxel model and the pseudo-wireframe resp. between the sets V_1 and V_2 from Section 2 is also an interesting subject to further theoretical investigations.

References

1. Ganesan, R.; Devarajan, V.: *Intersecting features extraction from 2D orthographic projections.* Computer-Aided Design 30(11), pp. 863-873, 1998.
2. Kuo, M. H.: *Reconstruction of Quadric Surface Solids from Three-View Engineering Drawings.* Computer Aided Design, 30(7), pp. 517-527, 1998.
3. Mäntylä, M.; Nau, D.; Shah, J.: *Challenges in Feature-Based Manufacturing Research.* Communications of the ACM, 39(2), pp. 77-85, 1996.
4. Masuda, H.; Numao, M.: *A Cell-Based Approach for Generating Solid Objects from Orthographic Projections.* Computer-Aided Design 29(3), pp. 177-188, 1997.
5. Meeran, S.; Taib, J.M.: *A Generic Approach to Recognising Isolated, Nested, and Interacting Features from 2D Drawings.* Computer Aided Design 31(14), pp. 891-910, 1999.
6. Requicha, A.A.G.: *Geometric Reasoning for Intelligent Manufacturing.* Communications of the ACM, 39(2), pp. 71-76, 1996.
7. Shum, S. S. P.; Lau, W. S.; Yuen, M. M. F.; Yu, K. M.: *Solid Reconstruction from Orthographic Views Using 2-Stage Extrusion.* Computer-Aided Design 33(1), pp. 91-102, 2001.
8. Wesley, M. A.; Markowsky, G.: *Fleshing out projections.* IBM Journal of Research and Development, 25(6), pp. 934-954, November 1981.
9. Xu, L.; Oja, E.: *Randomized Hough Transform (RHT): Basic Mechanisms, Algorithms, and Computational Complexities.* CVGIP: Image Understanding, 57(2), pp. 131-154, March 1993.
10. Zhang, A.; Xue, Y.; Sun, X.; Hu, Y.; Luo, Y.; Wang, Y.; Zhong, S.; Wang, J.; Tang, J.: *Reconstruction of 3D Curvilinear Wireframe Model from 2D Orthographic Views.* Lecture Notes in Computer Science, Vol. 3037, pp. 404-411, 2004.

Plane Segmentation from Two Views in Reciprocal-Polar Image Space

Zezhi Chen, Nick E. Pears, Bojian Liang, and John McDermid

Department of Computer Science, University of York, YO10 5DD, UK
{chen,nep,bojian,jam}@cs.york.ac.uk

Abstract. We present a new method of segmenting planar regions when an uncalibrated camera undergoes (near) pure translation. We show that, for pure translation parallel to a plane, the relation between the two views, when expressed in a reciprocal-polar $(\frac{1}{r}, \theta)$ space, is a pure shift in the $(\frac{1}{r})$ dimension for a given value of θ. Furthermore, we show that the magnitude of these shifts follows a sinusoidal form along the θ direction over a maximum of π radians. This allows planar image motion to be accurately detected and recovered by 1D correlation. Simultaneous planar pixel grouping and recovery of the plane homography, thus amounts to a robustly fitting a sinusoid to shifts of maximum correlation in reciprocal-polar space. The phase of the recovered sinusoid corresponds to the orientation of the vanishing line of the plane and the amplitude is related to the magnitude of the camera translation.

1 Introduction

In this work[1], a new method of segmenting planar regions when a camera undergoes pure translation is presented. It has been demonstrated as a practical approach to monocular mobile robot obstacle detection or, equivalently, mobile robot ground plane segmentation. The essence of the paper is the reciprocal-polar image (and the use of sinusoid models to segment planes within that space), which is new, although related ideas have been presented in the literature. For example, Pollefeys et al [3] suggested a polar rectification (r, θ) to aid stereo matching. Wolberg and Zokai [4] have used the well-known log-polar transformation to aid affine motion recovery. Both of these allow more general motion than our transformation, but do not give the main benefit of the reciprocal-polar transformation, which allows correlation based matching over large camera motions. The use of intensity correlation means that there can be low density (or even zero) corner features on the plane of interest (eg ground plane). Thus our method gives good performance compared to other plane segmentation schemes particularly when the camera motion is large and the density of corner features on the plane of interest is low. We summarise the advantages as follows:

[1] The authors acknowledge the support of the UK DTI Aeronautics Research Programme and the EPSRC

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 638–646, 2004.
© Springer-Verlag Berlin Heidelberg 2004

1. It renders image motion, for a given radial direction on a planar surface, into a pure shift. This allows image motion to be recovered by direct correlation methods (in addition to the usual feature correspondences), which fail in the original image space for large image motions, due to perspective distortion. In reciprocal-polar space, for a given epipolar line (i.e radial direction relative to the focus of expansion), distant points on a plane (i.e. points close to the focus of expansion) move the same amount as points on the same plane, which are close to the camera.
2. The amount of radial image motion expressed in reciprocal-polar space, follows a sinusoidal variation, where the phase of the sinusoid is the orientation of the vanishing line of the plane (plane horizon) and the magnitude of the sinusoid is dependent on the magnitude of the translation, camera intrinsic parameters and the proximity of the plane. This allows planar grouping down to pixel level, using a simple sinusoid model, given that there is some local texture around the pixel. In terms of (robust) least squares extraction of models (here, homographies), you get better solutions if you can use more data associated to that model. This is why our method can improve on methods based purely on feature matching approaches, particularly in scenes where the features on the plane of interest (ground plane) are sparse.
3. The method works even in the absence of feature correspondences on the ground plane, although feature correspondences are not precluded from being classified as ground plane or obstacle using the sinusoid model. Indeed, in the absence of local texture, we cannot segment on a pixel basis, and contours of smooth regions can be matched (using the focus of expansion, or epipole) and classified using the sinusoid model.
4. Compared to other approaches in mobile robot visual navigation, it is simple to distinguish the ground plane from other planes in the scene, using some a-priori knowledge of the camera roll orientation with respect to the ground plane. Usually the roll angle is very close to zero, so the vanishing line and phase of the sinusoid are close to zero. Planes at other orientations relative to the ground plane have different phase sinusoids. Planes parallel to the ground plane (eg tops of obstacles) and below the camera have the same phase but all have a larger amplitude sinusoid than the ground plane. Planes above the camera and parallel to the ground plane (eg the ceiling, if visible) have a phase shift of π relative to the ground plane's sinusoid model.

2 The Relation Between Planar Homography (H) and Fundamental Matrix (F) Under Pure Translation

In pure translation of the camera, one may consider an equivalent situation in which the camera is stationary, and the world undergoes a translation -$\mathbf{t}$. In this situation, points in 3D space move on straight lines parallel to $\mathbf{t}$, and the imaged intersection of this parallel lines is the focus of expansion (FOE) $\mathbf{v}$ in the direction of $\mathbf{t}$. It is evident that $\mathbf{v}$ is also the epipole $\mathbf{e}$ and $\mathbf{e}'$ for both views, and the imaged parallel lines are epipolar lines.

$$\mathbf{F} = [\mathbf{e}']_\times = [\mathbf{e}]_\times = [\mathbf{v}]_\times. \tag{1}$$

where $[\bullet]_\times$ is a skew-symmetric matrix corresponding to the vector.

Let $\mathbf{X}_i$ be a set of points which are coplanar in 3-D Euclidean space. The images of $\mathbf{X}_i$ from two view points are related by a plane to plane projectivity or homography $\mathbf{H}$, such that, $\lambda \mathbf{x}_{i2} = \mathbf{H}\mathbf{x}_{i1}$. where λ is a scalar, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are homogenous image coordinates of the images of point $\mathbf{X}_i$, $\mathbf{H}$ is a 3×3 matrix representing the homography. Note that two corresponding point pairs fully determine the H-matrix under pure translation parallel to the plane. Suppose the cameras are calibrated with the world origin at the first camera and the intrinsic parameters constant.

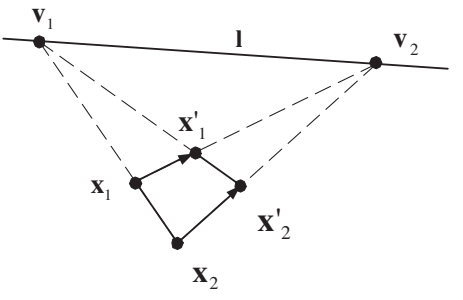


Fig. 1. Two corresponding point pairs fully define the FOE and vanishing line

$$\mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}] \qquad \mathbf{P}' = \mathbf{K}[\mathbf{R}|\mathbf{T}]. \tag{2}$$

and the world plane π_E has coordinates $\pi_E = (\mathbf{n}^T, d)^T$ so that $\mathbf{H} = \mathbf{K}(\mathbf{R} - \lambda \mathbf{t}\mathbf{n}^T)\mathbf{K}^{-1}$. Where $\lambda = \frac{1}{d}$, For a pure translation, $\mathbf{R}$=$\mathbf{I}$, and so $\mathbf{H}$ has the form $\mathbf{H} = \mathbf{I} - \lambda(\mathbf{K}\mathbf{t})(\mathbf{K}^{-T}\mathbf{n})^T$. We note that $\mathbf{K}\mathbf{t}$ is the FOE $\mathbf{v} = \eta(x_f \quad y_f \quad 1)^T$, and $\mathbf{K}^{-T}\mathbf{n}$ is the vanishing line $\mathbf{l} = v(a_v \quad b_v \quad 1)^T$ corresponding to plane π_E . Thus we have

$$\mathbf{H} = \mathbf{I} - k\mathbf{v}\mathbf{l}^T. \tag{3}$$

Where k is a constant scalar. Because two corresponding point pairs fully define the FOE and vanishing line, so H-matrix can, in theory, also be fully determined by the two corresponding matches (Fig. 1). However, in our robust method, we recover the FOE using the algorithm given in the following section. We then determine the vanishing line orientation using the phase of a sinusoid fitted to feature matches and/or correlation maxima in reciprocal-polar space, using RANSAC and least-squares (SVD).

3 Recovering the Focus of Expansion

We can detect (near) pure translation by intersecting all lines defined by *all* corner correspondences from the image pair and if a large percentage of intersections lie in a small area (such as 90% of intersections should lie within a 50 pixels radius), then pure translation is assumed and the FOE is computed. The question now is: how can we calculate the FOE with high accuracy and high stability? We have developed a robust method which uses a corner matching procedure and then uses random sample consensus (RANSAC) and standard linear methods (SVD) to solve the overdetermined system of linear equations $\mathbf{x}_i \times \mathbf{x}'_i) \bullet \mathbf{v} = 0$, where $\mathbf{x}_i \longleftrightarrow \mathbf{x}'_i$ $(i = 1, 2, \cdots, n)$ is any pair of matching points

in two images and $\mathbf{v}$ is the required FOE. After this linear process an iterative procedure is used to minimise the epipolar distance, $f(\mathbf{v})$, given by:

$$f(\mathbf{v}) = \left(\frac{1}{\sqrt{(\mathbf{F}\mathbf{x}_i)_1^2 + (\mathbf{F}\mathbf{x}_i)_2^2}} + \frac{1}{\sqrt{(\mathbf{F}^T\mathbf{x}'_i)_1^2 + (\mathbf{F}^T\mathbf{x}'_i)_2^2}} \right) \left| \mathbf{x}_i'^T \mathbf{F} \mathbf{x}_i \right|$$

where $\mathbf{F}$ is the fundamental matrix.

4 Plane Segmentation and Vanishing Line Recovery

Once the FOE has been computed, we shift image coordinates so that each image is centred on the FOE, using a centering tranlslation matrix, $\mathbf{T}_c$. The FOE is then at homogenous coordinates $\mathbf{v}' = (0, 0, 1)^T$ and vanishing line becomes $\mathbf{l}' = \mathbf{T}_c^{-T}\mathbf{l} = (a_v, b_v, \mathbf{v}^T\mathbf{l})^T$. The homography relating points in FOE centred coordinates is

$$\mathbf{H}' = \mathbf{I} - k\mathbf{v}'\mathbf{l}'^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -ka_v & -kb_v & (1 - k\mathbf{v}^T\mathbf{l}) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ s & \mu & q \end{bmatrix}. \tag{4}$$

where $q = 1 - k\mathbf{v}^T\mathbf{l}$, $s = -ka_v$ and $\mu = -kb_v$. If the robots translation direction is parallel to the ground, $q = 1$, since the FOE lies on the vanishing line for this special motion. In this specialisation, original H-matrix has four *dof*, and is sometimes termed an elation[2]. Under the new homography, we have

$$(sx'_1 + \mu y'_1 + q)^2 (x'^2_2 + y'^2_2) = (x'^2_1 + y'^2_1). \tag{5}$$

If we define $\rho = \frac{1}{r}$, where r is the Euclidean distance between an image point and the FOE in a frame, then taking square roots of Eq.(5)

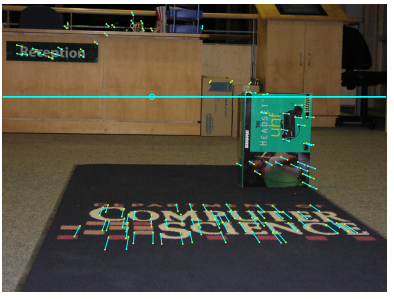
$$f(\theta) = \rho_2 - q\rho_1 = s\frac{x'_1}{r_1} + \mu\frac{y'_1}{r_1} = k_{s\mu}\sin(\theta + \alpha). \tag{6}$$

where θ is the angular position of a pixel in a frame centred on the FOE.

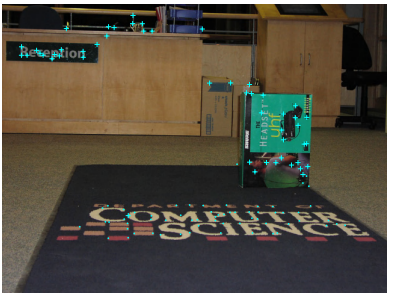
$$k_{s\mu} = \sqrt{s^2 + \mu^2}, \sin\alpha = \frac{s}{k_{s\mu}}, \cos\alpha = \frac{\mu}{k_{s\mu}}, \tan\alpha = \frac{a_v}{b_v}. \tag{7}$$

Eq.(6) indicates that we need to find three constants $(q, k_{s\mu}, \alpha)$ in order to recover the homography and that the computation should be implemented in (ρ, θ) image space (note that a planar homology has five *dof*, but two have been recovered in the FOE computation). We call this reciprocal-polar image space. Thus, after computing the FOE, a cubic-spline based interpolation procedure is used to generate a reciprocal-polar image for each image in the image pair.

For each pixel in image 1, its position in reciprocal-polar image space is computed, and a 1D window is created around this position along the $q\rho$ dimension ($q = 1$ for translation parallel to the ground plane, for other translations, q is easily computed). We then correlate this window along the ρ in reciprocal-polar image 2, at the same value of θ. This correlation process is possible because of

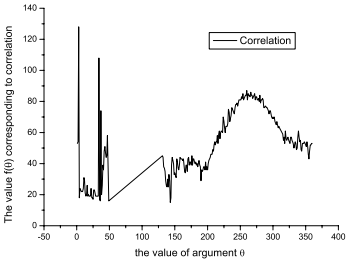


(a)

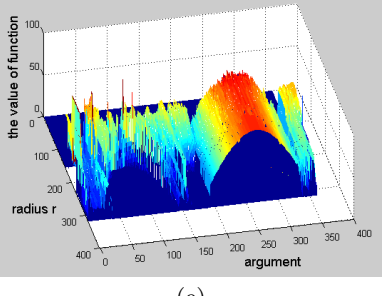


(b)

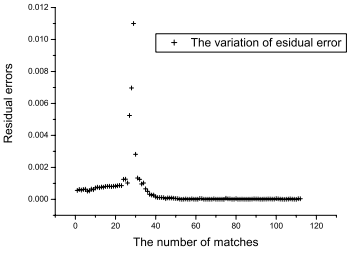
(c)



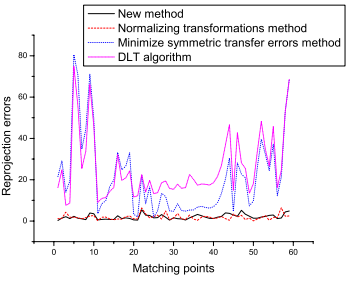
(d)



(e)

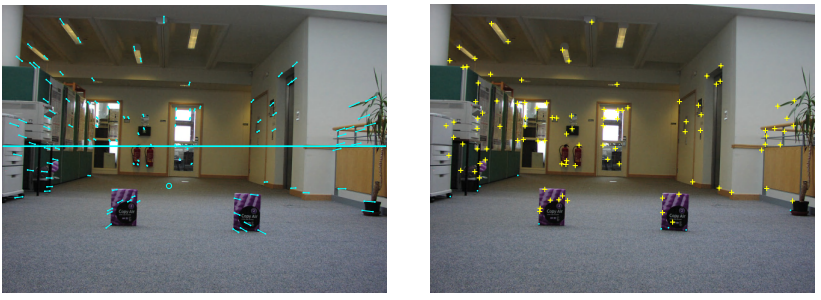


(f)



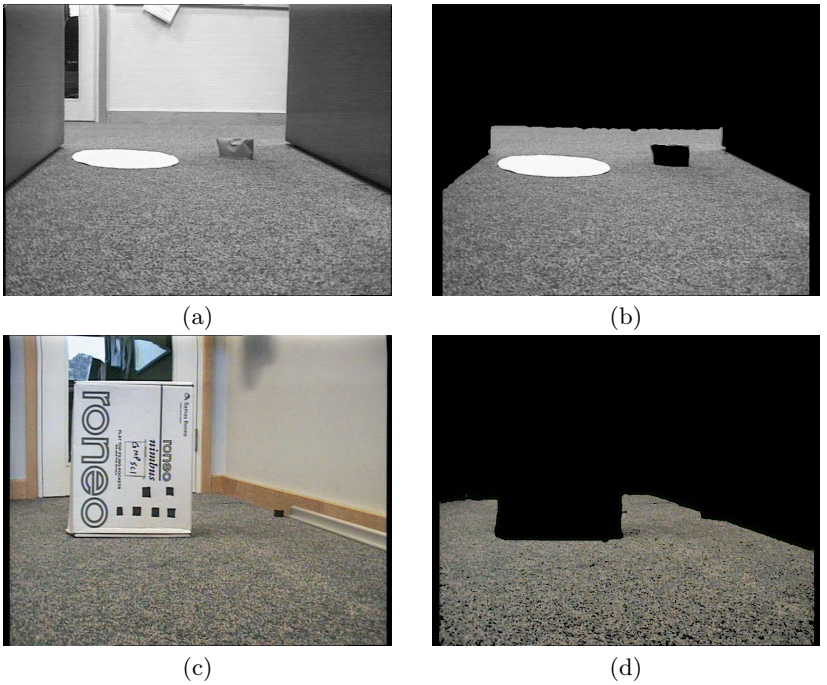
(g)

Fig. 2. (a) one of the original images with their matching points (•), feature tracks, FOE and vanishing line. (b) ground plane points (•) and obstacle points (+). (c) reciprocal-polar image corresponding to (a). (d) The value of $f(\theta)$. (e) 3D sinusoid. (f) The variation of residual errors (the number of coplanar points is between 54 and 112). (g) Reprojection errors of several alternative methods.



(a) (b)

Fig. 3. (a) Correspondences (b) The segmented ground plane points (•) and obstacle points (+)



(a) (b)

(c) (d)

Fig. 4. (a) Original image 1 (b) The segmented ground plane region (c) Original image 2 (b) The segmented ground plane region

the 'pure-shift' relation between ρ_2 and $q\rho_1$, expressed in Eq.(6) and the position of the maximum value of the correlation is related as a value of $f_i(\theta)$.

Eq.(6) indicates that correlation maxima and feature correlations in reciprocal-polar space, which are associated with a planar surface, lie on a sinusoid and the constants $(k_{s\mu}, \alpha)$ may be recovered by fitting a sinusoid to the

data for $f(\theta)$. Suppose that we have two value of $f(\theta)$, $f_{i,j}$ measured at two angles, $\theta_{i,j}$, so that

$$f_i = k_{s\mu} \sin(\theta_i + \alpha), f_j = k_{s\mu} \sin(\theta_j + \alpha). \tag{8}$$

collecting terms in $\tan \alpha$ and rearranging gives

$$\tan \alpha = \frac{f_j \sin \theta_i - f_i \sin \theta_j}{f_i \cos \theta_j - f_j \cos \theta_i}. \tag{9}$$

Thus, in theory, a pair of f values, at different angular positions, for pixels belonging to the same plane, allows us to estimate the orientation of the vanishing line of that plane. Then, given the phase angle, α, corresponding to the orientation of the vanishing line, we can compute $k_{s\mu}$ from the Eq.(8).

In order to robustly and accurately estimate the vanishing line orientation from many correlation maxima and feature correspondences in reciprocal-polar space, many of which will not be associated with the ground plane, a random sample consensus (RANSAC) [1] and least-squares process is used. Pairs of (θ, f) values are selected across random values of ρ and θ, with the constraint that there must be a minimum angular separation between the two θ values. The magnitude and phase of a sinusoid that passes through these two (θ, f) coordinates is computed and the number of inliers stored. An inlier is defined as some image motion (correlation maxima or feature correspondence) that is within threshold of the putative sinusoid. Thus $r < threshold$, where r is the residual error: $r = |(\rho_{P'} - q\rho_P) - k_{s\mu} \sin(\theta + \alpha)|$ and $\mathbf{P}$ and $\mathbf{P'}$ are correspondences. The sinusoid with the maximum number of inliers is used to initialise an iterative procedure where a standard linear least-squares estimate of the sinusoid parameters and the associated set of inliers are computed until the inlier distribution stabilises or the maximum number of iterations is reached. In this way, co-planar pixels may be grouped without explicit construction of a homography matrix, although this is easily recovered in the FOE centered frame from the FOE and the two parameters (amplitude and phase) of the sinusoid. Note that the FOE and phase of the sinusoid give the vanishing line, and the amplitude of the sinusoid gives the scale of the translation, related to k, hence we have all the parameters that we need in equation 3. The homography in the original frame, if required, can be computed as $\mathbf{H} = \mathbf{T}_c^{-1}\mathbf{H}'\mathbf{T}_c$.

5 Experimental Results and Applications

5.1 Results Using Corner Correspondences

In the first experiment, the camera was moved parallel to the ground plane. Fig. 2 (a) shows one of the original images with their matching points (•), feature tracks, FOE and vanishing line. After the FOE was obtained, the images were then converted to reciprocal-polar (ρ, θ) form, as shown in (c).(d) shows us the one of the sinusoidal forms of $f(\theta)$ for a fixed ρ (and hence fixed r). The partial sinusoidal curve ($\theta \in [193.25, 315.00]$) is clearly shown and represents the

motion of the ground plane in reciprocal-polar space. (The phase is shown close to 180 degrees rather than 0 degrees because the direction of y in the image is directed upwards from the FOE rather than downwards). The 3D sinusoidal form of reciprocal-polar ground plane motion, which is obtained when ρ (and hence r) is varied, is shown in (e). In this set of experiments, the number of coplanar matching points varied from 54 to 112. The mean of residual errors of ground plane points is 1.558E-5, the standard deviation is 1.393E-5 and the maximum value is 6.3E-5. But the mean of residual errors of non ground plane points is 0.00101, the standard deviation is 0.00183 and the maximum value is 0.011. The variation of residual errors of matching points is shown in (f), those less than 0.00025 are classified as ground plane points. The reprojection errors of the recovered planar homographies using our new method and several methods described by Hartley and Zisserman [2] are shown in (g), which shows us that the accuracy of the new method is very similar to the normalizing transformation method, and much better than the direct linear transform (DLT) method and the method of minimisation of symmetric transfer error. Finally, all the coplanar points and the points that lie on obstacles can be segmented using the sinusoid model. The result is shown in (b), where 'ground-plane points' and 'obstacle points' are marked as (•) and (+),respectively.

In the second experiment, the camera was moved in a forward translation mode, but in a direction not parallel to the ground. The camera axis and motion is inclined downwards towards the ground plane. In this situation, at least two matching points are needed to calculate q. We compute the FOE as $\mathbf{v} = (277.73, 302.41)$, and the vanishing line as $\mathbf{l} = (0, 0.0005, -0.13)$. Some correspondences, feature tracks, vanishing line and FOE (o) are shown in Fig. 3 (a) The segmented ground plane points (•) and obstacle points (+) are shown in Fig. 3 (b).

5.2 Experiment Using Correlations and Contour Matching

The final experiment presented in this paper, uses correlations within locally textured regions and contour matching to determine whether smooth (textureless) regions should be grouped with the ground plane. Note that an additional process is required, not described in this paper, which is our own quadtree split-merge region segmentation algorithm, which extracts homogenous regions of colour-texture. Smooth (textureless and featureless) regions cannot be classified as ground plane or non-ground plane as they cannot be matched across an image pair. Their boundaries, however, can be and, in the case of pure translation, this matching is easily done by 'casting' rays from the FOE recovered from *all* corner matches (note that there may be little or even no corner correspondences on the ground!).

Fig. 4 (a) shows an image with two regions on the floor which have little texture. The first is a circular piece of white paper, which can be driven over, and the second is a small cardboard box, which can't. The boundaries of these regions are extracted and the FOE is used to cast a ray in order to match points along corresponding boundaries. If the motion of all matched boundary points

falls within the threshold of the sinusoid model in reciprocal-polar space, it is classified as belonging to the ground plane, otherwise the region is classified as an obstacle. Fig. 4 (b) shows the extracted ground region, where the textured carpet has been classified on a pixel by pixel basis, and the the textureless white paper region has been included by virtue of its boundary motion being consistent with ground plane motion in reciprocal-polar space. A second example of pixel based segmentation is shown in figures 4 (c) and (d). Note that there are some 'drop outs' in the foreground of the image, but the shape of the segmentation is excellent, to the extent that even the small black doorstop to the centre right of the original image 4 (c) has been correctly classified as an obstacle and removed in image 4 (d).

6 Conclusions

Firstly, we have presented a novel reciprocal-polar (ρ, θ) image rectification, which transforms planar image motion under pure translation into a pure shift, irrespective of the degree of perspective distortion of the planar surface. Hence correlation can be done over large translations for the class of planar homographies described as 'elations', when correlation in the original image space would fail. Secondly, we have shown that it is possible to group/segment co-planar regions by matching all co-planar pixels which have local intensity variations to the same half-wave sinusoid. This can be done without explicit computation of the homography. A nice feature of the algorithm is that *all* pixels on a plane (that have some local intensity variation) contribute to the ground plane grouping process and (if required) associated homography estimation. Thus you get a dense (i.e. pixel based) ground plane segmentation. Our results show that our algorithm performs very well to outliers and noises and the stability, accuracy and robustness performs favourably to other methods in terms of the reprojection errors of the recovered homography.

References

1. Fischler M. A. and Bolles R. C. Random sample consensus: apardigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 24:381–395, 1981.
2. R Hartley and A Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2001.
3. Reinhard Koch Pollefeys M and Luc Van Gool. A simple and efficient rectification method for general motion. In *ICCV*, pages 496–501, 1999.
4. Wolberg and Zokai. Robust image registration using log-polar transform. In *Proc IEEE Int. Conf. Image Processing*, 2000.

Tracking of Points in a Calibrated and Noisy Image Sequence

Domingo Mery[1], Felipe Ochoa[2], and René Vidal[3]

[1] Departamento de Ciencia de la Computación
Pontificia Universidad Católica de Chile
Av. Vicuña Mackenna 4860(143), Santiago de Chile
Tel. (+562) 354-5820, Fax. (+562) 354-4444
dmery@ing.puc.cl
http://www.ing.puc.cl/∼dmery
[2] Departamento de Ingeniería Informática, Universidad de Santiago de Chile
felipe_ochoa@entelchile.net
[3] Center of Imaging Science, Johns Hopkins University
rvidal@cis.jhu.edu

Abstract. In this paper an algorithm that performs the tracking of points in a calibrated and noisy image sequence is presented. The candidate points to be tracked must satisfy certain constraints that can be deduced from the multiple view geometry. The idea is to consider as noisy points those candidates which cannot be tracked in the sequence. The robustness of the algorithm has been verified on simulated data using different constraints. The methods are assessed in several cases where the number of noisy points and the noise in the measurement of the points to be tracked are varied. Using this study, it is possible to know the performance of the tracking method. An example that shows a perfect tracking of 8 points in a sequence of 10 images with 500 noisy points per image is shown.

Keywords: Tracking, computer vision, multiple view geometry, epipolar geometry, trifocal tensors.

1 Introduction

Tracking is posed as a matching problem between correspondences of features in an image sequence [1]. In this problem, corresponding features, i.e., features in different views that are representation of the same 3D entity, must be matched. Generally, the problem is solved in two steps. The first step is the extraction of features obtained from the (2D) images of the sequence, e.g. points, lines or curves. The second step is to find correspondence between them. The most difficult problem is the correspondence of features. Since many tracking algorithms have been developed to date, we present now, in order to make this topic more clear, the following dichotomies of tracking. We do not intend to make here a survey of all previous work. However, the references presented in this review give a good overview to the reader.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 647–654, 2004.
© Springer-Verlag Berlin Heidelberg 2004

i) Tracking by short or long sequences: in short sequences, the tracking is performed using epipolar or trifocal constraints [2]. The use of a larger number of images allows significant smoothing to be achieved, where a long sequence of frames is taken so that there is only a small change between adjacent frames [1].

ii) Tracking in monocular or stereo frames: with one camera 2D information is used to establish the correspondence in the sequence [3], whereas stereo frames gives the possibility to infer 3D information [1].

iii) Tracking from geometric features (points, lines, contours) or from grey values features (regions, windows): feature based methods rely on extraction of discrete object features in successive images, the image coordinate of which are used to estimate the object motion or to find the trajectories of the objects in the sequence. Geometric feature tracking is pose as a matching problem between points [4], lines [5] or contour curves [6] constructed out of edges in the image. Grey value feature tracking uses correlation techniques between regions or windows in order to establish the matching in two frames [1].

iv) Tracking using a motion model or based on a dense motion field: object motion is computed based on some model of the evolution of 3D motion parameters (e.g. kinematics model with Kalman filter [1]). In the other hand, optical flow methods represent motion in the image plane as sampled continuous velocity fields that can be used to track features [7].

v) Single object motion or multiple object motion: in the first case there is a single moving object in the sequence [8], whereas in the second case there are multiple objects with independent motions [4].

vi) Rigid or deformable objects: when the objects are rigid, the 3D shape of the object is invariable over the time, i.e., the shape can be easily modelled and stays relatively constant over the sequence [8]. In contrast, there are cases where the region of interest changes shape throughout the sequence, corresponding to deformable objects [9].

In this paper we deal with the problem of tracking of points of a single rigid object in long calibrated image sequences with monocular frames, where the density of noisy points is high. In a *calibrated* image sequence the model 3D $\rightarrow$ 2D is a priori known because it was obtained in an off-line process called calibration [2]. The idea of the tracking algorithm is to consider as noisy points those candidates which cannot be tracked in the sequence. The tracking is performed using geometrical multiple view constraints. The original algorithm was developed in [10] for automated visual inspection analysing multiple views with simple geometric constraints. However, no detailed study of the performance was made in [10]. In our paper, i) we demonstrate that the consideration of the geometric constraints based on the Sample distance [2] achieves a better performance than the simple constraints; and ii) we show how robust is the tracking algorithm under several noisy conditions. The rest of the paper is organised as follows: Section 2 gives an overview of the tracking algorithm. In Section 3 the different criteria used to establish the correspondence between the points of the views are described. Section 4 shows the experiments and the results obtained on synthetic data. Finally, Section 5 gives concluding remarks.

2 Tracking Algorithm

After the candidate points in each image of the sequence are segmented, the coordinates of the points are normalised using the Cholesky factorisation [2]. Then the attempt is made to track them in the sequence in order to separate the noise from the object points. The tracking algorithm consists of three steps: *matching in two views*, *tracking in three views* and *tracking in four views*.

2.1 Matching in Two Views

Matching requires the position of each detected point. In this work, $\mathsf{a} = (a, p)$ will denote the identified point a in image p. It is assumed that the image sequence has N images ($1 \leq p \leq N$) and n_p points were identified in image p ($1 \leq a \leq n_p$). The position of point $\mathsf{a} = (a, p)$ is arranged in a *position vector* $\mathbf{m}_p^a$. One obtains then the position vector $\mathbf{m}_p^a = (x_p^a\ , y_p^a)$. This step matches two points (of two views), point $\mathsf{a} = (a, p)$ with point $\mathsf{b} = (b, q)$, for $p \neq q$, if they fulfil all following *matching conditions*:

• **Constraint in two views:** $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$ must satisfy the constraint of correspondence between two views, i.e., the epipolar constraint.

• **Correct location in 3D:** the 3D point reconstructed from the position of the points must belong to the space occupied by the object. From $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$ the corresponding 3D point $\hat{\mathbf{M}}$ is estimated. It is necessary to examine if $\hat{\mathbf{M}}$ resides in the volume of the object, the dimensions of which are usually known a priori.

In addition, a *similarity condition* can be used, if certain features of the candidate points and their neighbourhood are available. The matching is established if the points are similar enough.

The matching conditions in both identified points $\mathsf{a} = (a, p)$ and $\mathsf{b} = (b, q)$ are evaluated in 3 consecutive frames, for $p = 1, ..., N - 3$; $q = p + 1, ..., p + 3$; $a = 1, ..., n_p$ and $b = 1, ..., n_q$. If a point is not matched with any other one, it will be considered as noise. Multiple matching, i.e., a point that is matched with more than one point, is allowed. Using this method, problems like non-segmented points or occluded points in the sequence, can be solved by the tracking if a point is not identified in consecutive views.

2.2 Tracking in Three Views

A match between two points a and b will be denoted by $\mathsf{a} \leftrightarrow \mathsf{b}$ or $(a, p) \leftrightarrow (b, q)$. A $N_2 \times 4$ matrix $\mathsf{A} = [\mathsf{a}_k\ \mathsf{b}_k] = [a_k\ p_k\ b_k\ q_k]$, $k = 1, ..., N_2$, is defined, where N_2 is the number of all matches determined in Section 2.1.In the tracking problem, it is required to find *trajectories* of points in different views. To establish the correspondence of points in three images one seeks all possible links of three points in matrix A that satisfy the condition of correspondence in three views.

The procedure is as follows: first, one looks for all two rows i and j of A ($i, j = 1, ..., N_2$ and $i \neq j$) that satisfy $\mathsf{b}_i = \mathsf{a}_j$. Supposing rows i and j fulfil this condition, in other words, the last two elements of row i are equal two the first

two elements of row j, e.g. $\mathsf{A}_i = [a\ p\ b\ q]$ and $\mathsf{A}_j = [b\ q\ c\ r]$, one finds three points $(a,p) \leftrightarrow (b,q) \leftrightarrow (c,r)$ with coordinates $\mathbf{m}_p^a$, $\mathbf{m}_q^b$ and $\mathbf{m}_r^c$ respectively that could be corresponding points. Finally, to examine if they really correspond to each other, the condition of correspondence between three views must be evaluated. If the condition is fulfilled, then it is assumed that the points are corresponding. The non tracked points are eliminated, while the N_3 linked triplets are arranged in a new matrix B.

2.3 Tracking in Four Views

A $N_3 \times 6$ matrix $\mathsf{B} = [\mathsf{a}_k\ \mathsf{b}_k\ \mathsf{c}_k] = [a_k\ p_k\ b_k\ q_k\ c_k\ r_k]$, $k = 1, ..., N_3$, is defined, where N_3 is the number of all triplets determined in Section 2.2. Now, the attempt is made two find two triplets in B that correspond to the same 3D point. It is well known that given four points (in four views), if the first three are corresponding points and the last three are corresponding points too, then all of them are corresponding points. In this case, it is not necessary to evaluate a condition of correspondence between four views, because it is redundant. That means, to seek quadruplets that satisfy the condition of correspondence in four views, it is necessary to look for all rows i and j of B for $(i, j = 1, ...N_3$ and $i \neq j)$ that satisfy $\mathsf{b}_i = \mathsf{a}_j$ and $\mathsf{c}_i = \mathsf{b}_j$. Supposing rows i and j fulfil this condition, e.g. $\mathsf{B}_i = [a\ p\ b\ q\ c\ r]$ and $\mathsf{B}_j = [b\ q\ c\ r\ d\ s]$, four corresponding points $(a,p) \leftrightarrow (b,q) \leftrightarrow (c,r) \leftrightarrow (d,s)$ (with coordinates $\mathbf{m}_p^a$, $\mathbf{m}_q^b$, $\mathbf{m}_r^c$ and $\mathbf{m}_s^d$ respectively) are found. The N_4 detected quadruplets are placed in a new matrix C. Finally, a tracking in more views can be achieved by linking quadruplets of matrix C having common elements.

3 Correspondence and 3D Reconstruction Criteria

In this investigation different criteria of correspondence of points between two and three views and 3D reconstruction from two views are used. Normally, the epipolar constraint, the trilinear constraint and the triangulation approach are used for these tasks. However, a typical observation consists of a noisy point correspondence which does not in general satisfy the correspondence constraints [2]. For this reason, other criteria should be used. In this Section, we present alternative criteria that take into account the noise in the imaged points.

3.1 Correspondence in Two Views

Two methods are used to investigate if $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$ can be corresponding points.
• **Simple:** the constraint is fulfilled if the perpendicular Euclidean distance from the epipolar line of the point $\mathbf{m}_p^a$ to the point $\mathbf{m}_q^b$ is smaller than ε [2].
• **Sampsom:** the constraint is fulfilled if the first-order geometric error of the epipolar constraint is smaller than ε [2].

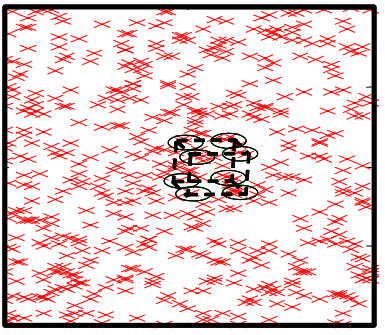


Fig. 1. Simulation of a projected cube into an image with $N_p = 500$ random points.

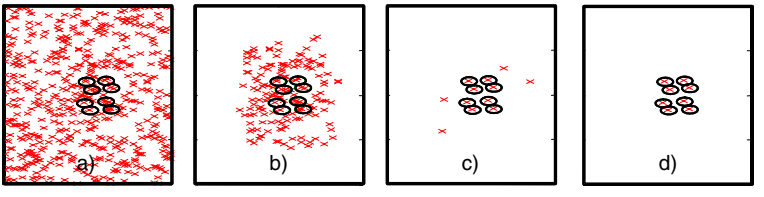


Fig. 2. Tracking of the vertices in a sequence of $N = 10$ images with $N_p = 500$ noisy points per image: a) original image of the sequence, and points after b) matching in two views, c) tracking in three views, and d) tracking in four views.

3.2 Correspondence in Three Views

Two methods are used to investigate if $\mathbf{m}_p^a$, $\mathbf{m}_q^b$ and $\mathbf{m}_r^c$ satisfy the correspondence in three views:

- **Simple:** The position of third point is estimated from $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$ using a reprojection approach based on trifocal tensors. The constraint is fulfilled if the Euclidan distance between estimated point and $\mathbf{m}_r^c$ is smaller than ε [10].
- **Sampsom:** the constraint is fulfilled if the first-order geometric error of the trilinear constraints is smaller than ε [2].

3.3 3D Reconstruction

Two methods are used to estimate a 3D point $\hat{\mathbf{M}}$ that may have produced the imaged points $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$.

- **Simple:** Point $\hat{\mathbf{M}}$ is estimated using normalised projection matrices and a projective 3D transformation [11].
- **Sampsom:** Point $\hat{\mathbf{M}}$ is estimated using the triangulation of corrected points $\hat{\mathbf{m}}_p^a$ and $\hat{\mathbf{m}}_q^b$ that are computed from the first-order geometric correction of $\mathbf{m}_p^a$ and $\mathbf{m}_q^b$ [2].

4 Experiments and Results

In this Section we present the results obtained recently using synthetic data[1]. The data was simulated as follows: A cube was defined and located in N different positions. For each position the 8 vertices of the cube were projected using a perspective transformation into an image of $M \times M$ pixels. Thus, a sequence of N binary images of $M \times M$ pixels was obtained. Additionally, N_p points were randomly superimposed in each image. Fig. 1 illustrates one of these images with $N_p = 500$ and $M = 400$, where the doted lines and the circles were intentionally added to distinguish the projection of the cube. In order to simulate a more realistic projection, the coordinates of the projected vertices were altered adding normal distributed noise with standard deviation σ and mean zero.

The effectiveness of the tracking algorithm for 500 noisy points is illustrated in Fig. 2. Only one image (see Fig. 2a) of a whole sequence of 10 images is presented. This image corresponds to Fig. 1. After matching in two views, the remaining points are shown in Fig. 2b. The other points were eliminated because they were not matched. The dimensions of the cube were $2 \times 2 \times 2$, and the space used in the correct location in 3D criterion was $8 \times 8 \times 8$ centred in the cube, i.e., 64 times bigger than the cube. For this reason, the points that satisfied the matching conditions are approximately located in an area of $4h \times 4h$, where h is the distance between two adjacent projected vertices. After tracking in three and four views (see Fig. 2c and 2d), all noisy points were eliminated without discriminating the points of the cube.

The performance of the algorithm was evaluated for different conditions varying the number of noisy points ($N_p = 10, 25, 50, 100, 200, 500$); the standard deviation of the projected points ($\sigma = 0, 0.05, 0.1, 0.2, 0.35, 0.5$); the tolerance used in correspondence constraints ($\varepsilon = 0.0005, 0.001, 0.005, 0.01$); and the correspondence and 3D reconstruction criteria (*simple* or *Sampson*). In these experiments, the size of the images was 400×400 pixels, and the number of images in the sequence was 10. Each situation was simulated 10 times and the average of the following variables was computed: the number of true positives (TP), i.e., points of the cube correctly tracked (ideally, $TP = 8$); the number of false positives (FP), i.e., tracked points that do not correspond to the vertices of the cube (ideally, $FP = 0$); the computational time (t) required for the tracking process. The simulation environment was programmed in MATLAB using a PC based on a CPU Pentium 4, 2.6 GHz, 512 MB RAM, and operating system Microsoft Windows XP.

In order to evaluate the performance of the tracking, the true positive rate ($TPR = TP/8$) and the false positive rate ($FPR = FP/N_p$) were computed for different noise conditions (N_p, σ) varying parameter ε in the geometric constraints. Ideally, $TPR = 1$ and $FPR = 0$, i.e., all object points are tracked without flagging false alarms. The best operation point (TPR^*, FPR^*) at $\varepsilon = \varepsilon^*$ is chosen as the point with the smallest Euclidean distance to the ideal point $(1, 0)$. The obtained results are summarised in Table 1, where the performance

[1] A real experiment is available in `http://www.ing.puc.cl/~dmery/sequences.htm`.

Table 1. Performance (TPR^*, FPR^*) for different noise conditions (N_p, σ).

σ/M	method	$Np = 10$	$Np = 25$	$Np = 50$	$Np = 100$	$Np = 200$	$Np = 500$	average
0.000000	simple	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	
	Sampson	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	
0.000125	simple	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	
	Sampson	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.001$	
0.000250	simple	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(0.97,0.00)	(0.97,0.00)	(0.99,0.00)
		$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.001$	$\varepsilon^* = 0.001$	
	Sampson	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	
0.000500	simple	(1.00,0.00)	(1.00,0.00)	(0.99,0.00)	(1.00,0.00)	(1.00,0.00)	(0.95,0.00)	(0.99,0.00)
		$\varepsilon^* = 0.005$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	
	Sampson	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)
		$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.005$	
0.000875	simple	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(0.97,0.00)	(0.98,0.00)	(0.99,0.00)
		$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	$\varepsilon^* = 0.005$	
	Sampson	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(1.00,0.00)	(0.92,0.00)	(0.99,0.00)
		$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.005$	
0.001250	simple	(1.00,0.00)	(0.96,0.01)	(0.97,0.00)	(0.95,0.01)	(0.95,0.00)	(0.88,0.00)	(0.95,0.00)
		$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.005$	
	Sampson	(0.93,0.00)	(0.97,0.00)	(1.00,0.00)	(0.99,0.00)	(1.00,0.00)	(0.91,0.00)	(0.97,0.00)
		$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	$\varepsilon^* = 0.010$	

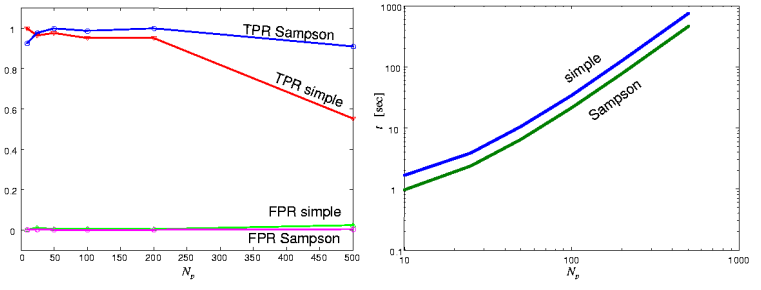


Fig. 3. Evaluation of the methods depending on the number of noisy points N_p: a) TPR and FPR at $\varepsilon = 0.01$ and $\sigma/M = 0.00125$; and b) computational.

was calculated for different (N_p, σ) combinations. As shown in Table 1, parameter ε^*, i.e., the parameter ε that gives the best operation point, depends strongly on the inherent noise of the projected points (σ) and the number of noisy points (N_p). The larger is σ, the larger should be set ε if one wants to track all object points. However, the larger is ε, the larger the false alarm rate when N_p is large, because several noisy points can be tracked. As example, Fig. 3a shows TPR and FPR for $\sigma/M = 0.001250$ and $\varepsilon = 0.01$. In addition, the computational time is presented in Fig. 3b. We observe that the performance and the computational time of the Sampson criteria is better. On the other hand, the results of average column Table 1 show the high robustness of the tracking algorithm in both discrimining noisy points and detecting object points.

5 Conclusions

In this paper an algorithm that performs the tracking of points in a calibrated and noisy image sequence was presented. The idea is to consider as noisy points those candidates which cannot be tracked in the sequence. The tracked points satisfy certain constraints of the multiple view geometry. The robustness of the algorithm has been verified on simulated data using two different constraints: simple and Sampson. The methods were assessed in several cases where the number of noisy points and the noise in the measurement of the points to be tracked are varied. The obtained results are: the robustness of the algorithm is very high, and the best performance was achieved using the Sampson distance.

Acknowledgments. This work was supported by FONDECYT – Chile under grant no. 1040210.

References

1. Zhang, Z., Faugera, O.: Estimation of displacements from two 3D frames obtained from stereo. IEEE Trans. Pattern Analysis and Machine Intelligence **14** (1992) 1141–1156
2. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
3. Shirai, Y.: Estimation of 3-D pose and shape from a monocular image sequence and realtime human tracking. In: Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling, Ottawa, Canada (1997) 130–140
4. Chetverikov, D., Verestóy, J.: Tracking feature points: a new algorithm. In: Proceedings of the 14th International Conference on Pattern Recognition, ICPR-1998, Brisbane, Australia (1998) 2:1436–1438
5. P.E. López-de Teruel, A.R., García, J.: A parallel algorithm for tracking of segments in noisy edge images. In: Proceedings of the 15th International Conference on Pattern Recognition, ICPR-2000, Barcelona (2000) 1051–1055
6. Freedman, D.: Effective tracking through tree-search. IEEE Trans. Pattern Analysis and Machine Intelligence **25** (2003) 604–615
7. Mae, Y., Shirai: Tracking moving object in 3-d space based on optical flow and edges. In: Proceedings of the 14th International Conference on Pattern Recognition, ICPR-1998, Brisbane, Australia (1998) 2:1439–1444
8. Broida, T., Chellapa, R.: Estimating the kinematics and structure of a rigid object from a sequence of monocular images. IEEE Trans. Pattern Analysis and Machine Intelligence **13** (1991) 497–513
9. Mansouri, A.: Region tracking via level set PDEs without motion computation. IEEE Trans. Pattern Analysis and Machine Intelligence **24** (2002) 947–961
10. Mery, D., Filbert, D.: Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. IEEE Trans. Robotics and Automation **18** (2002) 890–901
11. Hartley, R.: A linear method for reconstruction from lines and points. In: 5th International Conference on Computer Vision (ICCV-95), Cambridge, MA (1995) 882–887

Multiresolution Approach to "Visual Pattern" Partitioning of 3D Images*

Raquel Dosil[1], Xosé R. Fdez-Vidal[2], and Xosé M. Pardo[1]

[1] Dep. de Electrónica e Computación, Univ. de Santiago de Compostela, Campus Universitario Sur, s/n, 15782, Santiago de Compostela, Spain
rdosil@usc.es, pardo@dec.usc.es
http://www-gva.dec.usc.es/grupo/grupo.htm
[2] Escola Politécnica Superior, Univ. de Santiago de Compostela, Campus Universitario, s/n, 27002, Lugo, Spain
faxose@usc.es
http://www.lugo.usc.es/~ffeeps

Abstract. This paper deals with the problem of low level representation of 3D image contents. The presented solution makes use of multiresolution techniques to recover the so-called *visual patterns* or *integral features* that form images. It consists of decomposing the image into a set of elementary image features, representing frequency channels, using a filter bank, and grouping them by means of clustering analysis. The method introduces a novel design of the bank of oriented scaled filters. In addition, a new measure of dissimilarity between pairs of features is applied to the hierarchical clustering technique.

1 Introduction

Many image analysis tasks involve the manipulation of low level image representations. Much effort has been made in the field of image analysis to achieve robust feature detection techniques –see Kovesi [1] for a review. A good representation should meet the requirements of being invariant to contrast and scale changes. It also must be robust in the presence of noise.

A 2D representation model that fulfills these requirements is the RGFF model developed by Rodríguez-Sánchez et al. [2]. This model is based on the criterion of Field [3], who states that meaningful structures in an image present some degree of alignment in its spectral domain representation. The features that constitute the representation of the image in the RGFF are called *visual patterns* and are defined as patterns with phase alignment along wide frequency ranges. To isolate visual patterns, the RGFF first decomposes the image into elementary features using a bank of log Gabor filters tuned at different scales and orientations. Then, these components are grouped together by cluster analysis according to some measure of dissimilarity among them.

* The authors desire to acknowledge the Xunta de Galicia for their financial support of this work by means of the research project PGIDT01TIC20601PN.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 655–663, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The method has been recently redesigned to be applied on 3D data from video sequences [4]. In this application the goal is to detect motion patterns, but it could also be employed in the analysis of volumetric data, like for example in the fields of medical imaging and non-invasive inspection of materials.

In this paper, a tool for the decomposition of volumetric data in visual patterns is presented, introducing some novelties. We have designed a new 3D filter bank which enhances the behavior of previous approaches to 3D oriented filters [5]. Moreover, an alternative measure of feature dissimilarity, based on the normalized mutual information of pairs of filter responses, has been studied.

In next section the method for visual pattern extraction is described. The 3D filter bank definition is introduced subsection 2.1 and the selection of its most relevant channels in 2.2. The measure of dissimilarity between filter responses, based on mutual information, is defined in subsection 2.3. Subsection 2.4 presents the hierarchical clustering method employed to group visual patterns. Some results of this method are shown in section 3 and the conclusions derived from them are presented in section 4.

2 Visual Pattern Detection

As mentioned before, the method for visual pattern extraction consists of the decomposition of the image in a set of frequency channels and their grouping according to some dissimilarity measure. The process can be described as a sequence of four steps:

1. Selection of *active* filters in the 3D filter bank, i.e., channels with high information content, by analyzing the spectral energy map and generation of its responses.
2. Measure of dissimilarity between the energy maps of pairs of filter responses based on their normalized mutual information
3. Hierarchical clustering of the features based on the dissimilarity matrix

In the next subsection the filter bank employed to represent the elementary features of an image is presented. The first step of the sequence of processes, the technique to select the most relevant filters in the bank, is described in subsection 2.2.

2.1 3D Filter Bank

Visual patterns are represented as a combination of the responses of a set of band-pass filters tuned in different scales and orientations. The filters' transfer function T is designed as the product of separable factors R and S in the radial and angular components respectively. The radial term R of the filter is given by the log Gabor function

$$R(\rho;\rho_i) = \exp\left\{-\log^2(\rho/\rho_i)\Big/\left(2\log^2(\sigma_\rho/\rho_i)\right)\right\}, \tag{1}$$

where σ_ρ is the standard deviation and ρ_i the central radial frequency.

To achieve orientation selectivity, the angular component of the filter usually represents a scattering function centered at the filter's direction. A common approach [4, 5] is to define S like the product of two Gaussians in each of the spherical angular coordinates, azimuth ϕ and elevation θ

$$S(\phi,\theta;\phi_i,\theta_i) = \exp\left\{-(\phi-\phi_i)^2/(2\sigma_\phi^2) - (\theta-\theta_i)^2/(2\sigma_\theta^2)\right\}, \tag{2}$$

where (ϕ_j, θ_j) is the filter orientation and σ_ϕ and σ_θ are the standard deviations of the Gaussian. Considering $\sigma_\phi = \sigma_\theta$, this expression represents an isotropic attenuation regards the central direction, as long as ϕ and θ are treated as Cartesian coordinates. However, they actually are spherical coordinates, so that S is projected over a curved plane and the Gaussian is no longer isotropic.

A better choice is to define S as a Gaussian on the angular distance α between the directions of the filter and the position vector of a given point **f** in the spectral domain

$$\alpha(\phi_i,\theta_i) = \operatorname{acos}\left(\mathbf{f}\cdot\mathbf{v}/\|\mathbf{f}\|\right), \tag{3}$$

where $\mathbf{v} = (\cos\phi_j \cos\theta_j, \cos\phi_j \sin\theta_j, \sin\phi_j)$ is a unit vector in the filter's direction and **f** is in Cartesian coordinates. Then, for a given angular standard deviation σ_α

$$S(\phi,\theta;\phi_i,\theta_i) = S(\alpha) = \exp\left\{-\alpha^2/(2\sigma_\alpha^2)\right\}. \tag{4}$$

The shape of S from equations (2) and (4) is illustrated in Fig. 1. In can be seen that expression (2) has rotational symmetry when represented in the plane, but not in the sphere, while expression (4) behaves the opposite

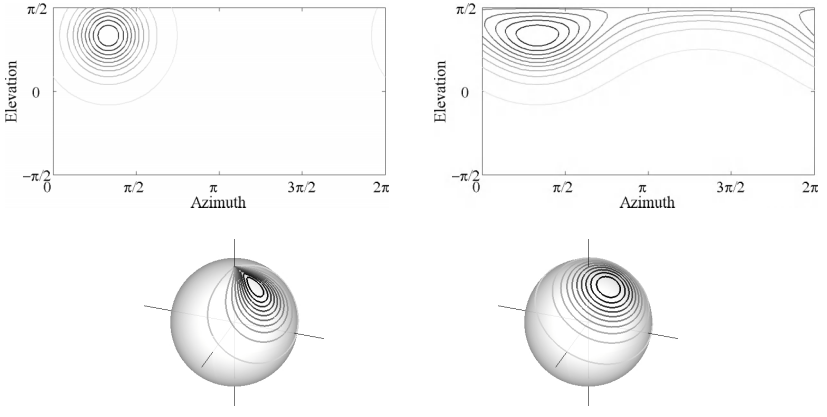


Fig. 1. Representation of $S(\pi/3, \pi/4)$ *Top*: in the (ϕ, θ) plane as isolevel curves *Bottom*: over a unit radius sphere as isolevel curves *Left*: S from equation (2) *Right*: S from equation (4)

The complete 3D bank is composed of a number of the above described filters to tile the frequency domain, selecting a number of wavelengths and orientations and tuning the bandwidths to cover the spectrum properly. The usual solution to sample direction is to sample the angular coordinates uniformly [4, 5], but this produces an overrepresentation of the region near $\theta = \pm\pi/2$. An alternative is to choose the direc-

tions of vectors pointing at the vertices of regular polyhedrons centered at the origin [6], but this solution lacks of flexibility in the selection of the number of filters.

In our configuration, elevation is sampled uniformly while azimuth is non-uniformly sampled, in order that "density" of filters at each elevation value is conserved. This is achieved by maintaining equal arc-length between adjacent azimuth values over the unit radius sphere. The filter bank configurations resulting from uniform and non-uniform sampling of the angular frequency space are shown in Fig. 2.

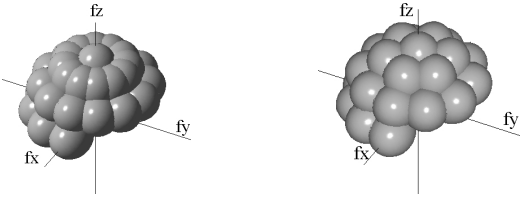


Fig. 2. Filter bank representation by isosurfaces verifying $H_i = e^{-1/2}$ *Left*: filter bank with uniform angular frequency sampling *Right*: non-uniform sampling

Following this criterion, the filter bank has been designed using 4 elevations –only one hemisphere is needed due to symmetry– and 6 azimuths to sample half the $z = 0$ plane, yielding 23 orientations with angular bandwidth of 25º. In the radial axis, 4 values have been taken with wavelengths 4, 8, 16 and 32 and 2 octave bandwidth

2.2 Selection of Active Bands

In the most general case, the filter bank consists of 92 filters. Their responses to the input data must be obtained and their pairwise dissimilarities calculated. To decrease the high computational cost of such tasks, the number of filters can be reduced. Firstly, filters with wavelengths greater than half the image size are discarded, as they approximately represent the average intensity level. Secondly, filters with low information content can be eliminated. To this end, the transformed spectral energy density $E = \log(|F| + 1)$ is used as a measure of the information associated to a channel, where F is the Fourier transform of the image.

The energy density map E reflects the preferred orientations and frequency bands plus a background of spectral noise. A 2D example can be seen in Fig. 3.a. In [2], the classification of the bands into active and non-active is realized applying cluster analysis to the total energy comprised by each channel. The problem with this solution is that the energy values of the activated filters range a wide interval, following a geometric decay as frequency increases [3]. This makes the definition of a class of activated patterns very difficult.

A simpler solution has been chosen here, involving the characterization, detection and elimination of non active channels, instead of active ones, through statistical methods. Non active bands are associated to the background noise. A band is active if it comprises any energy value over the estimated maximum level of the spectral noise. The maximum noise level is estimated as $m + x\sigma$, where m is the mean noise

energy, σ is its standard deviation and $x > 0$. Here, $x = 3$ has been considered enough to eliminate the main noise contribution and m and σ have been measured in the band of frequencies greater that double the largest of the bank's central frequencies.

Still, subtracting the maximum noise level from the energy map is not sufficient to "clean" the background. Some spurious values remain, as shown in Fig. 3.b. To eliminate them, some kind of filtering can be applied that takes into account the characteristics of spectral energy maps. Relevant features in the spatial domain are translated to the spatial-frequency domain into alignment through scales [3]. Therefore this kind of structures must be preserved in the filtering process.

Standard median filters produce the elimination of thin lineal structures. To avoid this, we have designed a radial median operator. The difference with an ordinary median filter is that, given a certain pixel in the energy map, it only considers neighbors that are anterior or posterior in the radial direction to calculate the median. This eliminates isolated peaks but preserving the continuity of structures along scales. The expression of the radial median filter mask M of size $N \times N \times N$ is as follows

$$\mathrm{M}(\mathbf{q},\mathbf{p}) = \begin{cases} 1 & \text{if } [(\mathbf{q}-\mathbf{p})/\|\mathbf{q}-\mathbf{p}\|] = [\mathbf{p}/\|\mathbf{p}\|] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{p}$ and $\mathbf{q}$ are points in the image and mask domains respectively, $[\cdot]$ represents rounding to the nearest integer and the origin is in the image center. In this work the mask size is taken to $N = 3$. The behavior of this filtering is illustrated in Fig. 3.c.

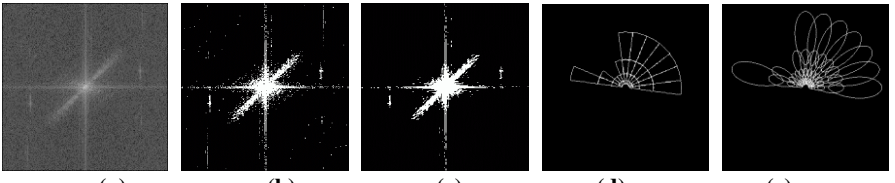


Fig. 3. 2D example of radial median filtering (b) $E = \log(|F| + 1)$, where $|F|$ is the spectral energy (c) Noise subtraction on E, non null values depicted in white (d) Radial median filtering, non null values in white. (e) Bands comprising non null values (f) Associated active filters

2.3 Dissimilarity Measure Between Energy Maps

In this work the measure of dissimilarity between pairs of energy maps has been defined as a function of their mutual information MI. MI captures the amount of cooccurrence of values in the two signals, giving an idea of the amount of shared information. Since it depends on the individual information amounts of each energy map, it must be normalized to represent the shared information in relation to the total information quantity. Hence, the normalized mutual information NMI is

$$\mathrm{NMI}(R_i, R_j) = 2 \cdot \frac{\mathrm{MI}(R_i, R_j)}{\mathrm{H}(R_i) + \mathrm{H}(R_j)} = 2 \cdot \frac{\mathrm{H}(R_i) + \mathrm{H}(R_j) - \mathrm{H}(R_i, R_j)}{\mathrm{H}(R_i) + \mathrm{H}(R_j)} \quad , \tag{6}$$

where H stands for entropy and R_i and R_j are the energy maps of the responses of the filters *i* and *j*.

NMI values range from zero, which means no information in common, to one, representing images with the same information content. To transform NMI into a measure of dissimilarity instead of similarity, its range must be inverted. It is also convenient to apply transformations to equalize its range. The distance measure used here is

$$D(R_i, R_j) = \sqrt{1 - \mathrm{NMI}(R_i, R_j)^2} \quad . \tag{7}$$

This transformation enlarges high differences and shortens low ones. This improves the performance of clustering, since generally *D* tends to be very small due to the fact that most of the energy values correspond to the background.

An important advantage of NMI is that it is invariant to contrast changes. This is convenient because high frequency components of visual patterns have higher contrast responses than low frequency ones [3], but they should be grouped together if they have the same location, i.e., there is phase congruency. Other advantages of the NMI are its simplicity compared with the measure used in the RGFF, where decisions are to be taken about the values of a great amount of parameters, and its lesser time consume –taking a small bin size in the histogram calculations.

2.4 Feature Clustering

To group filter responses a hierarchical clustering method has been chosen. Other clustering techniques, like k-means, are not adequate due to the nature of our dissimilarity data. Since it is not a Euclidean distance, the verification of the triangular inequality is not guaranteed. Hence, the data can not be embedded in an N-dimensional vector space. Hierarchical clustering has been applied using a complete-link algorithm. The distance between clusters is defined as the maximum of all pairwise distances between features in the two clusters, thus producing compact clusters.

The number of clusters N_c that a hierarchical technique generates is an input parameter of the algorithm. The usual strategy to determine the N_c is to run the algorithm for each possible N_c and evaluate the quality of each resulting configuration according to a given validity index. In this work, the modified Davies-Boulding index introduced in [7] has proved to produce good results. It is a graph-theory based index that measures the compactness of the clusters in relation to their separation.

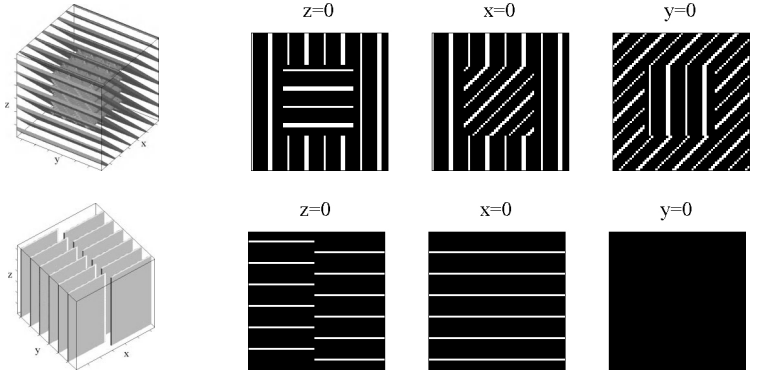


Fig. 4. *Top:* Data set D1 *Bottom*: Data set D2 *Left*: Isosurface representation of to half the maximum intensity level *Right*: Volume cross sections showing slices normal to the image axis

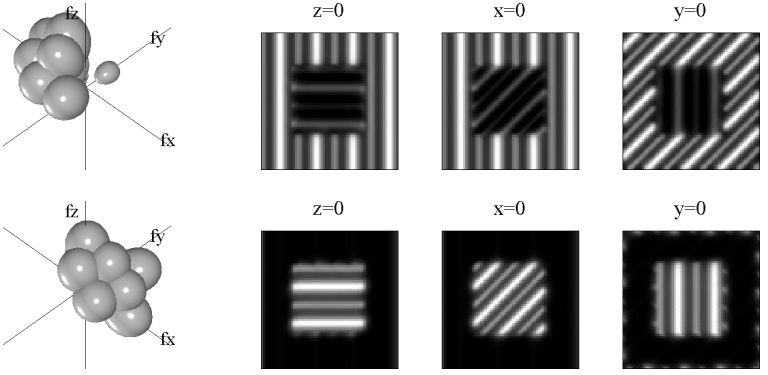


Fig. 5. Results obtained for image D1 *Top*: Cluster 1 (*left*) and cross sections of the volume reconstructed from it (*right*) and *Bottom*: Cluster 2 (*left*) and its reconstructed volume (*right*)

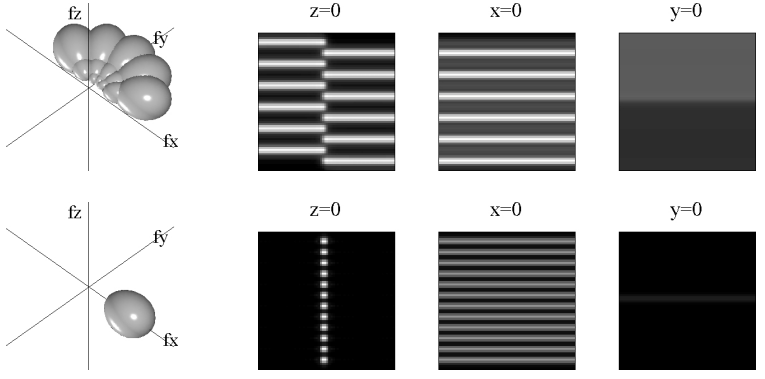


Fig. 6. Results obtained for image D2 *Top*: Cluster 1 (*left*) and cross sections of the volume reconstructed from it (*right*) and *Bottom*: Cluster 2 (*left*) and its reconstructed volume (*right*)

3 Results

The method stated in this paper has been tested using synthetic 3D images. In this section two different examples are presented to illustrate the capabilities of the system in isolating diverse types of relevant low level structures. In the first example the data set, named D1, shown in Fig. 4 is used. It is a 64×64×64 image showing two grating patterns with different orientations in different volume regions. The method classifies the filters in the bank in two categories or clusters. The sum of the responses of the filters in each cluster is represented in Fig. 5. As can be seen, the method separates the bands correspondent to the inner and outer regions.

A second example has been used to study the behavior of the method when applied to features different to orientation. In this case, the image D2 of size 64×64×64 represents two grating patterns with the same frequencies and orientations but with different phase, as seen in Fig. 4. The results are shown in Fig. 6. In this case the method yields two clusters, the first of which represents the gratings themselves. The second cluster represents the perceived but fictitious plane separating the two gratings.

4 Conclusions

In this paper we have presented a method for the isolation of visual patterns from 3D images. Although other similar methods have already been developed by other authors, in this work the extension to 3D has been accomplished introducing two remarkable novelties: a new filter bank design and a new dissimilarity measure.

In our filter bank, a new definition for a 3D log Gabor filter has been employed that enhances previous approaches, for it presents rotational symmetry. In addition, we have introduced a new bank design with non-uniform sampling of the frequency angular coordinates to achieve a more isotropic sampling of the orientation space, in the sense that there are no orientations more densely tiled than others. The selection of active filters has also been modified to avoid the use of ad hoc parameters.

The dissimilarity measure used here is based in the calculus of the normalized mutual information of a pair of filter response energy maps. The advantage of using this measure is that it is less parameterized, less complex and less computationally expensive. In future work, the system must be tested with more sophisticated textures and, especially, with real volumetric data.

References

1. Kovesi, P.D.: Invariant Measures of Image Features from Phase Information, The University or Western Australia, (1996), http://www.cs.uwa.edu.au/pub/robvis/theses/PeterKovesi/

2. Rodríguez-Sánchez, R., García, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R.: The RGFF Representational Model: A System for the Automatically Learned Partition of "Visual Patterns" in Digital Images, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 21(10) (1999) 1044-1073
3. Field, D.J.: Scale–Invariance and self-similar "wavelet" Transforms: An Analysis of Natural Scenes and Mammalian Visual Systems. In: Farge, M., Hunt, J.C.R., Vassilicos, J.C. (eds.): Wavelets, fractals and Fourier Transforms, Clarendon Press, Oxford (1993) 151-193
4. Chamorro-Martínez, J., Fdez-Valdivia, J.A., García, J.A., Martínez-Baena, J.: A frequency Domain Approach for the Extraction of Motion Patterns, in IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, Vol. 3 (2003) 165-168
5. Yu, W., Sommer, G., Daniilidis, K.: Three dimensional orientation signatures with conic kernel. Image and Vision Computing, Vol. 21(5) (2003) 447-458
6. Granlund, G.H., Knutsson, H.: Signal Processing for Computer Vision. Kluwer Academic Publishers, Boston (1995)
7. Pal, N.R., Biswas, J.: Cluster Validation Using graph Theoretic Concepts. Pattern Recognition, Vol. 30(6) (1997) 847-857

Visual Cortex Frontend: Integrating Lines, Edges, Keypoints, and Disparity

João Rodrigues[1] and J.M. Hans du Buf[2]

[1] University of Algarve - Escola Superior Tecnologia, Faro, Portugal
[2] University of Algarve - Vision Laboratory - FCT, Faro, Portugal

Abstract. We present a 3D representation that is based on the processing in the visual cortex by simple, complex and end-stopped cells. We improved multiscale methods for line/edge and keypoint detection, including a method for obtaining vertex structure (i.e. T, L, K etc). We also describe a new disparity model. The latter allows to attribute depth to detected lines, edges and keypoints, i.e., the integration results in a 3D "wire-frame" representation suitable for object recognition.

1 Introduction

During the last decade, the modeling of processes in the visual cortex has become a mature research topic. Models of cells, i.e. simple, complex and end-stopped, have been developed, e.g. [5,17]. In addition, models of bar and grating cells [12,13], line/edge detection [4,7,8,16] and disparity [3,11] have become available. Hence, it is now possible to develop a vision frontend that integrates all types of processing and that can be used to explore higher-level tasks like object recognition.

The basic syntax in the primary cortex seems to consist of lines (bars) and edges, also keypoints, in scale space. However, there is more going on: because of the ocular dominance columns in the primary cortex [6], which bring retinotopic, orientation-specific projections of the left and right eye closely together such that neural dendritic fields can cover both, we must assume that disparity estimation already starts at the first processing layers [18].

We present a 3D "wire-frame" representation based on an integrated frontend (Fig. 1 left) which allows to detect positive and negative lines and edges, keypoints, vertex structure and depth. We also present a new disparity model that is not based on phase [3] nor amplitude summations [11]. The basic idea is extremely simple: once we have line/edge detection, we also have access to the central, linear part of the Gabor responses. Below we first introduce keypoints with stabilizations and the new method for underlying vertex structure, then line/edge extraction with stabilizations, and finally the new disparity model.

2 Simple, Complex, and End-Stopped Cells

Line, edge and keypoint detection are based on the responses of simple, complex and end-stopped cells. Gabor quadrature filters provide a model of cortical simple

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 664–671, 2004.
© Springer-Verlag Berlin Heidelberg 2004

cells [9]. In the spatial domain they consist of a real cosine and an imaginary sine, both with a Gaussian envelope. Since all filtering is done in the frequency domain, we apply polar-separable transfer functions

$$G_{i,j}(f,\rho) = \exp\left(-\left(f-f_j\right)^2/2\sigma_{f_j}^2 - \left(\rho-\rho_i\right)^2/2\sigma_\rho^2\right),$$

with $i = 0,\ldots N_\theta - 1$ and $j = 0,\ldots N_s - 1$, σ_f and σ_ρ being the bandwidths in the frequency and orientation, and N_s and N_θ the number of scales and orientations ($N_\theta = 8$ in our case). We apply a linear scaling between $f_{\min}$ and $f_{\max}$, with the possibility of using many contiguous scales or a few noncontiguous scales with neighboring micro-scales.

In the spatial domain, the responses of even and odd simple cells, which correspond to the real and imaginary parts of the Gabor filters, are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale and i the orientation (i.e. $\theta_i \perp \rho_i$ and $\theta_i = i\pi/(N_\theta - 1)$). In order to simplify the notation, and because the same processing is done at all scales, we drop the subscript s. The responses of complex cells are modelled by the modulus $C_i(x,y) = [\{R^E_i(x,y)\}^2 + \{R^O_i(x,y)\}^2]^{1/2}$. There are two types of end-stopped cells [5,17], i.e. single (S) and double (D). If $[\cdot]^+$ denotes the suppression of negative values, $\mathcal{C}_i = \cos\theta_i$ and $\mathcal{S}_i = \sin\theta_i$, then

$$S_i(x,y) = \left[C_i(x + d\mathcal{S}_i, y - d\mathcal{C}_i) - C_i(x - d\mathcal{S}_i, y + d\mathcal{C}_i)\right]^+;$$

$$D_i(x,y) = \left[C_i(x,y) - \frac{1}{2}C_i(x + 2d\mathcal{S}_i, y - 2d\mathcal{C}_i) - \frac{1}{2}C_i(x - 2d\mathcal{S}_i, y + 2d\mathcal{C}_i)\right]^+.$$

The distance d is scaled linearly with the filter scale s, i.e. $d = 0.6s$. All end-stopped responses along straight lines and edges need to be suppressed, for which we use tangential (T) and radial (R) inhibition:

$$I^T(x,y) = \sum_{i=0}^{2N_\theta - 1} \left[-C_{i \bmod N_\theta}(x,y) + C_{i \bmod N_\theta}(x + d\mathcal{C}_i, y + d\mathcal{S}_i)\right]^+;$$

$$I^R(x,y) = \sum_{i=0}^{2N_\theta - 1} \left[C_{i \bmod N_\theta}(x,y) - 4 \cdot C_{(i+N_\theta/2) \bmod N_\theta}(x + \frac{d}{2}\mathcal{C}_i, y + \frac{d}{2}\mathcal{S}_i)\right]^+,$$

where $(i + N_\theta/2) \bmod N_\theta \perp i \bmod N_\theta$. Instead of applying the inhibition to individual end-stopped cells [5,17], slightly better results are obtained by applying $I = I^T + I^R$ once to the pooled activity:

$$K^S(x,y) = \sum_{i=0}^{N_\theta - 1} S_i(x,y) - gI(x,y) \quad \text{and} \quad K^D(x,y) = \sum_{i=0}^{N_\theta - 1} D_i(x,y) - gI(x,y),$$

with $g \approx 0.4$, after which the keypoint map is obtained: $K(x,y) = \max\{K^S(x,y), K^D(x,y)\}$, i.e. at each filter scale s: $K_s(x,y)$.

This cell model leads to the detection of many spurious events. The main reason for this problem lies in the Gabor filtering itself: at an L junction the

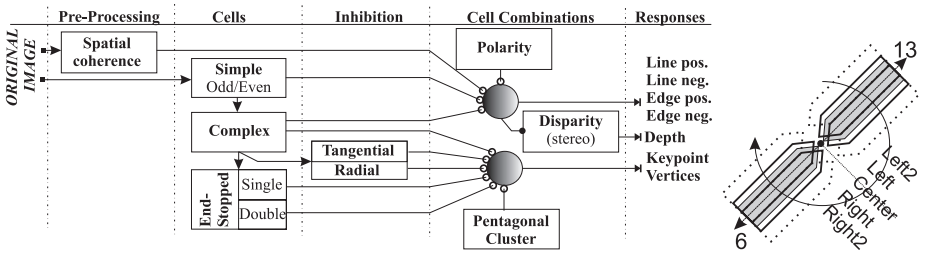


Fig. 1. Left: integrated frontend (see text); right: pentagonal cell clusters.

filters respond beyond the line/edge parts, and non-ideal edges lead to shifts of the even and/or odd responses parallel to the "edge center." For this reason the accuracy must be improved by postprocessing, and a further stabilization can be achieved by combining detection at multiple scales.

3 Keypoint Stabilization and Vertex Classification

For pattern recognition applications we want to obtain a clean, single-pixel keypoint map, and classify the keypoints according to the underlying vertex structure, i.e. K, L, T, + etc. All postprocessing of $K(x, y)$ is done in four steps, for each scale, after which different scales are combined.

First, local maxima of $K(x, y)$ in x and y are detected. If there is a small cluster of connected points with equal values, the centroid will be computed.

Second, for each local maximum (centroid) the responses of the complex cells $C_i(x, y)$ are analyzed in order to keep the *dominant orientations* θ_i: we mark all orientations for which (a) $C_i > C_{i-1} \wedge C_i > C_{i+1}$, where i is cyclic over 0 and $N_\theta - 1$, and (b) C_i exceeds a threshold value of $0.05 \cdot C_{\mathrm{PA}}$. C_{PA} is the average of all complex cell responses pooled over a 3×3 neighborhood and all orientations: $C_{\mathrm{PA}} = 1/9 \cdot \sum_{3\times3} C_{\mathrm{P}}$ and $C_{\mathrm{P}} = \sum_{i=0}^{N_\theta - 1} C_i$. Then, the number of dominant orientations are counted in a 9×9 neighborhood, and those with a count less than 0.05 of the total count are discarded.

Third, in the case of e.g. T and L vertices the complex cells respond beyond the keypoints in dominant orientations, so now we need to analyze opposite *directions*. Insignificant directions are eliminated by probing the dominant orientations on a line in opposite directions starting at the local maximum (centroid), until a distance of d_1 which is linearly scaled with the filter scale. If the number of dominant orientations found is below $d_1/2$, directions are rejected.

Fourth, probed and passed directions are further confirmed by analyzing the dominant orientations, now with three pentagonal cell clusters used for each probed direction (width w and length d_1, we use 3 and 6 respectively).

All the constants and mask sizes depend on filter scale (i.e. are scaled linearly). Figure 1 (right) shows the central pentagonal cluster (shaded) and shifted ones (orthogonally ± 1 pixel, dark outline) for directions 6 and 13. Only those directions are kept which have consistent dominant orientations in at least one

cluster. Keypoints which have no consistent directions in at least one cluster are eliminated. A final correction of ± 1 pixel of the position is applied by using again shifted pentagonal clusters (shift is ± 2). Figure 1 shows these with dotted outlines. Hence, in direction 13 there are five masks: L2, L, C, R and R2. If the dominant orientations are consistent in L2, L and C, but not in R and R2, the keypoint is shifted one pixel to the upper-left. The final keypoint shift equals the summed shifts of all directions, but never more than 1 pixel up/down and left/right. Now only keypoints at vertices between lines and/or edges remain, and these are labelled with the line/edge directions, i.e., the vertex type (K, L, T, + etc.) is also available (see Fig. 2 bottom-left).

In a final step the keypoint stability in scale space is confirmed by considering, around each scale, a small scale interval: 4 *micro*-scales, i.e. two scales slightly finer and two slightly coarser than the actual scale. In the case of the smallest (largest) scale, four coarser (finer) scales are applied. Only keypoints which are consistent over 3 neighboring micro-scales are accepted.

4 Line/Edge Stabilization and Classification

Van Deemter and du Buf [16] presented a scheme for line and edge detection based on the responses of simple cells. A positive line is detected where R^E shows a local maximum in the orthogonal filter orientation and R^O shows a zero crossing. In the case of an edge the even and odd responses must be swapped. This gives 4 possibilities for positive and negative events: local maxima/minima plus zero crossings. Here we combine the responses of simple and complex cells, i.e. simple cells serve to detect positions and event types, whereas complex cells are used to increase the confidence. Since the use of Gabor modulus (complex cells) implies some loss of precision at vertices [2] we increase precision by considering multiple scales.

For each orientation the simple cells responses (R_i^E or R_i^O) orthogonal to the orientation are checked for a local maximum (or minimum), until a distance of $\pm\lambda/4$, λ being the wavelength of the Gabor filter. All positions that do not show a local maximum (minimum) are discarded. Then complex cells (C_i) are checked, using the same process. Only at positions that pass the previous tests the quadrature filter is checked for a zero crossing on $\pm\lambda/4$. If so, the position is accepted and the event type has been determined (Fig. 2 top-right, where four different colors represent the line/edge events).

Finally, polarity and spatial coherence are checked. Event polarity must be corrected when, due to interference effects, the Gabor responses are distorted. We correct individual points of lines and edges if their polarity differs from that of the neighboring points. Spatial coherence is improved by suppressing all events at positions where the local variance of the *input* image is too low. This is necessary to suppress line/edge events beyond keypoints in insignificant directions (e.g. at L and T junctions). As is done in the case of keypoints, the coherence is also improved by checking the results at 4 neighboring micro-scales.

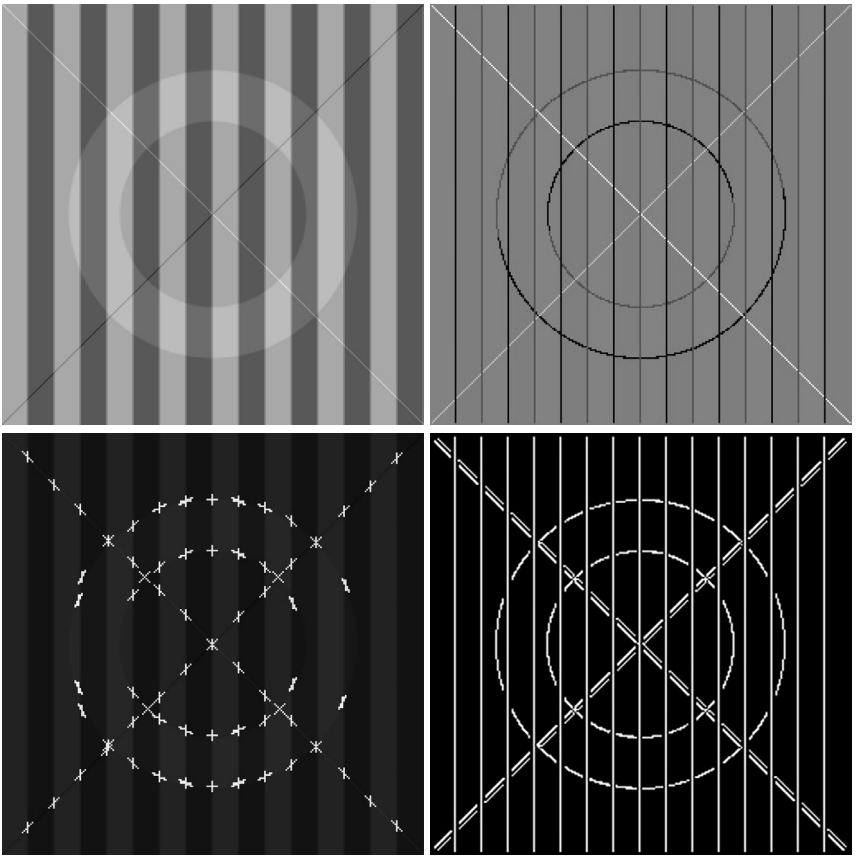


Fig. 2. Ledge image (top-left), line/edge detection (top-right), keypoints with vertex structure (bottom-left). For comparison, the result of Canny's edge operator is shown bottom-right.

5 Disparity Estimation

Our new disparity model is based on the central, linear part of the Gabor responses, i.e. the sinusoidal part with $\sin x \approx x, |x| < \pi/4$. Assuming ideal events, i.e. lines with a Dirac profile and edges with a Heaviside step profile, or nonideal ones obtained by Gaussian filtering, and complex Gabor filters with the same orientation, the responses are (scaled) Gabor functions and complex errorfunctions. It has been shown that the latter can be approximated by scaled Gabor functions [2]. In other words, both line and edge responses are essentially scaled Gabor functions with the sinusoidal part, real or imaginary, being linear on $\pm\lambda/8$; see Fig. 3. One step in line/edge detection consists of checking the Gabor response $R_i^O(x,y)$ (the odd, imaginary part in the case of a line), or $R_i^E(x,y)$ (the odd, real (!) part in the case of an edge) for a zero crossing on $\pm\lambda/4$ (Fig. 3 right). Here, for disparity, we apply the same event detection steps to two images, left

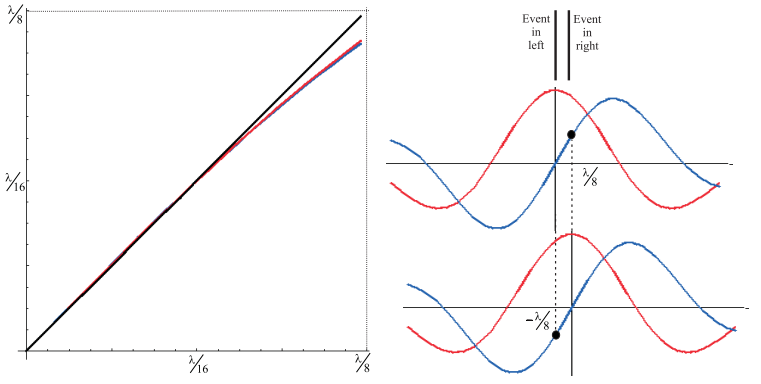


Fig. 3. Left: linear Gabor responses on $[0, \lambda/8]$, at a line (red) and at an edge (blue); right: disparity detection (see text).

and right. In the case of the *left* image, we (1) check the existence of an event of the same type in the *right* image on $\pm\lambda/8$, and (2) if so, we take $\pm R^O$ or $\pm R^E$ of the *right* image at the event (zero crossing) position in the *left* image. The sign depends on the event polarity and, in order to obtain values which do not depend on the event amplitude, $\pm R^O$ or $\pm R^E$ is divided by the modulus (complex cell response) of the *left* image, which is maximum at the event position. After this normalization yet another one is applied: the response is divided by the scale s of the filter. Hence, the slope of the linear response part will not depend on the event amplitude nor on the filter scale, i.e. disparity estimates obtained at different scales will be the same. The same processing can be done in the case of the right image, by exchanging *left* and *right*. Of course, the disparity estimates need to be calibrated once using real data, like the way babies need to learn in the first months. One problem we encountered were small fluctuations of the disparity estimates, especially at the finest scales. These are due to the fact that we need to work at discrete pixel positions, and the maximum of the modulus used in the first normalization is therefore not the theoretical maximum. We solved this by averaging disparity estimates over neighboring micro-scales.

6 Results and Discussion

Figure 5 shows the application of line/edge and keypoint detection to traffic signs. Single- and multi-scale stabilization have eliminated many spurious keypoints, one of which is shown by the small diamond of 4 pixels (zoomed image). All keypoints of the van have been detected, but three directions are still missing (encircled). Here the structures have a size of 2 to 4 pixels: we are at the limit of what can be achieved by using Gabor wavelets. At the moment we are experimenting with image zooming in order to be able to work with structures of smaller size. Disparity estimation is shown in Fig. 4. The stereo images were obtained by shifting left, in one image (Fig. 2 top-left) of a pair, the first

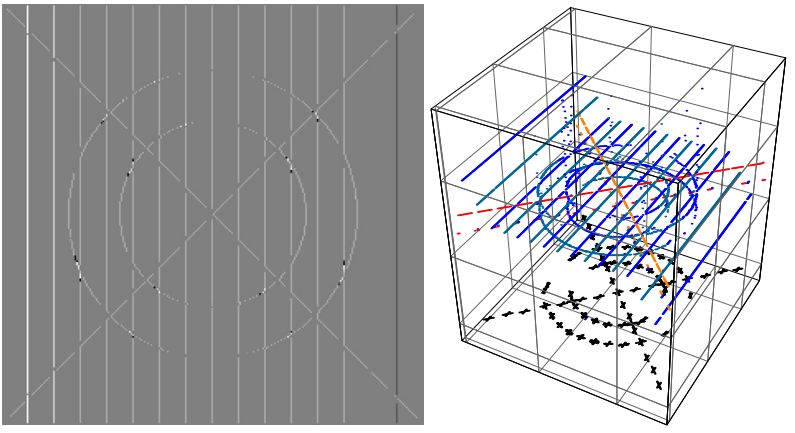


Fig. 4. Left: ledge disparity; right: 3D representation of lines, edges, depth and vertices.

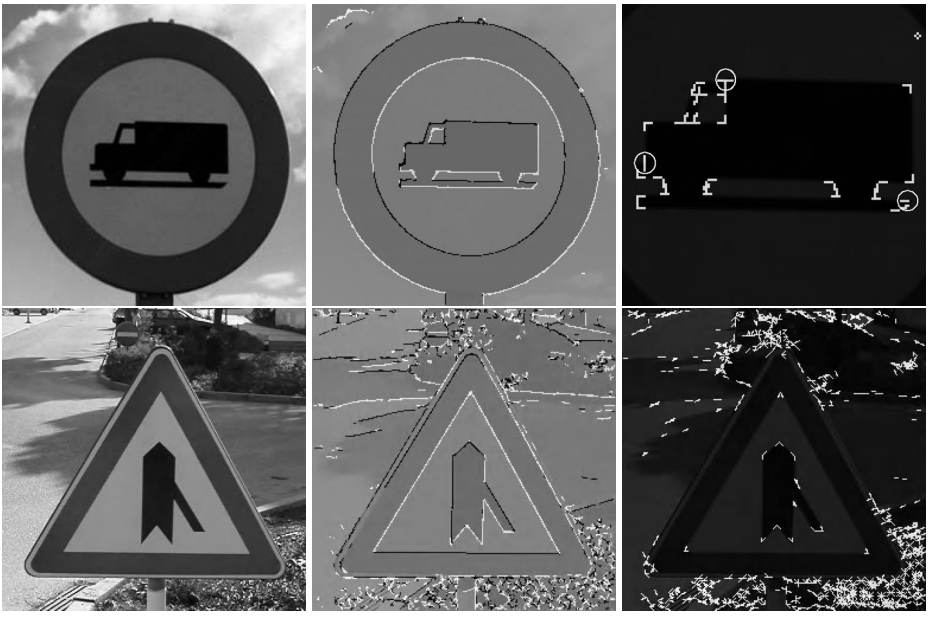


Fig. 5. Top: sign2 image, line/edge detection, keypoints with vertex detection after multi-scale stabilization (zoomed). Bottom: sign6.

vertical edge 3 pixels, the following edge 2, and the next edges 1 pixel. The second-last edge was not changed, whereas the last one was shifted right. The diagonal lines and the ring were shifted left 1 pixel. Different colors in 2D (Fig. 4 left) represent depth, which can be projected in 3D (right). There are still some problems around keypoints, and experiments with real images showed that the interval $\pm\lambda/8$ of the filters that we use is too small, even of the biggest filters.

The reason is that these filters are the smallest ones in the frequency domain, and a reasonable approximation of a Gaussian function requires a few samples. This problem is being solved by creating bigger filters, doing the filtering by convolution in the spatial domain.

The main conclusion is that it is now possible to create a 3D "wireframe" representation (Fig. 4 right) in which lines, edges, keypoints, vertex structure and disparity are all integrated by one extraction "process". This will simplify 3D object recognition. The same might occur in our visual cortex, although this is still speculative. Finally, to the best of our knowledge there does not exist, also in "non-biological" work, the extraction of all attributes that we can achieve in one process, see "SUSAN" [14], "Lowe's SIFT" [10] and e.g. [1,15].

References

1. R. Bergevin and A. Bubel. Object-level structured contour map extraction. *Comp. Vis. and Image Underst.*, 91:302–334, 2003.
2. J.M.H du Buf. Responses of simple cells: events, interferences, and ambiguities. *Biol. Cybern.*, 68:321–333, 1993.
3. D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.
4. C. Grigorescu, N. Petkov, and M.A. Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE Tr. Im. Proc.*, 12(7):729–739, 2003.
5. F. Heitger and et al. Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Res.*, 32:963–981, 1992.
6. D.H. Hubel. *Eye, brain and vision.* Scientific American Library, 1995.
7. M. Kolesnik and A. Barlit. Iteractive orientation tuning of simple cells in V1: A comparative study of two computational models for contrast detection in images. *Proc. Symp. Biol. Mach. Vis., AISB Convention*, pages 114–127, 2003.
8. P. Kovesi. Image feactures from phase congruency. *J.C. Vis. Res.*, 1(3):2–27, 1999.
9. T.S. Lee. Image representation using 2D Gabor wavelets. *IEEE Tr. PAMI*, 18(10):pp. 13, 1996.
10. K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based feactures. *Brit. Mach. Vis. Conf., Norwich*, 2003.
11. I. Ohzawa, G.C. DeAngelis, and R.D. Freeman. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.*, 18(77):2879–2909, 1997.
12. N. Petkov and P. Kruizinga. Computational models of visual neurons specialised in detection of periodic and aperiodic visual stimuli. *Biol. Cybern.*, 76:83–96, 1997.
13. L.M. Santos and J.M.H du Buf. Computational cortical cell models for continuity and texture. *Biol. M. Comp. Vis. Work., Tuebingen*, 2002.
14. S.M. Smith and J.M. Brady. Susan - a new approach to low level image processing. *Int. J. Comp. Vis.*, 23(1):45–78, 1997.
15. A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Tr. PAMI*, 22(9):1226–1238, 2002.
16. J.H. van Deemter and J.M.H. du Buf. Simultaneous detection of lines and edges using compound Gabor filters. *Int. J. Patt. Recogn. Artif. Intell.*, 14:757–777, 1996.
17. R.P. Würtz and T. Lourens. Corner detection in color images by multiscale combination of end-stopped cortical cells. *Artif. N. Net. - ICANN'97*, 1997.
18. K. Yoshiyama, T. Uka, H. Tanaka, and I. Fujita. Architecture of binocular disparity processing in monkey inferior temporal cortex. *Neur. Res.*, 48:155–167, 2004.

Estimation of Directional and Ambient Illumination Parameters by Means of a Calibration Object*

Alberto Ortiz and Gabriel Oliver

Department of Mathematics and Computer Science,
University of the Balearic Islands, Spain
{alberto.ortiz,goliver}@uib.es

Abstract. Assuming a model composed of both parallel directional lighting and ambient illumination, this paper presents a method using a spherical calibration object to estimate the orientation of directional lighting together with the strengths of both types of light source. A set of experimental results showing the performance of the method proposed is included and discussed at the end of the paper.

1 Introduction

The study of the detection of the illumination distribution from images can be traced back to the 1980's. Since then, several light source estimators have been proposed (see [1,2, 3], among many others). The primary motivation behind the work was to extend shape from shading algorithms to operate with less *a priori* information, although only single point light source cases were considered. More recently, researchers have developed techniques for dealing with multiple point light sources, which extend the applicability of the estimations to other areas such as image-based computer graphics or photo-realistic image synthesis (see [4,5,6,7,8], among the most recent ones).

The method presented in this paper considers one directional point light source and ambient illumination and is based on the use of a calibration sphere. Despite its limitations as for the number of light sources, several advantages make it attractive in single-source cases against multi-source methods: (1) in general, its complexity is lower than the multi-source methods; (2) a previous geometric calibration of the camera is not required and the estimation of the surface normal vectors of the calibration sphere is embedded within the method itself; (3) all the image pixels belonging to the sphere projection are used in the estimation of the lighting parameters; (4) no threshold or parameter needs to be set up. Besides, the mathematical formulation of the method has been developed taking into account a model of the operation of CCD cameras [9], which has allowed identifying the noise sources corrupting digital pixel values in order to counteract their effects throughout the estimation process; furthermore, such a formulation has revealed that any estimation of the strength of ambient illumination always includes dark current, a parameter related to the performance of the camera, so that it must be estimated apart to get the real strength of ambient illumination. To finish, this lighting estimation method has been successfully used within a strategy for calibrating CCD cameras at the

* This study has been partially supported by project CICYT-DPI2001-2311-C03-02 and FEDER funds.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 672–679, 2004.
© Springer-Verlag Berlin Heidelberg 2004

radiometric level, to estimate, precisely, camera performance parameters such as dark current or the distribution of the fixed pattern noise (see [10] for the details). In this case, as in other applications, particularly shape from shading, an illumination model consisting of one directional point light source together with ambient illumination is enough.

The rest of the paper is organized as follows: section 2 describes the model of image formation; sections 3 and 4 describe the lighting parameters estimation method; section 5 presents experimental results; and, finally, conclusions appear in section 6.

2 Interaction of Light with Matter

It is generally accepted that objects reflection is an additive composition of body or diffuse reflection and interface or specular reflection [11]. Besides, this model is enhanced by a further term accounting for non-directional or ambient lighting which interacts with the scene increasing objects radiance irrespectively of local surface geometry. All in all, radiance at a scene point p and for a wavelength λ can be summarized as indicated in equation 1:

$$L(p;\lambda) = \overbrace{L_a(\lambda)\rho_a(p;\lambda)}^{L_a(p;\lambda)} + \overbrace{m_b(p)\left[L_d(\lambda)\rho_b(p;\lambda)\right]}^{L_b(p;\lambda)} + \overbrace{m_i(p)\left[L_d(\lambda)\rho_i(p;\lambda)\right]}^{L_i(p;\lambda)}, \quad (1)$$

where: (i) $L_a(\lambda)$ represents light coming from all directions in equal amounts while $L_d(\lambda)$ is for directional lighting; (ii) ρ_a, ρ_b and ρ_i are surface material reflectances expressing the fraction of the incoming light which is conveyed by the corresponding reflection component (ambient, body and interface, respectively), being ρ_a assumed a linear combination of the body and interface reflectances, ρ_b and ρ_i; (iii) m_b is a term dependent on local surface geometry, which is most times given as $m_b(p) = \cos\theta(p)$, where $\theta(p)$ is the angle between the unit direction towards the light source at p, $\boldsymbol{s}(p)$, and the unit surface normal at p, $\boldsymbol{n}(p)$, and, thus, $m_b(p) = \boldsymbol{s}(p) \cdot \boldsymbol{n}(p)$; and, finally, (iv) m_i is a geometrical term for interface reflection.

3 Estimation of Lighting Parameters from Single Band Images

General Procedure. In case of a uniformly coloured matte object ($\rho_b(\lambda) = \rho_b(p;\lambda) = \rho_a(p;\lambda)$, $\rho_i(p;\lambda) = 0, \forall p$ in the object), equation 2 results from equation 1 for colour channel c after adding the corresponding camera noise terms from the model proposed in [9]:

$$D^c(i,j) = K(i,j)(L_a\rho_b)^c A^c + E_{DC}A^c + m_b(i,j)K(i,j)(L_d\rho_b)^c A^c + N^c(i,j)\,, \quad (2)$$

where: (i) D^c is the final digital value given by the camera at image location (i,j); (ii) K represents a random variable of mean 1 which accounts for the site-to-site nonuniformities among image cells due to processing errors during CCD fabrication, also called *fixed pattern noise*; (iii) E_{DC} is the expected *dark current* generated by thermal energy

at every CCD cell; (iv) A^c is the camera gain for colour channel c, and (v) N^c is a zero-mean additive term noise accounting for the so-called *shot*, *readout* and *quantization* noises. (In case gray-level calibration images were considered, superscript c would be removed from the formulation.)

The relevance of equation 2 is that it reveals a (noisy) linear relationship under the aforementioned circumstances between D^c and m_b for all the object pixels. Therefore, if enough m_b values can be related to the corresponding D^c values, so that some pairs (m_b, D^c) can be obtained, those (noisy) pairs can be fitted by a straight line $D^c = \alpha^c + m_b\beta^c$ which allows estimating $\alpha^c = (L_a\rho_b)^c A^c + E_{DC}A^c$ and $\beta^c = (L_d\rho_b)^c A^c$. The noise in the pairs (m_b, D^c), which comes from N^c and the spatial variation in K, can be significantly removed if the D^c values for those pixels corresponding to the same m_b are averaged and this average $\overline{D}^c$ is used in the fitting. If the matte object under consideration is white ($\rho_b(\lambda) = 1$), once the straight line parameters α^c and β^c are known, β^c is an estimation of the strength of the directional lighting of the scene $L_d^c A^c$, while, once known $E_{DC}A^c$, $\alpha^c - E_{DC}A^c = L_a^c A^c$ is an estimate of the strength of the ambient illumination, both for colour channel c.

The following sections show that surface normal vectors $\boldsymbol{n}$ and the lighting direction $\boldsymbol{s}$, which are essential for computing m_b, can be determined from the elliptical region which corresponds to the projection of the calibration sphere over the image plane.

Estimation of Surface Normal Vectors. The surface normal vector $\boldsymbol{n}(x, y, z)$ for a sphere centered at (x_0, y_0, z_0) and having radius R is given by equation 3:

$$\boldsymbol{n}(x, y, z) = \left(\tfrac{x-x_0}{R}, \tfrac{y-y_0}{R}, -\sqrt{1 - \left(\tfrac{x-x_0}{R}\right)^2 - \left(\tfrac{y-y_0}{R}\right)^2} \right)^T . \tag{3}$$

Given $(x, y) = \frac{z}{f}(u, v)$, where (u, v) is the projection of (x, y, z):

$$n_x = \frac{\frac{z}{f}u - x_0}{R} = \frac{u}{f}\frac{z}{R} - \frac{x_0}{R}\,,\; n_y = \frac{\frac{z}{f}v - y_0}{R} = \frac{v}{f}\frac{z}{R} - \frac{y_0}{R}\,. \tag{4}$$

In order for these equations to be useful, f, $\frac{z}{R}$, $\frac{x_0}{R}$ and $\frac{y_0}{R}$ should be known beforehand. In the following, f and the ellipse fitting the contour of the sphere projection are used to derive expressions to determine the other values.

On the one hand, given the focal distance f, $\frac{x_0}{R}$, $\frac{y_0}{R}$ and $\frac{z_0}{R}$ can be put as a function of f and the parameters of that ellipse. First of all, given the relationship between the center of the sphere and the center of the ellipse $(x_0, y_0) = \frac{z_0}{f}(u_0, v_0)$, if both equations are divided by R, expressions for $\frac{x_0}{R}$ and $\frac{y_0}{R}$ are obtained in terms of $\frac{z_0}{R}$:

$$\frac{x_0}{R} = \frac{u_0}{f}\frac{z_0}{R}, \quad \frac{y_0}{R} = \frac{v_0}{f}\frac{z_0}{R}\,. \tag{5}$$

Secondly, $\frac{z_0}{R}$ can be determined from some geometrical relationships between the sphere and its projection over the image plane. These relationships are established in figure 1, being the plane depicted in figure 1(b) the one containing the triangle $(0, 0, 0) - (0, 0, f) - (u_0, v_0, f)$ of figure 1(a). Over this plane, angle ψ happens to be:

$$\tan\psi = \frac{R}{\sqrt{x_0^2 + y_0^2 + z_0^2}} = \frac{r}{\sqrt{u_0^2 + v_0^2 + f^2}}\,, \tag{6}$$

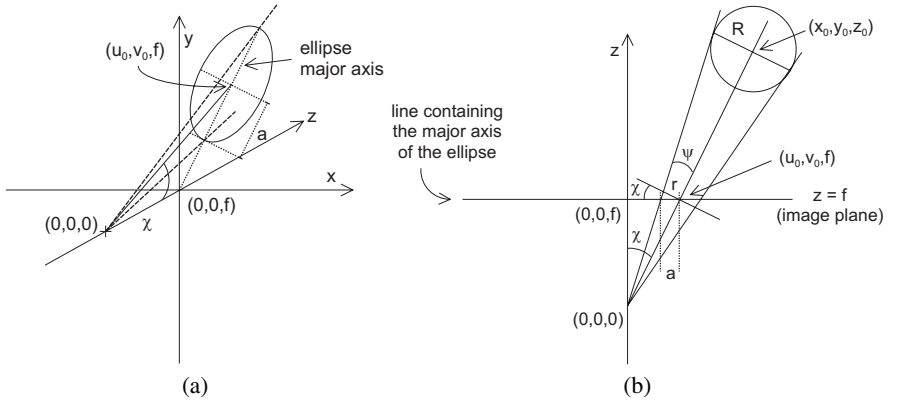


Fig. 1. Graphs used to derive $\frac{z_0}{R}$: (a) sphere projection over the image plane; (b) view of (a) after rotating the viewpoint so as to put it on top of the ellipse major axis.

from which, substituting (x_0, y_0) by $\frac{z_0}{f}(u_0, v_0)$, $\frac{z_0}{R} = \frac{f}{r}$ results. Now, since $r = a\cos\chi$ (see figure 1(b)) and $\cos\chi = \frac{f}{\sqrt{u_0^2+v_0^2+f^2}}$ (see figure 1(a)), $\frac{z_0}{R}$ is definitely given by:

$$\frac{z_0}{R} = \frac{\sqrt{u_0^2 + v_0^2 + f^2}}{a}\,. \tag{7}$$

On the other hand, $\frac{z}{R}$ is a data structure of values the same size as the sphere projection which can be determined using the equation of the sphere surface $(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2$, substituting $x = u\frac{z}{f}$ and $y = v\frac{z}{f}$ and dividing the resulting equation by R^2, to achieve equation 8:

$$\left(\frac{u}{f}\frac{z}{R} - \frac{x_0}{R}\right)^2 + \left(\frac{v}{f}\frac{z}{R} - \frac{y_0}{R}\right)^2 + \left(\frac{z}{R} - \frac{z_0}{R}\right)^2 = 1\,, \tag{8}$$

from which, after a reordering of terms, a second-order polynomial in $\frac{z}{R}$ is obtained:

$$\left[\left(\frac{u}{f}\right)^2 + \left(\frac{v}{f}\right)^2 + 1\right]\left(\frac{z}{R}\right)^2 - 2\left[\frac{u}{f}\frac{x_0}{R} + \frac{v}{f}\frac{y_0}{R} + \frac{z_0}{R}\right]\left(\frac{z}{R}\right) + \left[\left(\frac{x_0}{R}\right)^2 + \left(\frac{y_0}{R}\right)^2 + \left(\frac{z_0}{R}\right)^2 - 1\right] = 0\,. \tag{9}$$

Estimation of the Lighting Direction. If the light coming from the directional light source is assumed distant, $\boldsymbol{s}(i,j) \approx \boldsymbol{s} = (s_x, s_y, s_z)$ throughout the scene. In this way, using the expressions developed for n_x and n_y (equation 4), and defining p and q as $p = \frac{u}{f}\frac{z}{R} - \frac{x_0}{R}$ and $q = \frac{v}{f}\frac{z}{R} - \frac{y_0}{R}$, $\cos\theta$ results to be:

$$\cos\theta = ps_x + qs_y - \left(\sqrt{1 - p^2 - q^2}\right)s_z\,. \tag{10}$$

$\boldsymbol{s}$ can now be determined from the intensity pattern of the sphere projection. In effect, assuming for the moment $N^c(i,j) = 0$ and $K(i,j) = 1$, the isophote curve of intensity L of the image (i.e. $D^c = L$), without considering the background, is given by:

$$L = \alpha^c + \left[ps_x + qs_y - \left(\sqrt{1-p^2-q^2}\right) s_z\right] \beta^c \,, \tag{11}$$

where $\alpha^c = (L_a \rho_b)^c A^c + E_{DC} A^c$ and $\beta^c = (L_d \rho_b)^c A^c$.

Equation 12 can now be obtained reordering equation 11:

$$\left(s_x^2 + s_z^2\right) p^2 + \left(s_y^2 + s_z^2\right) q^2 + 2 s_x s_y pq - 2\mathcal{C} s_x p - 2\mathcal{C} s_y q + \mathcal{C}^2 - s_z^2 = 0 \,, \tag{12}$$

where $\mathcal{C} = \frac{L-\alpha^c}{\beta^c}$. Clearly, equation 12 defines a rotated conic in terms of the variables p and q. The eigenvalues of the corresponding quadratic form are $\lambda_1 = s_x^2 + s_y^2 + s_z^2 = 1$ and $\lambda_2 = s_z^2$. Since both are positive (i.e. have the same sign), the conic results to be an ellipse. On the other hand, the coordinates of the center of the ellipse (p_0, q_0) are given by $p_0 = \mathcal{C} s_x$ and $q_0 = \mathcal{C} s_y$, while the slopes m_1 and m_2 of the axes of the ellipse are $m_1 = s_y/s_x$, for the shortest axis (i.e. the one related to λ_1), and $m_2 = -s_x/s_y$, for the largest axis. Therefore, if (p,q) points corresponding to a given isophote curve are fitted by an ellipse, $s_z = -\sqrt{\lambda_2/\lambda_1}$ (curve fitting methods typically introduce a scale factor into the curve parameters which also affects λ_1 and λ_2, so that s_z cannot be directly recovered from λ_2), while the tilt of the illumination $\tau = s_y/s_x$ is given by $\tau = \tan^{-1} m_1$, being the precise angular quadrant determined by the signs of $(p_0, q_0) = (\mathcal{C} s_x, \mathcal{C} s_y)$, since $\mathcal{C} \geq 0$. Taking into account that the slant of the illumination σ is $\cos^{-1} s_z$, then the lighting orientation can be completely recovered as $\boldsymbol{s} = (\sin\sigma\cos\tau, \sin\sigma\sin\tau, \cos\sigma)$. The accuracy of the estimation of $\boldsymbol{s}$ can be enhanced if the isophote curves for colour channel c and several intensities L are considered and the corresponding estimates $\widehat{\boldsymbol{s}}^{L,c}$ of $\boldsymbol{s}$ are properly aggregated [10].

Finally, as for the noise terms N^c and K: (1) given the fact that N^c is a zero-mean noise term, N^c vanishes if several images of the calibration sphere are taken and the estimation is performed over the average image; and (2) the effect of the pattern noise K can be counteracted by means of the ellipse fittings using the sets of points whose intensities lie within $[L - \Delta/2, L + \Delta/2]$, for small Δ, instead of just using the isophote curve of intensity L. $\Delta = 4$ has resulted a good election in all the experiments performed.

Final Considerations. Since the parameters of the sphere projection —i.e. the ellipse center, (u_0, v_0), and the length of its main axis, a—, determine the value of $\boldsymbol{n}(i,j)$, and consequently of $m_b(i,j)$, the accuracy of the estimation of the lighting parameters depends on how well the sphere projection data have been calculated. The joint effect of spatial discretization and aliasing in real cameras make uncertain where the real contour of the sphere projection lies, what suggests using a sort of optimization strategy by which those parameters leading to the best estimation are chosen. Given the fact that, in the end, the sphere parameters are needed to estimate the lighting strengths through a linear model which relates m_b with the observations $\overline{D}^c$, it seems reasonable to look for the ellipse parameters leading to the lowest fitting error to that model. Accordingly, the user is asked to point out two circumferences so that one is completely included into the other and the right contour is most likely to be among both. Besides, a first guess about

the ellipse corresponding to the real contour is also required. Next, an elitist genetic strategy is applied. That is to say, several ellipses are generated by Gaussian mutation of the parameters of the first guess. The one leading to the best fitting of points $(m_b, \overline{D}^c)$ is kept for the next generation, where again other ellipses are produced by mutation. After some generations without improvement, the best ellipse is considered to be the best estimation of the real projection contour.

4 Lighting Parameters Estimation from Colour Images

If the procedure described in section 3 is applied to several colour channels, a slightly different estimation of vector $\boldsymbol{s}$ can be obtained for every colour channel, together with the corresponding estimates of $L_a^c A^c$ and $L_d^c A^c$. In order to obtain an only $\boldsymbol{s}$, it is proposed to aggregate the estimations $\widehat{\boldsymbol{s}}^{L,c}$ of all the colour channels, instead of aggregating over every channel separately.

On the other hand, the availability of colour data gives rise to several groups of pairs $(m_b, \overline{D}^c)$, one for every colour channel, which are fitted to determine the illumination strengths within every colour band. As a consequence, a number of fitting errors appear, as many as colour channels, for every ellipse considered as contour of the sphere projection. In order to ensure the best matching with the theoretical linear relationships between m_b and $\overline{D}^c$ for all c, the objective function is redefined as the maximum of the fitting errors of the different colour channels available. In this way, the same optimization procedure described in section 3 can be applied, without other changes, to colour calibration images.

5 Experimental Results

In the following, the results of a number of experiments which have been performed in order to evaluate the robustness of the estimation procedure described in this paper are discussed.

In the first experiment, the lighting estimation method was faced against synthetic images with increasing levels of zero-mean additive Gaussian noise and different light direction tilts, but the same strengths for ambient and directional lighting. In figure 2, the uppermost row of plots shows the angle in degrees between true and estimated $\boldsymbol{s}$, true and estimated L_a^c and true and estimated L_d^c. The lowermost row corresponds to the absolute difference between the norms of the true and the estimated lighting strengths. As it can be seen, the estimates of $\boldsymbol{s}$ are quite accurate even for high levels of noise, being the error angle below 1 degree in all cases. The same applies for the angle between true and estimated L_a and L_d, which means the colour of ambient and directional illumination is accurately estimated even under noisy conditions. As for illumination strengths, the estimates appear slightly more affected by noise.

In the second experiment, 20 images were taken, varying the F-number of the camera from one image to the next, but without altering the geometrical relationship between camera and calibration object, so that $\boldsymbol{s}$ kept constant. Therefore, the application of the estimation method to these 20 images should yield very similar estimates for $\boldsymbol{s}$ in the different cases, although $L_a^c A^c + E_{DC} A^c$ and $L_d^c A^c$ would vary accordingly. A histogram of the difference angles between the estimations of $\boldsymbol{s}$ for every possible pair

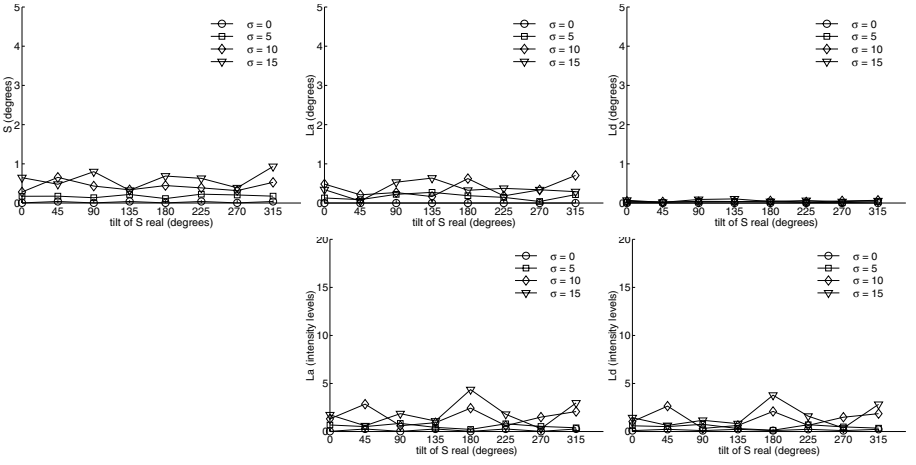


Fig. 2. Results for robustness experiment 1. (σ is the standard deviation, in intensity levels, of the zero-mean Gaussian noise added to images.)

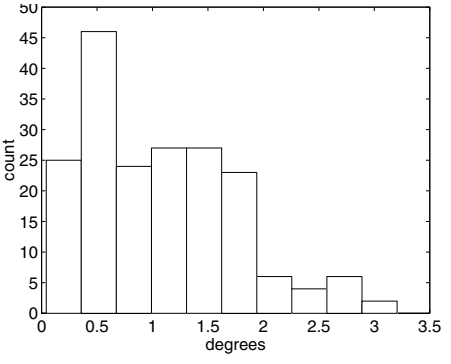


Fig. 3. Results for robustness experiment 2.

of images of the set is given in figure 3. From this distribution, the mean difference angle results to be 1.08°, while the maximum difference angle reaches 3.20°.

Finally, in the third experiment, the camera was rotated and moved eight times, without changing the parameters of the point light source, in order to check the robustness of the estimates of $L_a^c A^c + E_{DC} A^c$ and $L_d^c A^c$ against changes in $\boldsymbol{s}$. The corresponding estimates are given in table 1(top), while some statistics are provided at the bottom in the same table, with: $\alpha^c = L_a^c A^c + E_{DC} A^c$, $\beta^c = L_d^c A^c$, $\boldsymbol{\alpha} = (\alpha^R, \alpha^G, \alpha^B)$, $\boldsymbol{\beta} = (\beta^R, \beta^G, \beta^B)$, $|\cdot|$ is the vectors norm, avg$\{\cdot\}$ and $\sigma.$ stand for, respectively, mean and standard deviation, $255\sqrt{3}$ is the maximum achievable norm of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{v}, \boldsymbol{w}}$ is the angle between vectors $\boldsymbol{v}$ and $\boldsymbol{w}$.

Table 1. Results for robustness experiment 3.

# image	$\boldsymbol{s}$	$\boldsymbol{\alpha} = (\alpha^R, \alpha^G, \alpha^B)$	$\boldsymbol{\beta} = (\beta^R, \beta^G, \beta^B)$
1	(-0.72,-0.25,-0.65)	(21.65,17.75,15.42)	(177.61,172.49,164.66)
2	(-0.75,-0.13,-0.65)	(21.00,17.43,15.23)	(175.15,169.27,161.33)
3	(-0.76,0.09,-0.64)	(19.91,16.76,14.82)	(173.46,167.49,160.47)
4	(-0.72,0.23,-0.65)	(23.62,20.47,18.46)	(173.54,168.07,161.12)
5	(-0.74,-0.30,-0.60)	(21.48,17.54,15.18)	(182.54,176.74,169.12)
6	(-0.74,-0.35,-0.57)	(24.22,19.93,17.11)	(185.64,180.22,173.28)
7	(-0.87,-0.16,-0.47)	(19.49,14.98,12.22)	(163.74,158.84,152.62)
8	(-0.75,0.35,-0.56)	(21.26,18.07,15.83)	(168.51,162.50,155.69)

$\frac{\sigma_{|\boldsymbol{\alpha}|}}{\text{avg}\{|\boldsymbol{\alpha}|\}} \times 100 = 8.98\%$ $\frac{\sigma_{|\boldsymbol{\alpha}|}}{255\sqrt{3}} \times 100 = 0.65\%$ $\text{avg}\{\widehat{\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j}\} = 1.67°$ $\max\{\widehat{\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j}\} = 5.08°$

$\frac{\sigma_{|\boldsymbol{\beta}|}}{\text{avg}\{|\boldsymbol{\beta}|\}} \times 100 = 4.10\%$ $\frac{\sigma_{|\boldsymbol{\beta}|}}{255\sqrt{3}} \times 100 = 2.72\%$ $\text{avg}\{\widehat{\boldsymbol{\beta}_i, \boldsymbol{\beta}_j}\} = 0.14°$ $\max\{\widehat{\boldsymbol{\beta}_i, \boldsymbol{\beta}_j}\} = 0.31°$

6 Conclusions

A method for estimating scene lighting parameters has been presented. The method provides an orientation for the directional lighting and measures of the strength of ambient and directional illumination. Results about its robustness have shown it reasonably stable as for lighting orientation against changes in lighting strength, and vice versa. Results for synthetic noisy images have also been provided. A complete set of experiments, including a comparison with other well-known estimation methods, which have not been included here due to lack of space, can be found in [10].

References

1. Lee, C., Rosenfeld, A.: Improved methods of estimating shape from shading using the light source coordinate system. Artificial Intelligence **26** (1985) 125–143
2. Pentland, A.: Local shading analysis. PAMI **6** (1984) 170–187
3. Zheng, Q., Chellappa, R.: Estimation of illuminant direction, albedo and shape from shading. PAMI **13** (1991) 680–702
4. Bouganis, C.S., Brookes, M.: Multiple light source detection. PAMI **26** (2004) 509–514
5. Sato, I., Sato, Y., Ikeuchi, K.: Illumination from shadows. PAMI **25** (2003) 290–300
6. Wang, Y., Samaras, D.: Estimation of multiple illuminants from a single image of arbitrary known geometry. In: Proceedings of ECCV. (2002) 273–288
7. Zhang, Y., Yang, Y.: Multiple illuminant direction detection with application to image synthesis. PAMI **23** (2001) 915–920
8. Zhou, W., Kambhamettu, C.: Estimation of illuminant direction and intensity of multiple light sources. In: Proceedings of ECCV. (2002) 206–220
9. Healey, G., Kondepudy, R.: Radiometric CCD camera calibration and noise estimation. PAMI **16** (1994) 267–276
10. Ortiz, A., Oliver, G.: Scene lighting parameters estimation and radiometric camera calibration. Technical Report A-2-2003, (DMI, Univ. de les Illes Balears) (2003)
11. Shafer, S.: Using color to separate reflection components. COLOR Research and Application **10** (1985) 210–218

Environment Authentication Through 3D Structural Analysis

Toby P. Breckon and Robert B. Fisher

Institute of Perception, Action and Behaviour
School of Informatics, University of Edinburgh
Edinburgh, UK
toby.breckon@ed.ac.uk / rbf@inf.ed.ac.uk
http://www.ipab.inf.ed.ac.uk

Abstract. We address the validation of the current sensed environmental state against a model known from earlier perception. Our surface based approach compares 3D range data of a built environment against an *a priori* scene model with the analysis of identified differences facilitating the hypothesis of causal structural changes within the scene. Experimental results show good success rates in identifying and analysing realistic structural changes introduced across example industrially themed scenes.

1 Introduction

Environment authentication is used to describe the validation of the current built environment against some *a priori* model in order to ascertain where changes may have occurred. Here we look at this technique with reference to building interiors where several applications, notably in the domain of mobile robotics and the nuclear/chemical safeguards industry, call for robust and generalised change identification and analysis methods [1]. In the latter case, the primary concern is to detect covert changes that may compromise safety or security [2].

Previous work in this area falls into two related but distinct categories - firstly the general topic of scene and object recognition and secondly the more specific area of scene change identification. The former, with specific relation to range data, has been well established through previous studies [3] of which [4] brought several of these techniques together to present a complete range-based 3D recognition system dealing with each stage of acquisition, segmentation and recognition. This work embodied a clear architecture for use in this field, primarily dividing the process into two tiers - the model independent tier involving the acquisition and segmentation of a 3D scene image and the model based reasoning tier handling issues of successful model to scene surface matching. Here we adopt this architecture for scene recognition and extend it by adding an additional post-process of structural change reasoning.

Work on the analysis of structural change within scenes is reasonably limited. Fillatreau et al [5] considered feature mis-placement with regard to validating the correctness of an *a priori* model based on two stage tight and then relaxed

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 680–687, 2004.
© Springer-Verlag Berlin Heidelberg 2004

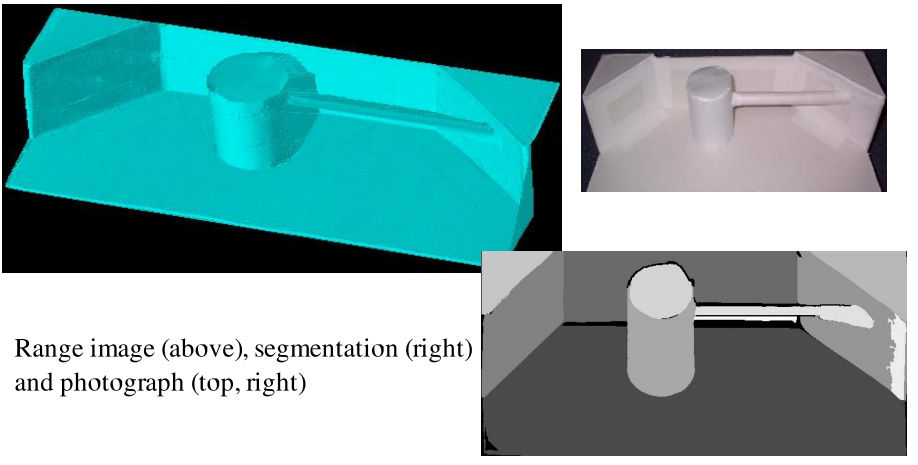


Fig. 1. Scene segmentation example

constraint based matching. However, in this work [5] perfect segmentation is assumed and only localised feature movement is considered within the scene. Additionally, a significant body of work also exists within the remit of remote sensing (e.g. [6,7]) but this fails to address structural analysis within the localised environment concentrating instead on aerial or satellite image analysis. Similar work [8,9,10] has concentrated more on image differencing than on structural change detection, whilst work specifically with the remit of environment authentication [11] has shown the limitations of range occupancy grids in addressing problems in this domain.

Here we examine an alternative approach that targets full scene understanding through the use of classical surface matching techniques, augmented with additional structural reasoning, to provide a causal hypothesis for structural changes present in a possibly mis-segmented environment. Our results show that a surface based approach combined with later reasoning, based on spatial and geometric scene awareness, can provide a successful approach to 3D environment authentication.

2 Structural Analysis

Here structural analysis is based upon the comparison of a range image of the current environment against a known scene model. For this work we limit ourselves to the consideration of simple, industrially themed building interiors - where the geometric scene nature lends itself well to established model registration and comparison techniques [12].

Initially segmentation is performed using mean and Gaussian curvatures, with additional region growing and surface fitting techniques to provide a surface map of the sensed environment (Fig. 1) together with a parameterised

surface description - surface type[1], orientation, position and also radii where appropriate [13]. Due to the potential effects of noise on range imaging some mis-segmentation is to be expected and must be isolated from true cases of structural change.

For simplicity, our *a priori* scene model is represented using a basic scheme of generalised polyhedral, circular and cylindrical surfaces although a VRML or CAD model could similarly be employed.

The initial stages of processing follow the common invocation, matching and verification architecture of [4,12]. Our work here, however, focuses on a later stage of post-processing - *match analysis* of environment to model differences with a view of hypothesising causal structural changes within the scene.

Prior to detailing this aspect of our work in depth we briefly describe our classical approach in these earlier stages of processing. Firstly, invoked matching is used to produce a 'coarse match which is fast and inexpensive' based on the common position invariant surface attributes readily available from our representations. This lightweight match is then refined using a standard interpretation tree matching approach to provide a set of mutually consistent scene model to data surface matches. Here matching is considered through the consideration of unary and N-ary surface consistency - the matching of the individual and relative surface positions and orientation (within defined noise tolerances identified from ground truth environmental data) [3,12].

Once this matching process has identified a consistent set of surface matches, with a subset containing mutually non-parallel and independent surface orientations, the established SVD least squares fitting method of [14] is used to calculate the model to scene registration.

This registration is then used to both verify existing surface matches and find further matches in the scene based on a process of surface reprojection. Each scene surface is projected onto the model using the known registration and geometrically tested for correspondence against *a priori* model surfaces [15, 12]. Once this verification process is complete all possible surface matches that can be found under normal strict matching conditions have been identified.

The next stage, an augmentation to the classical recognition architecture of [4,12] and extending the earlier work of [5], considers the analysis of the remaining unmatched surfaces by using relaxed matching to target full scene understanding through explanation of these occurrences - *match analysis*.

Match Analysis

The match analysis stage has two goals: 1) to isolate unmatched surface cases occurring due to occlusion and mis-segmentation; 2) to form a structural change hypothesis for remaining unmatched surfaces.

[1] {plane — cylinder — cone — sphere — general quadric} - however, only planar and cylindrical surfaces are fully considered in this work.

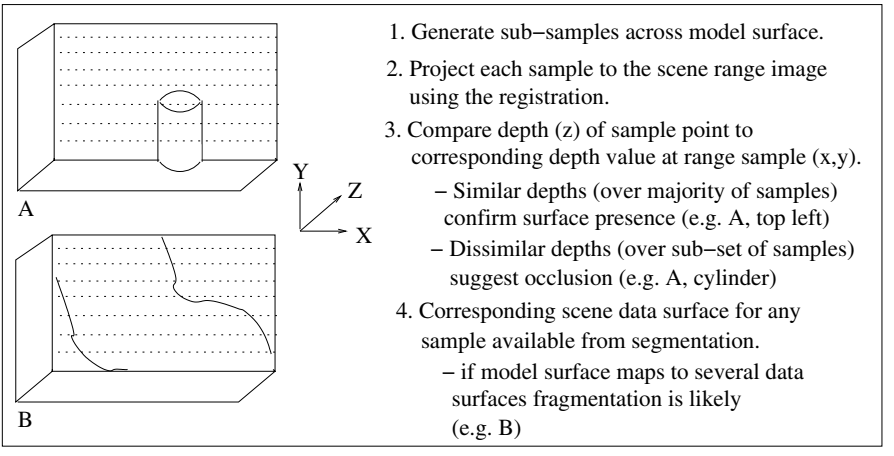


Fig. 2. Fragmentation / occlusion identification

Initially the remaining unmatched *a priori* model surfaces are screened to eliminate back facing (i.e. hidden) surfaces given the known model registration (back face culling, [16]). Next, a stage of occlusion and fragmentation analysis identifies surface matches missed by earlier processing because either they are a) partially occluded within the scene or b) they have been mis-segmented into one or more smaller **surface fragments**. Here these occurrences are detected by sub-sampling each unmatched model surface and projecting it onto the scene using the known model registration. By comparing the expected depth (model) and the measured depth (range image) at each sample, the true presence of these surfaces in the scene can be ascertained and instances of partial surface occlusion and fragmented surface presence identified (Fig. 2). This allows surfaces to be successfully matched to an occluded partner or a set of occurring surface fragments.

The final stage of this pre-classification analysis is to perform a 'mop-up' of remaining segmentation errors occurring within the scene data by identifying data fragments within the segmented scene. A **data fragment** is a surface considered to be too small to be of realistic significance and most likely to have been created erroneously due to the effects of noise on surface segmentation. Such surfaces are identified, based on size, and discarded from the remainder of the analysis.

Once these cases occurring due to mis-segmentation and occlusion have been identified and removed, the proper structural analysis can commence. The remaining unmatched scene and model surfaces are now classified as structural change based on four possible instances:

– **Movement:** when a surface match can be found by relaxing the matching constraints to consider only the position invariant surface attributes and

there exists further evidence, based on relative surface position, that the surface has moved within the scene.
- **Shape Change:** when a surface match can be found within the corresponding locale of the scene but differences in the shape attributes indicate the surface has changed shape within its original position.
- **Missing:** when no surface can be found to match an *a priori* model surface.
- **New:** when no model surface can be found to match an surface present in the scene environment.

In practise this classification uses the following top-down rule-set, which is evaluated for each remaining unmatched scene surface in terms of how it can be matched to a corresponding surface present in the model:

- *IF* match(size / shape attributes) *AND NOT* match(position)
  - **Movement**
- *ELSE IF* match(type, orientation and position)
  - **Shape Change**
- *ELSE*
  - **New**

Any remaining unmatched model surface is classified as **Missing**. All matching is performed based on identified noise tolerances which are derived empirically from measuring scene noise levels in ground truth environment to model matches.

In summary, our match analysis process operates sequentially as two components - firstly it eliminates possible causes of erroneous structural change detection by identifying occlusion and mis-segmentation instances, and secondly it classifies the remaining structural changes based on available evidence within the scene. Once this process is complete, every surface within the scene has been accounted for as either present, hidden, mis-segmented or forming part of a hypothesis explaining the structural change present in the scene. Overall our target of full scene understanding has been achieved.

3 Results

The match analysis process described above was tested over example structural change scenarios containing both single and multiple structural change occurrences (Table 1).

These were constructed using range scans of scale building interior models scanned with a 3D Scanners Reversa 25 laser range scanner[2] (e.g. Fig. 1) . The type of changes introduced varied from subtle changes in surface positions and sizes to more complex and realistic changes such as alterations to plant within a scene (e.g. Fig. 3) and the detection of false walls / sealed doorways (e.g. Fig. 3 / Fig. 4). Figure 3 shows the successful detection of a false back wall (A), occurring in the scene forward of its expected model position, as well as movement (D)

[2] x/y resolution: 0.4mm, depth accuracy (z): 50 microns

Table 1. Experimental Results

	Scenario	% Success	Surface Examples Tested
G	Ground truth match	94%	104
S	*Single changes within the scene*	87.5%	56
M	*Multiple changes within the scene*	87%	46
F	*Surface fragmentation within the scene*	88%	17

and change (B) in the central 'water tank' configuration. Additional surface fragments (E / F), hidden surfaces (G / H) and a missing support cylinder (C) are similarly identified. Figure 4 shows the successful detection of a sealed doorway in the central rear wall of the scene (A) and again successful surface fragment detection (B/C).

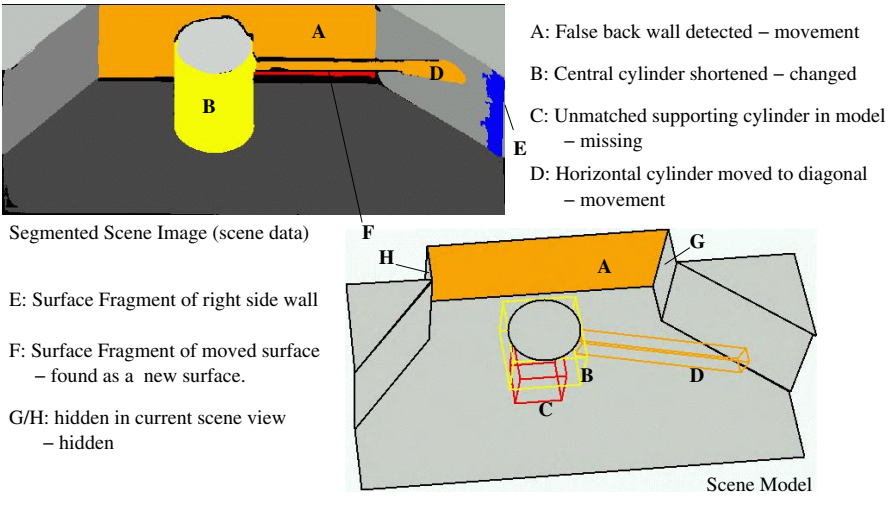


Fig. 3. Successful false wall, plant changes and surface fragment detection.

Overall the system was successful in detecting 87% of the 102 structural changes (cases S and M from Table 1) introduced to the test scenes. In the remaining cases a combination of poor segmentation and surface fitting, due to scene noise in the range capture, reduced sensitivity. Lesser sensitivity is equally apparent when surfaces move subtly within their defining plane and when surfaces are heavily occluded within the scene. This is due to the limited availability of structural change evidence outwith the tolerances of regular scene noise. Additionally, the process is susceptible to any errors present in achieving a ground truth match and to identifying movement cases in mis-segmented surfaces (Fig. 3, item F).

Segmentation related errors were also successfully detected in testing with all data fragments and 88% of surface fragments being correctly identified (case F from Table 1, e.g. Fig. 3, item E). In all cases where surface fragments were missed a high level of scene noise caused poor segmentation making successful surface matching difficult. Further advances in noise tolerant segmentation and 3D data acquisition may help to counter this issue.

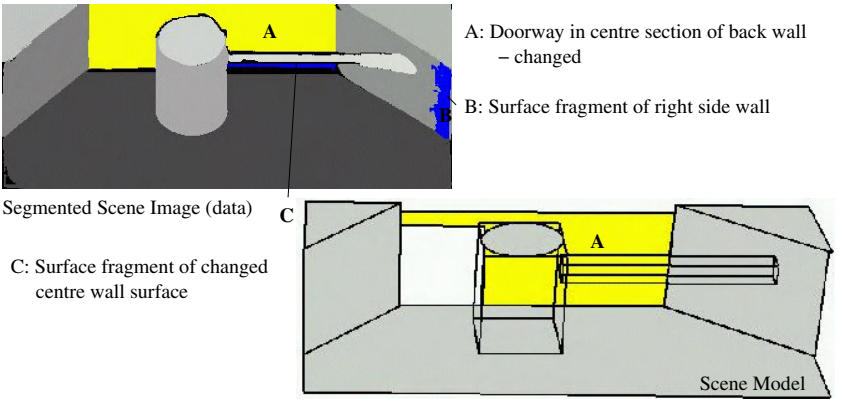


Fig. 4. Detection of sealed doorway and surface fragments.

4 Conclusions

We have presented a method for environment authentication and structural change analysis through full scene understanding of a segmented environment range image. Here we overcome the limitations of earlier work [11,8,5] by identifying individual surface changes, handling mis-segmentation and occlusion based issues and performing analysis to form an explanative hypothesis for identified structural changes.

Although the results quoted show a significant level of success for these techniques, further work is still required in a number of areas. Notably, improvements in range image segmentation and surface extraction together with those in occluded surface reconstruction [17] offer possibilities for improving accuracy and reducing the reliance on current mis-segmentation and occlusion handling techniques. Similarly, improvements in match analysis techniques to consider more advanced best fit based hypothesis construction, the use of further scene analysis evidence metrics and also the use of probability based classification may be beneficial in future work. Additionally, the extension of this or similar techniques, to consider the matching of meshes representing curved surfaces and the consideration of higher-level 'changed structures' as consistent groupings of underlying surface changes may be of future interest in advancing work in this area.

Acknowledgements. This work was partially supported by studentship funding from the EPSRC and CASE sponsorship by QinetiQ PLC *(http://www.qinetiq.com)*.

References

1. Latombe, J.: Robot Motion Planning. Kluwer (1991)
2. Government, H.M.: Confidence, security and verification - the challenge of global nuclear weapons arms control. Technical report, Atomic Weapons Establishment, UK (2000)
3. Grimson, W.: Object Recognition by Computer: The Role of Geometric Constraints. MIT Press (1991)
4. Fisher, R., Fitzgibbon, A., Waite, M., Trucco, E., Orr, M.: Recognition of complex 3D objects from range data. In Impedovo, S., ed.: Proc. 7th Int. Conf. on Image Analysis and Processing, Monopoli, Bari, Italy (1993) 509–606
5. Fillatreau, P., Fisher, R.: Non-polyhedral landmark recognition using 3D depth images and partial correct models. In: Proceedings 3rd Int. Symp. Intelligent Robotic Systems. (1995) 271–278
6. Ridd, M., Liu, J.: A comparison of four algorithms for change detection in an urban environment. Remote Sensing of Environment **63** (1998) 95–100
7. Huertas, A., Nevatia, R.: Detecting changes in aerial views of man-made structures. Image and Vision Computing **18** (2000) 583–596
8. Benjanin, M., Huertas, A., Medioni, G., Nevatia, R.: Model validation for change detection. In: DARPA Image Understanding Workshop, Monterrey, CA, USA. (1994)
9. Rosin, P.: Thresholding for change detection. In: Int. Conf. on Computer Vision. (1998) 274–279
10. Murakami, H., Nakagawa, K., Hasegawa, H., Shibata, T., Iwanami, E.: Change detection of buildings using an airborne laser scanner. ISPRS J. of Photogrammetry and Remote Sensing **54** (1999) 148–152
11. Campos, G., Goncalves, J.: Environment authentication using range occupancy grids. JRC Technical Note I93.130, EC JRC, Institute for Systems Engineering and Informatics, Ispra, Italy (1993)
12. Faugeras, O.: Three-dimensional computer vision: a geometric viewpoint. MIT Press (1993)
13. Fisher, R., Fitzgibbon, A., Eggert, D.: Extracting surface patches from complete range descriptions. In: Proc. Int. Conf. on Recent Advances in 3D Digital Imaging and Modelling, Ottawa, Canada (1997)
14. Arun, K., Huang, T., Blostein, S.: Least-squares fitting of two 3D point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence **9** (1987) 698–700
15. Haines, E.: Point in polygon strategies. In Heckbert, P., ed.: Graphics Gems IV, Academic Press (1994) 24–46
16. Foley, J., van Dam, A., Feiner, S., Hughes, J.: Computer Graphics: Principles and Practice. Addison-Wesley (1996)
17. Castellani, U., Livatino, S., Fisher, R.: Improving environment modelling by edge occlusion surface completion. In: Proc. Int. Symp. on 3D Data Processing Visualisation and Transmission, Padova, Italy (2002)

Camera Calibration Using Two Concentric Circles

Francisco Abad*, Emilio Camahort, and Roberto Vivó

Universidad Politécnica de Valencia, Camino de Vera s/n, Valencia 46021, Spain
{fjabad, camahort, rvivo}@dsic.upv.es,
http://www.sig.upv.es

Abstract. We present a simple calibration method for computing the extrinsic parameters (pose) and intrinsic parameters (focal length and principal point) of a camera by imaging a pattern of known geometry. Usually, the patterns used in calibration algorithms are complex to build (three orthogonal planes) or need a lot of features (checkerboard-like pattern). We propose using just two concentric circles that, when projected onto the image, become two ellipses. With a simple mark close to the outer circle, our algorithm can recover the full pose of the camera.
Under the perfect pinhole camera assumption, the pose and the focal length can be recovered from just one image. If the principal point of the camera has to be computed as well, two images are required. We present several results, using both synthetic and real images, that show the robustness of our method.

1 Introduction

In the past two decades, several methods have been proposed for calibrating a camera by taking images of a pattern with known geometry. First in photogrammetry and then in computer vision, researchers have developed methods to recover a camera's extrinsic parameters (position and orientation) and intrinsic parameters (focal length and principal point). Those methods usually require expensive laboratory settings, or use complex fiducials [1].

In order to take the computer vision from the laboratory to the home user, robust, inexpensive and effective techniques are needed. In this paper, we present an algorithm that easily recovers the pose and the focal length of a camera by taking a single photo of a simple calibration pattern. We use a pattern made of two concentric circles of known radii, usually printed on a sheet of paper. We show how this pattern can be used in a simple setup to recover the camera parameters. Our method can be applied to camera tracking and related problems like robotics, entertainment and augmented reality.

This paper is organized as follows. The next section presents previous work in the field of camera calibration with circular markers. Section 3 presents the

* This work was partially funded by the Programa de Incentivo a la Investigación of the Polytechnic University of Valencia, and by project TIC2002-04166-C03-01 of Spanish Ministry of Science and Technology

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 688–696, 2004.
© Springer-Verlag Berlin Heidelberg 2004

theoretical model and mathematical foundations of our work. In the following section we present some results of our method, and discuss the tests we run with both synthetic and real data. Our paper finishes with some conclusions and directions for future work.

2 Previous Work

Early work that used conics for computer vision applications was reported in [2, 3,4]. Circular markers have been extensively used in tracking applications due to their robustness properties [6,7]. Kim et al. [8,9] proposed a calibration method using two concentric circles. Their algorithm requires some initial information about the camera to get an initial value for the intrinsic matrix. They define a cost function on the calibration parameters and minimize it. This method only recovers the normal of the marker's supporting plane.

Another method that recovers the supporting plane of the circles was proposed in [10]. The method computes the plane's normal and a point on it expressed in camera coordinates. The method assumes that the principal point is at the center of the image.

Unlike the previous methods, our algorithm does not require any a priori information about the camera parameters to calibrate it. Furthermore, we recover the pose (the full rotation matrix and the translation vector) using a simple marker. Finally, we also compute the position of the principal point.

3 Calibrating the Camera

3.1 Detecting the Marker

Our marker is composed of two concentric circles or radii r_1 and r_2, and an exterior mark that intersects with a circle of radius r_3 (see Fig. 1).

The ellipses can be automatically recovered from an image by applying standard methods in Computer Vision. Pixel chains are extracted from the image and ellipses are fitted with, e.g., Fitzgibbon's algorithm [5]. See for example [7] for an explanation of an automatic extraction algorithm. To find the X axis mark, a circle of radius r_3 has to be projected using the same (unknown) camera as the other two. In an Appendix we explain how to project a circle of arbitrary radius concentric to two circles whose projections are known.

3.2 Pinhole Projection of a Circle

The pinhole camera configuration (assuming zero skew and square pixels) is usually described using an intrinsic parameter matrix (A), that describes the focal length and principal point (see Fig. 2), and an extrinsic parameter matrix (M), that establishes the camera pose (position and orientation) from a given global coordinate system:

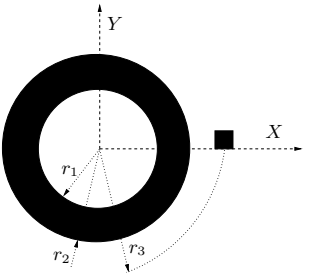


Fig. 1. Design of our fiducial

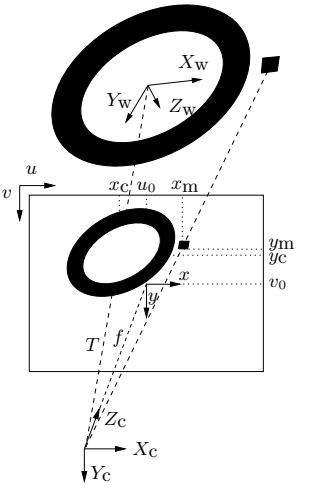


Fig. 2. Pinhole camera in the scene

$$A = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \qquad M = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_X \\ R_{21} & R_{22} & R_{23} & T_Y \\ R_{31} & R_{32} & R_{33} & T_Z \end{bmatrix} \quad . \tag{1}$$

In Fig. 2, the world coordinate system (WCS) has its origin at the center of the concentric circles. Those circles are in the $\boldsymbol{X}_W\boldsymbol{Y}_W$ plane of the WCS, so the $\boldsymbol{Z}_W$ axis is perpendicular to them. The projection operator (P), that computes the image pixel coordinates that corresponds to a 3D point in WCS is $P = AM$.

Given a point $\boldsymbol{X}$ in WCS, equation $\lambda \boldsymbol{x} = P\boldsymbol{X}$ computes its homogeneous coordinates in the image coordinate system. The two circles are located on the plane $Z_W = 0$, so we can write:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11}X_W + p_{12}Y_W + p_{14} \\ p_{21}X_W + p_{22}Y_W + p_{24} \\ p_{31}X_W + p_{32}Y_W + p_{34} \end{bmatrix} \quad .$$

If we assume that the image coordinate system is centered at the principal point of the image, then $u_0 = 0$ and $v_0 = 0$ in (1) and we can write (see [10]):

$$X_W = \frac{\boldsymbol{x}^t\,(\boldsymbol{R_2} \times \boldsymbol{T})}{\boldsymbol{x}^t \boldsymbol{R_3}} \quad \text{and} \quad Y_W = \frac{\boldsymbol{x}^t\,(\boldsymbol{T} \times \boldsymbol{R_1})}{\boldsymbol{x}^t \boldsymbol{R_3}} \quad , \tag{2}$$

where $\boldsymbol{x} = \begin{bmatrix} u\ v\ f \end{bmatrix}^t$, and $M = \begin{bmatrix} \boldsymbol{R_1}\ \boldsymbol{R_2}\ \boldsymbol{R_3}\ \boldsymbol{T} \end{bmatrix}$.

In the WCS, the exterior circle of radius r_2 has coordinates $C(X_W, Y_W) = X_W^2 + Y_W^2 - r_2^2 = 0$. Substituting (2) in this equation and factoring, we can express the exterior circle in terms of the image coordinate system as follows:

$$C'(x, y) = Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \ . \tag{3}$$

3.3 Recovering the Circles' Projected Center

Under perspective projection any conic is transformed into another conic. Specifically, circles are transformed into elipses when imaged by a camera. The projected center of the original circle, however, does not generally coincide with the center of the ellipse in the image. The projected center of the circles has to be computed in order to recover the normal to the supporting plane. The direction of the $\boldsymbol{Z}_{\mathrm{W}}$ axis in the camera coordinate system (or $\boldsymbol{R_3}$) can be computed as follows [2,3]:

$$\begin{bmatrix} R_{13} \\ R_{23} \\ R_{33} \end{bmatrix} = \pm N \left\{ Q \begin{bmatrix} x_{\mathrm{C}} \\ y_{\mathrm{C}} \\ f \end{bmatrix} \right\} \quad , \tag{4}$$

where $(x_{\mathrm{C}}, y_{\mathrm{C}})$ are the coordinates of the projected circle center in the image coordinate system (see Fig. 2), N represents the normalization to a unit vector, and Q is the matrix that describes the ellipse, as defined in [3]:

$$Q = \begin{bmatrix} A & B/2 & D/2f \\ B/2 & C & E/2f \\ D/2f & E/2f & F/f^2 \end{bmatrix} \quad . \tag{5}$$

Parameters A to F are those defined in (3) and f is the focal length. Two methods to recover the projected center of two concentric circles can be found in [8] and [9]. In the Appendix we present our own original method.

3.4 Recovering the Pose

Each parameter of the ellipse in (3) can be expressed in terms of f^2, f^2/T_{Z}^2 and a constant term by substituting (5) in (4) [10]. This derivation uses the properties of the rotation matrices and the following relations derived from the pinhole camera model in Fig. 2:

$$T_{\mathrm{X}} = \frac{T_{\mathrm{Z}} x_{\mathrm{C}}}{f} \quad \text{and} \quad T_{\mathrm{Y}} = \frac{T_{\mathrm{Z}} y_{\mathrm{C}}}{f} \quad . \tag{6}$$

The result is

$$Q' = \begin{bmatrix} \alpha_1^2 & -\alpha_1^2 r_2^2 & (\alpha_1^2+\alpha_2^2)y_{\mathrm{C}}^2 + 2\alpha_2\alpha_3 y_{\mathrm{C}} + \alpha_3^2 \\ 2\alpha_1\alpha_2 & -2\alpha_1\alpha_2 r_2^2 & -2(\alpha_2\alpha_3 x_{\mathrm{C}} + (\alpha_1^2+\alpha_2^2)x_{\mathrm{C}}y_{\mathrm{C}} + \alpha_1\alpha_3 y_{\mathrm{C}}) \\ \alpha_2^2 & -\alpha_2^2 r_2^2 & (\alpha_1^2+\alpha_2^2)x_{\mathrm{C}}^2 + 2\alpha_1\alpha_3 x_{\mathrm{C}} + \alpha_3^2 \\ -2\alpha_1(\alpha_1 x_{\mathrm{C}} + \alpha_2 y_{\mathrm{C}}) & -2\alpha_1\alpha_3 r_2^2 & -2\alpha_3(\alpha_3 x_{\mathrm{C}} + \alpha_2 x_{\mathrm{C}} y_{\mathrm{C}} - \alpha_1 y_{\mathrm{C}}^2) \\ -2\alpha_2(\alpha_1 x_{\mathrm{C}} + \alpha_2 y_{\mathrm{C}}) & -2\alpha_2\alpha_3 r_2^2 & -2\alpha_3(-\alpha_2 x_{\mathrm{C}}^2 + \alpha_1 x_{\mathrm{C}} y_{\mathrm{C}} + \alpha_3 y_{\mathrm{C}}) \\ (\alpha_1 x_{\mathrm{C}} + \alpha_2 y_{\mathrm{C}})^2 & -\alpha_3^2 r_2^2 & \alpha_3^2(x_{\mathrm{C}}^2 + y_{\mathrm{C}}^2) \end{bmatrix} \tag{7}$$

where:

$$\alpha_1 = \frac{2Ax_{\mathrm{C}} + By_{\mathrm{C}} + D}{2} \quad \alpha_2 = \frac{Bx_{\mathrm{C}} + 2Cy_{\mathrm{C}} + E}{2} \quad \alpha_3 = \frac{Dx_{\mathrm{C}} + Ey_{\mathrm{C}} + 2F}{2} \quad .$$

Therefore, (3) can be expressed as:

$$C'(x,y) = \left(Q' \begin{bmatrix} f^2 \\ f^2/Tz^2 \\ 1 \end{bmatrix} \right)^t G = 0 \ , \tag{8}$$

where $G = \begin{bmatrix} x^2 & xy & y^2 & x & y & 1 \end{bmatrix}^t$. The unknowns to be computed are f^2 and f^2/T_{Z}^2, so we rearrange (8) to leave the constant terms on the right-hand side of the expresion:

$$\left(\begin{bmatrix} q_{11} & q_{21} & q_{31} & q_{41} & q_{51} & q_{61} \\ q_{12} & q_{22} & q_{32} & q_{42} & q_{52} & q_{62} \end{bmatrix} G \right)^t \begin{bmatrix} f^2 \\ \frac{f^2}{T_{\mathrm{Z}}^2} \end{bmatrix} = - \begin{bmatrix} q_{13} & q_{23} & q_{33} & q_{43} & q_{53} & q_{63} \end{bmatrix} G \ , \tag{9}$$

where q_{ij} is the element of row i, column j of matrix Q' in (7). Given N points of the ellipse in the image we can build an N-degree over-determined system $WX = B$:

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ \vdots & \vdots \\ W_{N1} & W_{N2} \end{bmatrix} \begin{bmatrix} f^2 \\ \frac{f^2}{T_{\mathrm{Z}}^2} \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{bmatrix} \ , \tag{10}$$

where W_{i1}, W_{i2} and B_i are computed using (9) with (x, y) replaced by the coordinates (x_i, y_i) of the i-th point on the ellipse.

This system can be solved using the least square pseudo-inverse technique:

$$\begin{bmatrix} f^2 \\ \frac{f^2}{T_{\mathrm{Z}}^2} \end{bmatrix} = \left(W^t W\right)^{-1} W^t B \ .$$

Solving the system leads to f and T_{Z}. The components of $\boldsymbol{R_3}$ can be computed by replacing f in (4). T_{X} and T_{Y} can be recovered from (6).

Following the previous steps we recover the normal to the plane that contains the circles ($\boldsymbol{R_3}$) and the position of the origin of the WCS in camera coordinates ($\boldsymbol{T}$) (see Fig. 2). Fremont [10] proposed a calibration pattern that uses three orthogonal planes to recover the other two axes ($\boldsymbol{X}_{\mathrm{W}}$ and $\boldsymbol{Y}_{\mathrm{W}}$). Instead, we use a single mark on the exterior circle that defines the $\boldsymbol{X}_{\mathrm{W}}$ direction, an idea that has been used before in marker detection [7].

Given the pixel coordinates of the $\boldsymbol{X}_{\mathrm{W}}$ axis mark in the image, we reproject it onto the plane of the concentric circles. That plane is completely defined by its normal ($\boldsymbol{R_3}$) and a point on it ($\boldsymbol{T}$). Let the X axis mark position be $(r_3, 0, 0)$ in WCS, $(x_{\mathrm{m}}, y_{\mathrm{m}})$ in image coordinates, and $\boldsymbol{X}_{\mathrm{m}}$ in camera coordinates. Then $\boldsymbol{X}_{\mathrm{m}} = \mu \begin{bmatrix} x_{\mathrm{m}} & y_{\mathrm{m}} & f \end{bmatrix}^t$ where

$$\mu = \frac{-D}{R_{13}x_{\mathrm{m}} + R_{23}y_{\mathrm{m}} + R_{33}f} \ ,$$

and $D = -\boldsymbol{R_3}^t\boldsymbol{T}$. Having the 3D coordinates of the $\boldsymbol{X}_\mathrm{W}$ axis mark given in camera coordinates, and the 3D coordinates of the origin of the WCS, given in camera coordinates as well, the $\boldsymbol{X}_\mathrm{W}$ axis (or $\boldsymbol{R_1}$) is defined by $\boldsymbol{X}_\mathrm{W} = N\{\boldsymbol{X}_\mathrm{m} - \boldsymbol{T}\}$ where N is a normalization operator. Obviously, in a right-handed coordinate system, $\boldsymbol{Y}_\mathrm{W} = \boldsymbol{Z}_\mathrm{W} \times \boldsymbol{X}_\mathrm{W}$, or $\boldsymbol{R_2} = \boldsymbol{R_3} \times \boldsymbol{R_1}$.

3.5 Recovering the Principal Point

So far we have assumed that the optical axis of the camera is perfectly centered at the image (i.e., the principal point is the center of the image). In this section we remove this assumption and compute the principal point using the results of the previous sections.

Due to the error in the estimation of the principal point, reprojecting the original circle using the parameters computed in the previous sections does not produce the ellipses in the image. This misalignment is proportional to the error incurred in the estimation of the position of the principal point. By minizating that error, the principal point can be recovered. When processing a video stream with multiple frames, the principal point can be recovered once and kept fixed for the remaining frames. This is true as long as the internal camera settings are not changed.

Once the parameters that define the projection have been recovered, we can reproject the circle of radius r_2 onto an ellipse in the image. By minimizing the error in the reprojection, a good approximation to the principal point can be computed. We have found that the error of reprojection can be defined as the distance between the center of the ellipse used for the calibration and the center of the reprojected ellipse. Alternatively, we can define the error in terms of the angle between the principal axes of those two ellipses. The algorithm would be:

1. Start with an initial guess of the principal point (i.e., the center of the image).
2. Define the ellipses and the X axis marker of the image with respect to that principal point.
3. Calibrate the camera.
4. Reproject the original circle (of radius r_2) using the parameters obtained in the previous step.
5. Compute the reprojection error and update the working principal point accordingly.

Optimization methods like Levenberg-Marquardt [11] (implemented in MINPACK) can efficiently find the 2D position of the principal point that minimizes the error of reprojection.

4 Validating Our Method

We have validated our method using both synthetic and real data. We use synthetic data to determine how robust is our method in the presence of noise.

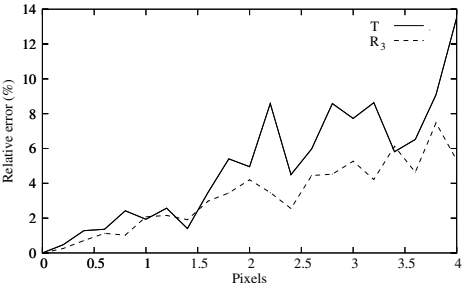


Fig. 3. Relative errors in the estimations of $\boldsymbol{T}$ and $\boldsymbol{R_3}$ ($\boldsymbol{Z}_{\mathrm{W}}$)

4.1 Robustness

To check the robustness of the algorithm, we project two concentric circles using a known synthetic camera configuration.Then, we perturb the points of the projected circles by adding random noise to their coordinates. We fit an ellipse to each set of perturbed points using Fitzgibbon's algorithm [5]. Finally, we compute the camera parameters using these two ellipses. Figure 3 shows the errors that the added noise produces in the recovered normal of the supporting plane ($\boldsymbol{R_3}$) and the translation vector ($\boldsymbol{T}$). Note that the error incurred is relatively small.

We have found that the system is very robust in the presence of systematic errors, i.e., when both ellipses are affected by the same error (for instance, with a non-centered optical axis). On the other hand, if the parameters of the ellipses are perturbed beyond a certain limit, the accuracy of the results decreases dramatically.

4.2 Experimental Results

In order to validate the computed calibration with real images, we have applied our algorithm to several images taken with a camera. Figure 4 shows an example of the process. First, the ellipses were recovered from the image and the camera parameters were computed. By using those parameters, we can draw the WCS axes on the image. Furthermore, the marker has been reprojected using the same parameters. The marker seen in the image has the following properties $r_1 = 2.6$ cm, $r_2 = 5$ cm and $r_3 = 6.5$ cm.

5 Conclusions and Future Work

In this paper we introduce a camera calibration technique that uses a very simple pattern made of two circles. The algorithm obtains accurate intrinsic and extrinsic camera parameters. We show that our method behaves in a robust manner in the presence of different types of input errors. We also show that the algorithm

Fig. 4. Reprojecting the marker and the coordinate system in the images

works well with real world images as long as good ellipse extraction and fitting algorithms are used.

Our work has a lot of applications, particularly in camera tracking and related fields. Our marker is easy to build and use. This makes it particularly well suited for augmented reality and entertainment applications. We are currently working on applications in these two areas. We are also trying to extend our camera model to take into account skew and lense distortion, in order to better approximate the behavior of a real camera. We are exploring the working limits of our algorithm and we are studying techniques to make the results more stable in the presence of noise.

References

1. Zhang, Z.: A Flexible New Technique for Camera Calibration. IEEE Trans. Patt. Anal. Machine Intell., vol. 22, no. 11, (2000) 1330–1334
2. Forsyth, D., Mundy, et al.: Invariant Descriptors for 3-D Object Recognition and Pose. IEEE Trans. Patt. Anal. Machine Intell., vol. 13, no. 10, (1991) 971–991
3. Kanatani, K., Liu, W.: 3D Interpretation of Conics and Orthogonality. *CVGIP: Image Undestanding*, Vol. 58, no. 3, (1993) 286–301
4. Rothwell, C.A., Zisserman, A., et al.: Relative Motion and Pose from Arbitrary Plane Curves. Image and Vision Computing, vol. 10, no. 4, May (1992) 250–262
5. Fitzgibbon, A.W., Pilu, M., Fisher, R.B.: Direct Least Squares Fitting of Ellipses. IEEE Trans. Patt. Anal. Machine Intell., vol. 21, no. 5, (1999) 476–480
6. Ahn, S.J., Rauh, W., Kim, S.I.: Circular Coded Target for Automation of Optical 3D-Measurement and Camera Calibration. Int. Jour. Patt. Recog. Artificial Intell., vol. 15, no. 6, (2001) 905-919
7. López de Ipiña, D., Mendonça, P.R.S., Hopper, A.: TRIP: a Low-Cost Vision-Based Location System for Ubiquitous Computing. Personal and Ubiquitous Computing Journal, Springer, Vol. 6, no. 3, (May 2002) 206–219
8. Kim, J.S., Kweon, I.S.: A New Camera Calibration Method for Robotic Applications. Int. Conf. Intelligent Robots and Systems, Hawaii, (Oct 2001) 778–783
9. Kim, J.S., Kim, H.W., Kweon, I.S.: A Camera Calibration Method using Concentric Circles for Vision Applications. 5^{th} Asian Conf. Computer Vision (2002)
10. Fremont, V., Chellali, R.: Direct Camera Calibration using Two Concentric Circles from a Single View. 12^{th} Int. Conf. Artificial Reality and Telexistence (2002)
11. More, J.J.: The Levenberg Marquardt algorithm: implementation and theory. Numerical Analysis, G.A. Watson ed., Springer-Verlag (1977)

Appendix

Given two concentric circles whose projections are known, we show how to project a third circle of known radius using the same projection. A circle C of radius r centered at the origin and located in the plane $Z = 0$ is defined by:

$$X^t C X = \begin{bmatrix} X\ Y\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -r^2 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \quad .$$

A projection matrix P projects a circle C onto an ellipse Q by $\lambda Q = P^{-t} C P^{-1}$. We compute the difference between the projection of a circle of radius $r + \alpha$ and the projection of a circle of radius r:

$$\lambda_1 Q_{r+\alpha} - \lambda_2 Q_r = P^{-t}(C_{r+\alpha} - C_r)P^{-1} = \alpha(\alpha + 2r)M \quad , \tag{11}$$

where $M = \boldsymbol{q}^t \boldsymbol{q}$ and $\boldsymbol{q}$ is the third row of matrix P^{-1}. Therefore, we can write:

$$\begin{cases} Q_3 = Q_1 - \alpha_1(\alpha_1 + 2r_1)M \\ Q_3 = Q_2 - \alpha_2(\alpha_2 + 2r_2)M \end{cases} ,$$

where $\alpha_1 = r_3 - r_1$ and $\alpha_2 = r_3 - r_2$. Using these two equations we can express Q_3 as:

$$Q_3 = kQ_2 - Q_1 \frac{r_3^2 - r_2^2}{r_3^2 - r_1^2} \quad , \tag{12}$$

where k is a scale correcting factor of Q_1 and Q_2. That factor can be computed applying the rank 1 condition to the ellipses (note that as $C_{r+\alpha} - C_r$ in equation (11) has rank 1, $\lambda_1 Q_{r+\alpha} - \lambda_2 Q_r$ should have rank 1, too) [9]. Solving for k in $Q_1 - kQ_2$ to have a rank 1 matrix leads to a scale correcting factor. Therefore, equation (12) allows us to project a circle of any radius given the projection of two circles, all of them concentric.

This process can be used to find the projected center of the concentric circles as well. As the projected center of a circle is always enclosed in its projected ellipse, if we project circles of smaller and smaller radii, we will be reducing the space where the projected center can be. In the limit, a circle of radius zero should project onto the projected center of the circles. Applying equation (12) to a circle of radius $r_3 = 0$ results in an ellipse (of radius zero) whose center is at the projected center of the concentric circles (x_c, y_c). The center of an ellipse in matrix form is given by [8]:

$$x_c = \frac{Q_{(2,2)}Q_{(1,3)} - Q_{(1,2)}Q_{(2,3)}}{\left(Q_{(1,2)}\right)^2 - Q_{(1,1)}Q_{(2,2)}} \quad \text{and} \quad y_c = \frac{Q_{(2,3)}Q_{(1,1)} - Q_{(1,2)}Q_{(1,3)}}{\left(Q_{(1,2)}\right)^2 - Q_{(1,1)}Q_{(2,2)}} \quad .$$

Three-Dimensional Object Recognition Using a Modified Exoskeleton and Extended Hausdorff Distance Matching Algorithm

Rajalida Lipikorn[1], Akinobu Shimizu[2], and Hidefumi Kobatake[2]

[1] Mathematics Department, Faculty of Science, Chulalongkorn University, Phatumwan, Bangkok 10330, Thailand
Rajalida.L@chula.ac.th
http://www.springer.de/comp/lncs/index.html
[2] Graduate School of BASE, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan
{simizu, kobatake}@cc.tuat.ac.jp

Abstract. This paper presents a simple but efficient matching methodology for three-dimensional object recognition using a modified exoskeleton (mES) and Hausdorff distance transform. The aim of this paper is to propose a method that can be used in computer vision system for robots or in digital image retrieval. The most significant part of this method is that it is less susceptible to distortion caused by noise and it is also rotation and translation invariant. The effectiveness of the proposed method is evaluated by implementing the method on three-dimensional object recognition where the mES of an observed object is compared against the mES of each training object in the database and the similarity between the two objects is measured using the extended Hausdorff distance matching algorithm. ...

1 Introduction

It is known that object recognition or object matching is one of the most important processes in computer vision, especially in the field of robotics, artificial intelligence, and bio-information [3],[6]. The tasks of visual object recognition and classification are essential for biological visual systems, but exceedingly difficult to replicate in computer vision systems. The process of biological visual systems is very complicate and requires the integration of several systems. As a scientist, simulation of such systems is a challenge and the question is, "Is it possible for us to develop a computer vision system for a robot to recognize objects regardless of their orientation, size, and color as our visual system does?" Beside computer vision systems, object recognition has been widely used in ditigal image retrieval system where searching of images from a large collection of image data or from a remote database is needed. In order to recognize an object, the object has to be represented by some certain features that are more compact yet retain meaningful information of the object. Shape is an essential feature used to describe the object and one of the most extensively used shape features is called

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 697–704, 2004.
© Springer-Verlag Berlin Heidelberg 2004

a skeleton. The disadvantage of a skeleton is that it is very sensitive to noise. This problem can be overcome by an alternative shape representation called a modified exoskeleton (mES) proposed in [4]. A modified exoskeleton is a special type of skelton that possesses measurable characteristics to be used in object recognition. A mES has been verified to be less susceptible to distortion caused by rotation and noise [5]. By using a mES, it is possible to find noise affected shapes or discard distorted components when needed. However, shape representations in two dimensions is not as efficient as those in three dimensions. This is because when a 3D object is projected onto a 2D image plane, one dimension of the object is lost. Thus, it is more desirable to develop a compute vision system for three dimensions. This motivates us to develop an effective recognition methodology for three-dimensional object recognition in order to improve the performance of computer vision systems in terms of accuracy and complexity.

The proposed matching methodology for three-dimensional objects uses a modified exoskeleton as an object representation and extended Hausdorff distance transform to measure the similarity between objects. The main processes of our proposed method consist of extracting a mES from each object and applying the extended Hausdorff distance matching algorithm to the mESs of each pair of objects.

The rest of this paper is organized as follows. Section 2 briefly reviews previous work. Section 3 provides a description of the proposed method. The experimental results are then presented in Section 4, and conclusions are given in Section 5.

2 Previous Work

Many shape representations and matching methods have been developed in the past. This section gives brief reviews on related techniques that are used in our proposed method.

2.1 3D Modified Exoskeleton

A modified exoskeleton mES is a skeleton of a sphere, with an original object embedded in the center, whose radius can be calculated from

$$r = \max_{(l,m,n)\in mES} \{d((l_c, m_c, n_c), (l, m, n))\} + \varepsilon, \tag{1}$$

where (l_c, m_c, n_c) is the centroid, (l, m, n) is any voxel on the boundary of the object, $d(.,.)$ is the Euclidean distance between two voxels, and ε is a constant value added to ensure that the sphere covers the original object thoroughly.

Definition Let $X = \{(x, y, z)\}$ be a 3D object in a binary image F, $\bar{X}$ be the complement of X, and $C = \{(x, y, z)\}$ be the intersection of $\bar{X}$ and a sphere that circumscribes the object X as shown in Fig. 1 where a cross shape represents X, the shaded area represents a sphere, and the shaded area outside the cross shape represents C. The mES is defined as a locus of voxels of C, or a skeleton of C,

whose distance to any voxel of $\bar{C}$ (the complement of C) is the shortest, whereas none of its neighbors has greater distance value. The mES and its exoskeleton function $mesf$ can be expressed as

$$\begin{aligned} mES(X) &= \{(x,y,z)|(x,y,z) \in C, \\ D(x,y,z) &= \min_{(l,m,n)\in\bar{C}}(d((x,y,z),(l,m,n))), \\ D(x,y,z) &\geq \max_{(p,q,r)\in N(x,y,z)}(D(p,q,r))\} \end{aligned}$$

$$mesf(x,y,z) = \{D(x,y,z)|(x,y,z) \in mES(X)\} \tag{2}$$

where $d((x,y,z),(l,m,n))$ represents the Euclidean distance between voxels (x,y,z) and (l,m,n), and $N(x,y,z)$ is a set of neighbors of (x,y,z).

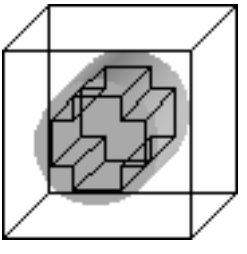


Fig. 1. Illustration of an object embedded in a spherical background (Cross section of a 3D image).

Since the mES is a skeleton of a sphere and the sphere is rotation and translation invariant, it has been verified that the mES inherits these properties from the sphere and thus it is also rotation and translation invariant.

2.2 Hausdorff Distance Matching Algorithm for 2D Objects

Hausdorff distance (HD) is a nonlinear operator, it computes distance values between two sets of pixels extracted from the objects. By given two finite point sets $A = \{a_1, a_2, ..., a_p\}$ and $B = \{b_1, b_2, ..., b_q\}$, the Hausdorff distance is defined as

$$H(A,B) = \max(h(A,B), h(B,A)) \tag{3}$$

where

$$h(A,B) = \max_{a\in A}\min_{b\in B}||a-b|| \tag{4}$$

and $||.||$ denotes some norm on the points of A and B. The functions $h(A,B)$ and $h(B,A)$ are called the directed Hausdorff distances from A to B and B to

A, respectively. Function $h(A, B)$ identifies the point $a \in A$ that is farthest from any point of B and measures the distance from a to its nearest neighbor in B (using the given norm $||.||$) [2]. The Hausdorff distance $H(A, B)$ is the maximum of $h(A, B)$ and $h(B, A)$.

2.3 Modified Hausdorff Distance Matching Algorithm for 2D Objects

The problem with conventional Hausdorff distance transform is that it is sensitive to distortion caused by noise. Various techniques have been proposed to solve this problem, among them the modified Hausdorff distance matching (MHD) for two-dimensional objects proposed by Dubission and Jain measures the similarity between two objects based on the average distance value [1].

The directed MHD, $h_{MHD}(A, B)$ is defined as

$$h_{MHD}(A, B) = \left(\frac{1}{N_A}\right) \sum_{a \in A} d_B(a) \tag{5}$$

where $d_B(a)$ denotes the smallest distance of point set B at point a and N_A is the number of points in the set A.

3 Extended Hausdorff Distance Matching Algorithm for 3D Objects

The problem with the MHD is that the similarity is measured from the average distance which can be affected by pixels that represent distortion caused by noise and it is not rotation and translation invariant. For two dimensions, we overcame these problems by developing a new matching algorithm for two-dimensional objects that computes the similarity between two objects from pixels on their mESs and their exoskeleton functions [5]. This exoskeleton function acts as a weighting factor to help determining the matching result because the exoskeleton functions of pixels that represent the same components tend to have similar values, whereas those pixels that represent different components tend to have different exoskeleton function values. Therefore, this matching methodology can provide more accurate results since two pixels on the mESs are considered to represent the same component if they locate at similar locations and their exoskeleton functions also contain similar values. We then extend this method to 3D object recognition. The extended Hausdorff distance matching algorithm is rotation and translation invariant because it first computes the closest distance $d_B(a)$ of voxel set on B at a voxel a. This distance is measured from the centroid of the mES to each voxel on the mES. It then computes the difference between the exoskeleton functions $E(a, b)$ of the corresponding voxels. The extended Hausdorff distance transform $h_{EHD}(A, B)$ is obtained by finding the average of the sum of $d_B(a)$ and $E(a, b)$. Finally, the degree of similarity between

two objects $H_{EHD}(A, B)$ is equal to the maximum of the extended Hausdorff distance transform.

In case of noisy object, voxels that represent noisy components typically have small exoskeleton function values, thus the exoskeleton function can also be used to eliminate distortion by removing any voxel whose exoskeleton function is less than the threshold value t. This threshold value can be set according to the requirement of each system. Moreover, we can also detect noisy components by scanning through voxels whose exoskeleton functions are less than the threshold value.

As a consequence, the similarity measure between an observed object A and each training object B in a database is equal to the extended Hausdorff distance transform with respect to either A or B, whichever has the maximum value. This extended Hausdorff distance matching algorithm is defined as follows:

$$H_{EHD}(A, B) = \max(h_{EHD}(A, B), h_{EHD}(B, A)) \tag{6}$$

where

$$h_{EHD}(A, B) = \left(\frac{1}{N_A}\right) \sum_{a_t \in A} (d_B(a_t) + E(a_t, b_t)), \tag{7}$$

$$E(a_t, b_t) = |mesf(a_t) - mesf(b_t)| \tag{8}$$

and a_t is any voxel on A, b_t is any voxel on B, $d_B(a_t)$ denotes the closest distance of voxel set on B at voxel a_t, and N_A is the number of voxels on A.

Further improvement can be performed by adjusting the size of the mES so that the matching result will not be affected if the sizes of the two objects are too much different. Before performing the matching, the mES and $mesf$ of a smaller object are enlarged according to the scaling factor which is calculated from the ratio of the radii of the two objects. Suppose that the mES A is larger than the mES B, then the scaling factor, ρ, can be calculated from

$$\rho = \frac{r_A}{r_B} \tag{9}$$

where r_A and r_B are the radii of objects A and B, respectively. This enlargement can reduce time complexity because we only need to enlarge the mES of a smaller object not the whole object itself.

4 Experimental Results

The experiments were conducted on a set of 3D geometric shape objects. The database of training objects consists of 20 individual objects in an image of size 80 x 80 x 60. Some examples of training objects are illustrated in Fig. 2. The experiments were divided into two cases. In the first case, each training object was transformed by rotating, translating, and scaling to produce 200 test objects.

We then used the proposed method to search for the most similar shape for each transformed object from the database. The matching process started by first extracting the mESs from the objects and then finding the scaling factor. Next the mES of the smaller object was enlarged and finally the extended Hausdorff distance matching algorithm was applied to the mESs. The training object with minimum similarity measure was returned as a result. In order to evaluate the performance of our proposed method, we also implemented the matching algorithm on skeletons of the same set of objects. The experimental results show that the matching algorithm can performed well on both representations and the recognition rates from both experiments are 100%.

In the second case, we used the proposed method to search for the most similar shape when an object was corrupted by noise. Each training object was distorted by noise to produce 100 deformed objects which were used as test objects. For each noisy object, uniform noise was added to a training object by randomly selecting voxels on the training object and converting them to background voxels. We then applied the same matching process on these deformed objects.

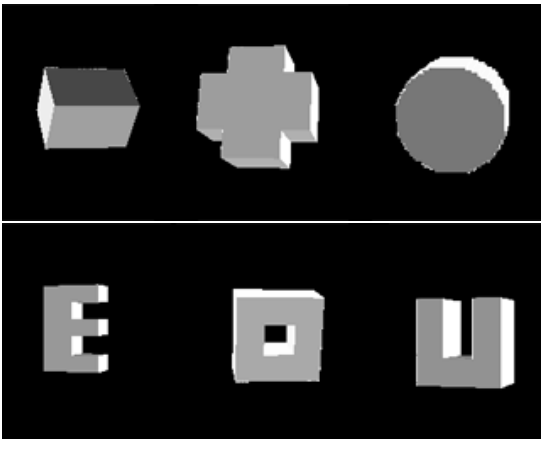
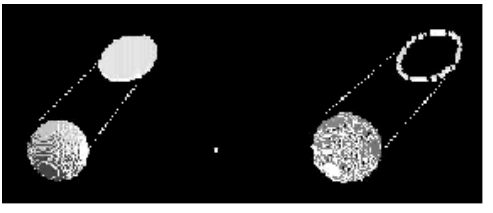


Fig. 2. Examples of training objects.

The effectiveness of our proposed method was evaluated against the matching on skeletons of the same objects. Example of an original sphere and its noisy version with their skeletons and their mESs are illustrated in Fig. 3. It can be seen that the skeleton is very sensitive to noise, only subtle changes can cause the structure of the skeleton to deform. The skeleton of the original sphere is merely a dot(voxel) (Fig. 3(a)) but after the sphere was corrupted by noise, the skeleton of the deformed sphere is as shown in Fig. 3(b). These two figures demonstrate the effect of noise on the structure of skeleton, whereas the mESs of the deformed sphere does not change significantly. The deformation of a skeleton

has a great impact on the matching result. The experimental results reveals that the matching on the mESs achieved a 96% recognition rate, whereas the matching on the skeletons achieved a 90% recognition rate.



(a)An original sphere(left), a skeleton(middle), a mES(right).

(b) A deformed sphere(left), a skeleton(middle), a mES(right).

Fig. 3. Central slides of an original sphere and its deformed version (segmented from the 3D objects).

5 Conclusions

This paper presents a matching methodology for three-dimensional object recognition that is robust to noise, rotation, and translation. It integrates the robustness of the mES with the robustness of the extended Hausdorff distance matching algorithm. The advantages of the proposed method are that the mES is less sensitive to noise and orientation and the extended Hausdorff distance matching algorithm is translation and rotation invariant. The reason that the extended Hausdorff distance matching algorithm is translation and rotation invariant is because it uses the relative distance with respect to the centroid of the mES to find the position of each voxel, thus it is not needed to superimpose one mES onto another mES when the matching is performed. These properties can significantly improve the performance of the proposed method. The experimental results reveal that the proposed method can improve the accuracy by 6%.

References

1. Dubisson, M.P., Jain, A.K.: A Modified Hausdorff Distance for Object Matching. *Proc. of the 12th International Conference on Pattern Recognition* (1994) 566–568

2. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing Images Using Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9 (1993) 850–863
3. Kaneko, T., Gu, L., Fujimoto, H.: Abdominal Organ Recognition using 3D Mathematical Morphology. *Proc. of ICPR*, Vol. 2 (2000) 263–266
4. Lipikorn, R., Shimizu, A., Hagihara, Y., Kobatake, H.: A Modified Exoskeleton and Its Applications to Object Representation and Recognition. *IEICE Trans. on INF. & SYST*, vol. E85-D, no. 5 (2002) 884–896
5. Lipikorn, R., Shimizu, A., Kobatake, H.: A Modified Exoskeleton and a Hausdorff Distance Matching Algorithm for Shape-Based Object Recognition. *Proc. of CISST*, Vol. 2 (2003) 507–511
6. Naf, M., Szekely, G., Kikinis, R., Shenton, M.E., Kubler, O.: 3D Voronoi Skeletons and Their Usage for the Characterization and Recognition of 3D Organ Shape. *Computer Vision and Image Understanding*, Vol. 66, No. 2 (1997) 147–161

Recognition of 3D Object from One Image Based on Projective and Permutative Invariants

J.M. González[1], J.M. Sebastián[1], D. García[1], F. Sánchez[2], and L. Angel[1]

[1]Departamento de Automática, Ingeniería Electrónica e Informática Industrial (DISAM)
Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid
C/ José Gutiérrez Abascal, 2, 28006 Madrid, Spain
{jgonzale, jsebas, dgarcia, langel}@etsii.upm.es
[2]Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Facultad de Informática, Universidad Politécnica de Madrid
28660 Boadilla del Monte, Madrid, Spain
fsanchez@fi.upm.es

Abstract. We present a new approach to recognize 3D objects in an image taking the available information in it. To reach this aim a method based on a two step algorithm will be developed. In the first step, each 3D model will be converted in a set of 2D images using an aspect graph algorithm in which the concept of visual event is defined as the moment when the projective invariants got from the model have changed. In the second step, we use these invariants to recognize the object situated in the scene. The main contribution to this work is the combination of two typical techniques used to object recognition like are aspect graph and projective invariants. We have obtained better results with this combinations than by using them separately.

1 Introduction

The object recognition has been studied by many authors but they have been few the works that deal with the recognition of any 3D object from one image. We define the recognition as the process followed by a system to identify and localize the objects observed in the scene. There are different methods to do it, but we can group them in two: based on graph and based on indexing tables. In the first group, the algorithm developed by Grimson [1] is well known. He designs a system in which each model is represented by a graph where the nodes are the features of the model and the arcs are the relations between them. In the on-line stage the algorithm builds a graph of the object present in the scene which is compared with the model graphs. The object is recognised when the two graphs match correctly. The most significant features are used to reduce the time required to cover all the search space, and with their properties restrictions to prune this space are built. The features utilized are 2D or 3D, and the properties are invariants in a Euclidean space. The drawback of this method is the high computational time required to build and to match the graph in the on-line process from the information available in the image of the scene. However, this has been often studied by other authors who propose an improvement of the process by means of measurements of significant and conditional probability [2],[3],[4].

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 705–712, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Another alternative to the system proposed by Grimson is presented by Pope and Lowe [5], who propose a method focussed in the camera instead of in the object. They present a different approach to the problem; instead of building a model based on the 3D features of the object, their system converts the problem into a 2D-2D recognition, which is possible by means of an aspect graph that represents the whole of the predicted appearances from the object. This method, however, has the same drawback because it is necessary to build and to match the graph of the object in the scene in the on-line stage.

The problem of time requirement in the on-line stage are resolved by the method based on indexing tables [6],[7]. In this one, the features of the models are utilized to build n-dimensional tables with the properties associated to these features. In the on line stage is only necessary to get the properties vector and to index in the tables to obtain the object present in the scene. The systems based on this method differ on the type of the chosen properties and their invariance for a defined transformation. One of most studied the system is the use of properties, which are invariant to projective transformation, because this is the group of transformations that better describes the projection of a 3D object in a 2D image. Nevertheless, this kind of invariants does not exist when the transformation takes place between spaces of different dimensions [8]. This is the reason why these algorithms have been restricted to planar object or 3D object with a specific configuration [9],[10],[11],[12],[13]. To give a solution to these limitations, other authors look for invariant functions, which relate the 3D features of the object with their representation in the 2D image plane [14],[15],[16],[17]. But these solutions increase the time required to recognise the objects.

What we propose is a new approach to the problem with two main aims: to keep the advantages found in the use of absolute projective invariants relating to the time required to recognise the object, and to not impose any restriction to the objects to be recognised. To get these objectives we develop a recognition system in two steps. In one hand, we process every object to obtain a set of 2D views, with a method similar to the aspect graph method**.** The main difference is that we define two views as different, when they generate different invariant vectors. For every model an invariant vector will be built and for each invariant we obtain a indexing table, as we will describe in next section. On the other hand, to recognise the object situated in the scene, we only need to get the projective invariants and look for those models which have their value inside the established threshold in the indexing tables.

Before the proposed algorithm will be detailed, we will review the concepts utilized in their implementation. Therefore in section 2, we describe how to get the aspect of a 3D object and how to measure their detectability. In section 3 the projective and permutative invariants used in the invariant vector and their uncertainty are studied. In section 4 we review the algorithm. Finally in section 5 we describe the experiments, analyse the results and outline the conclusions of this article.

2 Set of Views of an Object

The first stage in our system consists in transforming the 3D object in a set of 2D images, from which we extract the needed features to obtain the projective invariants. For its generation, we use a technique based in the aspect graph, which can be defined as the set of 2D views that can represent completely a 3D object. We define two

views as different, when the sets of invariants extracted from the significant features of the image have different values.

To get the set of aspects, which will describe an object it is necessary to study the viewpoint space, that it is approximated by a circumscribed sphere to the object. The infinite points of the space will be the infinite points of this sphere. The sphere is centred on the object to model and its radio has a value similar to the distance from the object to the point of view. The entire sphere instead of a half is considered because we allow the object to be supported by any face. In the practice a discrete sphere depending on the details of the object to model is used.

There are two algorithms to get the viewpoints: the uniform or object-independent algorithm and the object-dependent algorithm. In the first one the twenty points obtained from a regular icosahedron inscribed in the sphere are used. If more points were necessary, we will make a tessellation of every face obtaining four times the number of viewpoints in each iteration. If the object were simple to use a geometric figure with less faces as the hexahedron could be sufficient. With this method, an established number of appearances are got. It is chosen by its easy implementation, but it is necessary to reduce the redundant information by means of grouping of appearances when these have the same invariant vector.

The second algorithm used is the object-dependent algorithm, based on visual events, defined as the change of the appearance of an object. The objective of this implementation, to divide the sphere of viewpoints in regions without redundant information, is reached with the arcs that represent the visual event. These are obtained by studying the viewpoints in which occlusions of the features used to obtain the invariant vector are produced. Associated with each appearance, we obtain a measurement of its detectability according to the region of the viewpoint space from which this aspect is seen. This could have a value in the interval (0,1).

3 Invariants and Their Uncertainty

An important task of computer vision is to determine the geometric properties of objects from perspective images and to use these properties in recognition of 3D objects. The vision problem, so stated, immediately suggests that it is essential to derive properties of the object geometry which are invariant to perspective projection. An image invariant I(P) with respect to this group G acting on a geometric structure described by a parameter vector P, subject to a linear transformation T of the coordinates x = XT, is the expression that has its value following the next relation $I(p) = I(P)|T|^w$ where, if w = 0, the invariant is called absolute invariant. It is known that the number of invariants associated to a set of features are defined by the difference *m*-*n*, where *m* are the degrees of freedom of the features' set and *n* are the degrees of freedom of the group G, in our case, the plane projective group, what means *n*=*8*. In this way, we deduce that it is necessary a set of features with 9 dof to get an invariant. If we consider that a conic introduces 5 dof, a line 2 dof and a point also 2 dof, we need two conics, one conic and two points or lines and five points or lines to build some invariants.

In our system we use the projective invariants widely studied and for all of them the associated uncertainty was calculated by following first an analytic calculation, and second an empiric studied to verify the obtained results.

For a five lines set we have the invariants presented in (1), where the elements of the matrix are the coefficients of the line equation. A disadvantage of this invariant is that the value depends on the order we take the lines.

$$I_1 = \frac{\det(m_{431}) \cdot \det(m_{521})}{\det(m_{421}) \cdot \det(m_{531})}; \quad I_2 = \frac{\det(m_{421}) \cdot \det(m_{532})}{\det(m_{432}) \cdot \det(m_{521})} \qquad m_{ijk} = \begin{pmatrix} a_i & a_j & a_k \\ b_i & b_j & b_k \\ c_i & c_k & c_k \end{pmatrix} \tag{1}$$

Our aim is to get an invariant expression not only for the projective space but also the permutative. The p^2-invariants for 5 lines, is a five-dimensional J vector with two independent components (2), which elements are ordered.

$$J^{(1)} = I_1 \quad J^{(2)} = I_2 \quad J^{(3)} = \left(\frac{I_1}{I_2}\right) \quad J^{(4)} = \left(\frac{I_2 - 1}{I_1 - 1}\right) \quad J^{(5)} = \left(\frac{I_1(I_2 - 1)}{I_2(I_1 - 1)}\right) \tag{2}$$

The uncertainty of the five line invariants is proportional to their value. When we get two conics as features of our image, we obtain other two different projective invariants as we show in (3). We need to normalize the conics to (unit determinants).

$$I_1 = Trace\left[c_1^{-1} c_2\right] \quad I_2 = Trace\left[c_2^{-1} c_1\right] \tag{3}$$

We have two permutative invariants: a vector with the two projective invariants ordered in decrease order, or a symmetric function as the addition or the product of both invariants. In this case the uncertainty has quadratic dependence of the coefficients of the conics. To find the invariants of two points and one conic, we use the expression (4), which is related with the non-Euclidean distance.

$$I = \frac{\left(x_1^T c x_2\right)^2}{\left(x_1^T c x_1\right)\left(x_2^T c x_2\right)} = \begin{cases} \cosh^2(\delta(x_1, x_2)) \text{ if } |\mathrm{I}| > 1 \\ \cos^2(\delta(x_1, x_2)) \text{ if } |\mathrm{I}| < 1 \end{cases} \tag{4}$$

Finally we also calculate the invariants when we get two lines and two points from the image where the uncertainty is also proportional to their absolute value.

$$I = \frac{l_1 x_1}{l_2 x_1} \circ \frac{l_2 x_2}{l_1 x_2} \tag{5}$$

We calculate two measures: the conditional probability and the weight associated with each invariant value. The conditional probability is built inversely proportional to the number of object with the same invariant value. Its value is inside the interval (0,1). The weight is a combination of two factors; the first one, which value is between 0 and 1, is directly proportional to the ratio between the uncertainty of the invariant and the maximum uncertainty obtained for the whole of invariants; the other is a measurement of the uniqueness of the invariant based on the histogram of each invariant. We associate to the uniqueness of the invariant the values 1, 0.5, 0.2 and 0.1.

4 Recognition Algorithm

The proposed algorithm is based on a two stages process, as Fig. 1 shows. In the first stage (off-line), we obtain the tables and the invariant vectors from the views of the models and in the second stage, we recognise the object by indexing in this table.

In the first step of the off-line stage, we get the views of the models to be processed (Fig. 2). These are obtained depending on an existing CAD model of the object. If we have the CAD model we use the dependent-object algorithm described in section two. In other case we use a table with three degrees of freedom, which allows us to control the orientation of the object in relation to the camera and we implement the uniform algorithm also described in section two. In the next steps the edges of the image are got using the canny filter and the main features, lines, conics and intersection points, are extracted. Afterwards we calculate the values of the projective invariants with these features. Four different invariants are used: invariant of five lines, invariant of two conics, invariant of a conic and two lines (non-Euclidean distance) and the two lines and two points invariant. The disadvantage of the projective invariants is that the value differs depending on the arrangement of the features that we use to calculate them. In order to avoid this problem, we obtain the permutative invariants as we mentioned in section three. This operation is repeated so many times as different views of the object we have. All the process is repeated for every model of our data base.

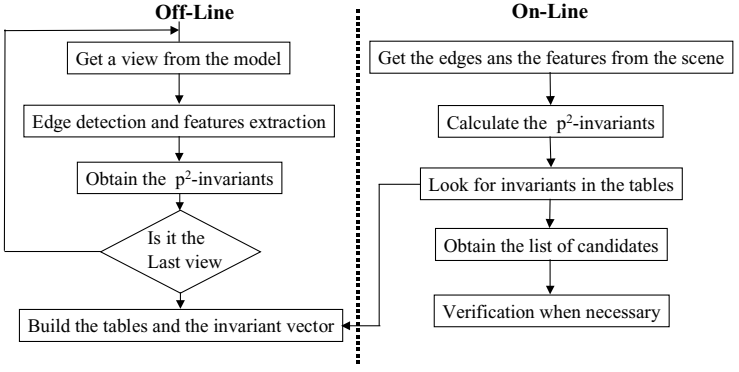


Fig. 1. Flow chart of the algorithm

We obtain the detectability of each invariants vector and we get the conditional probability of every component of the n-vector related to the number of models that have the same invariant value, considering the associated uncertainty. We also calculate the weight associated to the uncertainty of the invariant.

It is important to say that when we incorporate a new model in the data base, we need to recalculate the conditional probability of all the affected invariant values. This requirement has no effect on the efficiency of the system, because this task is processed off-line, and it does not increase the process time in the recognition step. Before the indexing tables are built, we get a measurement of the uniqueness of the invariant type, by means of an histogram of the conditional probability, in which we divide abscises axis in ten interval of likelihood. In this way, the invariant that gives more information of our set of objects is evaluated. This invariant will have more weight in

the recognition stage. Finally in the off-line stage, we build the indexing tables where we saved all this information.

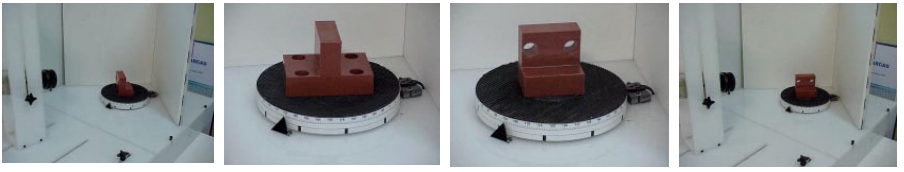


Fig. 2. Pictures of the table setup used to model the object

The second stage, which takes place on-line, is the recognition stage. Initially we have an image with an object. The first three steps are similar to the off-line stage: extract the edges of the object, find the features and calculate the p^2-invariants with these features. At this point, we have transformed the input image in an invariant vector. With the values of the invariant we look for the objects, which have the invariant value inside the interval established by the uncertainty in the indexing tables. Now we have a list of chosen objects with a priority. This priority is obtained from the weight of the invariant, the conditional probability and the detectability factor associated with each chosen object and saved in the indexing tables together with the invariant value. Its value will vary between 0 and 15. The relation, which we have adapted according to the results got from the experiments, gives more importance to the weight of the histogram and uncertainty, than to the conditional probability and to the detectability (6). The coefficients have also been empirically adjusted.

$$Priority = (5*weight+3*conditional_probability+2*detectability_factor) \quad (6)$$

Depending on the priority associated to the first and second objects in the list, we perform a verification step. If the values differ in more than five, the verification step is not necessary; we conclude that the object present in the scene is the first of the list. But when the values are similar and it is difficult to discriminate between them, we do a verification step, which confirm or not the established hypothesis.

5 Experiments and Results

The system was firstly probed with synthetic images got from a CAD program with added Gaussian noise, to observe the robustness of the used invariant. The results are shown in Table 1.

Table 1. Stability of the invariants obtained from synthetic images with Gaussian noise. The measure is the percentage of invariants whose values are inside of the margin established by the uncertainty

	Gaussian noise			
	$\sigma = 0.0$	$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.6$
Invariant of five lines	1.00	0,99	0,95	0,90
Invariant of two conics	0.98	0,93	0,89	0,83
Invariant of one conic and two lines	0.99	0,95	0,91	0,86
Invariant of two lines and two points	1.00	0,97	0,93	0,88

Next the system was probed with a twenty pieces set; a group of them are shown in Fig. 3. The trials were carried out with scenes, where only the object to be recognised was present. It could be thought, that a limitation exits if we compare this technique with other systems already mentioned in the article, which are capable to recognise an object in a scene where more items are present. The fact is that when this procedure is needed, we convert the unexpected object recognition algorithm into an expected object one by repeating our program so many times as the number of available models. The first part of the experiment, done with a handy grouping of primitives, obtained very good results, as it is showed in the second column of the Table 2. It is also presented the percentage of cases where it was necessary a verification stage, because the difference of priority between the first and second selected objects was less than five. Afterward we implement an algorithm to group the features automatically. The results are shown in the third and fourth columns of the Table 2.

Table 2. Results from the experiment with the object available in our data base. We show the percentage of successful recognition in two cases: grouping by hand and automatically. We also remark the percentage of cases in which a verification stage is necessary

	Percentage of successful recognition			
	Hand grouping		Automatic grouping	
	All cases	Cases with verification stage	All cases	Cases with verification stage
Piece 1	0,91	0,09	0,81	0,23
Piece 2	0,93	0,13	0,83	0,34
Piece 3	0,89	0,12	0,78	0,36
Piece 4	0,96	0,08	0,84	0,19

We observe that the grouping of features is the most critical step in the process. If we consider the situations when the system was not able to recognise properly the object, we conclude that in a 90% this was due to singular configurations of the features needed to get the invariants.

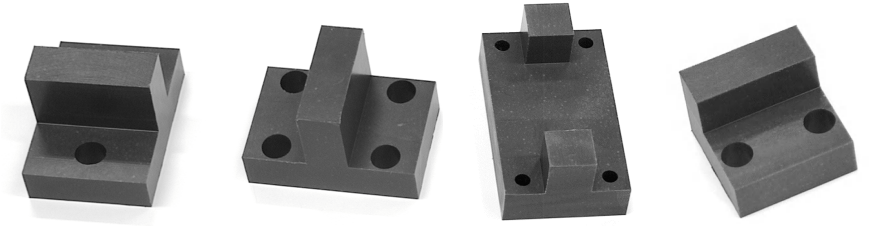


Fig. 3. Pieces from the model set.

6 Conclusions

The recognition system, which has been presented here, tries to take advantage of the projective invariant systems by doing extensive its use to all possible 3D objects. To reach this objective, we have developed an algorithm in two steps. The first step, previous to the recognition phase, converts the 3D object in a set of 2D views, using for it aspect graph techniques. In the second step, we only have to process an invariant recognition task. This system does not have the auto-occlusion problems referenced

by other authors, because they are studied and evaluated in the off-line phase. The method was developed to recognise only one object in a scene, following the unexpected object recognition theory, but it could be as well be used for those cases where different objects are present in the same scene. It is in these last situations, where the occlusion problems appear, but if there is an invariant, our system is not affected. Finally is important to remark, that the obtained results and the low computational costs make this technique available for real time industrial processes.

This work was supported by the Comisión Interministerial de Ciencia y Tecnología of the Spanish Government under the Project DPI2001-3827-C02-01

References

1. Grimson, W. Eric L.: Object Recognition by Computer. The Role of Geometric Constraints. The MIT Press, Cambridge (1990).
2. Mao, J., Flynn, P.J., Jain, A.K.: Integration of Multiple Feature Groups and Multiple Views into a 3D Object Recognition System. Computer Vision and Image Understanding, Vol. 62, No. 3 (1998) 309–325
3. Tan, T.N., Sullivan, G.D., Baker, K.D. : Model-Based Localisation and Recognition of Road Vehicles. International Journal of Computer Vision 27(1) (1998) 5-25
4. Häusler, G., Ritter, D.: Feature-Based Object Recognition and Localization in 3D-Space, Using a Single Video Image. Computer Vision and Image Understanding 73(1) (1999) 64-81
5. Pope, A.R., Lowe, D.G.: Probabilistic Models of Appearance for 3D Object Recognition. International Journal of Computer Vision 40(2) (2000) 149-167
6. Beis, J.S., Lowe, D.G.: Indexing Without Invariants in 3D Object Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(10) (1999) 1000-1015
7. Costa, M.S., Shapiro, L.G.: 3D Object Recognition and Pose with Relational Indexing. Computer Vision and Image Understanding 79 (2000) 364-407
8. Burns, J.B., Weiss, R.S., Riseman, E.M.: The Non-existence of General-Case View-Invariants in Geometric Invariance in Computer Vision. The MIT Press (1992)
9. Suk, T., Flusser, J.: Point-based projective invariants. Pattern Recognition 33 (2000) 251-261
10. Colios, C.I., Trahanias, P.E.: A framework for visual landmark identification based on projective and point-permutation invariant vectors. Robotics and Autonomous Systems 35 (2001) 37-51
11. Song, B.S., Yun, I.D., Lee, S.U.: A target recognition technique employing geometric invariants. Pattern Recognition 33 (2000) 413-425
12. Tien, S.C., Chia, T.L., Lu, Y.: Using cross-ratios to model curve data for aircraft recognition. Pattern Recognition Letters 24 (2003) 2047-2060
13. Lo, K.C., Kwok, S.K.W.: Recognition of 3D planar objects in canonical frames. Pattern Recognition Letters 22 (2001) 715-723
14. Weiss, I.: Model-Based Recognition of 3D Curves from One View. Journal of Mathematical Imaging and Vision 10 (1999) 175-184.
15. Roh, K.S., Kweon, I.S.: 3-D object recognition using a new invariant relationship by single-view. Patter Recognition 33 (2000) 741-754.
16. Song, B.S., Lee, K.M., Lee, S.U., Yun, I.D.: 3D target recognition based on projective invariant relationships. Journal of Visual Communication & Image Representation 14 (2003) 1-21
17. Song, B.S., Lee, K.M., Lee, S.U.: Model-Based Object Recognition Using Geometric Invariants of Points and Lines. Computer Vision and Image Understanding 84 (2001) 361-383

Wide Baseline Stereo Matching by Corner-Edge-Regions

Jun Xie and Hung Tat Tsui

Visual Signal Processing and Communications Lab
Electronic Engineering Department
The Chinese University of Hong Kong, Shatin, Hong Kong
{jxie, httsui}@ee.cuhk.edu.hk

Abstract. Most matching methods assume that all pixels of a correlation window have equal depths and find correspondences by comparing this local window of pixels. However, this assumption is not valid at depth discontinuities and wide baseline images. In this paper, a new method to construct the local planar patches called corner-edge-regions (CERs) is introduced. Based on this adaptive plane identification method, a novel similarity measure for estimating tentative correspondence is proposed. Using this measure makes the matching system robust to severe occlusions and shrinking. The proposed matching method is tested on several wide baseline image pairs. Experimental results show our method can find more correct correspondences than other correlation window based matching methods.

1 Introduction

Stereo vision systems determine depth from two or more images which are taken at the same time from different view-points. The most important and time consuming task for a stereo vision system is the identification of corresponding pixels. Depth estimation can be improved by increasing the baseline between images while wide baseline increases the likelihood of mismatching points. There is a trade-off between correctness of matching and precision of depth estimation.

Stereo matching for short baseline systems have been explored by many researchers (e.g. [1,2,3]). Numerical features have been proposed including Harris interest points [4], affine-invariant or scale-invariant measurement regions with Harris interest points, line segments and local homographies using parallelograms. Correlation is used over a significantly sized search window to match features in two images. A comparison of dense, two frames stereo matching methods can be found in [5].

The drawback of approaches for short baseline stereo is the camera configuration or the camera motion is heavily constrained. One would like to reconstruct 3D objects from a small number of still images, taken from very different viewpoints. The characters of wide baseline images include a large translation, scaling, or rotation between two frames with severe occlusions and shrinking.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 713–720, 2004.
© Springer-Verlag Berlin Heidelberg 2004

A wide-baseline system should be able to cope with large occlusions and depth discontinuities, as well as with intensity changes in the images.

Pritchett and Zisserman [6] proposed generating sets of local planar homographies to provide a better affinity measure between potentially matching features and to restrict the search for potential feature matches. Its main drawback is that the generation of possible homographies relies on suitable structures (parallelograms and large planar regions) being present in the scene. In [7], Tell and Carlsson applied feature vectors obtained by computing scale invariant Fourier coefficients for the intensity profile of the line between each pair of points. However, it is less local and does not produce local shape descriptions.

Tuytellars and Gool [8] constructed small local regions around corners using the extreme of a 2D affine invariant function and the texture covered by the regions. Because the local regions are parallelogram type regions, this method fails for wide baseline stereo where border errors and occlusions are significant. An improvement of the above method is proposed in [9] where corners were replaced by local intensity extreme as anchor points and parallelograms were replaced by intensity extreme along rays emanating from the anchor points. Unfortunately, this method is sensitive to illumination effects. The same weakness is suffered by the method in [10] where a similarity measure for establishing tentative correspondences is proposed based on the maximally stable extremal region (MSER).

Most recently, Xiao and Shah [11] proposed an affine invariant feature called edge-corner which can locate more precisely on the junction of edges than the Harris corner feature. Correspondences were found by comparing two $n \times n$ windows around each feature point in each image and a reasonable initial matching was applied before the affine invariant matching to improve the robust of the match. The authors demonstrated this method could find more correct inliers than other methods from wide baseline image pairs even without epipolar geometry refinement. A disadvantage of this method is it cannot cope with large depth variations because of the application of correlation windows. Moreover, the ignorance of occlusions leads this method not be able to find the feature correspondences on the boundaries of severely occluded or shrunk parts. There are also some researchers [12,13] using multiple baseline stereo (MBS) to make trade-off between correctness of matching and precision of depth estimates by simultaneously computing correspondences among pairs of images from multiple cameras with multiple baselines. However, in practice, only a few correspondences can be found over multiple wide baseline views.

In this paper, we focus on the matching problem between two gray level images with a wide baseline. To overcome the drawbacks of correlation window features, a novel matching feature is proposed in Section 2. A matching method based on the proposed feature is presented in Section 3. Several experimental results of our method on different pairs of images are shown and discussed in Section 4. Our conclusions are presented in Section 5.

2 Corner-Edge-Regions

One way to find correspondences between a pair of images is to compare a small window of pixels from one image to corresponding windows in another. However, using the correlation windows method, object borders are blurred and small details or objects are removed, depending on the size of the correlation window. Small correlation windows reduce the problem, but increase the influences of noise, which leads to a decrease of correct matches. The effect of noise is reduced by increasing the size of the correlation window. However, bigger correlation windows are more likely to cover areas where depth varies significantly. A change in depth results in a change of disparity so that only parts of the windows correspond to each other. This leads to errors at object boundaries. In [3], Hirschmuller proposed several improvements to tackle specific problems of correlation by the adaptive selection of the size of the correlation window. But adaptive windows do not always remove false matches especially for wide baseline stereo.

In our method, the corner detection procedure proposed in [11] is applied to get the edge-corners. Edge intersection points are employed to find edge junctions using local Hough transformation. Then the final feature points are identified via Harris corner operator.

After getting edge-corners on each image, we use the edges connecting each feature point to define the corner-edge-region feature as shown in Fig. 1. Let $C = (x, y)$ be a corner point. Set $E = \{e_1, e_2, \cdots, e_n\}$ contains all edges emanating from C and they are ordered anticlockwise. Let P_i be the second endpoint of edge e_i and θ_i the angle between two neighboring edges e_{i-1} and e_i. If $\theta_i < 180^o$, the two points $P_{i-1} = (x_{i-1}, y_{i-1})$ and $P_i = (x_i, y_i)$ together define a region A_{i-1} for corner C. If $\theta_i \geq 180^o$, we search for the nearest edge point $P_\theta = (x_\theta, y_\theta)$ along its angle bisector as

$$x_\theta = \frac{\sqrt{\lambda}\cos\frac{\theta}{2}[\sqrt{d_i}(y_{i-1} - y) - \sqrt{d_{i-1}}(y_i - y)]}{(y_{i-1} - y)(x_i - x) - (x_{i-1} - x)(y_i - y)} + x \tag{1}$$

$$y_\theta = \frac{\sqrt{\lambda}\cos\frac{\theta}{2}[\sqrt{d_i}(x_{i-1} - x) - \sqrt{d_{i-1}}(x_i - x)]}{(x_{i-1} - x)(y_i - y) - (x_i - x)(y_{i-1} - y)} + y \tag{2}$$

where d_i is the length of $\overrightarrow{CP_i}$ and λ is the length of $\overrightarrow{CP_\theta}$. Then points P_{i-1}, P_θ and P_i together define another region for corner C. Because each induced region has a uniform intensity, it is more likely to represent a planar patch. In this way, we are able to extract affine invariant corner-edge-regions around corner C such that each of those regions covers a physical part of a 3D surface, respectively. Then affine moment invariants can be computed on different subregions to match the two wide baseline images.

3 Wide Baseline Matching

Let $CER_1 = \{A_1, A_2, \cdots, A_i, \cdots, A_n\}$ and $CER_2 = \{B_1, B_2, \cdots, B_j, \cdots, B_m\}$ the corresponding set of corner-edge-regions of corner C_1 and C_2, respectively.

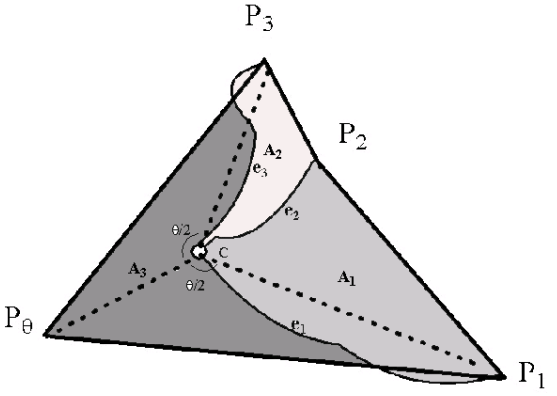


Fig. 1. An Illustration of Corner Edge Region. Corner C is the central feature point and regions with different intensities represent its neighboring corner-edge-regions.

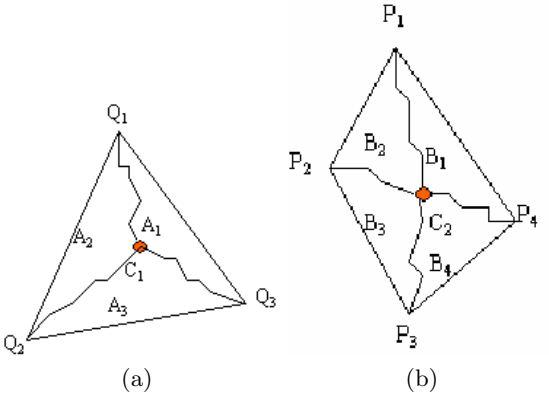


Fig. 2. Correspondence between two corners based on the corner-edge-region feature.

A_i is a corner-edge-region of C_1 and B_j is a corner-edge-region of C_2 as shown in Fig. 2. Let u_k denote the average intensity of corner-edge-region R_k defined as

$$u_k = \frac{\sum_{R_k} I(\mathbf{x})}{|R_k|} \tag{3}$$

Without loss of generality we assume $n \leq m$. Set P is then defined as

$$P = \{A_i | \ \exists B_j \in CER_2, |u_i - u_j| \leq \rho\} \tag{4}$$

where parameter ρ is a given maximum distance between two regions that represent a same plane in the real world. For each region A_i in set P, its corresponding region $f(A_i)$ in set CER_2 is defined as $f(A_i) = B_j$ where

$$|u_i - u_j| = \min_{B_k \in CER_2} |u_i - u_k| \tag{5}$$

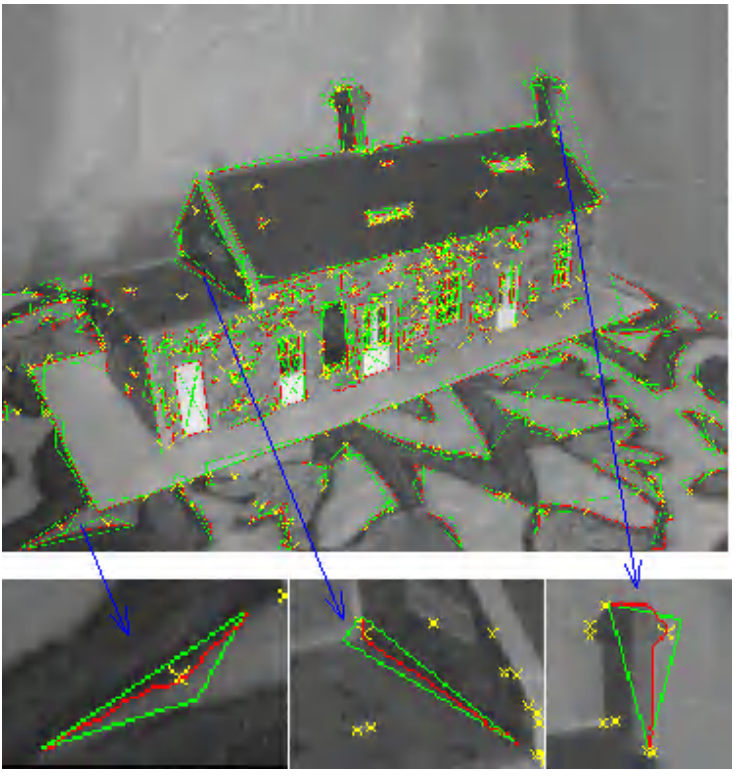


Fig. 3. Several samples of corner edge regions.

In [14], Shi and Tomasi demonstrated that using an affine model can effectively compensate the errors over video frames. Hence, we use an affine motion to compensate the geometric transformation between two corresponding corner-edge-regions. Following the method in [11], we use the light source reflection parameter μ and camera gain parameter δ to eliminate the illumination effect as follows

$$\mu_i H_i A_i + \delta_i = B_j \tag{6}$$

where H_i is an affine transformation matrix. Then the best match between corresponding regions A_i and B_j is obtained by minimizing the residual

$$\chi_{A_i} = \sum_{A_i} [\mu_i I_1(H_i \mathbf{x}) + \delta_i - I_2(\mathbf{x})]^2 \tag{7}$$

Finally, the similarity between corners C_1 and C_2 is measured as

$$\epsilon = \sum_{P} \frac{\chi_{A_i}}{|A_i|} \tag{8}$$

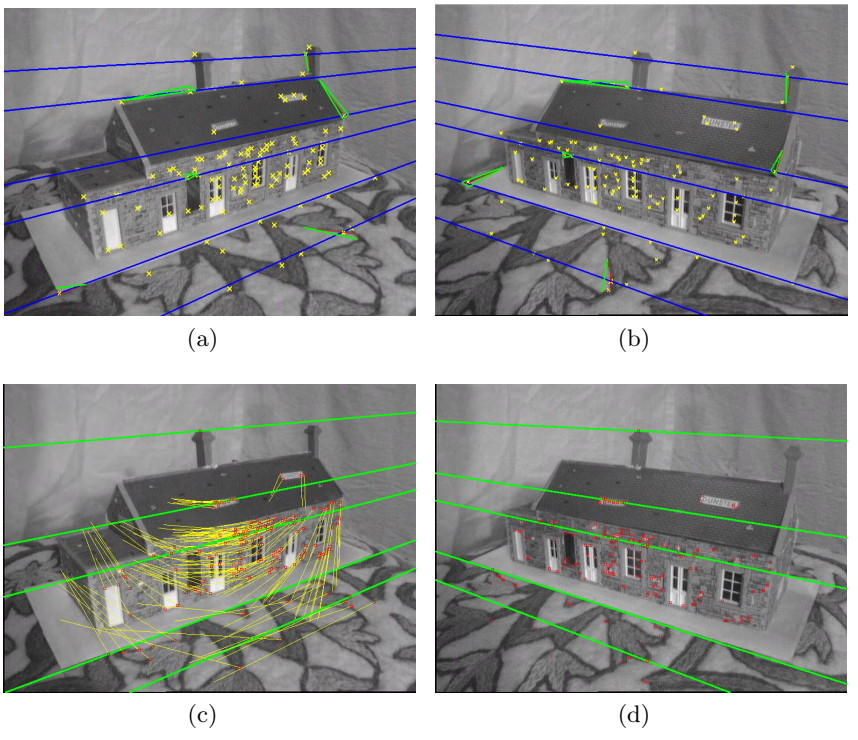


Fig. 4. Experimental results on Oxford house (frame0 and frame6). (a) and (b) are results of our method. (c) and (d) are adopted from Ref. [11].

4 Experiments and Discussions

Our method was implemented using C code. All of our experiments used grey level images and we set $\lambda \leq 10$, $\rho = 5$ and $\delta_i \leq 2$ in all those experiments. Fig. 3 shows several samples of corner-edge-regions on a real scene. The image at the top is frame0 of the Oxford house sequence with all obtained CER features. The yellow crosses present the corner points and red curves are the intensity edges detected using Canny edge detector. Each green line segment connects a pair of endpoints. We can observe that the surrounding region of each corner point was divided into several sub-regions based on the intensity similarity. This makes each subregion more likely to represent a patch of a 3D plane.

The matching results of the house frames are shown in Fig. 4. The results of our method are shown in Fig. 4(a) and Fig. 4(b). The test images are frame0 and frame6 of the Oxford house sequence, respectively. The included angle between those two images is 60^o. Without the refinement using epipolar geometry, there were totally 106 inliers found by using our CERs based matching method. Several corner edge regions and their epiloplar lines are shown on each image. Compared with the results of [11] in Fig. 4(c) and Fig. 4(d), where 83 correct

(a) (b)

Fig. 5. Experimental results our method on Oxford Univ. "Valbonne" (frame2 and frame9).

correspondences were found, the results of our method obtained more correct correspondences located on the edges of occluded planes and severely shrank parts. The Root Mean Square (RMS) error of our method in this experiment was 0.543 pixels.

Fig. 5 shows the experimental results of an outdoor scene. The pair of images are frame2 and frame9 of the "Valbonne" from Oxford University. There were 91 correct correspondences and the RMS error in experiment was 0.621 pixels, before using epipolar geometry refinement.

5 Conclusion

We have presented a novel method for wide baseline stereo matching. One major novelty of the paper is the proposed corner-edge-regions feature defined by using corner points and the neighboring edges. Because the intensities are alike within each CER, a sub-region is more possible to represent a patch of a same 3D plane. Thus, this feature is more capable to deal with occlusions and shrinking between wide baseline images. Based on this novel feature, we proposed a stereo matching method and applied it on several pair of wide baseline images of indoor and outdoor scenes with occlusions and illumination changes. The experimental results show our method can find more correspondences on the edges of

parts with severe occlusions and shrinking than other correlation window based methods (e.g. [11]). However, because the similarity between two corners is measured using each pair of sub-regions, a limitation of our current method is the mismatches introduced by the CERs feature. A possible approach to overcome this drawback is using topological constraints. This will be studied in our future work.

References

1. Deriche, R., Zhang, Z., Luong, Q., Faugeras, O.: Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In: Proc. European Conf. Computer Vision, Stockhom, Sweden (1994) 567–576
2. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521623049 (1998)
3. Hirschmuller, H.: Real-time correlation-based stereo vision with reduced border errors. Int. J. Computer Vision **47** (2002) 229–246
4. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of the 4th Alvey Vision Conf., Manchester,England (1988) 147–151
5. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Computer Vision **47** (2002) 7–42
6. Pritchett, P., Zisserman, A.: Wide baseline stereo matching. In: Proc. Int. Conf. Computer Vision, Bombay, India (1998) 754–760
7. Tell, D., Carlsson, S.: Wide baseline point matching using affine invariants computed from intensity profiles. In: Proc. European Conf. Computer Vision, Dublin,Ireland (2000) 814–828
8. Tuytelaars, T., Gool, L.: Content-based image retrieval based on local affinely invariant regions. In: Proc. Third Int'l Conf. on Visual Information Systems, Amsterdam, The Netherlands (1999) 493–500
9. Tuytelaars, T., Gool, L.: Wide baseline stereo based on local, affinely invariant regions. In: Proc. British Conf. Machine Vision, London, UK (2000) 412–425
10. Matas, J., Chum, O., Martin, U., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. British Conf. Machine Vision. Volume 1., London, England (2002) 384–393
11. Xiao, J., Shah, M.: Two-frame wide baseline matching. In: Proc. Int. Conf. Computer Vision. Volume 2., Nice, France (2003) 603–609
12. Okutomi, M., Kanade, T.: A multiple-baseline stereo. **15** (1993) 353–363
13. Ferrari, V., Tuytellars, T., Gool, L.V.: Wide-baseline multiple-view correspondences. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Madison, Wisconsin (2003) 718–725
14. Shi, J., Tomasi, C.: Good features to track. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Seattle, Washington (1994) 593–600

Gradient Based Dense Stereo Matching

Tomasz Twardowski, Bogusław Cyganek, and Jan Borgosz

AGH - University of Science and Technology
Department of Electronics
Al. Mickiewicza 30, 30-059 Kraków, Poland
cyganek@uci.agh.edu.pl

Abstract. This paper contains a study on image matching techniques based on image gradients. Gradients of the intensity signal play the very important role in the human visual system and also in computer vision. Since the onset of computer image processing, features, in a form of gradients, have been used for matching. However, the second part of a matching process is a choice of the matching criterion, called a comparison measure. The main contribution of this paper is a survey of the gradient comparison measures, ended with a proposition of the two new measures that we believe are more appropriate for gradient stereo matching. The paper presents results of many experiments performed to validate our methods.

1 Introduction

In this paper we present an overview of the comparison measures appropriate for matching of the gradient images. The work is an outcome of the study on stereo methods working in the domain of gradients formed from the intensity signals [1][3]. Results of this research can be also used for image retrieval in database systems or web technologies.

It is known that gradients fields play an important role in the human visual system [7]. It is also well known that image gradients form a very distinctive field of features if only there are any changes of the input signal and sufficiently above the noise level. Gradient based techniques paved the way for edge detection, motion analysis, and stereo correspondence. The latter application is just a subject of this report.

The starting point was the publication by Scharstein [9] devoted to the problem of view synthesis based on stereo. The two stereo methods were presented there that produce dense disparity maps; one of them computes disparity directly from gradient fields. Scharstein proposed a measure of evidence for or against a match at a certain location under a certain displacement that was based on gradient fields.

This paper starts with an overview of the gradient matching techniques. Based on this study the new comparison measure – competitive to the one proposed by Scharstein – is proposed and then compared with the other. The experimental results of this measure applied to the stereovision are also presented. They show great usefulness of the presented methods.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 721–728, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Overview of Matching Techniques Based on Image Gradients

Stereo matching by means of the gradients fields requires computation of gradient vectors from the initial intensity signals. Difference (or similarity) between those vectors is then accumulated. In result a measure of similarity among compared neighborhoods of gradients is obtained.

The principal problem in matching based on image gradients is a way the two vectors are compared. One of the most popular definitions of the cost function E is a sum of absolute differences (SAD) of coordinates of gradient vectors, as follows:

$$E = \left|w_{xl} - w_{xr}\right| + \left|w_{yl} - w_{yr}\right| \tag{1}$$

where w_{ab}, mean a coordinate of vector b.

Another commonly known measure is a sum of squared differences (SSD):

$$E = (w_{xl} - w_{xr})^2 + (w_{yl} - w_{yr})^2 \tag{2}$$

where symbols have the same meaning as in (1).

Scharstein [9] proposed an evidence measure which is a sum of the following two components:

$$m = (\left|w_l\right| + \left|w_r\right|) \tag{3}$$

$$d = \left|w_l - w_r\right| \tag{4}$$

where $\left|w_a\right|$ stands for a module of a vector a. Then Scharstein defined an evidence measure for a match as a weighted sum of these two terms:

$$e = \overline{m} - \alpha d = \frac{(\left|w_l\right| + \left|w_r\right|)}{2} - \alpha\left|w_l - w_r\right| \tag{5}$$

where α is the weight parameter. In this evidence measure, larger value of e exhibits "more similarity" between compared vectors.

3 Novel Matching Measure for Gradients

The new idea of a matching measure is to compare two coordinates (m,φ) of the gradient vectors, where m – is the module of a vector, φ – is a phase of the vector. Under these assumptions, the proposed vector matching measure can be expressed in the following form:

$$E = \left|m_r - \cos(\varphi_r - \varphi_l)m_l\right| + \alpha\left|\sin(\varphi_r - \varphi_l)m_l\right| \quad \wedge \quad m_r, m_l \neq 0 \tag{6}$$

where m_r, m_l are modules of right and left gradient vectors, φ_r, φ_l are angles of right and left gradient vectors, α is a weight parameter.

Modulating by $\cos(\Delta\varphi)$ in the first part of (6) of the left vector module and in the second part by $\sin(\Delta\varphi)$ splits the left vector into two parts, parallel and perpendicular to the right vector. This situation is depicted in Fig. 1(a). The parameter α constitutes

the weight of both deviations. By changing this parameter we can amplify difference between angles or modules as it is shown in Fig. 1(b).

The contour lines shown in Fig. 1 (c)-(d) are locations of end-points of vectors for which the measure (6) gives the same value of E. The weight parameter α was set during experimental to 0,2. For that value of the weight parameter for the exemplary vectors $(m,\varphi)_1 = (1,0^\circ)$ and $(m,\varphi)_2 = (1,45^\circ)$ their endpoint contours are presented in Fig. 1.

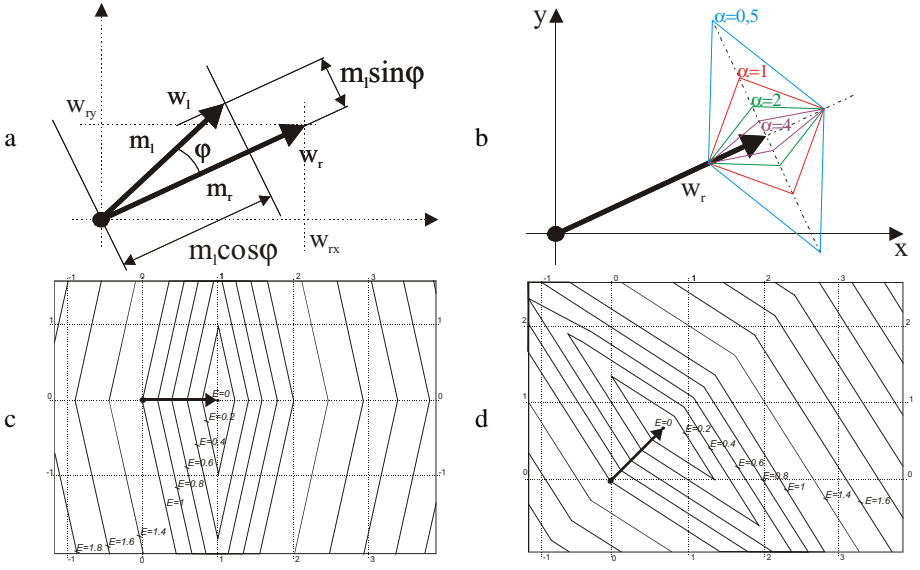


Fig. 1. (a) Geometrical interpretation of the measure (6), (b) An influence of the parameter α on shape of the contour lines. (c) Contour lines of E for a match with the vector (1,0°), $\alpha = 0,2$ (d) Contour lines measure E for a match with vector (1,45°), $\alpha = 0,2$

The advantage of proposed measure lies in separating parallel and perpendicular coordinates of compared vectors, what allows for an analysis of the two differences separately. Experiments show that better results of matching can be obtained if we put more attention to parallel element of compared gradient, and amplify that difference. Another desirable feature of the novel measure comes from using modules and angles of the compared vectors – i.e. the polar representation. The contour lines shown in Fig. 1b., 1c., and 1d are rotating in dependence on the vector phase. This property causes that the measure (6) estimates "similarity" of the compared vectors as follows: The same value is assigned if vectors have the same modules and angles between them. The measures (1) and (2) do not follow this scheme.

Table 1 contains values of different cost functions for hypothetical pairs of vectors. To interpret presented results it is necessary to remember that for measure (5) the larger value of the function e and c is assigned to the more similar vectors. In other cases the situation is just opposite, the cost function E equal to zero means that vectors are identical, larger values of this cost mean that vectors are more different.

The proposed gradient comparison measure was used and verified in the gradient based stereo algorithms. The algorithm for each pixel finds a value of a cost function

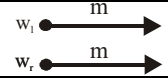
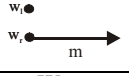
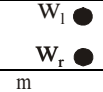
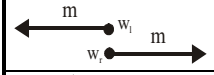
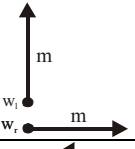
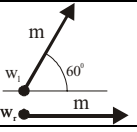
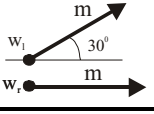
for potential pixels from the reference (right) image. At this stage gradient fields are compared in a $w_x x\ w_y$ window with the novel measure. Simultaneously the results are accumulated by low pass filter. In experiments size of the matching kernels was set in the range from *1x1* to *17x17* pixels. The final disparity value was asset in the "winner takes all" fashion. Then the disparity map was cross-checked to tackle occlusions. Also the sub-pixel correction was applied, according to the following formula:

$$correction = \frac{\cos t[disparity-1]-\cos t[disparity+1]}{\cos t[disparity-1]+\cos t[disparity+1]-2\cos t[disparity]} \tag{7}$$

where *cost[a]* is a value of the cost function for disparity *a*.

It is worth mentioning that this matching algorithm can be further improved by the fast block matching techniques, such as the winner-update or three-step-search [2].

Table 1. Different comparison measures used for gradient matching

Test vectors	Measure (1)	Measure (2)	Measure (5)	Measure (6)
	$E=0$	$E=0$	$e=+m$	$E=0$
	$E=m$	$E=m^2$	$e=-\frac{1}{2}m$	$E=m$
	$E=0$	$E=0$	$e=0$	$E=0$
	$E=2m$	$E=4m^2$	$e=-m$	$E=2m$
	$E=2m$	$E=2m^2$	$e=(1-\sqrt{2})m$	$E=m+\alpha m$ $for:\alpha=0,2$ $E=1,2m$
	$E=\frac{1+\sqrt{3}}{2}m$	$E=m^2$	$e=0$	$E=\frac{1}{2}m+\frac{\sqrt{3}}{2}\alpha m$ $for:\alpha=0,2$ $E\approx 0,7m$
	$E\approx 0,38m$	$E\approx 0,3m^2$	$e\approx 0,48m$	$E\approx 0,13m+\alpha m$ $for:\alpha=0,2$ $E\approx 0,33m$

The disadvantage of the measure (6) is a computation complexity, especially when compared with (5). The algorithm was implemented in C++, and was tested on IBM PC with 500 MHz clock, 128 Mb and Windows XP. Running times for measures (1), (2), (5) and (6) are presented in Table 2.

Table 2. A summary of running times for different stereopairs and kernels

Image size	Matching criterion	Execution time [s] for different sizes of kernels:			
		1x1	*3x3*	*5x5*	*7x7*
198x159	(1)	1	5	11	21
	(2)	1	5	11	19
	(5)	2	8	21	40
	(6)	2	10	27	52

4 Experimental Results

Fig. 2 depicts disparity maps for the *"Trees"* image pair, computed by the stereo algorithm using the proposed gradient comparison measure (6).

Sizes of gradient fields that are compared during matching have a direct influence on performance and quality of results. To small dimensions of kernels produce many false matching, what is caused by insufficient amount of information. At the other hand, increasing size of the matching windows to some extend improves quality of disparity maps but exceeding a certain size (e.g. for the pair *"Trees"* it was found to be *7x7 – 9x9*) there is an inevitable loss of details. For example in Fig. 2 (i) tress visible in the center of image start to join together. A similar phenomenon is noticed for the *"Pentagon"* pair in Fig. 3. Increasing size of matching windows decrease amount of false matching but above a characteristic size (*5x5 – 7x7*) the negative influence on performance of matching results is visible. On maps showed in Fig. 3 (h)-(i) corridors inside the Pentagon are not distinguishable as it was in Fig. 3 (d)-(e). The gradients in all experiments were computed with filters proposed by Simoncelli [8] that were also used by Scharstein [9]. These filters are strongly recommended since they exhibit superior spectral characteristics over a commonly used ones. As a result it is possible to obtain gradient signals that are rotation invariant.

Fig. 4 depicts two disparity maps of the artificial image from the University of Tsukuba. The first disparity map (Fig. 4c) is a result of matching with *3x3* windows and with the measure (6). The second map (Fig. 4d) was obtained with *7x7* masks.

Fig. 5 presents two disparity maps before and after left-right validation for elimination of false matches.

One way of comparing quality of the matching algorithm is a comparison of the output disparity maps with the ground truth maps. Unfortunately, such reference maps are rarely available [10]. Another possibility is to compare one of the input images with an image reconstructed from the disparity map. Such verification method was chosen for testing of the presented algorithm. The results of this verification presents Fig. 6 where the *PSNR* measure has been taken to express differences between images. An average difference between images in Fig. 3(a) and 6(c) is *PSNR = 37.5dB*; between Fig. 3(b) and 6(f) *PSNR = 34.9 dB*.

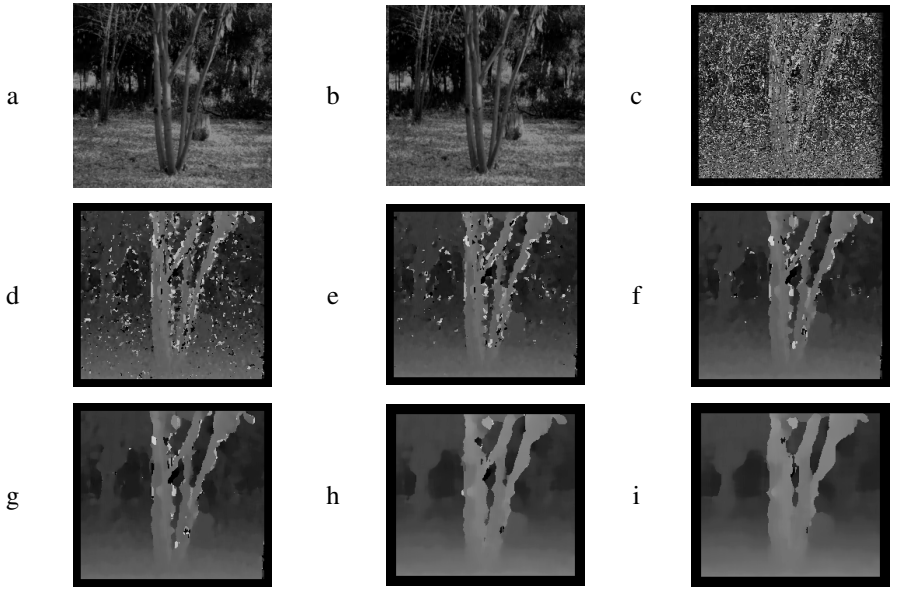


Fig. 2. Disparity maps for the *Trees*. (a) left image (b) right image (c) disparity map, kernel: *1x1* (d) disparity map, kernel: *3x3* (e) disparity map, kernel: *5x5* (f) disparity map, kernel: *7x7* (g) disparity map, kernel: *9x9* (h) disparity map, kernel: *13x13* (i) disparity map, kernel: *17x17*

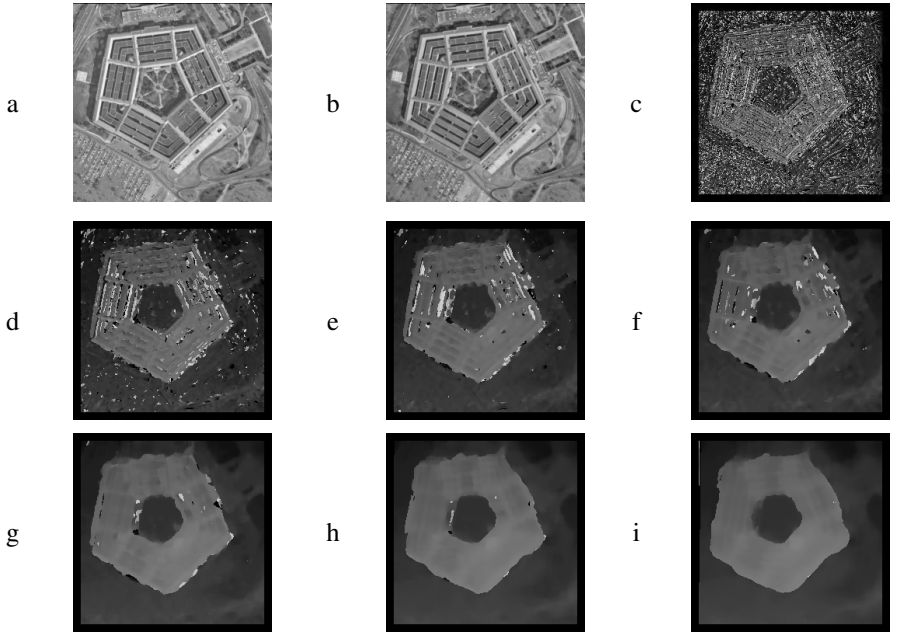


Fig. 3. Disparity for *Pentagon*. (a) left image, (b) right image, (c) disparity map, kernel: *1x1,* (d) disparity map, kernel: *3x3,* (e) disparity map, kernel: *5x5*, (f) disparity map, kernel: *7x7*, (g) disparity map, kernel: *9x9*, (h) disparity map, kernel: *13x13*, (i) disparity map, kernel: *17x17*

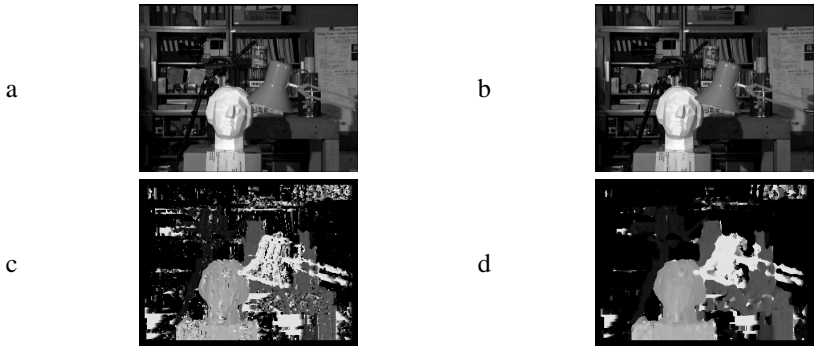


Fig. 4. Disparity maps for the *"Tsukuba"* stereo pair. (a) left image, (b) right image, (c) disparity map, kernel size: *3x3,* (d) disparity map, kernel size: *7x7*

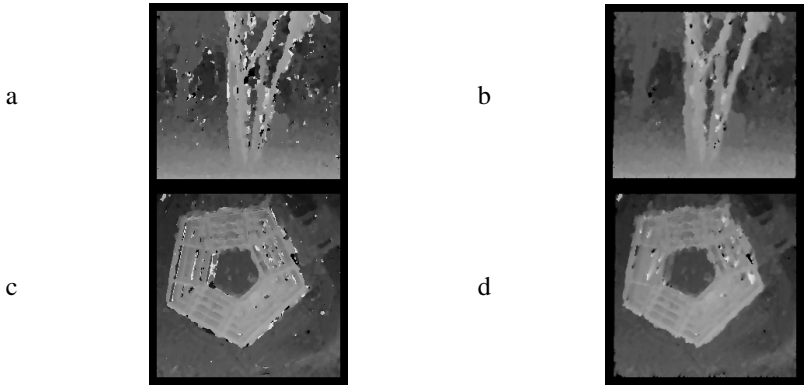


Fig. 5. Disparity maps before and after cross-checking. (a) the disparity for *"Trees"*, (b) *"Trees"* after cross-checking, (c) the disparity for *"Pentagon"*, (d) cross-checked

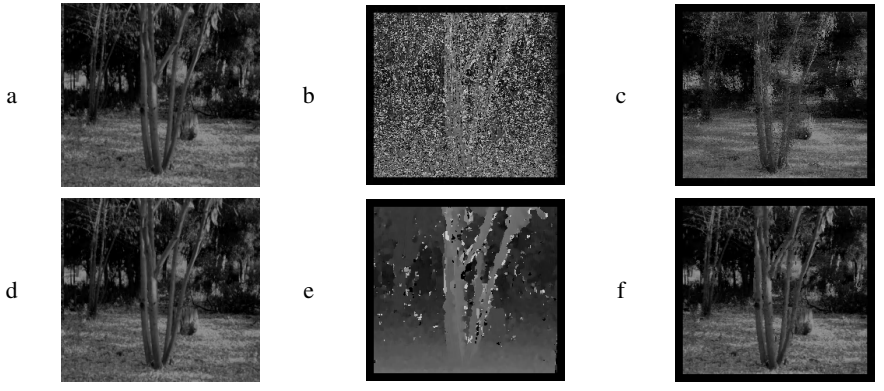


Fig. 6. (a) Original right image of the pair *"Trees",* (b) disparity map, kernel size *1x1*, (c) reconstruction of right *"Trees"* image on the basis of the disparity map. (d) The original right image, (e) the disparity map, kernel size *5x5*, (f) the reconstructed image

5 Conclusions

This paper presents an overview of different matching methods based on gradient fields of images. These techniques are tested in the stereo vision applications but can be also used e.g. for motion detection, optical flow, etc. Starting from the already known gradient comparison measures, such as the one proposed by Scharstein [9], the novel comparison measure for gradients (6) was proposed. It was shown that this measure allows for better representation of the mutual relations of the gradient vectors. Moreover, the proposed measure can be extended to the case of any vector field.

The proposed gradient comparison measure was tested in stereo matching that computes dense disparity maps from monochrome images. The influence of occlusions was alleviated by the cross-checking. An additional sub-pixel correction allowed for smoothing of the output disparity maps. The experimental results show very promising results for many types of input images. The main drawback of using gradients is in a case of images with constant intensity or images with increased level of noise. Performance of the method is greatly affected by a choice of a maximum expected disparity. However, this can be done to some extend automatically [4]. The presented matching technique can be considered in many other image processing areas such as motion analysis or segmentation. The latter application has been exploited in the road sign detection system for background segmentation.

Acknowledgement. This paper has been sponsored by the Polish Committee of Scientific Research (KBN) grant number: 3T11C 045 26.

References

1. Brown, M.Z., Burschka D., Hager G.D.: Advances in Computational Stereo. IEEE Trans. PAMI, Vol.25, No. 8, (2003) 993-1008
2. Chen, Y-S., Hung, Y-P., Fuh, C-S.: Fast Block Matching Algorithm Based on the Winner-Update Strategy. IEEE Trans. On Image Processing, Vol. 10, No. 8, (2001) 1212-1222
3. Cyganek, B.: Three Dimensional Image Processing, (in Polish) EXIT Warsaw (2002)
4. Cyganek, B., Borgosz, J.: An Improved Variogram Analysis of the Maximum Expected Disparity in Stereo Images, LNCS 2749, Springer-Verlag (2003) 297-303
5. Egnal, G., Wildes R.P.: Detecting Binocular Half-Occlusions: Empirical Comparisons of Five Approaches. IEEE PAMI, Vol. 24, No. 8, (2002) 1127-1132
6. Hartley, R.I., Zisserman A.: Multiple View Geometry in Computer Vision. CUP (2000)
7. Howard, I.P., Rogers, B.J.R.: Binocular Vision and Stereopsis. Oxford Univ. Press, 1995
8. Simoncelli, E.,P.: Design of Multi-Dimensional Derivative Filters. IEEE International Conference on Image Processing, (1994)
9. Scharstein, D.: View Synthesis Using Stereo Vision. LNCS 1582, Springer-Verlag (1999)
10. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV, Vol. 47,l No. 1-3 (2002) 7-42

Accelerating Multimedia Search by Visual Features

Grzegorz Galinski, Karol Wnukowicz, and Wladyslaw Skarbek

Warsaw University of Technology, Faculty of Electronics and Information Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland
{G.Galinski, K.Wnukowicz, W.Skarbek}@ire.pw.edu.pl

Abstract. Visual features used for search of visual material usually have computationally complex similarity functions. Therefore for large databases to get real time response for queries by examples is necessary avoiding their full search. In this paper we show efficiency of selected techniques for accelerating visual object retrieval. They belong to three independent groups: filtering, partial similarity computing, and tree based data structures. We show on description examples of motion trajectory, face recognition, and distributed color image temperature that different types of visual features require different accelerating techniques.

1 Introduction

Visual features are extracted to analyze visual content of multimedia material. The form of meta-data for images and video sequences depends on the goal the analysis is performed for. For instance in order to implement an efficient search in MPEG-7 standard [1] in its visual part there are defined dozen descriptions to be used for indexing of visual objects. Generally the MPEG-7 descriptions are based on color, texture, shape, and motion features.

The generic search scheme is based on exhaustive search in meta-data database and selection the most similar descriptions to the description of the query object. Each type of description has its own similarity function which is called at each new description accessed in the meta-data database. If the database is large and the similarity function is computationally complex then the performance of searching engine is far from real time. Therefore in such cases avoiding full search with full similarity computation is necessary condition to reduce the time of system response.

There are three generic techniques for accelerating the search in meta-data database: filtering, partial similarity computing, and tree searching.

Filtering is used when feature vectors are long and the technique defines a filtering attribute used for rejection of database objects before computing the similarity value. This filtering feature is precomputed and stored in the database along the regular feature vector. During the search the algorithm updates the interval of filtering features which correspond to K nearest, so far found objects. Descriptions with filtering attribute falling outside of the current filtering interval are not compared with the description of the query object.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 729–736, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Partial similarity computing is also used for long feature vectors, when the similarity function can be calculated in incremental way for consecutive components of the feature vector. It is checked, after each iteration, whether such actual partial similarity is above the current maximum distance to K nearest, so far found objects. If it is the case, the distance calculation is interrupted and the next description is accessed for the comparison.

The tree searching is based on appropriate tree-like index structure. The goal is to recursively subdivide the data space into cells and assign the feature vectors into appropriate cell. There are various commonly known and used tree structures [6], which differ in features and algorithms used for space clustering, as well as their internal structure. We consider here for comparison only two types: X-tree [3] and M-tree [5], [7].

In section 2 we show that in case of motion trajectory, using the simple concept of centroid and dispersion filtering features we can skip many unnecessary feature comparisons.

Section 3 demonstrates that in case of advanced face recognition descriptor defined in MPEG-7 while tree approach gives marginal acceleration (due to curse of multidimensionality) and filtering attempt fails (due to source gaussianity), the simple partial similarity leads to significant acceleration of the searching process.

Section 4 considers a new type descriptor which assigns to the image a set of color image temperatures corresponding to dominant colors. In this case the M-tree data structure is a good choice to improve the search efficiency.

2 Trajectory Search Acceleration

Motion trajectory descriptor represents motion of the single object occurring in the scene. It is realized by specifying coordinates for some representative point of the object (e.g. its centroid) in specified time points (called "key points"). Between key points object's coordinates are given by the interpolation function, which can be a polynomial of the first or the second order.

For long trajectories, the matching of them by using distance function:

$$\rho\left(Tr_1, Tr_2\right) = \sqrt{\frac{1}{t_e - t_s}\int_{t_s}^{t_e}\left(\left[f_{x1}(t) - f_{x2}(t)\right]^2 + \left[f_{y1}(t) - f_{y2}\right]^2\right)dt} \quad (1)$$

can be too slow for many applications (Tr denotes motion trajectory, $[t_s, t_e]$ is the range of time, and f_x and f_y are interpolation function for x and y coordinate respectively). Therefore we propose two acceleration techniques reducing significantly the number of trajectory comparisons: distance of trajectory centroids and distance of trajectory dispersion vectors.

Centroid of Motion Trajectory

Most of indexing structures require fixed size of feature vectors [2]-[6]. For motion trajectory such a feature is for instance the centroid, defined as follows:

$$\overline{Tr} = \begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix} = \left[\frac{1}{t_e - t_s} \int_{t_s}^{t_e} f_x(t)dt \qquad \frac{1}{t_e - t_s} \int_{t_s}^{t_e} f_y(t)dt \right]^T \tag{2}$$

For such feature vectors various indexing structures can be built, in particular X-tree [3]. In the next section we experimentally demonstrate, that X-tree built for centroids of motion trajectories is more efficient than the M-tree [5] (which does not require fixed size of feature vectors) built directly for motion trajectories.

Centroids of motion trajectories can be also used for approximation of distance between trajectories (or their descriptors). From Schwarz inequality:

$$\left(\frac{1}{t_e - t_s} \int_{t_s}^{t_e} f(t)dt \right)^2 \leq \frac{1}{t_e - t_s} \int_{t_s}^{t_e} f^2(t)dt \tag{3}$$

we have:

$$\left(\frac{1}{t_e - t_s} \int_{t_s}^{t_e} \left(f_{x1}(t) - f_{x2}(t) \right) dt \right)^2 + \left(\frac{1}{t_e - t_s} \int_{t_s}^{t_e} \left(f_{y1}(t) - f_{y2}(t) \right) dt \right)^2 \leq$$

$$\leq \frac{1}{t_e - t_s} \int_{t_s}^{t_e} \left(\left[f_{x1}(t) - f_{x2}(t) \right]^2 + \left[f_{y1}(t) - f_{y2}(t) \right]^2 \right) dt \tag{4}$$

Hence

$$\rho\left(\overline{Tr_1}, \overline{Tr_2}\right) = \|\overline{Tr_1} - \overline{Tr_2}\|_2 = \sqrt{\left(\overline{x}_1 - \overline{x}_2\right)^2 + \left(\overline{y}_1 - \overline{y}_2\right)^2} \leq \rho\left(Tr_1, Tr_2\right) \tag{5}$$

As a conclusion, distance between motion trajectories cannot be less than the distance between their centroids. If the distance between centroids of some trajectories is greater than the given threshold (e.g. distance between trajectory given as a query and its actual K-th nearest neighbor) the currently matching trajectory can be rejected without need to calculate the actual value of trajectory distance function. Experiments shows, that this method allows to reject above 80% of trajectories stored in database during the K-nearest neighbours (K-NN) query (for $K = 10$) for sequential scan of all stored data. The number of trajectory similarity function calculation can be significantly reduced for K-NN queries in case of using indexing structure (e.g. M-tree or X-tree), too.

Dispersion Method

It is also possible to use dispersion for an approximation of the distance between trajectories (e.g. for throwing out trajectories from nearest neighbors candidate set). Dispersion is given by the distance of the trajectory to its centroid:

$$V_{Tr} = \begin{bmatrix} V_x \\ V_y \end{bmatrix} = \left[\sqrt{\frac{1}{t_e - t_s} \int_{t_s}^{t_e} \left(f_x(t) - \overline{x} \right)^2 dt} \quad \sqrt{\frac{1}{t_e - t_s} \int_{t_s}^{t_e} \left(f_y(t) - \overline{y} \right)^2 dt} \right]^T \tag{6}$$

As $(\overline{x}, \overline{y})$ is the centroid of motion trajectory, then for any real value p we have:

$$\int_{t_s}^{t_e} \left(f_x(t) - \overline{x} \right)^2 dt \leq \int_{t_s}^{t_e} \left(f_x(t) - p \right)^2 dt \tag{7}$$

The same is true for y coordinate. Therefore from the triangle inequality it follows:

$$\begin{aligned}\rho\left(Tr_{x1},\overline{x_1}\right) &\leq \rho\left(Tr_{x_1},\overline{x_2}\right) \leq \rho\left(Tr_{x1},Tr_{x2}\right)+\rho\left(Tr_{x2},\overline{x_2}\right)\\ \rho\left(Tr_{x2},\overline{x_2}\right) &\leq \rho\left(Tr_{x_2},\overline{x_1}\right) \leq \rho\left(Tr_{x1},Tr_{x2}\right)+\rho\left(Tr_{x1},\overline{x_1}\right)\end{aligned} \tag{8}$$

Hence

$$\begin{aligned}-\rho\left(Tr_{x1},Tr_{x2}\right) \leq \rho\left(Tr_{x1},\overline{x_1}\right)-\rho\left(Tr_{x2},\overline{x_2}\right) &\leq \rho\left(Tr_{x1},Tr_{x2}\right)\\ \left|\rho\left(Tr_{x1},\overline{x_1}\right)-\rho\left(Tr_{x2},\overline{x_2}\right)\right| = \left|V_{x1}-V_{x2}\right| &\leq \rho\left(Tr_{x1},Tr_{x2}\right)\end{aligned} \tag{9}$$

Therefore, the distance between motion trajectories can be approximated by the distance between their dispersion vectors:

$$\begin{aligned}\rho\left(Tr_1,Tr_2\right) &= \sqrt{\rho^2\left(Tr_{x1},Tr_{x2}\right)+\rho^2\left(Tr_{y1},Tr_{y2}\right)} \geq\\ &\geq \sqrt{\left(V_{x1}-V_{x2}\right)^2+\left(V_{y1}-V_{y2}\right)^2} = \left\|V_{Tr1}-V_{Tr2}\right\|\end{aligned} \tag{10}$$

Formula (10) shows, that distance between trajectories cannot be less that the distance between their dispersion vectors. Hence, the distance between dispersion vectors can be used − as in centroids case − for throwing out some data from K nearest neighbors candidate set without calculating trajectory similarity function (which is much more complex). Experiments shows, that using this method in addition to centroid one, allows to prune above 90% of trajectories stored in database when sequential scan of all stored data was performed (during K-NN search, for $K = 10$). When using indexing structure, it allows for further significant reduction of motion trajectory similarity function calculation.

Experimental Results

In experiments the efficiency of search using full search approach (FS), M-tree (MT) and X-tree (XT) built for trajectory centroids was compared. The number of distance function calculations was also observed while using centroids and dispersion rules for approximating the distance between trajectories.

The test set consist of real trajectories (from MPEG-7 test set) as well as computer generated ones. The latter were obtained from real data by affine transformations and by random polynomial generation.

In tests K-NN (K Nearest Neighbors) queries were performed for $K = 10$.

The number of distance function calculations for various database size are presented on Fig. 1. In case of not using any approximation method, the X-tree (XT) is much more efficient than M-tree (MT).

Using centroids for distance approximation significantly cuts the number of distance function calculation. In full search (FS C) more than 80% of data is thrown out without distance calculation. In the case of M-tree (MT C), the number of distance calculations is reduced about four times. For X-tree (XT C) the results are not so spectacular, but the number of trajectory comparison is also reduced significantly.

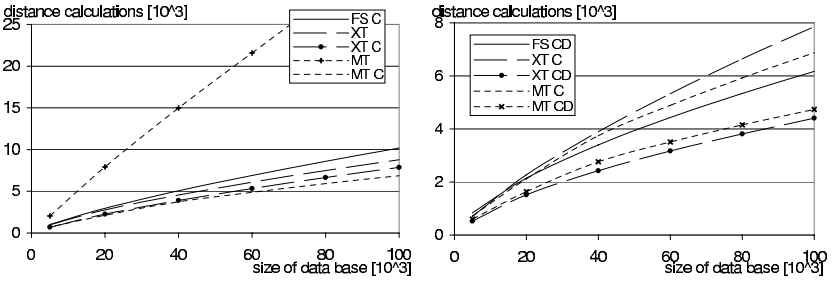


Fig. 1. The number of distance function calculation (in thousands) for various database size. FS − full search, XT − X-tree, MT − M-tree, C − centroid method applied, CD − centroid and dispersion methods applied

The second acceleration approach, the dispersion method, gives further reduction of trajectory distance calculations. In the full search (FS CD) nearly 95% of data is thrown out without distance function calculation. In case of M-tree (MT CD) about 30% of data, that remain after applying centroid method, are pruned after applying dispersion rule. In case of X-tree (XT CD) almost half of remaining data is rejected without distance computation, and in that case the X-tree becomes the most efficient indexing structure for motion trajectories databases.

3 Face Image Search Acceleration

The acceleration of distance function calculation is possible also for face recognition descriptor, which is a vector of dimension 48, and distance between two descriptors is given by l_2 norm.

In these case centroid lower bounding is given by a simple absolute difference between mean values of the descriptors, scaled by the square root of data dimension:

$$\rho\left(\mathbf{d}_1, \mathbf{d}_2\right) = \|\mathbf{d}_1 - \mathbf{d}_2\|_2 \geq \sqrt{N}|m_1 - m_2| \tag{11}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ represents two face recognition descriptors, N specifies their dimension, and m_1 and m_2 represents mean values of coefficients.

The dispersion bound is given by:

$$\rho\left(\mathbf{d}_1, \mathbf{d}_2\right) = \|\mathbf{d}_1 - \mathbf{d}_2\|_2 \geq |v_1 - v_2| \tag{12}$$

where v_i is the dispersion of vector $\mathbf{d}_i$, i.e. distance between vector and mean value of their coefficients:

$$v_i = \|\mathbf{d}_i - m_i\|_2 \tag{13}$$

It turns out, that for vectors of such a large dimension the efficiency of the above bound is very low. Namely, the centroid (mean) bound allows for discarding less

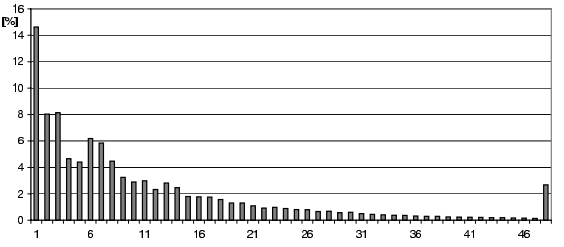


Fig. 2. Number of i-th partial distance calculation, relative to the number of descriptors (database size)

than 2% of descriptors without distance calculation during K-NN search. The dispersion bound discards as less as 0.1% of remaining data. The experiments was performed for face recognition descriptors, calculated for 8190 face images, taken from 1937 people (MPEG-7 test set) and for $K = 20$.

Some further acceleration can be achieved by reducing the computational cost of distance function calculation using partial distance method, where the i-th partial distance is calculated using only first i coefficient of vector $\mathbf{d}$:

$$\rho_i = \sum_{k=1}^{i} \left(\mathbf{d}_1^{(j)} - \mathbf{d}_2^{(j)} \right)^2 = \rho_{i-1} + \left(\mathbf{d}_1^{(i)} - \mathbf{d}_2^{(i)} \right)^2 \tag{14}$$

where $\mathbf{d}^{(i)}$ is the i-th coordinate of vector $\mathbf{d}$. After each dimension, the actual value of partial distance is compared with given threshold (maximum value of distance), and the calculation is stopped if it exceeds the threshold.

Figure 2 shows how many (on average) descriptors are rejected after i-th partial distance calculation. After first iteration nearly 15% of data are rejected from search. Then, the second and third iteration each discards 8% (relative to database size, not number of data remaining), and so on. Finally, after 48 iteration the exact value of distance between two vectors is calculated, which is the case for about 3% of descriptors stored in database.

Taking into consideration the average number of iteration needed for calculating the distance (or its bound sufficient for discarding the vector from search):

$$\overline{c} = \sum_{i=1}^{N} i \cdot p_i \tag{15}$$

where p_i is the probability that i-th partial distance is calculated, we can see, that on average less than 11 iterations are needed, compared to 48 iterations needed for exact distance calculation. As a result, the computational cost of calculation the distance between two descriptors can be reduced over 4 times, while executing the K-NN search.

Table 1. Results of the experiment for searching the M-tree dictionary

M-tree configuration	Distance calc.	Distance ratio
balanced, page: 4096, min. overlap	1107.07	16.32
balanced, page: 8192, min. overlap	1309.00	19.30
balanced, page: 4096, min. radius	1011.09	14,91
40% util., page: 4096, min. overlap	1061.56	15.65
20% util., page: 4096, min. radius	853.63	12.59

4 Dominant Color Temperature Distribution

The dominant color temperature distribution describes color temperature content of images. The descriptor contains up to 8 pairs: color temperature of dominant colors and its percentage contents in image. The descriptor is a discrete distribution of color temperature regarding their percentage values. We define the distance function as the integral of difference between the pdf distributions of two descriptors:

$$\rho\left(f_1, f_2\right) = \int_{T_{min}}^{T_{max}} |f_1(t) - f_2(t)|dt \tag{16}$$

where f denotes the color temperature distribution, and $[T_{min}; T_{max}]$ is the range of possible color temperatures.

Dominant Color Temperature Distribution is a complex feature, but it is not a vector, so the R-tree and its variants cannot be used here for construction of a search optimized data structure. The distance function (16) can be shown to be a metric, what means that the M-tree could be used in this case.

The experiments were carried out as follows. The descriptors for a set of 6782 images were extracted and added to a M-tree dictionary. Next, K-NN queries for $K = 10$ were performed for all of the 6782 images and the mean number of distance calculations and the mean number of read disk pages for all the queries was evaluated. The experiments were conducted for few configurations of M-tree such as node sizes and split method policies. The results are presented in table 1. The column "M-tree configuration" contains the configurations of the M-tree for which the test was conducted. There were three types of minimum node utilization after split: balanced (50%), 40% and 20%. There were two types of node split policy: minimum radius and minimum overlap between the nodes. There were two disk page sizes: 4096 and 8192 bytes. The column "Distance calculations" contains the number of average distance computations per one query, and in the column "Distance ratio" the computational gain of using the M-tree is presented in comparison to full sequential search as the ratio of the number of distance computations to the number of all features contained in the M-tree dictionary.

The results show that using the M-tree for storing the dominant color temperature distribution descriptors is beneficial for the search applications. The gain in the number of distance computations might be as many as 8 time less

than in the case of the sequential search. The best result was achieved for a tree with unbalanced nodes of 4096 bytes size, with low node utilizations, and minimum radius split policy.

5 Conclusions

Visual features used for search of visual material usually have computationally complex similarity functions. Therefore for large databases avoiding their full search is necessary to get real time response for queries by examples.

In this paper we showed efficiency of selected techniques for accelerating visual object retrieval. They belong to three independent groups: filtering, partial similarity computing, and tree based data structures. The experiments shows, that for various visual features these techniques allow to reduce computational cost of calculating the distance functions.

In case of the motion trajectory the combination of tree-like structure and filtering techniques allows to reject about 95% of data without calculating the distance function. In case of color temperature distribution the M-tree also allows to reduce number of distance calculation about 8 times.

The computational complexity of calculating distance function can be further reduced by calculating the partial similarity. For face recognition descriptor this technique allows to reduce the computational complexity of distance function over 4 times.

Acknowledgment. The work presented was developed within VISNET, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme

References

1. Information technology – Multimedia content description interface – Parts 1-8, ISO/IEC FDIS 15938-[1-8]:2002 (E) (2002)
2. Beckmann, N., Kriegel, H.-P., Schneider, R., Seeger, B.: The R*-tree: An efficient and robust access method for points and rectangles. Proc. ACM SIGMOD Int. Conf. Management of data (1990)
3. Berchtold, S., Keim, D. A., Kriegel, H.-P.: The X-tree: An Index Structure for High-Dimensional Data. Proceedings of the 22nd VLDB Conference, Bombay (1996)
4. Cha, G.-H., Zhu, X., Petkovic, D., Chung, C.-W.: An Efficient Indexing Method for Nearest Neighbour Searches in High-Dimensional Image Databases. IEEE Trans. on Multimedia, vol. **4**, no. **1** (2002)
5. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. Proceedings of the 23rd VLDB Conference, Athens (1997)
6. Gaede, V., Gunter, O.: Multidimensional Access Methods. ACM Computing Surveys, **30(2)**, (1998)
7. The M-tree Project: http://www-db.deis.unibo.it/Mtree/

Semantic Browsing and Retrieval in Image Libraries

Andrea Kutics[1] and Akihiko Nakagawa[2]

[1] Tokyo University of Technology, 1404 Katakura, Hachioji, Tokyo 192-0982, Japan
andi@media.teu.ac.jp

[2] Japan Systems Co. Ltd., 2-31-24, Ikejiri, Setagaya-ku, Tokyo 154-8541, Japan
ranaka@ppp.bekkoame.ne.jp

Abstract. In this paper, we address two main problems encountered in content-based image retrieval, namely the lack of image semantics that can be captured by extracting and indexing visual image features and the difficulty originating from the subjectivity and context dependency of user queries. This work proposes a new method for semantic browsing and retrieval of images by finding semantic coherence between words and image segments on three layers. The method is based on the matching of visual segment clusters with words on various levels of abstraction and is very promising for effective browsing and retrieval in large image databases. It supports various textual and/or visual query modes as well as both target- and category-type browsing and retrieval. Experiment conducted on a large set of natural images proved that step-by-step semantic inference on consecutive layers of image - word association helps to improve accuracy of retrieval and browsing.

1 Introduction

In recent years, limitations of content–based image retrieval methods based on extraction and indexing of visual features have become widely understood and the focus of research has shifted to bridging the semantic gap between low-level visual features and high-level semantic concepts. Several methods of combining textual and visual features have been proposed to overcome this problem. Relevance feedback-based algorithms [1] and various approaches drawn from the field of information retrieval, and document processing, such as methods using latent semantic indexing [2], statistical learning techniques based on mixture models [3], as well as HMM-based methods [4] and several neural network-based approaches [5] have been developed. Most of these models extract low-level visual features like color histograms, wavelet-based texture descriptions, etc. and compose feature vectors representing the whole image or evenly divided image blocks. These features are then directly mapped to words coming from various annotations and/or training data by using various learning approaches for estimating the joint distribution of the textual and visual features. One of the most promising approaches, described in [3], uses image segmentation and matches the detected objects with words by using a joint clustering approach on both feature spaces based on unsupervised learning using EM. Even though the reported results of these methods are very appealing, they have a number of limitations, namely, (1) they are very sensitive and dependent on training data which are sometimes difficult to provide, (2) they can learn only a limited number of semantic con-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 737–744, 2004.
© Springer-Verlag Berlin Heidelberg 2004

cepts, (3) they can not produce satisfactory results on an image domain different from the one they are originally trained on.

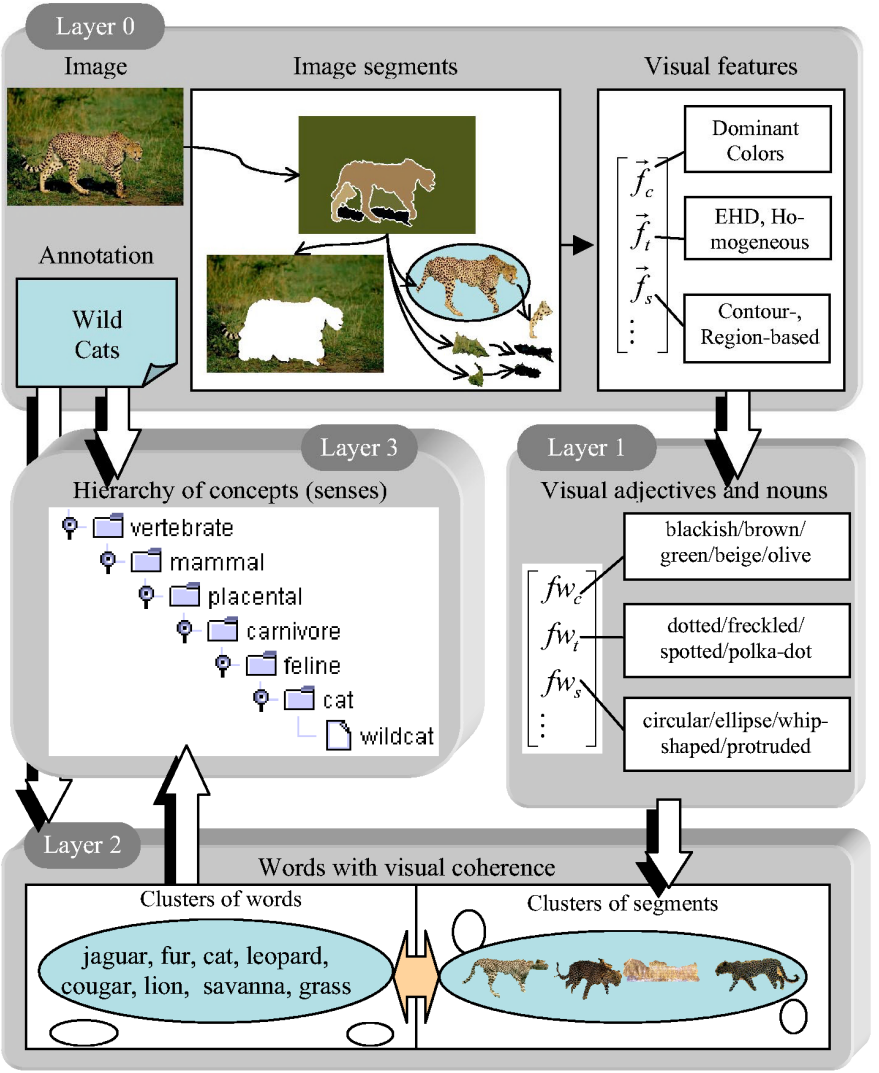


Fig. 1. Overview of the segment-word matching process

This paper presents a hierarchical method for associating words and image segments and thus capturing local and global image semantics by finding semantic coherence between these two cues on three different layers. The method is very promising for effective browsing and retrieval in large and unstructured image data sets. Our main purpose here is to ease two main problems encountered in content-based image retrieval, namely, lack of semantic information that is very hard to capture

using only visual feature-based indexing, and difficulty of handling the subjectivity and context dependency of user queries. We emphasize effective browsing and retrieval of images in large natural data sets without using any domain-specific information or annotations added for training purposes. A hierarchical model is very suitable for browsing, as shown in [3, 6], and effective browsing is an important function when searching in large image databases. Therefore, here we also apply a hierarchical approach, but we define it on a different basis. We associate image segments and words by matching clusters of visual segment features with textual ones in consecutive steps representing three semantic layers, namely, visual feature-related words, words with visual coherence and hierarchies of abstract concepts. As the spatial relationship among the segments is also determined, the joint segment–word hierarchies can be used for further high–level reasoning for more specific image categorization or topic determination.

2 Matching Image Segments with Words

2.1 Detecting Objects, Their Features, and Spatial Relations

As a first step, we carry out an image segmentation to detect prominent image objects by using a nonlinear inhomogeneous diffusion model based on both color and texture properties. This model is similar to the one we defined in one of our previous works [7], and it can be expressed as:

$$\frac{\partial}{\partial t} I(x,y,t) = div \cdot \left(c(x,y,t) grad\{I(x,y,t)\}\right), \; I(x,y,0) = I_0(x,y) \; . \tag{1}$$

where I indicates the feature vector, and function $c(x,y,t)$ describes the diffusivity, while div and $grad$ express the divergence and Gaussian-smoothed gradient ($DG_\sigma * I$), respectively. Diffusivity is defined as a function of the corresponding feature gradient in the combined, multi-valued texture and color diffusion processes in order to stop diffusion on region boundaries. It can be expressed as $c(x,y,t) = 1/\left\{1 + \left(\|grad\{I\}\|/K\right)^2\right\}$, where K indicates the conductance parameter. The reader is referred to [7, 8] for a detailed explanation of the above inhomogeneous diffusion model and that of the image segmentation process. In this paper, we use a simpler, new approach to integrate texture features (edge histograms and Gabor features) into the diffusion model and thus avoid over-segmentation. We apply a texture-dependent adaptive conductance parameter for this purpose, which is defined empirically as a decreasing function on the time axis. We estimate the texturedness by calculating the edge proportion in a given pixel neighborhood adjusted to the scale parameter (σ). On highly textured areas, we determine the conductance parameter on the basis of the texture gradient threshold, while on areas with low texturedness it is determined on the basis of the color gradient threshold. In this way, highly textured areas are handled as noisy regions and their texture is filtered out by evolving diffusion while preserving the texture boundaries. An online demonstration of the image segmentation can be found on the URL: http://www.rd-image.com/retrieval.

In further processing steps, we first select the most relevant image segments: mainly 4-6 segments per image. These are determined on the basis of their size, geometric properties, and layout features, or on the saliency of their features. Next, we create a visual description of each segment by calculating MPEG-7 compliant features for color and texture, such as scalable and dominant colors, edge histograms and homogeneous texture features for each segment. Moment invariants as well as contour-based and region-based descriptors are calculated to obtain a shape description. We also calculate layout properties such as segment areas, center coordinates, projections and length of the main axes. A hierarchical structure of the main segments considering the entire image as root is determined.

2.2 Determining Visual Feature-Related Words

On the first layer, we map the image segments to visual feature-related words, mainly adjectives and nouns, thus transforming their features to the textual domain. We accomplish this by applying clustering on each visual feature space such as color, texture, shape and layout extracted for the segments, and match the obtained clusters to psychophysically predetermined word clusters. In this process, we first determine dominant colors of the segments defined over a vector quantized HSV space and map them to color names by using a naming metric $D_{HSV}(s,i) = \left(\left(|V_s - V_i|\right)^2 + (S_s)^2 + (S_i)^2 - 2S_sS_i\cos\left(|H_s - H_i|\right)\right)^{1/2}$ to determine distances to prototypes of color name categories proposed in [9]. We also determine a hierarchy of texture feature clusters by using a self-organizing feature mapping (SOM), and vector quantizations on the texture features of the segments. Next, we determine texture-related adjectives by using the mapping of these texture clusters to texture-related keyword clusters (eleven main clusters) proposed as the texture lexicon in [10] and to several subclusters determined on the basis of psychophysical studies [11]. We also determine shape clusters by using agglomerative clustering on segment shape features (contour- and region-based) and map them to shape-related words determined heuristically. At the end of this process each segment is assigned by a couple of words: mainly adjectives and a few nouns. It has to be noted that this is not a one-to-one cluster matching of visual and textual clusters, except for texture features, as segments can possess multiple dominant colors especially when texture is present, and also a shape can be a member of multiple textual shape categories and vice versa. Examples of image segment–visual-related word mapping are presented in Fig. 1.

2.3 Matching Image Segments and Words with Visual Coherence

On the second layer, we determine joint clusters of segments mapped to visual adjectives and nouns, and words with visual coherence. It has already been noted by various researchers that some words are directly related or have stronger coherence with specific visual features like “rose, table, mountain, horse, etc.”, while others on a higher abstraction level like “religion, leisure, age, work, tradition etc.” have little or no visual coherence. In this process, we use only those words that were parsed from

naturally attached annotations and were not assigned for direct training purposes. Next, we try to select and eject the ones with more abstract meaning. This 'word selection' is simply done by using lexical definitions. In our case, we use the WordNet Lexical Database [12] and we drop words with a higher level of abstraction by determining their sense hierarchy (superordinate tree) and looking up visual or visual-related adjectives and nouns in their senses. As the WordNet provides only very short definitions, here we also utilized a lexical dictionary of pictures.

In order to obtain joint segment-word clusters, we apply a soft vector representation $(\vec{V}_{r_C}, \vec{V}_{r_T}, \vec{V}_{r_S})$ of the visual-related words where $v_{rj} \in [0,1]$ and define the probability of which word (j) expresses the image segment (r). These visual word probabilities can be defined by calculating the weighted normalized distance $(\alpha_C D_{j_C}, \beta_T D_{j_T}, \gamma_S D_{j_S})$ from the given visual cluster center. The weights $(\alpha_C, \beta_T, \gamma_S)$ express the relations between the corresponding visual and textual clusters. Next, we apply a vector quantization on these soft vectors to determine clusters of image segments that are similar in their visual properties. Then we estimate the conditional probability that a word with visual coherence ($item_i$) belongs to a given segment cluster ($cluster_j$). These probabilities are obtained by calculating the frequencies of these words or their direct superordinates, depending on the level of hierarchy, over the clusters and this can be expressed by the following equations:

$$P(item_i \mid cluster_j) = \frac{P(cluster_j \mid item_i) P(item_i)}{\sum_{l=1}^{I} P(cluster_j \mid item_l) P(item_i)} = \frac{\text{count}(item_i) \in cluster_j}{\sum_{l=1}^{I} (\text{count}(item_i) \in cluster_j)} \,. \tag{2}$$

Finally, we assign a maximum of eight words to segments determined on the basis of their probabilities ranked over the clusters.

2.4 Generating Concept Hierarchies

Finally, on the third layer we determine a hierarchy of abstract concepts. This can be simply accomplished by assigning global concept or superordinate hierarchies by using WordNet, and selecting a concept hierarchy on the basis of word frequency, co-occurrence and polysemy counts (word familiarity). Probabilities are also assigned to superordinate concepts on the basis of their frequencies and word familiarity. Another possible way to determine these concept hierarchies is by assigning the words with visual coherence together with their superordinate trees on the previous level. However, this would require a subtree matching or a very complicated training process. Concept hierarchies can also be determined by applying a hierarchical combination of asymmetric and symmetric clustering of words and segments on the second level as proposed in [3]. In this way, segments and words with higher probabilities are put on the higher level. However, here we argue that global semantic concept hierarchies are more suitable for browsing in general data sets, even if these are sometimes unbalanced. That is, there exist nodes with few or no membership. An overview of the segment-word matching process is shown in Fig. 1.

3 Browsing and Retrieval

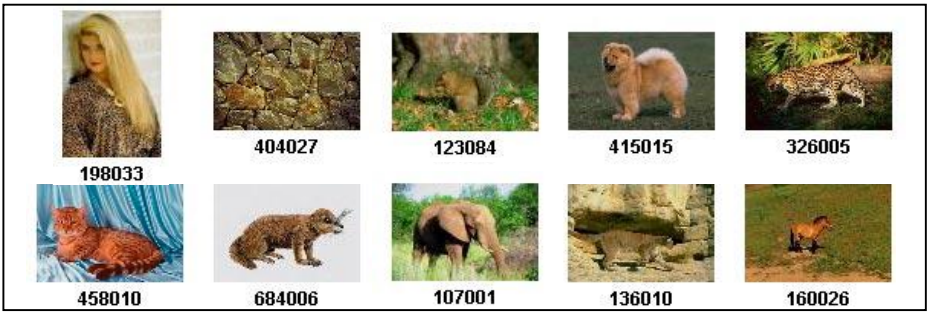
The method can easily handle both target- or category-type browsing and retrieval. The user can present a query by using keywords, phrases, example images or image segments by using the segmentation tool. Except for a very specific target-type search when the user has a given image or image object on his/her mind, users tend to present their queries for searching and also for browsing by specifying keywords as a starting step. In this case, the method retrieves a concept hierarchy and corresponding images, enabling the user to freely browse in the concept tree, find target images and execute a combined search. For queries ($qitem_{1...K}$) of both keyword(s) and/or image/segment example(s), segments/images ($seg_{1...I}$) are retrieved by calculating query-item probabilities over each cluster ($cluster_J$) weighted by the probability of the cluster with respect to the given segment(seg_I):

$$P(qitem_{1...K} \mid seg_I) = \sum_J \left(P(cluster_J \mid seg_I) * \prod_{k=1}^{K} P(qitem_k \mid cluster_J) \right) . \quad (3)$$

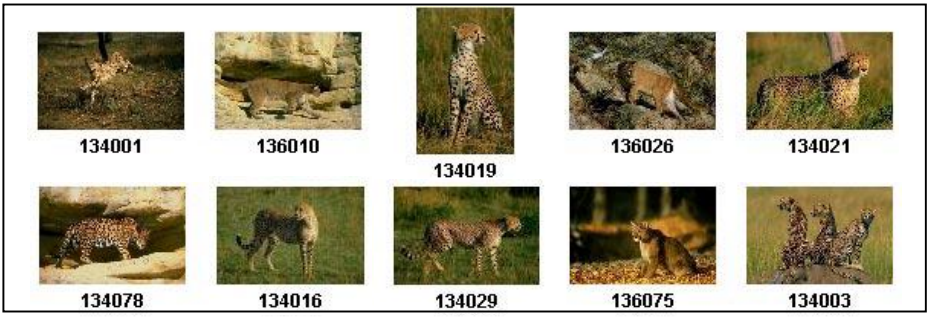
The user can also refine the retrieval results via relevance feedback, namely specifying relevant and/or irrelevant images among the retrieved images. In this process, not only the probabilities of both visual and conceptual words are updated, but user-specific query words are also saved in a "user dictionary" to enable user subjectivity to be handled more effectively. Retrieval result examples are illustrated in Fig. 2. The results shown in (a) were obtained via a traditional visual feature-based search by specifying a cheetah object query. Fig. 2(b) illustrates the results obtained by the proposed method by specifying the keyword "cheetah", and a "cheetah" segment (Fig. 1.) extracted from an example image selected via category-based browsing.

4 Experiments

To evaluate the method, we chose 12,500 natural images representing various categories (people, plants, animals, toys, vehicles, tools etc.) of the Corel Gallery collection. We did not provide any specific annotations for training purposes, but used only the group names (1 name/100 images) parsed from the image group titles that were assigned by Corel. These are very short, like "North American Deer", "Reflective Effects", etc., are determined very subjectively, and are redundant semantically. In order to produce a query image set, we randomly chose 100 test images from 10 different, additional category groups of Corel, regardless of the existence of well-defined relevant objects. We also used 6-12 test query words assigned to each test image by 5 persons representing different genders, generations (2) and cultural backgrounds (3 countries). We obtained the highest retrieval precision for combined queries (queries presented by both keyword(s) and image(s) or image segment(s)). About 70 percent of the most relevant objects or pictures were retrieved for each test person in the setof 24 result images in the first retrieval cycle and 95 percent retrieval precision was obtained after five relevance feedback cycles. We compared these results to those



(a) Visual feature-based search by the 'cheetah' segment query.



(b) Combined search by the proposed method using 'cheetah' segment and keyword queries

Fig. 2. Retrieval results

obtained by applying traditional visual feature-based searches carried out using the same test images or image segments, and the retrieval precision dropped to an aver age of 55 percent. Retrieval errors occur for two main reasons: (1) according to segmentation errors, which have the most severe effect on shape features, and mostly occur on images containing areas of inhomogeneous texture features and/or many small regions, (2) according to word matching errors on the second or third layers. In these latter cases, the method often fails to assign relevant words of visual coherence or fails to determine proper concept hierarchies. The number of these errors can be significantly decreased by applying more user interaction, i.e., by inventing more browsing steps into the retrieval or applying a more precise statistical model for concept matching, which is an ongoing research topic in our laboratory.

5 Conclusions

In this work, we presented a new hierarchical model for matching images and words by trying to establish semantic inference using visual and textual cues for effective image browsing and retrieval. The proposed method uses a novel non-linear segmentation algorithm to detect image objects in order to capture semantics on the visual layer. In this method, we match image segments with words on three layers, starting

by assigning visual adjectives and nouns, then creating joint clusters of segments and words with visual coherence, and finally determining a hierarchy of higher level concepts. The advantage of the method is that it supports both target- and category-type browsing and retrieval as well as various textual and/or visual query modes. By comparing the results of experiments conducted on a large natural image domain with reported results of traditional CBIR methods or other direct image-word matching approaches, it can be shown that higher retrieval precision and browsing effectiveness can be achieved by using step-by-step semantic inference on three consecutive layers of image word association. As a future work, it will be possible to further expand the method for high–level reasoning for more specific image categorization or topic determination by utilizing spatial relationships among segments.

Acknowledgement. The authors would like to thank Mr. Takaaki Matsumoto, Dr. Sakuichi Ohtsuka and the members of the Media Handling Technology Group of NTT Data Corporation for their valuable assistance in preparing this paper.

References

1. X. S. Zhou and T. S. Huang, Unifying Keywords and Visual Contents in Image Retrieval, IEEE Multimedia, Vol. 9, No. 2, (2002) , 23-33.
2. R. Zhao, et al., Negotiating the semantic gap: from feature maps to semantic landscapes, Pattern Recognition, Vol. 35, (2002), 593-600.
3. K. Barnard, et al., Matching Words and Pictures, Journal of Machine Learning Research, Vol. 3, (2003), 1107-1135.
4. J. Z. Wang and J. Li, 'Learning-based linguistic indexing of pictures with 2-D MHMMs,' Proc. ACM Multimedia, (2002), 436-445.
5. J-H. Lim, Q. Tian, P. Mulhem, Home Photo Content for Personalized Event-Based Retrieval, IEEE MultiMedia, Vol. 9, No. 2, (2003), 28-37.
6. T. Hofmann, Learning and Representing Topic. A Hierarchical Mixture Model for Word Occurrences in Document Databases, Proc. of the Conference for Automated Learning and Discovery (CONALD), Pittsburgh, (1998)
7. A. Kutics, et al., An object-based image retrieval system using an inhomogeneous diffusion model, Proc. of the ICIP'99, Vol. II, (1999), 590-594.
8. A. Kutics, et al., Image retrieval via the inhomogeneous diffusion of luminance and texture features, Journal of Electronic Imaging, Vol. 9(2), (2000), 159-169.
9. A. Mojsilovic, A method for color naming and description of color composition in images, Proc. of the ICIP2002, (2002).
10. N. Bhusnan, et al., The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images, Cognitive Science, Vol. 21(2), (1997), 219-246.
11. W. R. Hendee, P. N. T. Wells, The Perception of Visual Information, Springer, (1997).
12. C. Fellbaum, et al., WordNet: An Electronic Lexical Database, MIT Press, May 15, (1998).

Robust Shape Retrieval Using Maximum Likelihood Theory

Naif Alajlan[1], Paul Fieguth[2], and Mohamed Kamel[1]

[1] PAMI Lab, E & CE Dept., UW, Waterloo, ON, N2L 3G1, Canada.
{naif, mkamel}@pami.uwaterloo.ca
[2] System Design Dept., UW, Waterloo, ON, N2L 3G1, Canada.
pfieguth@uwaterloo.ca

Abstract. The most commonly used shape similarity metrics are the sum of squared differences (*SSD*) and the sum of absolute differences (*SAD*). However, Maximum Likelihood (*ML*) theory allows us to relate the noise (differences between feature vectors) distribution more generally to a metric. In this paper, a shape is partitioned into tokens based on its concave regions, invariant moments are computed for each token, and token similarity is measured by a metric. Finally, a non-metric measure that employs heuristics is used to measure the shape similarity. The desirable property of this scheme is to mimic the human perception of shapes. We show that the *ML* metric outperforms the *SSD* and *SAD* metrics for token matching. Instead of the *ML* metric based on histograms for PDF approximation, which suffer from being sensitive to choices of bin width, we propose a Parzen windows method that is continuous and more robust.

1 Introduction

In recent years, content-based image retrieval (*CBIR*) has become a major research area due to the increasing number of generated images every day [1]. *CBIR* uses generic image features such as color, texture, and shape to interpret the content of images. In this work, we are interested in using shape descriptors in *CBIR*. Given a query image, we would try to obtain a list of images from a database of shape images, which are most similar to the query image. This problem can be solved in two stages. Firstly, a feature vector represents the shape information of the image. Then, a similarity measure computes the similarity between corresponding features of two images.

A desirable property of a similarity measure is that it should mimic the human perception of shapes. In fact, it has been verified that metric distances between feature points are not suited to model perceptual similarity between shapes [2]. This fact is illustrated in Fig. 1, where shapes a and b are similar, i.e., $d(a, b)$ is small. Similarly, $d(b, c)$ is small. Whereas shapes a and c are very different, i.e., $d(a, c)$ is large. So, $d(a, b) + d(b, c) < d(a, c)$, which violates the triangular inequality. Therefore, the perceptual distance measure is non-metric. On the other hand, a metric distance has the desirable properties, i.e., symmetry,

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 745–752, 2004.
© Springer-Verlag Berlin Heidelberg 2004

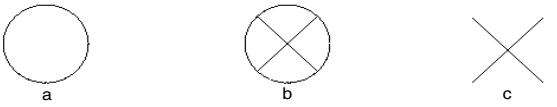


Fig. 1. Example of the triangular inequality failure. See text for explanation.

linearity, and triangularity, which make it efficient in determining the distance between two feature vectors.

In this paper we evaluate and compare shape retrieval efficiency using different metrics as the similarity measures for tokens. Each shape is partitioned into tokens in correspondence with its concave regions. Then, seven invariant moments are computed for each token, which are invariant to translation, scale, and rotation. A metric distance is used to measure the similarity between tokens. Three metric distances are considered, namely, *SSD*, *SAD*, and *ML*. A non-metric distance that employs heuristics is used to measure the similarity between shapes. It is chosen to be the majority vote, that is, a query shape is considered most similar to the shape in the database that shares the largest number of similar tokens with the query shape.

2 Shape Representation and Feature Extraction

Shape representation techniques can be categorized as structural versus global [3]. The main advantages of structural representations are the spatial localization of features and handling multi-object shapes. In the other hand, global representations are compact and, therefore, classical pattern recognition techniques can be applied. However, these global descriptors are imprecise to describe complex shapes. In order to take the advantages of both representations, a complex shape is decomposed into simpler shapes or tokens using a structural approach and global descriptors are obtained for the tokens.

2.1 Convex Hull

A non-convex shape can be analyzed by describing its concave regions. These can be identified by computing the difference between the convex hull of the shape and the shape itself. Borgefors and Baja proposed a technique to find the convex hull of a shape by repeatedly filling local concavities [4]. A good approximation of the convex hull can be achieved using 5×5 neighborhood of the shape's boundary elements. More precisely, the algorithm works as follows:

1. Each boundary pixel, i.e., a background pixel with at least one shape pixel neighbor, is labeled with the number of its shape pixel neighbors.
2. Boundary elements labeled more than 4, together with border elements labeled 4 and having at least one neighbor labeled more then 2 are changed to grey.

(a) (b)

Fig. 2. Result of the concavity filling algorithm: (a) original image, and, (b) the approximation of its convex hull.

The above algorithm is repeated until all concavity regions are filled. The resulted grey envelope that includes the shape represents the convex hull as shown in the example of Fig. 2.

2.2 Invariant Moments

The use of invariant moments to affine transformations (translation, scale, rotation, and skewness) is the most popular method for shape description. For a digital image, the moments are approximated by:

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{1}$$

Where the order of the moment is $(p+q)$, x and y are the pixel coordinates relative to some arbitrary standard origin, and $f(x,y)$ represents the pixel brightness. To make the moments invariant to translation, scale, and rotation, first the central moments are calculated:

$$\begin{aligned} &\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \\ &\bar{x} = \tfrac{m_{10}}{m_{00}}, \bar{y} = \tfrac{m_{01}}{m_{00}} \end{aligned} \tag{2}$$

Then, the normalized central moments are computed:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\lambda}}, \ s.t. \ \lambda = 1 + \frac{p+q}{2} \ \ and \ \ (p+q) \geq 2 \tag{3}$$

From these normalized parameters a set of invariant moments, found by Hu [5], can be calculated, which contain terms up to third order:

$$\begin{aligned} \phi_1 &= \eta_{20} + \eta_{02} \\ \phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\ \phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ \phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ \phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \left((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2 \right) + \\ &\quad (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right) \end{aligned}$$

$$\phi_6 = (\eta_{20} - \eta_{02}) \left((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right) + 4\eta_{11} (\eta_{30} + \eta_{12}) (\eta_{21} + \eta_{03})$$
$$\phi_7 = (3\eta_{21} - \eta_{30}) (\eta_{30} + \eta_{12}) \left((\eta_{30} + \eta_{12})^2 - 3 (\eta_{21} + \eta_{03})^2 \right) - \quad (4)$$
$$(\eta_{30} - 3\eta_{12}) (\eta_{21} + \eta_{03}) \left(3 (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right)$$

3 The Proposed Technique

An overview of the proposed system for shape retrieval is shown in Fig. 3. As can bee seen in the figure, two distinct measures of distance have been used: token (metric) and shape (non-metric) distances. The shape distance is obtained by combining token distances in order to derive a global measure of shape similarity. It is chosen to be the majority vote, that is, a query shape is considered most similar to the database shape that shares the largest number of similar tokens with the query shape. This measure is simple and, to some extent, mimics the human perception. Another desirable property of this scheme is that it provides means for partial matching.

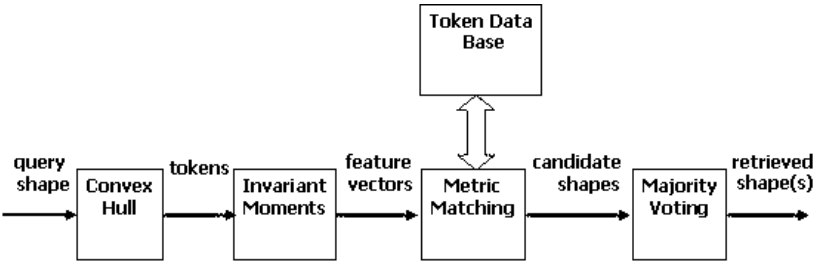


Fig. 3. The proposed system for shape retrieval.

All shapes in the database as well as a query shape are partitioned into tokens by subtracting each shape from its convex hull. To obtain a compact and discriminative description of these tokens, seven invariant moments are computed for each token. Then, for each query token, the database tokens are sorted using a metric. The metric matching in the tokens level results in a number of candidate shapes according to the user preference of the number of retrieved shapes. In the final stage, the candidate shape that shares the largest number of similar tokens with the query shape is considered the best match.

4 Maximum Likelihood Approach

In this section, three metrics that are used in the metric matching stage are viewed and explained. The difference vectors between each query token and all tokens in the database can be viewed as a noise with certain PDF. Sebe *et al.* showed how *ML* theory is used to relate the noise distribution to a metric [6].

Specifically, given the noise distribution, the metric that maximizes the similarity probability is:

$$\sum_{i=1}^{M} \rho\left(n_i\right) \tag{5}$$

Where n_i represents the i_{th} bin of the discretized noise distribution, M is the number of bins, and ρ is the maximum likelihood estimate of the negative logarithm of the probability density of the noise. In the case where the noise is Gaussian distributed, the PDF satisfies:

$$P\left(n_i\right) \propto \exp\left(-n_i^2\right) \tag{6}$$

Substituting (6) in (5) results in the so-called *SSD* or L_2 metric:

$$L_2(x,y) = -\sum_{i=1}^{M} \log\left(P\left(n_i\right)\right) = \sum_{i=1}^{M} \left(x_i - y_i\right)^2 \tag{7}$$

Similarly, for the two-sided exponential noise:

$$P\left(n_i\right) \propto \exp\left(-\left|n_i\right|\right) \tag{8}$$

Substituting (8) in (5) results in the so-called *SAD* or L_1 metric:

$$L_1(x,y) = \sum_{i=1}^{M} \left|x_i - y_i\right| \tag{9}$$

If the noise is Gaussian distributed, then (7) is equivalent to (5). Therefore, in this case the corresponding metric is *SSD*. In the same way, if the noise is exponential, then (9) is equivalent to (5) and the corresponding metric is *SAD*. However, if the noise distribution is neither Gaussian nor exponential, a metric can be extracted directly from the PDF of the noise, called the maximum likelihood metric, using (5):

$$L_{ML}(x,y) = -\sum_{i=1}^{M} \log\left(P\left(x_i - y_i\right)\right) \tag{10}$$

In practise, the probability density of the noise can be approximated as the normalized histogram of the differences between the corresponding feature vector elements [7]. For convenience, the histogram is made symmetric around zero by considering pairs of differences (e.g., $x-y$ and $y-x$). Nevertheless, the histogram approach for approximating the PDF of the noise suffers from being sensitive to the choice of the bin width (shift variant) and discontinuous. To overcome these drawbacks, Parzen windows method is employed where each noise point contributes linearly to the approximated PDF in the small proximity around that point using a given kernel function. Expressly, the approximated PDF is given by:

$$P(n) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{h_n} \phi\left(\frac{n - n_i}{h_n}\right) \tag{11}$$

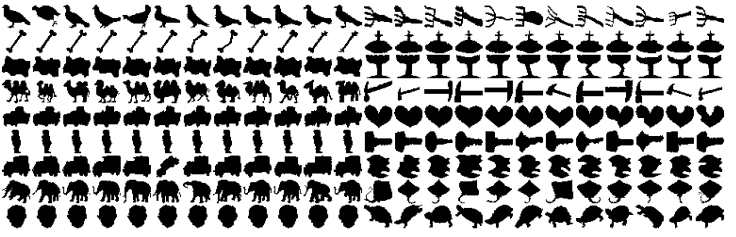


Fig. 4. The shape database used to test our system.

Where M is the number of training points or kernels, $\phi(.)$ is the kernel function, and h_n is the width of the kernel function. Too small width results in a noisy $P(n)$, where as too large width over-smoothes it.

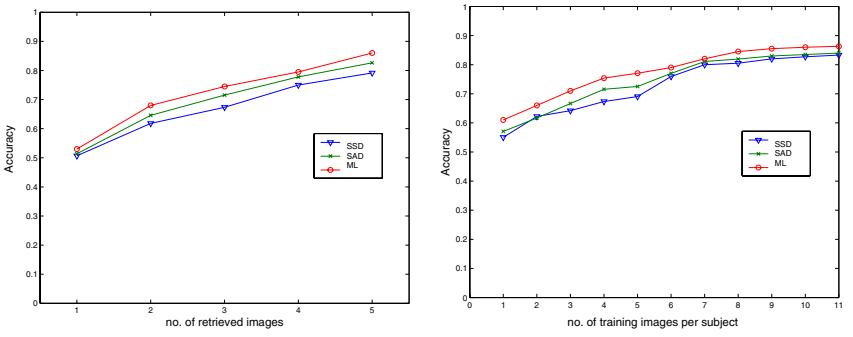
5 Results and Discussions

In the following, the retrieval efficiency of the proposed system is evaluated with more emphasis on role of the metric matching of tokens on the overall performance of the system. Each training image in the database is partitioned into tokens based on its concavity regions and the seven invariant moments are computed for each token. The result is a token database of labeled feature vectors of fixed size. Then, the query image is partitioned in the same way as training images and each query token is matched to its closest tokens from the database using a metric. Three metrics are used for token matching, *SSD*, *SAD*, and *ML*. The similarity between two shapes is measured based on the largest number of the shared similar tokens.

A database of 216 images of 18 shapes (12 images per subject) is used to test our system as shown in Fig. 4. Two experiments are performed to evaluate the retrieval accuracy using different metrics. In the first, the aim is to test the system ability to retrieve the correct shape among a certain number of retrieved database shapes. In other words, the precision is plotted versus the number of retrieved shapes. The intention in the second experiment is to evaluate the system's ability to learn from few examples, i.e., the retrieval accuracy as the number of training images per subject changes.

The results of the first experiment are shown in Fig. 5 (a). A retrieved shape is considered correct only if it belongs to the same subject of the query shape. So, similar shapes are not considered correct matches as long as they belong to different subjects. As can be seen in the figure, the *ML* metric outperforms both *SSD* and *SAD* metrics. Fig. 5 (b) shows the outcome of the second experiment. It can be noticed that the accuracy of the system does not improve significantly when more than five training images per subject are used, which means the system is able to learn from few examples. As in the first experiment, the *ML* metric does better than other metrics.

Finally, χ^2 test is used to prove whether the noise distribution follows Gaussian or exponential distributions or not. For χ^2 goodness-of-fit computation, the



(a) (b)

Fig. 5. Accuracy of retrieval using different metrics versus (a) the number of retrieved images using one-third of the database for training and two-thirds for testing, and, (b) the number of training images per subject using three retrieved images.

```
                CHI-SQUARED GOODNESS-OF-FIT TEST

NULL HYPOTHESIS HO:       DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA

       DISTRIBUTION:              GAUSSIAN

CHI-SQUARED TEST STATISTIC      =     716.1846
   DEGREES OF FREEDOM           =       33

   ALPHA LEVEL          CUTOFF              CONCLUSION
          10%         43.7452              REJECT HO
           5%         47.3999              REJECT HO
           1%         54.7755              REJECT HO

             DISTRIBUTION:             EXPONENTIAL

CHI-SQUARED TEST STATISTIC      =     210.7851
   DEGREES OF FREEDOM           =       33

   ALPHA LEVEL          CUTOFF              CONCLUSION
          10%         43.7452              REJECT HO
           5%         47.3999              REJECT HO
           1%         54.7755              REJECT HO
```

Fig. 6. Results of χ^2 test of the goodness-of-fit of the noise to the Gaussian and exponential distributions.

test statistic is defined as:

$$\chi^2 = \sum_{i=1}^{M} \frac{(O_i - E_i)^2}{E_i} \tag{12}$$

Where O_i is the observed frequency and E_i is the expected frequency for bin i. The hypothesis that the data are from a population with the specified distribution is rejected if $\chi^2 > \chi^2_{(\alpha,k)}$, where $\chi^2_{(\alpha,k)}$ is χ^2 percentage point function

with k degrees of freedom and a significance level of α. The results of χ^2 test are shown in Fig. 6. It can be deduced that the noise distribution is not Gaussian nor exponential, although the exponential fit is better than the Gaussian. These findings justifies the outcomes of the previous experiments where the *SAD* metric performed better than *SSD* metric and the *ML* metric outperforms both *SSD* and *SAD* metrics.

6 Conclusions

In this paper, a scheme for shape retrieval is proposed. The used shape similarity measure is non-metric and, roughly, mimics the human perception of shapes. However, it makes use of the advantages of the metric similarity measures in the tokens matching. The problem of finding the appropriate metric to use for token matching is addressed. From the experiments, the *SSD* and *SAD* metrics are not justified because the similarity noise distribution is not Gaussian nor exponential, respectively. The *ML* metric, extracted directly from the noise PDF, outperformed both *SSD* and *SAD* metrics. Parzen windows method was used to approximate the noise PDF. It is more robust than the histogram method, which is sensitive to the choice of the bin width. In the other hand, the main drawback of the *ML* metric is that, like most nonparametric approaches, it is computationally expensive. In applications where the speed is not a priority, the *ML* metric is a suitable choice.

References

1. Berretti, S., Bimbo, A., Pala, P.: Retrieval by shape similarity with perceptual distance and effective indexing. IEEE Transactions on Multimedia. **2** (2000) 225–239
2. Berretti, S., Bimbo, A., Pala, P.: Retrieval by shape using multidimensional indexing structures. ICIAP. (1999)
3. Zhang, D.S., Lu, G.: Review of shape representation and description techniques. Pattern Recognition **37** (2004) 1–19
4. Borgefors, G., Sanniti di Baja, G.: Analyzing non-convex 2d and 3d patterns. CVIU. **63** (1996) 145–157
5. Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. On Information Theory. **8** (1962) 179–187
6. Sebe, N., Lew, M.S., Huijsmans, D.P.: Toward improved ranking metrics. IEEE Trans. On PAMI. **22** (2000) 1132–1141
7. Sebe, N., Lew, M.S.: Maximum likelihood shape matching. ACCV. (2002) 713–718

A Novel Shape Feature for Image Classification and Retrieval

Rami Rautkorpi and Jukka Iivarinen

Helsinki University of Technology, Lab. of Computer and Information Science
P.O. Box 5400, FIN-02015 HUT, Finland
{rami.rautkorpi, jukka.iivarinen}@hut.fi

Abstract. In this paper a novel statistical shape feature called the edge co-occurrence matrix (ECM) is proposed for image classification and retrieval. The ECM indicates the joint probability of edge directions of two pixels at a certain displacement in an image. The ECM can be applied to various tasks since it does not require any segmentation information unlike most shape features. Comparisons are conducted between the ECM and several other feature descriptors with two defect image databases. Both the classification and retrieval performances are tested and discussed. The results show that the ECM is efficient and it provides noticeable improvement to the performance of our CBIR system.

1 Introduction

There are lots of different features available that are used in image classification and retrieval. The most common ones are color, texture and shape features [1]. During the last decade the content-based image retrieval (CBIR) systems have gained much popularity in many fields of industry and research [2,3]. The main idea in CBIR is to store and retrieve images based only on their visual content described by a set of features, not on manual annotations. Color and texture features are the most common ones in CBIR. Shape features usually require that images have segmentation information available which makes them useless for many tasks. However, different edge histograms, that can be considered as statistical shape features, do not require segmentation which has made them popular in CBIR applications.

In this paper a novel statistical shape feature called the edge co-occurrence matrix (ECM) is proposed. The ECM indicates the joint probability of edge directions of two pixels at a certain displacement in an image. Comparisons are conducted between the ECM, the MPEG-7 feature descriptors, and the simple shape descriptor with two defect image databases. Both the classification and retrieval performances are tested and discussed. Some earlier work with these databases and the MPEG-7 features are found eg. in [4,5].

2 Edge Co-occurrence Matrix (ECM)

The edge co-occurrence matrix (ECM) contains second order statistics of edge directions in an image. It resembles the classical gray level co-occurrence matrix

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 753–760, 2004.
© Springer-Verlag Berlin Heidelberg 2004

(GLCM) [6] but the main difference is that the ECM uses an edge image as its input when the GLCM uses a gray level image. The ECM is an extension of the co-occurrence matrix of edge directions that was proposed in [7]. For further discussion on the differences of these methods, see Section 5.

The first step when forming the ECM is to produce an edge image from the original gray level image. First the original image is filtered with eg. Sobel masks. The edges are detected in 8 directions and the direction of the strongest edge is selected for each pixel location. The edge image is thresholded, retaining only those edge pixels whose strength exceed a threshold value, defined as a percentage of the strongest edge value present in the image.

The edge co-occurrence matrix is then formed from the pairs of edge pixels separated by a given displacement. Let $\mathbf{I}$ be an edge image and let $\mathbf{d} = (d_x, d_y)$ be a displacement vector. Then the edge co-occurrence matrix $\mathbf{H}$ is defined as a matrix, the (i, j)th element of which is the number of appearances of edge directions i and j in the edge image $\mathbf{I}$ in the distance and direction determined by the displacement vector $\mathbf{d}$ from each other,

$$H_{ij} = \#\{x \mid I(\mathbf{x}) = i,\ I(\mathbf{x} + \mathbf{d}) = j\}, \tag{1}$$

where $\#$ is the number of elements in the set and $\mathbf{x} = (x, y)$ runs through the edge image $\mathbf{I}$. Since the edges were detected in 8 directions, the size of the ECM is 8×8.

The ECM entries are then used as elements in a feature vector. Multiple displacements can be considered by averaging the resulting matrices, although doing so will lose some of the extracted information, or by concatenating the feature vectors, which can result in unmanageably long feature vectors.

Some examples of ECMs are shown in Figure 1. The ECMs were calculated using a displacement $\mathbf{d} = (1, 1)$. The matrices are presented as bitmaps, with intensity representing bin values. For clarity, the values have been normalized so that the highest bin value in a histogram is shown as white.

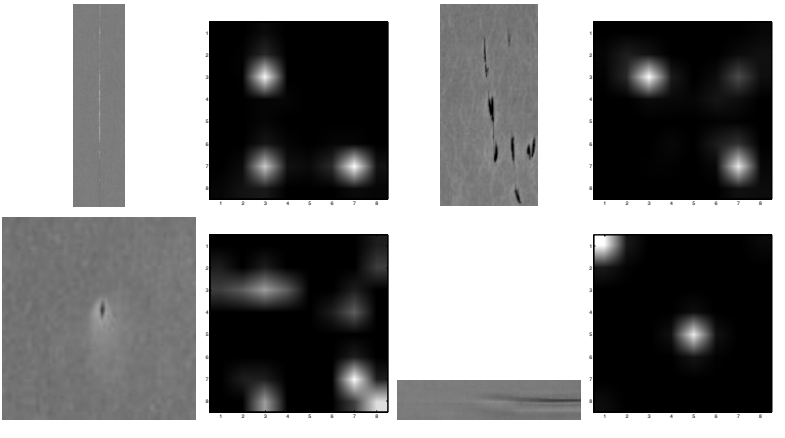


Fig. 1. Edge co-occurrence matrices for four images from the metal defect database.

3 Other Feature Descriptors

Other feature descriptors considered in this paper are taken from the MPEG-7 standard, formally named "Multimedia Content Description Interface" [8]. These descriptors were selected for several reasons. They are well standardized descriptors that are used in searching, identifying, filtering and browsing images or video in various applications. We have previously found them to perform well with our image databases [4,5]. In addition to MPEG-7 features we also tested a simple shape descriptor (SSD). The SSD was developed for surface defect description in our earlier project [9].

3.1 The MPEG-7 Standard Descriptors

The MPEG-7 standard defines several still image descriptors. We have used the following six descriptors. *Color layout (CL)* specifies a spatial distribution of colors. The image is divided into 8×8 blocks and the dominant colors are solved for each block in the YCbCr color system. Discrete Cosine Transform is applied to the dominant colors in each channel and the DCT coefficients are used as a descriptor. *Color structure (CS)* slides a structuring element over the image. The numbers of positions where the element contains each particular color are stored and used as a descriptor. *Scalable color (SC)* is a 256-color histogram in HSV color space, which is encoded by a Haar transform. *Homogeneous texture (HT)* filters the image with a bank of orientation and scale tuned filters that are modeled using Gabor functions. The first and second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor. *Edge histogram (EH)* calculates the amount of vertical, horizontal, 45 degree, 135 degree and non-directional edges in 16 sub-images of the picture, resulting in a total of 80 histogram bins. *Region-based shape (RS)* utilizes a set of 35 Angular Radial Transform (ART) coefficients that are calculated within a disk centered at the center of the image's Y channel.

From these descriptors the EH and RS are the most interesting ones since they are both shape descriptors. It is especially interesting to see how the ECM performs against the EH which describes first order statistics of edge directions.

3.2 Simple Shape Descriptor (SSD)

The SSD [9] consists of several simple descriptors calculated from an object's contour. The descriptors are convexity, principal axis ratio, compactness, circular variance, elliptic variance, and angle. They are not very efficient individually, but their combination produces good results with low computational costs.

4 Experiments

Experiments were carried out with two image databases containing defect images, one from a metal web inspection system and the other from a paper web

inspection system. All images were grayscale, supplied with mask images containing segmentation information, which was used to remove edge pixels from uninteresting areas of the images. The images have different kinds of defects and their sizes vary according to the size of a defect. Classification of defects is based on the cause and type of a defect, and different classes can therefore contain images that are visually dissimilar in many aspects. The paper defect database has 1308 images. They are preclassified into 14 different classes with 100 images in 12 of the classes, 76 images in class number 11 and 32 images in class number 12. The metal defect database has 2004 images. They are preclassified into 14 different classes, with each class containing from 101 up to 165 images. The databases were provided by ABB Oy.

The classification performance of different features was tested with the K-Nearest Neighbor leave-one-out cross-validation and the built-in CBIR analysis system of a PicSOM [10], a content-based image retrieval system developed at the Laboratory of Computer and Information Science at Helsinki University of Technology. The main features of PicSOM are efficient indexing based on tree-structured self-organizing maps and adaptive querying using relevance feedback. In our earlier work we have applied PicSOM successfully to similar defect image databases that are used in this paper, the largest one having approximately 45000 paper defect images [4,5].

Parameter values for the experiments were determined with some initial test runs. The optimum threshold value for these datasets was found to be approximately 15%, meaning that any detected edge pixels with a strength less than 15% of the strongest edge present in the image were discarded. Good results were obtained, while keeping the feature vector as short as possible, by using two edge pixel pair displacements, $\mathbf{d} = (1, 1)$ and $\mathbf{d} = (2, 2)$. The feature vector was formed by concatenating the ECMs written in vector form.

4.1 KNN Classification Results

Classification results with K-Nearest Neighbor leave-one-out cross-validation are shown in Tables 1 and 2. The results are compared with those obtained with the different MPEG-7 descriptors and the SSD using segmentation masks.

Tables 1 and 2 show also the results of a comparison using a set of features previously found to be good, including the edge histogram (EH). The classification results for a set of features were determined by voting. The comparison shows classification success rates with the earlier four features, then with the ECM added, and finally with the EH removed and replaced with the ECM. The average success rate is weighted with the number of images in each class.

The ECM performed very well as an individual feature, ranking as the second best feature in the metal database and the third best feature in the paper database. Overall it performed better than the EH, although the EH outperformed it in some individual classes, mostly in the paper database.

Adding the ECM to the set of features increases the classification success rate for most of the individual classes, and increases the average success rate a few percent in the metal database, while in the paper database there is no

noticeable change in the average success rate. Removing the EH feature results in a slight decrease in success rates, but in the metal database the results are still better than without the ECM, and in the paper database only slightly worse than without the ECM.

Table 1. KNN classification results with the metal defect database.

	Classification success rates (%)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg
CS	29	57	72	45	71	71	80	30	82	59	72	67	56	87	63
ECM	36	56	66	36	33	82	71	53	87	50	92	54	62	99	62
HT	33	53	38	44	64	73	57	31	64	59	79	39	57	95	57
EH	37	45	27	30	62	63	54	26	55	61	61	21	33	91	49
SSD	29	45	8	33	42	59	55	40	25	46	44	16	43	96	42
SC	22	25	40	10	44	22	43	20	51	28	27	21	23	67	32
CL	22	18	16	29	36	17	18	5	35	28	23	27	8	80	27
RS	16	13	14	8	29	38	24	9	17	13	16	7	11	65	20
CS,HT,EH,SSD	51	75	71	56	84	88	89	55	87	80	90	62	72	97	76
CS,HT,EH,SSD,ECM	52	79	82	54	81	90	88	60	95	74	99	78	75	98	79
CS,HT,SSD,ECM	49	77	76	54	77	90	89	60	95	72	97	77	78	99	78

Table 2. KNN classification results with the paper defect database.

	Classification success rates (%)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg
HT	96	54	71	89	74	82	97	65	91	94	75	0	64	93	79
CS	97	75	60	88	82	78	70	59	77	62	76	15	86	95	76
ECM	87	75	47	96	31	60	50	65	61	95	7	0	63	96	64
EH	61	25	68	91	55	23	76	37	64	69	79	11	78	91	61
SSD	50	44	47	88	37	24	41	77	65	72	14	4	30	91	52
SC	75	6	28	69	39	74	90	17	6	18	70	0	46	60	44
RS	54	27	37	52	35	26	43	34	45	57	38	0	35	88	42
CL	85	11	11	64	41	78	23	32	14	33	63	4	31	39	39
HT,CS,EH,SSD	99	84	83	100	80	87	84	96	95	93	35	0	88	96	86
HT,CS,EH,SSD,ECM	99	87	78	100	79	89	85	96	94	95	40	0	89	98	86
HT,CS,SSD,ECM	99	87	78	100	75	88	82	94	94	97	25	4	87	99	85

4.2 CBIR Results

Figures 2-3 show a similar comparison with PicSOM's CBIR analysis system, which emulates a human user making queries and giving feedback on the retrieved images in order to find images belonging to a specific class.

The results are recall and precision values for each iteration of image retrieval. Recall is the percentage of images belonging to the desired class that have been retrieved so far. A recall of 100% means that all desired images have been found. Precision is the percentage of desired images from all the images retrieved so

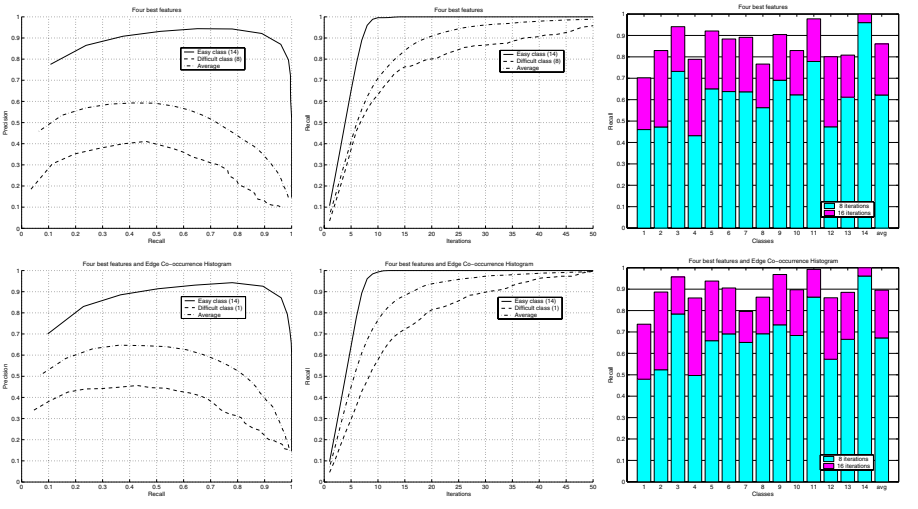


Fig. 2. PicSOM's CBIR results with the metal defect database.

far. Precision should be higher than the a priori probability of the desired class, otherwise the system's retrieval performance is worse than random selection.

First on the left are precision/recall graphs, illustrating the retrieval performance as the query progresses. The average graph shown is a result of calculating for each iteration the weighted average of the precision and recall values over all classes. The graphs in the middle show the recall value on each iteration. The bar graphs show the recall values for each class on iterations 8 and 16 for the metal database, and iterations 5 and 10 for the paper database. Nearly all classes in the metal database have between 140 and 160 images, and less than 100 in the paper database. Since 20 images are retrieved on each iteration, an ideal system would achieve 100% recall after 8 and 5 iterations on the metal and paper databases, respectively. To see how the recall value increases when additional iterations are used, the number is doubled.

Adding the ECM to the set of features increases recall results clearly, even in the paper database, which did not show any improvement in the KNN classification results. PicSOM's CBIR system is able to give more weight to a feature when it is performing well, so that a set of complementing features can be utilized efficiently. The results show that the ECM provides additional discriminatory power beyond that of the EH.

5 Discussion

Initial tests revealed the significant effect the choice of edge pixel pair displacements has on ECM results. Increasing the displacement distance reduced classification success rates, which is to be expected, because the further apart two

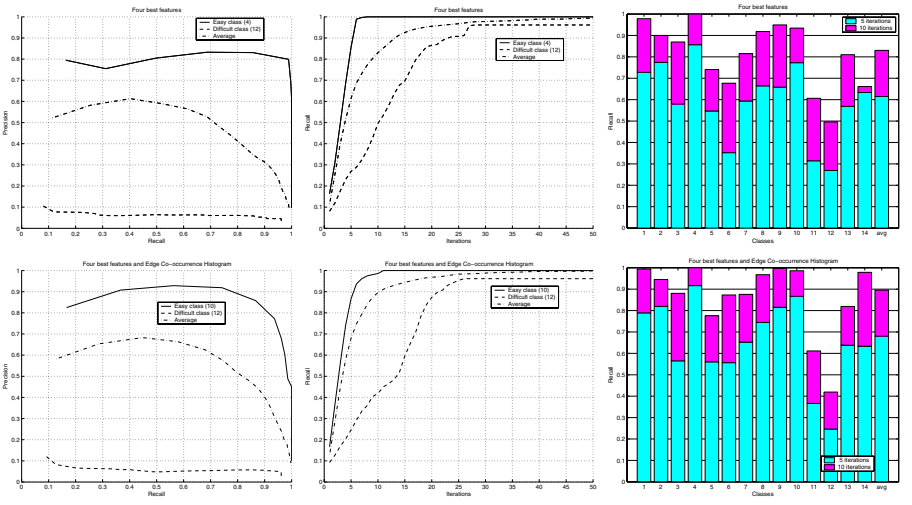


Fig. 3. PicSOM's CBIR results with the paper defect database.

pixels are, the less likely they are to have any meaningful relationship to each other, and also the less likely it is that they are both edge pixels, since a significant portion of the image contains no edges. Averaging multiple ECMs with different displacement directions was not very beneficial. Doing so introduces rotation invariance to the co-occurence histogram, and in this application the orientations of the defect shapes are significant.

The ECM differs from the GLCM in some crucial ways. Most of the contents of the GLCM is concentrated on the diagonal since a pixel in a natural image is usually surrounded by pixels with the same or very similar gray level. Also the upper right and lower left corners are usually very nearly empty, since it is unlikely that a pixel with the lowest possible gray level would be near a pixel with the highest possible gray level. However, in the ECM there are no "low" and "high" values for edge directions so the corners actually represent pairs with similar edge directions. The values are not usually distributed evenly along the diagonal but instead the general shape of an object in an image has a strong effect on the distribution. For example, a horizontally elongated object is likely to have most of the values concentrated around the elements (0,0) and (4,4) where 0 and 4 are the values for the horizontal edges.

As already noted, the ECM is an extension of the co-occurrence matrix of edge directions that was proposed by Brandt et al. [7]. While the former method collects edge pixel pairs separated by a given displacement, the latter method collects edge pixel pairs separated by a set of displacements. This set contains eight displacements that cover all eight neighboring edge pixel pairs for each pixel position in an image. So the co-occurrence matrix of edge directions is an average of eight ECMs that are formed with different displacement vectors.

6 Conclusions

In this paper a novel statistical shape feature called the edge co-occurrence matrix (ECM) was presented. The KNN classification performance and the retrieval performance in the PicSOM CBIR system were tested and compared with several other feature descriptors using two defect image databases. The results in all cases show the ECM to be quite efficient.

Using the entire ECM results in a long feature vector which can be unpractical. The work will be continued with experiments on the application of Haralick features, commonly used with GLCMs, on the ECM, in order to produce shorter and more efficient feature vectors.

Acknowledgments. The financial supports of the Technology Development Centre of Finland (TEKES's grant 40120/03) and our industrial partner ABB Oy (J. Rauhamaa) are gratefully acknowledged.

References

1. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis and Machine Vision. Chapman & Hall Computing, London (1993)
2. Del Bimbo, A.: Visual Information Retrieval. Morgan Kaufmann Publishers, Inc. (1999)
3. Johansson, B.: A survey on: Contents based search in image databases. Technical report, Linköping University, Department of Electrical Engineering, http://www.isy.liu.se/cvl/Projects/VISIT-bjojo/ (2000)
4. Pakkanen, J., Ilvesmäki, A., Iivarinen, J.: Defect image classification and retrieval with MPEG-7 descriptors. In Bigun, J., Gustavsson, T., eds.: Proceedings of the 13th Scandinavian Conference on Image Analysis. LNCS 2749, Göteborg, Sweden, Springer-Verlag (2003) 349–355
5. Rautkorpi, R., Iivarinen, J.: Content-based image retrieval of web surface defects with PicSOM. In: Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary (2004)
6. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics **SMC-3** (1973) 610–621
7. Brandt, S., Laaksonen, J., Oja, E.: Statistical shape features in content-based image retrieval. In: Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain (2000)
8. Manjunath, B.S., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley & Sons Ltd. (2002)
9. Iivarinen, J., Visa, A.: An adaptive texture and shape based defect classification. In: Proceedings of the 14th International Conference on Pattern Recognition. Volume I., Brisbane, Australia (1998) 117–122
10. Laaksonen, J., Koskela, M., Laakso, S., Oja, E.: Self-organising maps as a relevance feedback technique in content-based image retrieval. Pattern Analysis and Applications **4** (2001) 140–152

A Local Structure Matching Approach for Large Image Database Retrieval

Yanling Chi and Maylor K.H. Leung

School of Computer Engineering, Nanyang Technological University, 639798, Singapore
chiyanling@pmail.ntu.edu.sg, asmkleung@ntu.edu.sg

Abstract. The combination of a local structure based shape representation and a histogram based indexing structure is proposed to fast localize objects from large database. Four novel geometric attributes are extracted from each local structure. They are invariant to translation, scaling, rotation and robust to adverse distortions and noise. The search space is pruned by means of histogram intersection and the computation cost of the query is linear to the number of input features. The matching is performed by a non-metric similarity measure with regard to significance in reconstruction of query image and discrimination of different models. The concepts proposed were tested on thousands of images. The result manifests its efficiency and effectiveness.

1 Introduction

Recent years have seen a rapid increase in the size of digital image collections. Finding a moderate number of matches relevant to a given visual query is the key to a wide range of applications, such as, visual scene interpretation, object recognition, trademark registration, satellite image repositories and fingerprint identification. Among the research works in this area, the general-purpose shape matching technique is particularly challenging [1, 2]. On one hand, the *adverse environment and dimensionality curse* make the capabilities of current technology still fall short of most users' requirements. On the other hand, there is *little knowledge about the way in which humans perceive shape similarity* and there is enough evidence that metric distances are not suited to model shape similarity perception [1, 9]. This paper reports our investigations into the general-purpose shape matching approach which endeavors to overcome these two difficulties.

Many techniques for shape matching have been proposed these days. Boundary deformation [3] is to minimize a cost function that depends on a weighed sum of the mismatch of the two shapes and on the deformation energy. The complexity depends on the number of points on the contour. Hence, the optimization is computationally expensive and this, in spite of the greater precision, has limited its diffusion in image database. Some other techniques based on shape feature vectors are more suitable for effective indexing. These include Fourier descriptors [4], moment invariants [5], Zernike moments [6] and curvature scale-space [7]. However, they employ global shape descriptor and demonstrate metric distances (such as Euclidean) between feature vectors that do not model human visual perception, e.g. similarity evaluation in the case of partially occluded objects. The alternative solution is local feature de-

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 761–768, 2004.
© Springer-Verlag Berlin Heidelberg 2004

scriptors [8, 9], which is based on partition of shape boundary into tokens. This approach is suitable for the evaluation of similarity of partially occluded shapes.

A combination of a novel shape representation and new indexing structure is proposed as our investigation into ways of matching shape in a robust, efficient, perceptually significant manner. The local-structure-based shape representation makes use of the structural relationships inherent in a pattern's shape and is suitable to tackle the problem of partial object occlusion. Three aspects involved in local shape representation are primitives of local structure, selection rules, and compact feature extraction as shown in Fig. 1. Lines or arcs are selected as two primitives to construct the local structure due to their ease of detection. Gestalt laws are adopted as the selection rules to highlight perceptual significant shape features. Four novel geometric attributes are extracted from each local structure as local features. These attributes are invariant to translation, scaling and rotation and robust to adverse distortions and noise. A 4-D histogram indexing structure is proposed to conglomerate the local shape representations. In shape indexing, dimensionality curse is settled by means of histogram intersection and the computation cost of the query is linear to the number of input features. An efficient non-metric similarity measure is derived from the local shape representation to achieve a comprehensive shape measure since the similarity is with respect to significance in reconstruction of query image and discrimination of different models.

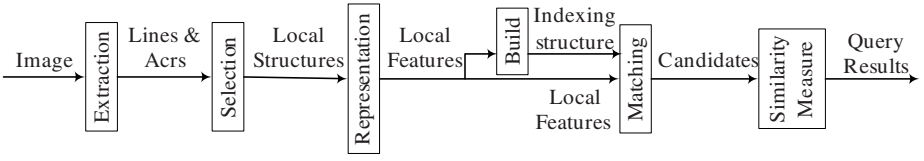


Fig. 1. Flowchart of the local structure shape matching.

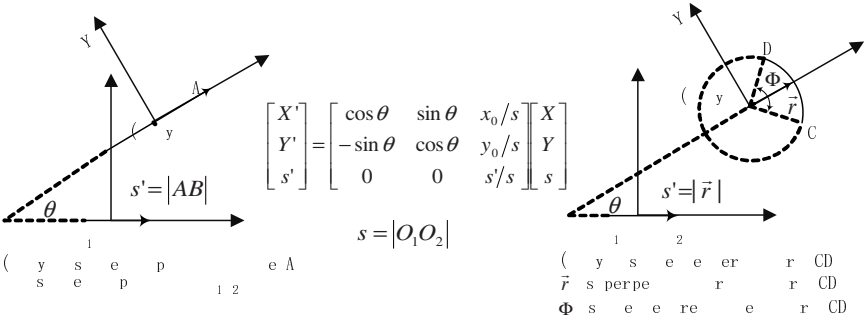
2 Shape Representation

Cognitive psychological studies [10] indicated that human beings recognize line drawings as quickly and almost as accurately as gray-level pictures. Line patterns have been proved to be highly effective for shape representation [11]. With the advantages of simple representation, low storage demanding and consistency in detection, line is one promising primitive feature for shape matching from large database. Arc is the complement of line when dealing with the circular object. Therefore, both line and arc are adopted as primitives to construct local structures. High level primitives such as corners are not considered since they require more complicate detection methods that usually fail when different types of distortions occur. The inherent structural relationship between a pair of primitives will then be extracted for representation.

2.1 Local Structure Selection

To follow human perception of significant shape features, the Gestalt psychology theory [12] is employed to guide the construction of local structure. Given a reference vector $\overrightarrow{O_1O_2}$, a line (AB) or arc (CD) can be related from reference frame X-Y to a

new frame X'-Y' by the matrix shown in Fig.2. The relationship parameters will be made use of later.



(a) Lines. (b) Arcs.

Fig. 2. The parameters of the transform matrix.

Four Gestalt theory laws [13, 14], i.e. law of focal point, law of proximity, law of similarity and law of symmetry, are adopted to select local structures. A local structure consists of one reference feature and its neighbors.

Reference features selection: According to Law of Focal Point, top longest lines and arcs with highest *Arcoid* (to be defined below) are selected as focal points, which will work as reference features and the local structure are constructed around each of the reference features. can be up to 100 percent, but usually 50 percent is sufficient. *Arcoid* is an evaluation of arc feature. It prefers long arc length and large angle (Fig. 2 (b)). It is defined as *ArcLength*/(1- /(2)).

Local neighbors' selection: Around each reference feature, a combination of the laws of proximity, similarity and symmetry (equation (1)), is applied to determine suitable neighbors. w_1, w_2, w_3 are weights to balance the impact of these three laws. Here, the three weights are set as 1, 1, and *s* empirically.

$$Dissimilarity = w_1 D_{pr} + w_2 D_{si} + w_3 D_{sy} \tag{1}$$

- D_{pr} is the distance measure of proximity. Primitive features near each other are preferred. D_{pr} is computed by rotating and transferring the primitive to the reference as

$$D_{pr} = \frac{1}{2} s|\theta| + \sqrt{x_0^2 + y_0^2} \tag{2}$$

- D_{si} compute the distance measure of similarity as

$$D_{si} = \left| s'^2 - s^2 \right| \tag{3}$$

- Let define another reference vector $\overrightarrow{O_1O_2}$ from midpoint of line (chord) CD to midpoint of line (chord) AB. Let T_{AB} and T_{CD} be the transform matrix relative to

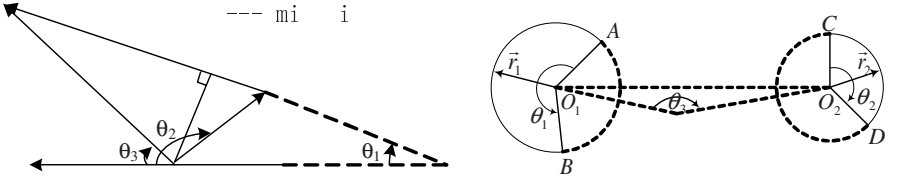
$\overrightarrow{O_1O_2}$ as $T_{AB} = \begin{bmatrix} \cos\theta_1' & \sin\theta_1' & x_1' \\ -\sin\theta_1' & \cos\theta_1' & y_1' \\ 0 & 0 & s \end{bmatrix}$ and $T_{CD} = \begin{bmatrix} \cos\theta_2' & \sin\theta_2' & x_2' \\ -\sin\theta_2' & \cos\theta_2' & y_2' \\ 0 & 0 & s' \end{bmatrix}$. If $s = s'$ and $\theta_1' + \theta_2' = \pi$, then AB and CD are axial symmetric; if $s = s'$ and $\theta_1' = \theta_2'$, then line segment AB and CD are central symmetric. D_{sy} can be defined according to these two symmetries as

$$D_{sy} = \min\left(\left|\theta_1' - \theta_2'\right|, \left|\theta_1' + \theta_2' - \pi\right|\right) \tag{4}$$

Since the unit of D_{sy} is radian here, w_3 is set as s, the scale of AB in T_{AB}.

2.2 Pair-Wise Geometric Attributes

In order to capture the local structural information honestly and achieve the desired robustness under adverse conditions such as different scale, different orientation, and broken lines effects, the selected geometric attributes of primitive features should be simple, invariant to translation, scaling and rotation, relatively robust to fragmentation and end-point erosion, and sufficient for discrimination.



(a) line (b) arc

Fig. 3. Pair-wise Geometric Attributes.

Line Pair-wise Geometric Attributes. A vector of four Euclidean invariant attributes, $GA_L = \left[\theta_1, \theta_2, \theta_3, d/|AB|\right]^T$ is computed to represent a pair of line segments. With these four geometric attributes, the pair of line segments can be completely reconstructed at a normalized scale and orientation. Let the line segments AB and CD be represented by the vector $\overrightarrow{AB}$ and $\overrightarrow{CD}$. The vectors are pointing away from their intersect point as shown in Fig. 3 (a). The range of angle θ_1 is [-π, π]. A positive sign is given to the relative angle θ_1 if the direction of the angle from the reference line AB to its neighbor CD is clockwise, and a negative sign is given if the direction is counter-clockwise. Similarly, the range of θ_2 and θ_3 are [-π, π]. d, normalized by the length of AB, works as the forth geometric attributes.

Arc Pair-wise Geometric Attributes. Similarly, a vector of four Euclidean invariant attributes $GA_C = \left[\theta_1 - \pi, \theta_2 - \pi, \theta_3, 1/\left(1 + |O_1O_2|/\left(|r_1| + |r_2|\right)\right)\right]^T$ is computed to describe a pair of arcs, shown in Fig.3 (b). The meaning of θ_1, θ_2, and θ_3 are straightforward, θ_1, is the angle of reference arc AB, θ_2 is the angle of neighbor arc CD, and θ_3 is the in-

tersecting angle between r_1 and r_2. The range of θ_1 and θ_2 is $[0,2\pi]$ while that of θ_3 is $[-\pi, \pi]$. In the forth geometric attribute, the component $|O_1O_2| / (|r_1|+|r_2|)$ is used to determine whether two circles intersect. It is normalized to the 4th attribute with range (0, 1]. If it is equal to 0.5, the two circles are tangent. If it is larger than 0.5, the two circles intersect. If it is less than 0.5, the two circles have no intersection point.

Arc-line Pair-wise Geometric Attributes. The four pair-wise geometric attributes between line segments and chords will be named as the arc-line pair-wise geometric attributes.

3 Shape Indexing

A 4-dimensional histogram (H) is built to conglomerate all the local structures of one image and work as the element of indexing structure. For an image with n pair features, let $\boldsymbol{x}$ be the pair-wise feature vector as $[x_1, x_2, x_3, x_4]$. x_1, x_2, x_3, and x_4 are the four geometric attributes from GA_L and GA_C. Let the normalized ranges of the four attributes be $[0, A_1]$, $[0, A_2]$, $[0, A_3]$ and $[0, A_4]$. When the attribute spans are divided into n_1, n_2, n_3, n_4 bins respectively, the histogram has $n_1 \times n_2 \times n_3 \times n_4$ entries. The frequency counts accumulated in $H(i, j, k, l)$ are

$$H(i,j,k,l) = \sum_{m=1}^{n} c_m \qquad where \quad c_m = \begin{cases} if: x_{m1} \in (A_{i-1,1}, A_{i,1}], x_{m2} \in (A_{j-1,2}, A_{j,2}], \\ \qquad x_{m3} \in (A_{k-1,3}, A_{k,3}], x_{m4} \in (A_{l-1,4}, A_{l,4}]. \\ 0 \quad otherwise. \end{cases}$$

$$here \quad A_{i,1} = \frac{i}{n_1} A_1 \quad A_{j,2} = \frac{j}{n_2} A_2 \quad A_{k,3} = \frac{k}{n_3} A_3 \quad A_{l,4} = \frac{l}{n_4} A_4$$

n is the number of pair-wise features. x_{m1}, x_{m2}, x_{m3}, and x_{m4}, are the four geometric attributes of the mth pair-wise feature, $A_{i,1}$, $A_{i,2}$, $A_{k,3}$, and $A_{l,4}$ are the upper limits of bin (i, j, k, l). For the purpose of saving disk space, only histogram entries not equal to zero are stored.

All the images in the database are integrated into one indexing structure ($\tilde{I}$) with the same dimensions and sizes of the previous histogram. Let $M = \{m_1, m_2, \ldots, m_D\}$ be a set of model images, $H = \{H_{m1}, H_{m2}, \ldots, H_{mD}\}$ be their corresponding image compact representations and $N = \{ N_{m1}, N_{m2}, \ldots, N_{mD} \}$ be their corresponding feature size. The bin entry of the indexing structure $\tilde{I}$ is defined by

$$I(i,j,k,l) = \left\{ \bigcup_m \left(m, H_m(i,j,k,l) \right) : m \in \mathcal{M}, H_m \in H, H_m(i,j,k,l) \neq 0 \right\}$$

Three separate steps are involved in the image indexing process. First, H_{input}, the compact image representation of the test image, is intersected with $\tilde{I}$ to identify non-empty entries from both structures to compose the candidate list. The score of each candidate model is then computed. Finally, the candidate results are ranked and the top most likely matches can be found. Since $\tilde{I}$ and H_{input} are of same dimensions and sizes, the intersecting process becomes a simple logical AND operation to isolate

common non-empty entries. The results from the intersecting are then combined into the set *X*.

$$\mathcal{X} = \left\{ \bigcup (I(i,j,k,l), H_{input}(i,j,k,l)) : H_{input}(i,j,k,l) \neq 0, I(i,j,k,l) \notin \Phi \right\}$$

The candidate models in X are ranked according to the following similarity score. To derive a measure of shape similarity from the indexed local structures, several factors are taken into consideration. The first is the number of corresponding local structures with respect to model *m* as $\sum \min (H_{input}(i, j, k, l), H_m(i, j, k, l))$. The second is the discrimination power of the local structures, i.e. how well these local structures can tell the difference among the models by dividing each count by $n(I(i, j, k, l))$ which is the number of images in entry $I(i, j, k, l)$ as shown in (5). The last is how significant these local structures are when used to construct the query example and the models. This is the normalization faction in (5) where N_{input} and N_m are the feature sizes of the test image and model respectively.

$$Score(m) = \frac{1}{2}\left(\frac{1}{N_{input}} + \frac{1}{N_m}\right)\left(\sum_{I(i,j,k,l)\in X} \frac{\min(H_{input}(i,j,k,l), H_m(i,j,k,l))}{n(I(i,j,k,l))}\right) \tag{5}$$

The on-line complexity of image indexing can be expressed in two parts; one is the number of geometric attributes of the input image that need to be computed. The other is the number of similarity measures of the model images that need to be computed. If C_X is the number of the candidates in set *X* after interception, there are $O(N_{input})$ geometric attributes and $O(C_X)$ scores need to be computed. The primitive feature extraction can be done in time linear in image size. Therefore the overall image filtering $O(N_{input} + C_X) \approx O(N_{input} + c\,N_{input})) = O((c+1)\,N_{input}) \approx O(|M|\,N_{input})$ can be done in time linear in the number of input image features.

4 Experiment Result and Conclusion

To assess effectiveness of the proposed method, the common performance measures such as precision and recall of the retrieval are adopted. The test trademark database is obtained from computer vision homepage [17]. 1296 test images are created and divided into 54 groups. The groups reflect general variations of shapes, such as occlusion, deformation, erosion, rotation, different scale, confusion of line and arc features, edge fraction, extra long line, noise, etc, as shown in Fig. 4.

Each of the 1296 images is used as a query in turn. For a given query, let *T* be the total number of relevant items available (the images in the group that the query belongs to), *r* the number of relevant items indexed, and *t* the total number of indexed items. Then precision *P* is defined as *r*/*t* and the recall as *r*/*T*. The results are displayed in Fig. 5 (a).

System robustness in case of partial object occlusion has been estimated by the query with some local structures deleted. 100% local structures of the 1101 model images are used for building the indexing structure. Experiments were carried out using these 1101 images with only 10%, 20% or 30% of local structures. System robustness is measured by the precision in the case of one image returned. Fig.5 (b) shows the precision figures for increasing degree of missing information.

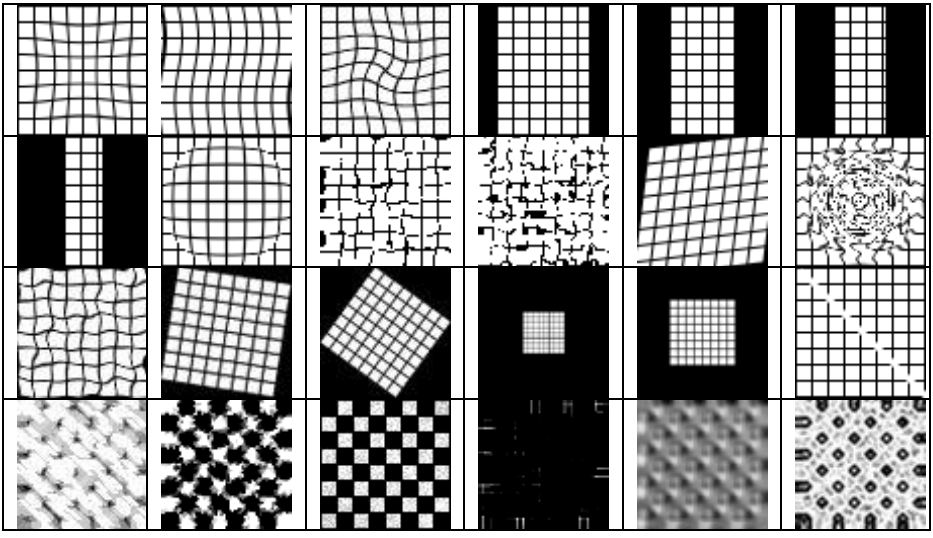


Fig. 4. Sample of distorted and noisy transforms.

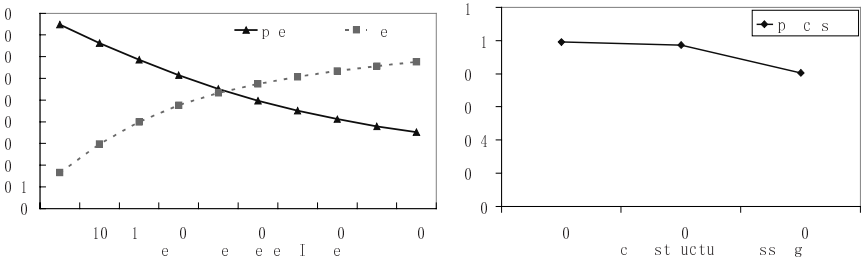


Fig. 5. (a)Precision and Recall versus number of retrieved Images, (b) precision for increasing degree of local structure missing.

From above results, the performance on classification is good and the recall is 70% with 50 images returned. Its performance on partial occlusion is distinctive. The system has almost 100% precision with only 20 percent information left. The indexing cost can be reduced further since it is linear to the number of input features. The method proposed is very robust to adverse conditions such as partial occlusion, reasonable noise and distortion. It degrades gracefully as the tolerances are exceeded. Such robustness makes its feasible to apply our approach to natural objects query since natural objects are usually in a complex environment which is similar to the simulations in this paper. Our future work is to test these concepts under more challenging conditions to generalize them naturally.

References

1. Arnold W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, 2000.
2. J.P. Eakins, K.J. Riley, J.D. Edwards, "Shape Feature Matching for Trademark Image Retrieval", *International conference on Image and video retrieval*, Illinois, USA, LNCS2728, pp. 28-38, July, 2003.
3. A.D. Bimbo, P. Pala, "Visual Image Retireval by Elastic Matching of User Sketches", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 121-132, 1997.
4. D. Zhang, G. Lu, "Shape-based image retrieval using generic Fourier descriptor", Signal Processing: image communication, vol 17, pp. 825-848, 2002.
5. C.-H. Teh, R.T. Chin, "On image analysis by the methods of moments", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 10, pp. 496-513, 1988.
6. Y.S. Kim, W.Y. Kim, "Content-based trademark retrieval system using a visually salient feature", Image and Vision Computing, vol. 16, pp. 931-939, 1998.
7. F. Mokhtarian and A.K. Mackworth, "The Curvature Scale Space Representation: Theory, Applications and MPEG-7 Standardization", *Kluwer Academic Publishers*, 2002.
8. R. Mehrota, J.E. Gary, "Similar-shape retrieval in shape data management", *IEEE Computer*, vol. 28, pp. 57-62, 1995.
9. S. Berretti, A.D. Bimbo, P. Pala, "Efficient Shape Retireval by Parts", *Computer analysis of images and patterns*, LNCS1689, pp. 57-64, 1999.
10. I. Biederman and J. Gu, "Surface versus Edge-Based Determinants of Visual Recognition," *Cognitive Psychology*, vol. 20, pp. 38-64, 1988.
11. B. Huet, E. R. Hancock, "Line Pattern Retrieval Using Relational Histograms", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1363-1370, 1999.
12. K.Koffka, "Principles of Gestalt Psychology, *Harcourt, Brace and Company*, New York, 1935.
13. M. Fisher, K. Smith-Gratto, "Gestalt theory: a foundation for instructional screen design". *Journal of educational technology systems* vol. 27, no. 4, pp. 361-371, 1998-99.
14. D. Chang, L. Dooley and J. E. Tuovinen, "Gestalt Theory in Visual Screen Design-A New Look at an Old Subject", *the Seventh World Conference on Computers in Education*, Copenhagen, Denmark. 2001.
15. B. Huet, E.R. Hancock, "Relational object recognition from large structural libraries", *Pattern Recognition*, vol. 35, pp. 1895-1915, 2002.
16. M.K.Leung and Y.H.Yang, "Dynamic two-strip algorithm in curve fitting", *Pattern Recognition,* vol. 23, no. 1/2, pp. 69-79, 1990
17. A.K. Jain and A. Vailaya, "Shape-Based Retrieval: A Case Study with Trademark Image Databases," *Pattern Recognition*, vol. 31, no. 9, pp. 1369-1390, 1998.
18. H. Kim, J. Kim, "Region-based shape descriptor invariant to rotation", scale and translation, *Signal Processing: Image Communication*, vol. 16, pp. 87-93, 2000.
19. Y. Rui, T. S. Huang, S.-F. Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues", *Journal of Visual Communication and Image Representation*, vol. 10, pp. 39-62, 1999.
20. A.K. Jain, R. P.W. Duin, J. Mao, "Statistical Pattern Recognition: A Review", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.
21. M. Brady, "Criteria for representations of shape", in *Human and Machine Vision, J. Beck, B. Hope and A. Rosenfeld, eds, Academic Press, New York*, pp. 39-84, 1983.

People Action Recognition in Image Sequences Using a 3D Articulated Object

Jean-Charles Atine

INSA,
135, Avenue de Rangueil,
31077 Toulouse cedex 4 France
jcatine@laas.fr

Abstract. This article deals with recognition of human action using motion tracking of articulated objects based on its 3D geometric modeling. We minimize a measure of difference between views of a real scene and the corresponding 3D ones. The minimization of the function uses colorimetric information and allows us to automatically estimate the position and the posture of the subject in the scene. We used an hybrid genetic algorithm to optimize the 3D model posture. The matching of views acquired by several video cameras displaced in the real scene and the synthesis views allows inferring the parameters that check the model. Afterward, the movement parameters obtain we will identify the people action using a fuzzy classification.

1 Introduction

The approach we use is marker-free. It is not constraining and appropriate to video database indexation. In the computer vision the approaches that do not use any markers are classified in two groups: the one that use a geometric 3D model and the one that do not. Generally, in the case where the goal is to determine perceived movements in the pictures and not to track a precise object, these methods come close to the optical flow.

We take for granted that we have access to a generic model of the subject and to its degree of freedoms to measure his antropometric data and to track the movement.

Some other method for tracking exist. Delamarre [1] presents a method for tracking based on forces. These forces guide the convergence of the model on the extracted silhouette of the reference picture. Ormoneit [4] presents an analysis method based on statistics and Bayesian tracking. In [5], the authors use a representation based on cards. This is an extension of the work presented in [6]on the recognition of facial expression to the problem of recognition and tracking human movement.

To avoid occultation problem we decided to use 4 cameras displaced in the corner of the room where the subject is moving. In the works of Kakadiaris [7], occultations are calculated dynamically and are based on the prediction of movements. The tracking is based on occultation boundaries. The visibility

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 769–777, 2004.
© Springer-Verlag Berlin Heidelberg 2004

criterion of the different parties of the body is an index that allows managing occultations.

Technically, the real scene is simulated using a 3D software. This simulation will allow us to compare the obtained postures to the initial posture of the model. The 3D parametric model is made with simple primitives. At each articulation of the hierarchical model, we associate a local mark whose y axe is located in the prolongation of the member. The hierarchy of the model makes the transformations spread from father to son. Our model has 13 degree of freedoms, see table 1 in the paper, that we need to know to estimate the body's pose. Once the anthropometric problem of scaling is solved, the study of the movement will be expressed in terms of parameters of rotation of the members and of global translation.

We minimize the measure of difference between the pictures of the respective view of the real scenes and the 3D ones. An example of similarity measure for video news indexation is given in [11]. Our function evaluates the correlation between the picture generated by projection of the 3D model from a given view, and the real picture of the view corresponding using colorimetric information. The matching of views acquired by several video cameras displaced in the real scene and the synthesis views allows inferring the parameters that check the model. The research of the parameters of the movement is made with a prior initialisation of the parameter near from the searched values.

Afterwards, the movement parameters obtain we will identify the people action using a fuzzy classification. To do this the obtained data will be compared to a data base to infer the model and the subject movement.

Initialisation and tracking can be helped by methods like segmentation. Our analysis is performed by segmenting character region from background region. Since datas about the scene are known, we can easily subtract background region from the initials regions.

2 The Tracking

2.1 Presentation of the System

First, we are interested in tracking a human moving in a 3D environment. The subject is filmed by several cameras which provide video sequences. The model is in a 3D scene containing OpenGL or Directx cameras (both have been tested) having the same position as the cameras of the real scene. In these conditions we can obtain the values of the parameters of movement. We will minimize this function by using genetic algorithm coupled to a simplex descent algorithm (or generalized simplex) presented by Nelder and Mead [8].

The Fig. 1 synthesizes the approach we have used, Fig. 2 show a global view of our approach, and table 1 show the degree of freedom of our model.

2.2 The Measure Used

The measure used makes it possible to quantify the apparent difference between images of an identical object, the ones being in real view and the others in

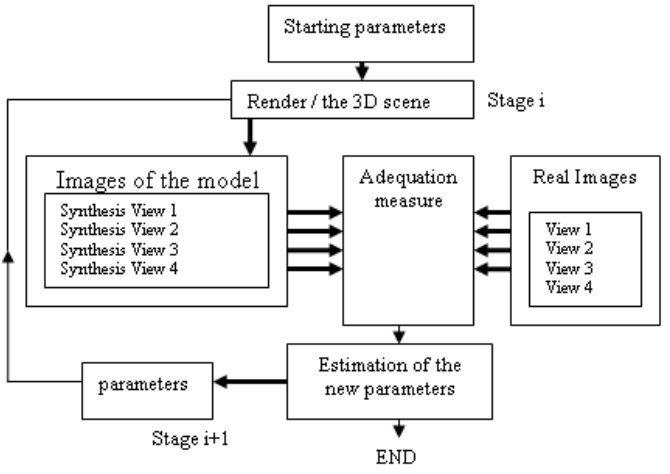


Fig. 1. Human feature extraction: parameters research relative to the basic images (real images of the subject). From the initial parameters, the syntheses view corresponding to each camera, are generated, and then compared to the real view to adjust the parameters values. This process is reiterated i-time to find the posture at time t, until it reaches the convergence criterion. The 4 cameras are calibrated and have 10 parameters: target, vertical, position in the scene, and the focal distance considered to be constant.

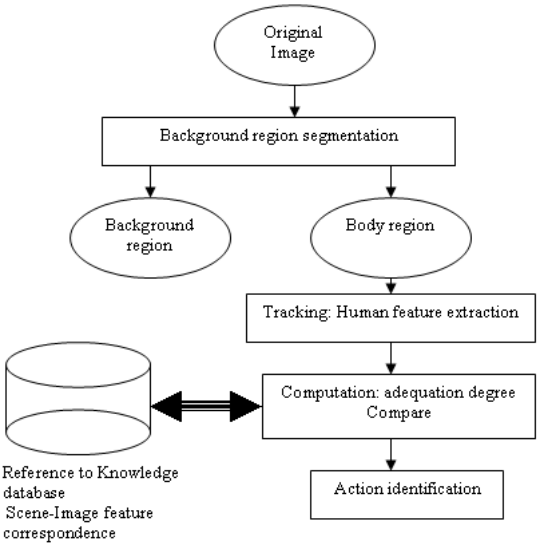


Fig. 2. Global view of our approach.

artificial view generated by projection of the real parametric model in the plan of the cameras. We want the distance to represent a minimum when the visual

Table 1. Degree of freedom of our model

organ	degree of freedom	axe	total
torso	translation	x,y,z	
left upper harm	rotation x,z	2	
right upper harm	rotation x,z	2	
left lower leg	rotation x	1	
right lower leg	rotation x	1	
left upper leg	rotation x	1	
right upper leg	rotation x	1	
left lower leg	rotation x	1	
right lower leg	rotation x	1	
	rotation x	1	
			13

agreement is maximum. Remember that the object in the synthesized images is relatively close to the real one.

The measure function is called ε and must be > 0 over $\Re$; ε must be fading when the similarity between the images increases, that mean is a correlation measurement between pair of images; ε is supposed to represent a minimum, and so $\varepsilon(I, I) = 0$.
Most of the differences between these images have an impact principally on the luminosity of the surfaces. That's why we chose a colorimetric space separating the luminosity characteristics (or luminance) from color (or chrominance). We choose the $YCrCb$ space where Y expresses the luminance, and Cr and Cb chrominance component.

We define thus two under distances, to separate the characteristics of chrominance and of luminance.

Let $c1$ and $c2$ be two colors. In the YCrCb space their values are respectively: $(y1, cr1, cb1)$ and $(y2, cr2, cb2)$.
The distance between two colors *dcoul* is:

$$d_{coul}(c1, c2) = \alpha.d_{chrom}(c1, c2) + (1 - \alpha).d_{lum}(c1, c2) \tag{1}$$

$$d_{chrom}(p1, p2) = \sqrt{(cr1 - cr2)^2 + (cb1 - cb2)^2} \tag{2}$$

$$d_{lum}(p1, p2) = |y1 - y2| \tag{3}$$

$$\varepsilon = \frac{1}{card(D)} \sum_{x=0,y=0}^{320,240} dcoul(I_1[x, y], I_2[x, y]) \tag{4}$$

To overcome brightness problems, we decrease luminance contribution by $\alpha = 2/3$. This measure is applied to each pair of real/synthesis images. These mea-

sures are next grouped together thanks to a linear combination:

$$D_m = \sum_{i=0}^{Nbcameras} \alpha_{cami}\varepsilon(I_{ireal}, I_{isynthesis}(p)) \quad (5)$$

α term is a weight parameter taking his value in the range $[0, 1]$, it lets you control the contribution of the two distances ; p is a vector of parameters ; and α_{cami} is a $camera_i$ weight, here we have $\alpha_{cami} = 1$.

To reduce the research domain around the subject, we use a bounding box so ε is reduced to the area defined by the box.

2.3 The Optimization Method

There exist a lots of minimisation algorithms which are presented in [10] [2]. In the domain of function minimisation, two important strategies exist. The first one tries to obtain a global minimum of the function to optimize, the second make local research, and can therefore lead to a local minima. It is difficult to find a global optima. Unless we have a special knowledge of the function to minimize, only a pass on all the domain parameters allows having a good solution. On the other side, local researches can stop on other local minima not expected. Usually, the functions to optimize have many local minima. To choose an optimization method, we must take into account the function's specificities that we want to minimize. In our case, we don't have any explicit expression of the derived function so we will not take into account a derivative based approach.

Genetic algorithm [3] allows us to find a solution to a problem from a set of randomly chosen element. Genetic algorithm (GA) comes from the evolution theory. We operate several operations from Genetic algorithm: Mutate, Crossover, and a Global Perturbation function... We will not detail the genetic algorithm, the interested reader can go to [3].

We adapted to our needs a robust method presented by Nelder and Mead [8] used in [9]. This method uses the principle of the simplex, adapted to non linear functions, and is named in the literature by his designers: the Nelder-Mead's method Tree operations are used by the simplex: Reflection, Contraction, and Expansion.

2.4 An Hybrid GA That Converges Faster and Avoid Simplex Problem

The main operations are: Crossover, Mutation on one element, Perturbation of a vector, and Premature Convergence. The premature Convergence is done using the simplex operation in our genetic algorithm. The permutation operation has not been taken into account in the data presented because to permute two random elements we have to verify if they have the same segment of value which is not a gain of time. Some other research has used a simplex coupled to a perturbation from genetic algorithm [9] to avoid local minima in the simplex.

In our algorithm the different vector are initialised randomly around the initial vector of parameters with the function *GenerateGeneValue* of our GA.During a mutate operation, the new parameter value is computed near the one of the vector which gives the best value for the fitness function for this we use a center and a radius. In our work we use an initialisation close to the search solution. With the genetic algorithm we can start far from this solution and the algorithm converge. The simplex method used lonely can converge to a local optima. With this combination we can do an automatic initialisation of the 3D model in the scene. Although this is not first goal of this article, we attempt to do an automatic research of the model position. We studied walking in the straight line, the subject is standing alone and we focus on finding the trunk. We use an approach based on the barycentre of the silhouette of the projection of the model and a back-projection in the 3D space in order to localize the pelvis. This method gives acceptable result but need improvement. Our algorithm stops if the number of errors is superior to the number of authorized error when the algorithm does not converge anymore.

The sample code in C♯ bellow shows the Hybrid GA operating:
1) We generate randomly N chromosomes or vector of parameters: GenerateGeneValue.
2) First use the genetic algorithm operators on the vector that fit best (GenomeMin).

```
genome Genome1 = (genome) GenomeMax.Reflect
Fitness1=CalculateErrorFitness ();...
if(Fitness1>GenomeMin.Fitness)
  genome Genome2 = (genome)
  Genome1.Dilation; Fitness2  = CalculateErrorFitness ();
  if(Fitness2 > GenomeMin.Fitness)
    Genome2.Clone((ArrayList)GenomeMax);//we save the new value
    else Genome1.Clone((ArrayList) GenomeMax);
else
    genome Genome2 = (genome) GenomeMax.Contraction;
    //GenomeMax represents the worst parameter for matching
    Fitness2  = CalculateErrorFitness ();
    if (Fitness2 > GenomeMin.Fitness)
    Genome2.Clone((ArrayList) GenomeMax);
```

3) If the GA operations don't succeed NumberOfError Time then use the simplex operation.
4)If neither the simplex operation nor genetic algorithm basic operation don't succeed after x consecutive error then apply a random perturbation to one or several vectors different from the vector that fit the best.

Table 2. Degree of freedom of our model

organ	degree of freedom	axe	total
Training set T1	Walk	Jump	Bend down
Agregation degree Pr	0.159037	0.133782	0.133782

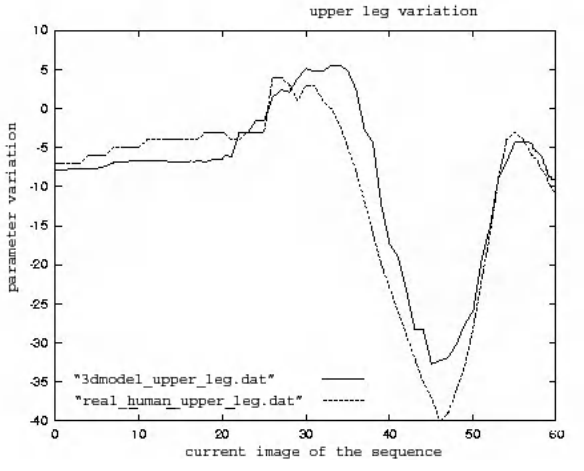


Fig. 3. Comparison of the result obtains from tracking the upper legs to the data base value.

2.5 Member's Identification

The priority images are introduced to favour the importance of a member during the research process. It let us introduce the priority notion. Priority images are constructed while attributing a different colour to every member of the 3D model. This colour picture allows us to identify each of the members of the projected subject and to attribute them a weight value during the treatment.

2.6 People Action Recognition in Image Sequences by Fuzzy Supervised Learning

For the recognition process we will compare signal movement of each member parameters during the video sequence to a database using a fuzzy classification method. The combination of the adequation degree to each class results in the people movement owing class.

In fact to test the approach, we will limit our test to 3 class identified by: $C = \{Towalk, Tojump, Tobenddown\}$.

We consider descriptors relative to the moving parameters of 3 members (torso, upper leg, lower leg) described in section 2.1 table 1.

To illustrate this method, let be one object $\overrightarrow{x}$ described by a finite and fixed number of descriptor $\{x_1, x_2, ..., x_n\}$, these descriptors are here quantitative, and a set of class $C = \{C_1, C_2, ..., C_k\}$ to make a possible confrontation between object and class. Given x_i and C_j, we compute for every x_i the value that descriptor takes over C_j and the combination of each value ρ_{ji} give us a global adequation degree Pr relative to class C_j. Pr is obtain by a basic membership function[11] [12] not detailed here due to lack of space. Pr is the possibility that one object has to belong to class C_j. During the treatment the maximum number of apparition of a class, all over the video sequence, let us assign the resulted movement. We can also use a correlation coefficient to deduce the movement.

3 Results

The genetic algorithm has been test with the hybrid version of this algorithm. Let's take the linear function $\sqrt{(\sum_{i=0}^{np} parameters_i)^2}$, were np is the number of parameters .We choose to optimize 12 parameters. The hybrid version converge to the value of 0.0113 in a mean of time about 45 second and our genetic algorithm converge in a random time between 25 and 55 sec to a value superior to 1 for a maximum of 600 errors allowed.

The combination of the two algorithms allows us to use the simplex speed of convergence and the fitness of a genetic algorithm. It allows us to avoid local minima due to the simplex method. The same observation has been lead with our 3D structure, but the times to proceed improve relative to the non linear system and the number of parameters. The Fig. 3 shows the result for the upper legs obtain from tracking.

Let's take a training set T1 among our data base where the subject is walking. The table 2 show a short execution of our algorithm.

4 Conclusion

In this article we present a system for automatic action recognition using a 3D articulated human model. We succeed in the classification of the people action done with the parameters obtained from tracking. But we should improve the method working with more complex movements.

Nevertheless other way in this work should be developed. We could recognize a subject toward several new objects in the scene. We would like to operate toward real-time tracking and indexation.

References

1. Quentin Delamarre et Olivier Faugeras, I.N.R.I.A RobotVis Project, " 3D articulated Models ", I.N.R.I.A - Projet Robot Vis, february 2000
2. Apprentissage artificiel, concepts et algorithmes, ed. EYROLLES
3. Darrell Whitley, A genetic algorithm Tutorial, Technical Report CS-93-103, March 10, 1993

4. D. Ormoneit , H. Sidenbladh, M.J. Black, T. Hastie, and D.J. Fleet, "Learning and Tracking Human Motion using Functional Analysis", proc IEEE Workshop "Human Modeling, Analysis and Synthesis ", page 2-9, June 2000
5. Shanon X. Ju, Michael J. Black, Yaser Yacoob, "Cardboard People : A parametrized Models of Articulated Image Motion",
6. M. J. Black and Y. Yacoob. Tracking and Recognizing Rigid and Non-rigid Facial Motion Using Local Parametric Models of Image Motion. ICCV, 1995, 374-381.
7. Ioannis Kakadiaris, Dimitris Metaxas, "Model-Based Estimation of 3D Human Motion", IEE transactions on patern analysis and machine Intelligence, vol. 22, no. 12, December 2000
8. J.A. Nelder and R. Mead. "A simplex method for function minimisation." The computer Journal, 7 : 308-313, Juillet 1965
9. Yannick Perret, "Suivi de parametres de modele géométrique à partir de séquences vidéo multi vues ", Univeristé Claude Bernard - Lyon1, Laboratoire d'informatique Graphique, Image et modélisation, 17 décembre 2001
10. Jean-Louis Amat, Gérard Yahiaoui, Traitement avancées pour le traitement de l'information, Réseaux de neurones, logique floue, algorithmes génétiques.
11. Reiko Hamada, Shuichi Sakai, and Hidehiko Tanaka, Scene identification in News video by character region segmentation, National Institute of Informatics
12. Sylvie Philipp-Foliguet, Marcello Bernardes Vieira, "Segmentation d'images en régions floues", Logique Floue et Applications, LFA 2000, La Rochelle, 2000.

CVPIC Compressed Domain Image Retrieval by Colour and Shape

Gerald Schaefer and Simon Lieutaud

School of Computing and Technology
The Nottingham Trent University
Nottingham, United Kingdom
gerald.schaefer@ntu.ac.uk

Abstract. Image retrieval and image compression have been pursued separately in the past. Only little research has been conducted on a synthesis of the two by allowing image retrieval to be performed directly in the compressed domain of images without the need to decode them first. In this paper we introduce a novel approach that provides such midstream content access [6]. Our work is based on the Colour Visual Pattern Image Coding (CVPIC) technique which represents a compression algorithm where the data in compressed form is directly visually meaningful. We present a compressed domain retrieval algorithm based on CVPIC that efficiently encapsulates the image content by colour and shape features. Retrieval results on the UCID dataset show good retrieval performance, outperforming methods such as colour histograms, colour coherence vectors, and colour correlograms.

Keywords: CBIR, midstream content access, compressed domain image retrieval, colour visual pattern image coding, CVPIC

1 Introduction

As computers become increasingly powerful, our expectations and desires grows rapidly. Machines are now not only used for complex computations as in the beginnings but are becoming the ultimate repository for all kinds of information. We are living in the decades of digital revolution. Virtually every piece of information is being transformed into digital data, thanks to the increasing availability of devices such as cameras, scanners and others. Motion picture is being transmitted through satellite links in digital form and often displayed on digital devices such as LCDs or projectors. Movies are being made using digital video cameras or are entirely rendered by computers. We can easily access any sequence in a movie and capture a frame because of the increasing availability of DVD and other technologies. It is undeniable that the amount of images available in the digital world has exceeded the boldest expectations of the past. It is equally undeniable that we are overwhelmed by the amount of information and no longer able to find anything useful in the endless ocean of digital imagery.

Therefore, it is obvious that effective content-based image retrieval (CBIR) techniques are desperately needed. Fortunately, this problem has been subject

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 778–786, 2004.
© Springer-Verlag Berlin Heidelberg 2004

of extensive research for over a decade starting with Swain and Ballard's pioneering work on colour indexing [12] which showed that low level image features can be exploited for image retrieval purposes. Although colour information is very important for object recognition and image retrieval, it is often not sufficient, especially if extracted globally for a whole image. This was soon realised and techniques which also address texture and shape properties investigated. Although these methods are usually not as effective as colour-based algorithms, incorporating several feature types provides improved performance.

While many CBIR methods have been suggested in the literature only few take into account the fact that - due to limited resources such as disk space and bandwidth - virtually all images are stored in compressed form. In order to process them for CBIR they first need to be uncompressed and the features calculated in the pixel domain. Often these features are stored alongside the images which is counterintuitive to the original need for compression. The desire for techniques that operate directly in the compressed domain providing, so-called midstream content access, seems therefore evident [6].

Colour Visual Pattern Image Coding (CVPIC) is one of the first so-called 4-th criterion image compression algorithms [9,8]. A 4-th criterion algorithm allows - in addition to the classic three image coding criteria of image quality, efficiency, and bitrate - the image data to be queried and processed directly in its compressed form; in other words the image data is directly meaningful without the requirement of a decoding step. The data that is readily available in CVPIC compressed images is the colour information of each of the 4×4 blocks the image has been divided into, and information on the spatial characteristics of each block, in particular on whether a given block is identified as a uniform block (a block with no or little variation) or a pattern block (a block where an edge or gradient has been detected). Furthermore, each pattern block is assigned to one of 14 universally predefined classes according to the orientation and position of the edge within the block.

In this paper we make direct use of this information and propose an image retrieval algorithm that utilises both colour and shape information. The colour information is summarised similar to colour coherence vectors introduced in [5] and the border/interior pixel approach in [11] which both show that dividing the pixels of an image into those that are part of a uniform area and those that are not can improve retrieval performance. In essence we create two colour histograms, one for uniform blocks and one for pattern blocks. For the shape information we exploit the fact that edge information is directly encoded in CVPIC and create an edge histogram. Integrating the three types of histograms allows for image retrieval based on (spatial) colour and shape features. Experimental results obtained from querying the UCID [10] dataset show that our approach not only allows retrieval directly in the compressed domain but that it also clearly outperforms popular techniques such as colour histograms, colour coherence vectors and colour correlograms.

The rest of this paper is organised as follows: in Section 2 the CVPIC compression algorithm used in this paper is reviewed. Section 3 describes our novel

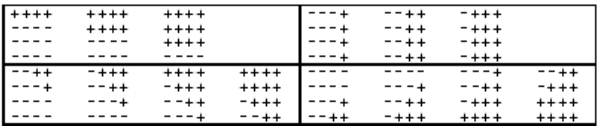


Fig. 1. The 14 edge patterns used in CVPIC

method of image retrieval in the CVPIC domain while Section 4 presents experimental results. Section 5 concludes the paper.

2 Colour Visual Pattern Image Coding

The Colour Visual Pattern Image Coding (CVPIC) image compression algorithm introduced by Schaefer et al. [9] is an extension of the work by Chen and Bovic [1]. The underlying idea is that within a 4×4 image block only one discontinuity is visually perceptible.

CVPIC first performs a conversion to the CIEL*a*b* colour space [2] as a more appropriate image representation. As many other colour spaces, CIEL*a*b* comprises one luminance and two chrominance channels; CIEL*a*b* however, was designed to be a uniform representation, meaning that equal differences in the colour space correspond to equal perceptual differences. A quantitative measurement of these colour differences was defined using the Euclidean distance in the L*a*b* space and is given in ΔE units.

A set of 14 patterns of 4×4 pixels has been defined in [1]. All these patterns contain one edge at various orientations (vertical, horizontal, plus and minus 45°) as can be seen in Figure 1 where + and - represent different intensities. In addition a uniform pattern where all intensities are equal is being used.

The image is divided into 4x4 pixel blocks. Determining which visual pattern represents each block most accurately then follows. For each of the visual patterns the average L*a*b* values μ_+ and μ_- for the regions marked by + and - respectively (i.e. the mean values for the regions on each side of the pattern) are calculated.

The colour difference of each actual pixel and the corresponding mean value is obtained and averaged over the block according to

$$\epsilon = \frac{\sum_{i \in +} \|p_i - \mu_+\| + \sum_{j \in -} \|p_j - \mu_-\|}{16} \tag{1}$$

The visual pattern leading to the lowest ϵ value (given in CIEL*a*b* ΔE units) is then chosen. In order to allow for the encoding of uniform blocks the average colour difference to the mean colour of the block is also determined according to

$$\sigma = \frac{\sum_{\forall i} \|p_i - \mu\|}{16} \text{ where } \mu = \frac{\sum_{\forall i} p_i}{16} \tag{2}$$

A block is coded as uniform if either its variance in colour is very low, or if the resulting image quality will not suffer severely if it is coded as a uniform rather than as an edge block. To meet this requirement two thresholds are defined. The first threshold describes the upper bound for variations within a block, i.e. the average colour difference to the mean colour of the block. Every block with a variance below this value will be encoded as uniform. The second threshold is related to the difference between the average colour variation within a block and the average colour difference that would result if the block were coded as a pattern block (i.e. the lowest variance possible for an edge block) which is calculated by

$$\delta = \sigma - min_{\forall patterns}(\epsilon) \tag{3}$$

If this difference is very low (or if the variance for a uniform pattern is below those of all edge patterns in which case σ is negative) coding the block as uniform will not introduce distortions much more perceptible than if the block is coded as a pattern block. Hence, a block is coded as a uniform block if either σ or δ fall below the thresholds of 1.75 ΔE and 1.25 ΔE respectively (which we adopted from [9]).

For each block, one bit is stored which states whether the block is uniform or a pattern block. In addition, for edge blocks an index identifying the visual pattern needs to be stored. Following this procedure results in a representation of each block as 5 bits (1 + 4 as we use 14 patterns) for an edge block and 1 bit for a uniform block describing the spatial component, and the full colour information for one or two colours (for uniform and pattern blocks respectively).

In contrast to [9] where each image is colour quantised individually, the colour components are quantised to 64 universally pre-defined colours (we adopted those of [7]). Each colour can hence be encoded using 6 bits. Therefore, in total a uniform block takes 7 (= 1 + 6) bits, whereas a pattern block is stored in 17 (=5+2∗6) bits. We found that this yielded an average compression ratio of about 1:30. We note, that the information could be further encoded to achieve lower bitrates. Both the pattern and the colour information could be entropy coded. In this paper however, we refrain from this step as we are primarily interested in a synthesis of coding and retrieval.

3 CVPIC Image Retrieval

We note from above that for each image block in CVPIC both colour and edge information is readily available in the compressed form: each block is coded either as a uniform block or as a pattern block. While for a uniform block only its colour needs to be stored, each pattern block contains two colours and belongs to one of 14 edge classes. We make direct use of this information for the purpose of image retrieval.

It is well known that colour is an important cue for image retrieval. In fact, simple descriptors such as histograms of the colour contents of images [12] have been shown to work well and have hence been used in many CBIR systems. Further improvements can be gained by incorporating spatial information as

techniques such as colour coherence vectors [5] and border/interior pixel histograms [11] have shown. Here the colour information is not summarised in one histogram but is represented in two separate histograms: one histogram of coherent pixels (i.e. pixels in uniform areas) and one histogram of scattered pixels for the colour coherence vector approach respectively one histogram of border pixels (i.e. those part of an edge) and one histogram of interior pixels in the border/interior pixel histogram technique. Our approach is fairly similar to these techniques but requires no explicit computation that provides the classification into the two categories. Rather we utilise the (pre-calculated) division into uniform and pattern blocks. Pixels that are part of a uniform area (i.e. 'coherent' or 'interior' pixels) will more likely be contained within a uniform block. On the other hand pixels that form part of an edge (i.e. 'border' pixels) will fall into pattern blocks. We can therefore immediately distinguish between these two types of pixels without any further calculation (as would need to be done for colour coherence vector or border/interior pixel calculation). We hence create two colour histograms: a uniform histogram H^u by considering only uniform blocks and a non-uniform histogram H^n calculated solely from edge blocks. While exact histograms could be calculated by simply adding the appropriate number of pixels to the relevant colour bins while scanning through the image we suggest a simpler, less computationally intensive, method. Instead of weighing the histogram increments by the relative pixel proportions we simply increment the affected colour bins (two for an edge block, one for a uniform block) by 1[1]. We also wish to point out that the resulting histograms are *not* normalised as is often the case with histogram based descriptors. The reason for this is that by not normalising we preserve the original ratio between uniform and pattern blocks - an image feature that should prove important for distinguishing between images with a similar colour content.

Having calculated H^u and H^n two CVPIC images can be compared by calculating a weighted sum of the L_1 norm between their histograms

$$d_{\text{colour}}(I_1, I_2) = \alpha \sum_{k=1}^{N} |H_1^u(k) - H_2^u(k)| + (1-\alpha) \sum_{k=1}^{N} |H_1^n(k) - H_2^n(k)| \quad (4)$$

where α can be set so as to put more or less emphasis on either of the two histograms. We set $\alpha = 0.5$, i.e. weigh the two histograms equally.

While image retrieval based on colour usually produces useful results, integration of this information with another paradigm such as texture or shape will result in an improved retrieval performance. Shape descriptors are often calculated as statistical summaries of local edge information such as in [4] where the edge orientation and magnitude is determined at each pixel location and an edge histogram calculated. Exploiting the CVPIC image structure an effective shape descriptor can be determined very efficiently. Since each (pattern) block contains exactly one (pre-calculated) edge and there are 14 different patterns we simply build 1×14 histogram of the edge indices[2]. We decided not to include a bin for

[1] We note that this puts more emphasis on the non-uniform histogram than on the uniform one.

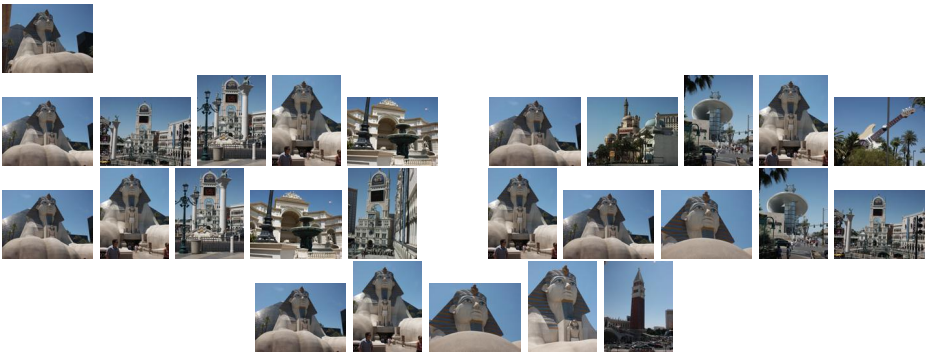


Fig. 2. Sample query together with 5 top ranked images returned by (from left to right, top to bottom) colour histograms, colour coherence vectors, border/interior pixel histograms, colour correlograms, CVPIC retrieval.

uniform blocks, since these give little indication of shape (rather they describe the absence of it). Edge histograms H_1^s and H_2^s are compared using

$$d_{\text{shape}}(I_1, I_2) = \sum_{k=1}^{N} |H_1^s(k) - H_2^s(k)| \tag{5}$$

Having calculated d_{colour} and d_{shape} for two images these two scores can now be combined in order to allow for image retrieval based on both colour and shape features which results in

$$d(I_1, I_2) = \alpha \sum_{k=1}^{N} |H_1^u(k) - H_2^u(k)| + \beta \sum_{k=1}^{N} |H_1^n(k) - H_2^n(k)| + \tag{6}$$
$$+ (1 - \alpha - \beta) \sum_{k=1}^{N} |H_1^s(k) - H_2^s(k)|$$

Again, the weights α and β can be adjusted so as to make either of the two colour features or the shape descriptor more dominant. In our experiment we opted for equal weights between colour and shape features and equal weights between uniform and non-uniform colour histograms, i.e. $\alpha = \beta = 0.25$.

4 Experimental Results

We evaluated our method using the recently released UCID dataset [10]. UCID, an Uncompressed Colour Image Database[3], consists of 1338 colour images all

[2] Again, no normalisation is applied. The edge histogram hence adds up to half the non-uniform colour histogram.

[3] UCID is available from `http://vision.doc.ntu.ac.uk/`.

Table 1. Results obtained on the UCID dataset.

	AMP
Colour histograms	90.47
Colour coherence vectors	91.03
Border/interior pixel histograms	91.27
Colour correlograms	89.96
CVPIC $\alpha = \beta = 0.25$	**94.24**

preserved in their uncompressed form which makes it ideal for the testing of compressed domain techniques. UCID also provides a ground truth of 262 assigned query images each with a number of predefined corresponding matches that an ideal image retrieval system would return.

We compressed the database using the CVPIC coding technique and performed image retrieval using the algorithm detailed in Section 3 based on the queries defined in the UCID set. As performance measure we use the modified average match percentile (AMP) from [10] defined as

$$\mathrm{MP}_Q = \frac{100}{S_Q} \sum_{i=1}^{S_Q} \frac{N - R_i}{N - i} \tag{7}$$

with $R_i < R_{i+1}$ and

$$\mathrm{AMP} = \frac{1}{Q} \sum \mathrm{MP}_Q \tag{8}$$

where R_i is the rank the i-th match to query image Q was returned, S_Q is the number of corresponding matches for Q, and N is the total number of images in the database.

In order to relate the results obtained we also implemented colour histogram based image retrieval ($8 \times 8 \times 8$ RGB histograms) according to [12], colour coherence vectors [5], border/interior pixel histograms [11] and colour auto correlograms [3]. Results for all methods can be found in Table 1. From there we can see that our novel approach is not only capable of achieving good retrieval performance, but that it actually clearly outperforms all other methods. While the border/interior pixel approach achieves an AMP of 91.27 and all other methods perform worse, CVPIC colour/shape histograms provide an average match percentile of 94.24, that is almost 3.0 higher than the best of the other methods. This is indeed a significant difference as a drop in match percentile of 3 will mean that 3% more of the whole image database need to be returned in order to find the images that are relevant; as typical image database nowadays can contain tens of thousands to hundreds of thousands images this would literally mean additionally thousands of images. The superiority of the CVPIC approach is especially remarkable so as it is based on images compressed to a medium compression ratio, i.e. images with a significantly lower image quality that uncompressed images whereas for all other methods the original uncompressed versions of the images were used. Furthermore, methods such as colour histograms,

colour coherence vectors and colour correlograms are known to work fairly well for image retrieval and are hence among those techniques that are widely used in this field. This is further illustrated in Figure 2 which shows one of the query images of the UCID database together with the five top ranked images returned by all methods. Only the CVPIC technique manages to retrieve four correct model images in the top 5 (with the next model coming up in sixth place) while colour correlograms retrieve three and all other methods only two.

5 Conclusions

In this paper we present a novel image retrieval technique that operates directly in the compressed domain of CVPIC compressed images. By exploiting the fact that CVPIC encodes both colour and edge information these features can be directly used for image retrieval. Two types of histograms are built: two colour histograms (one of uniform areas and one of edge areas) and one shape histogram. Both histograms are compared using histogram intersection and the resulting scores weighted to provide an overall similarity between two images. Experimental results on a medium-sized colour image database show that the suggested method performs well, outperforming techniques such as colour histograms, colour coherence vectors, and colour correlograms.

Acknowledgements. This work was supported by the Nuffield Foundation under grant NAL/00703/G.

References

1. D. Chen and A. Bovik. Visual pattern image coding. *IEEE Trans. Communications*, 38:2137–2146, 1990.
2. CIE. *Colorimetry.* CIE Publications 15.2, Commission International de L'Eclairage, 2nd edition, 1986.
3. J. Huang, S.R. Kumar, M. Mitra, W-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *IEEE Int. Conference Computer Vision and Pattern Recognition*, pages 762–768, 1997.
4. A.K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996.
5. G. Pass and R. Zabih. Histogram refinement for content-based image retrieval. In *3rd IEEE Workshop on Applications of Computer Vision*, pages 96–102, 1996.
6. R.W. Picard. Content access for image/video coding: The fourth criterion. Technical Report 195, MIT Media Lab, 1994.
7. G. Qiu. Colour image indexing using BTC. *IEEE Trans. Image Processing*, 12(1):93–101, 2003.
8. G. Schaefer and G. Qiu. Midstream content access based on colour visual pattern coding. In *Storage and Retrieval for Image and Video Databases VIII*, volume 3972 of *Proceedings of SPIE*, pages 284–292, 2000.
9. G. Schaefer, G. Qiu, and M.R. Luo. Visual pattern based colour image compression. In *Visual Communication and Image Processing 1999*, volume 3653 of *Proceedings of SPIE*, pages 989–997, 1999.

10. G. Schaefer and M. Stich. UCID - An Uncompressed Colour Image Database. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307 of *Proceedings of SPIE*, pages 472–480, 2004.
11. R.O. Stehling, M.A. Nascimento, and A.X. Falcao. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proc. 11th Int. Conf. on Information and Knowledge Management*, pages 102–109, 2002.
12. M.J. Swain and D.H. Ballard. Color indexing. *Int. Journal Computer Vision*, 7(11):11–32, 1991.

Automating GIS Image Retrieval Based on MCM

Adel Hafiane and Bertrand Zavidovique

Institut d'Electronique Fondamentale
Bat 220, Université Paris XI 91405 Orsay France
adel.hafiane@ief.u-psud.fr

Abstract. This paper describes the automation of a CBIR method based on image description by "motif co-occurrence matrix" (MCM). Motifs in that case refer to an optimal Peano scan of the picture. A prior segmentation into regions based on MCM by blocks of adapted size is shown to compete with the human region finding. Retrieval from the regions' MCM vector becomes then comparable for man and machine. Results on precision and recall support method comparison.

1 Introduction

Fast growth of geographic image-bases requires tools for efficiently manipulating and searching visual data. Features extraction is a crucial part for content based image retrieval (CBIR). Color, texture and shape are most used low level characteristics in CBIR [1]. Current advanced systems as QBIC, Virage and Photobook [2, 3, 4] tend to combine several, subject to the application, for more efficient CBIR. Aerial images contain different regions tied to various land properties, so texture proves to be an important visual primitive in that case, for both searching and browsing. For instance, works by Ma and Manjunath [5] show the effectiveness of texture in aerial photographs browsing. On an other hand, query by global image-features is limited in cases where there is no dominant semantics (e.g. no peculiar obvious object or situation to look for). Then some systems such as by Netra [6] and Blobword [7] segment images into meaningful regions to improve CBIR. They require users to point out a set of regions of interest of which the feature-vector is compared over the database. Yet the method leaves any context of the selected region in the user's mind. That is all the more penalizing as visual features alone cannot distinguish between similar images with different semantics. Spatial relationship provides for some basic image meaning. It gives access to relative positions of different image components with respect to one another. Several methods have been proposed: some key ones are 2D-string [8], R-strings based on directional relations [9], and the topological extension proposed in [10] aiming to increased robustness. Petrakis et al. [11] use the Attributed Relational Graph (ARG) that demonstrates better *precision* and *recall* performance in medical images.

In this paper we stress upon automating a new method of segmentation-based CBIR. The database images are segmented into several regions. In a first version was done by humans and a texture index named "Motifs Co occurrence Matrix" (MCM) was computed for each and every region. In a second version it is automatic based

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 787–794, 2004.
© Springer-Verlag Berlin Heidelberg 2004

again on MCM combined with Fuzzy C-means, and likewise the MCM by region is computed. MCM are bound to optimal Peano scans made of a sequence of 2x2 pixel grids over the image. That supports improving the search effectiveness by optimally scanning the image [12] and again by decomposing each image into a set of relevant segments. Images are represented by ARGs which nodes are attributed the region features –here MCM – and edges relate spatial relations. Thus, the distance between images is measured from both visual similarity and spatial relationship between pairs of regions when precision is required. Both main phases rely on MCM:

- Image segmentation: digital images are split based on MCM dissimilarity, and image regions are featured by their respective MCM.
- Query and retrieval: ARG is used to measure similarity by graph matching.

That is the reason why it appeared likely beneficial to automatically index images in view of retrieval in accordance.

Feature extraction and segmentation are discussed in section 2. Retrieval is developed in section 3. In section 4 we discuss results and compare retrieval efficiency by the same method but after human and machine based segmentation. The paper concludes in section 5.

2 Regions Extraction

The segmentation is a key step in image retrieval. We focus on satellite images where the texture is considered the main feature for representation. Our segmentation method utilizes MCMs to characterize textures combined with Fuzzy C-means to find regions.

2.1 Motifs Co-occurrence Matrix

The Peano space filling curves are used to traverse pixels with a specific local path made of the 6 primitive scans represented in Figure 1.

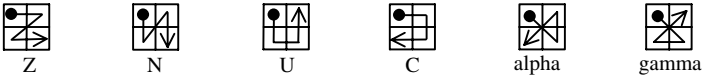


Fig. 1. Scan motifs to traverse a 2x2 grid

The image is divided into 2x2 pixel grids and each four pixels are traversed by Peano scans. One of 6 motifs represents the 2x2 pixel grid optimally wrt. a suitable criterion, upbringing the corresponding grid motif. For instance in the present application, the relevant motifs minimize the local intensity variation along the scan line.

Given $\begin{matrix} p_1 & p_2 \\ p_3 & p_4 \end{matrix}$, the optimal scan follows the permutation * that corresponds to

$$\textbf{\textit{min}} \; (|p_1 - p_2| + |p_2 - p_3| + |p_3 - p_4|)$$

Therefore the method fits local texture-property extraction. (Note that the image-result size is N/2 x N/2 from an original N x N). The Peano curves code the relation

between four neighbour pixels but it does not bring any more information about visual characteristics. One needs to consider the motifs spatial distribution, hence the motifs co-occurrence matrix [13] indicative of texture.

The way it was described above, the method would be very sensitive to translation. A shifted image is likely to be very different from the original one MCM-wise. A translation by one pixel causes the 2x2 neighborhood likely to vary significantly, as well as the optimal scan subjected to the intensity local variation. To compensate for translation effects we construct four MCM feature vectors, including shifted versions of the original image by one pixel horizontally, vertically and diagonally. Among four feature vectors per region, one of them would correspond to the query vector independent of the amount of translation, because of the motifs periodicity under translations by more than one pixel.

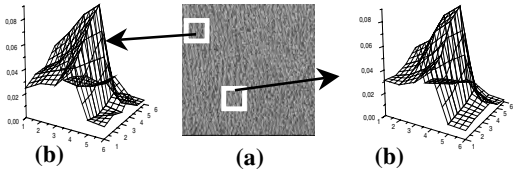


Fig. 2. a) Homogeneous texture image **b)** MCM value by rows and columns

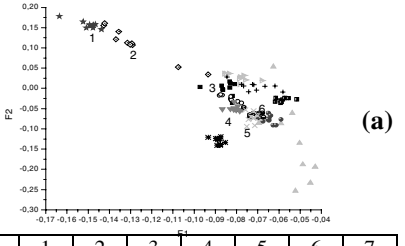
Figure 2 displays MCMs in highlighted windows. Considering that texture remains the same over the whole picture, respective MCMs in different windows should be similar. About 30 images randomly picked from Brodatz's have been tested against that conjecture and for different window sizes: MCM surfaces keep the same form and low L_2 inter-distance for textures with same visual aspect. Note that the significance of the motifs co-occurrence depends on the window size, through the texture granularity: relatively too small a window may not capture sufficient information. For instance, over the 300 aerial test images considered for retrieval, a 64x64 bloc size appears best suitable in the average, likely due to the coarseness distribution of textures.

To better investigate the separability between different texture classes through the MCM representation, a Principal Component Analysis (PCA) is needed since an MCM is a high – 36 – dimensional vector. Figure 3-a-b represent scatter diagrams of the 13 texture classes extracted from the Brodatz's album for study. 4 blocks are chosen randomly from each image and from its equalized histogram version. The MCM is computed on each block and on the images, resulting in a classification of 130 random samples into 13 different classes.

Classes are correctly distinct in 2D already (Fig 3-e), and almost fully separate in 3D. Increasing the block size tends to favor the class compaction, except for it hampers small regions or detail detection. Eventually, MCM proves efficient enough for capturing properties of this type of texture that will be confirmed by segmentation.

2.2 Segmentation

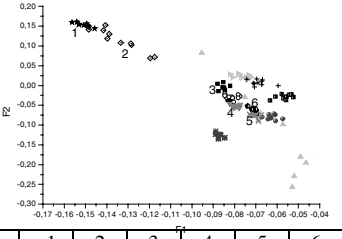
Targeting automatic extraction of meaningful regions, our method consists in partitioning the image into a grid of blocks. The MCMs per block are computed as



(a)

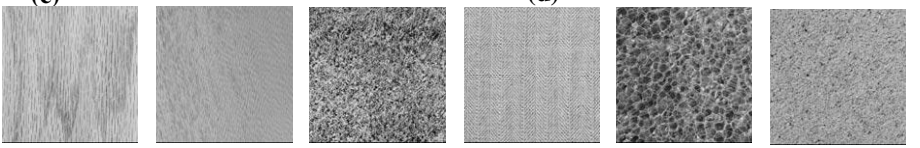
F	1	2	3	4	5	6	7
%	69	21,3	7,1	0,9	0,7	0,2	0,1

(b)



F	1	2	3	4	5	6	7
%	69,5	22,1	6,4	2,1	0,08	0,04	0,3

(c) (d)



1 2 3 4 5 6

(e)

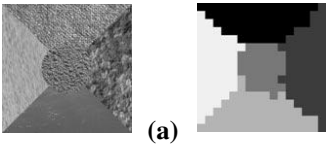
Fig. 3. a) 13x10 64x64 block-samples of textures grouped into 13 classes in the 2D MCM principal sub-space ; **b)** same for 128x128 blocks ; **c)** the 7 larger eigen-values in % for case **a** ; **d)** same for case **b** **e)** samples of texture classes as labeled in **a)** and **b)**

texture clue. Then similar blocks are classified into the same cluster using Fuzzy c-means (FCM) [14], that is an unsupervised method allowing subsets to be classified with a membership degree. As described in section 2.1 possible phase shifts are accounted for. The segmentation process considers the closest MCM out of the four computed.

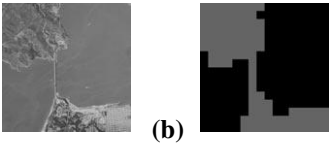
While the Euclidian distance is widely used, we found that other distances could bring more performance to calculate the distance of the feature vector x_j to the cluster centre $_i$: here we use the Bhattacharyya distance that provides better results.

$$d_{ij} = -\log \sum_p y_i(p) x_j(p) \tag{1}$$

Figure 4 shows segmentation results. The image in (a) made of Brodatz's textures is partitioned into a grid with blocks of 32x32 pixels. The number of classes is set to C=5, the fuzziness factor is set to m =3 and the difference between successive instances of the fuzzy partition matrix is set to ε = 0.001 for the termination criterion. In (b) the aerial photo, taken from the Brodatz's database again, is also divided into 32x32 pixel blocks. The cluster number is set to 2, m =3 and ε = 0.001 : the algorithm converges in 30 iterations. In this simple example, no reassignment of small regions was necessary, only isolated pixels had to be merged with the closest region.



(a)



(b)

Fig. 4. a) Segmentation result on a textured image **b)** segmentation result of an aerial image

3 Retrieval

As should appear in section 4, the above segmentation results into an image partition that improves the retrieval performance compared to the global method proposed in [13]. We focus on region-based retrieval; MCM is re-computed over each region as the feature vector. For more semantic and precision the spatial relationship between regions is introduced in the query process. Relative inter-region orientations are coded by the angle θ between the inter-centers segment (of length d) and the horizontal. Features MCM and spatial relationship (θ,d) in image P are jointly displayed into an ARG $GP(V,E)$, respectively in nodes $(V_1,V_2,...)$ and edges $(E_1,E_2,...)$. However, efficient ARG matching is an open problem which solutions are often tailored to a specific application. We use here a mere weighted Euclidian distance between nodes and edges, except it is arranged in two steps. Retrieval similarity is first evaluated on visual features, and then relationship is computed on couples of regions that output higher texture similarity. Note that several query options can be supported in this structure: query by region, by vector of regions and by organized regions with spatial relationship.The distance between nodes is defined by:

$$\mathfrak{D}r(R_i,R_j^{'})=\alpha_i\sum_p(V_i(p)-V_j^{'}(p))^2<th \tag{2}$$

and later subjected to threshold th. V_i, $V_j^{'}$ are the respective MCM of regions R_i and $R_j^{'}$. The weight is set to:

$$\alpha_i=\frac{\min(\omega_i,\omega_j^{'})}{\sum_i^M\omega_i}$$

with ω_i (resp ω'_j) the region surfaces. MCMs need to be normalized too before calculating any distance, for the region size not to bias the comparison that much.

Let denote by Q the query image and DB an image in the data base: Q and DB contain respectively M and N segments. Nodes are first compared through (3) resulting into m nodes in GQ that match m nodes in GDB ($m\leq min\{M,N\}$).

$$\mathfrak{D}(Q,DB)=\frac{1}{m}\sum_{i=1}^{M}\min_{j\in[1,N]}\left\{\mathfrak{D}r(R_i,R_j^{'})\bullet 1_{\{(i,j)/\mathfrak{D}r(R_i,R_j^{'})<th\}}\right\} \tag{3}$$

with $1_X(i,j)$ the set function of X : $1_X(i,j)=\begin{cases}1 & iff\ (i,j)\in X\\ 0 & else\end{cases}$

A more complete similarity measure including spatial constraints is then computed on the sub-graphs of matched nodes. The spatial similarity is measured by a weighted distance again, Gm:

$$Gm=\mathfrak{D}(Q,DB)+\frac{1}{m\beta_i}\sum\left|E_i-E_j\right| \tag{4}$$

with E_i and E_j the respective edges between nodes of corresponding pairs in Q and DB. Unlike α,β weigh more the larger regions in Q : $\beta_i=\frac{\omega_i}{\sum_i^M\omega_i}$

4 Experiments and Results

The experiments are completed on a set of 300 aerial images from *http://terraserver.microsoft.com* augmented with 30 images from the *Brodatz* data base http://sipi.usc.edu/services/database. In a previous series of tests the method was evaluated against the histogram technique, it proves better *precision* and *recall* performance in all tested cases (e.g. Figure 5-c). The present series of experiments aims at comparing retrieval results after human and automatic segmentations respectively. The feature representation for retrieval is MCMs in both cases. Figure 5-b shows an example of obtained curves "precision = *f*(recall)" for the set (group) of pictures Figure 6. Images were sorted into relevant groups by human experts (see Fig. 6 for an example of group). It is to be noted that these groups fit the geographic locations quite accurately. Groups contain 6 to 10 images. If a retrieved image belongs to the same group as the query image it is considered a "best match". In the sequel, "relevant" means that images represent similar parts of the same known geographic area (e.g. from their tag in the data base) and/or represent similar types of regions (e.g. country or urban environment) as being judged by experts or simply by any human being. It is a subjective evaluation but it was done prior to any experiment on the images of the data set. Let us underline that a proper framework for evaluation of the relevance of answers to queries is still to be found and is an open research problem. We use the common measures extensively used in the field called *recall* and *precision*.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \quad ; \quad \text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}}$$

The figure 5-a. shows the histogram of the normalized difference of surfaces between curves resulting from the human-segmentation and the machine-segmentation based retrieval respectively. The block size for machine segmentation is set to 64x64 pixels (32x32 motifs). Positive values indicate the program to perform better than the human expert and conversely for negative ones.

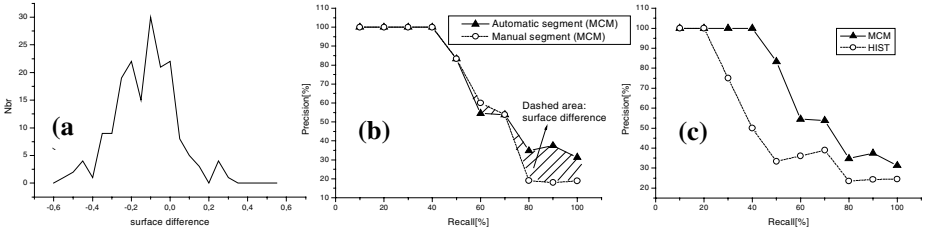


Fig. 5. a) histogram of the difference human/machine records in retrieving satellite images from regions **b)** precision = *f*(recall) after segmentation by man (-o-) and machine (▲) respectively **c)** same plot for Histogram and MCM based retrieval by the machine after automatic segmentation

The histogram is fairly Gaussian although more accurate test of populations (–0,05, and 0,05 or –0,15) could indicate a bi-gaussian distribution. Most differences belong to [–0,25,+ 0,05]. When checking the types of images where one or the other performs definitely better, it appears that, independent from other factors, the human segmentation makes the difference. It sticks to details exactly where it is required while otherwise keeping a low number of regions. Conversely when the image is coarse enough the programm may be more efficient. The automatic segmentation based on blocks with given size cannot adapt the trade of between texture evaluation and edge marking. Some small regions may vanish or some parts of regions be ill-classified that are important to retrieval. To confirm this hypothesis the segmentation was further tested with block sizes of 16x16 and 32x32. It can be observed as a general result that for all images that require accurate segmentation in some part (Figure7), increasing the number of blocks improves the retrieval up to making it sufficiently close to the human level although never catching back.

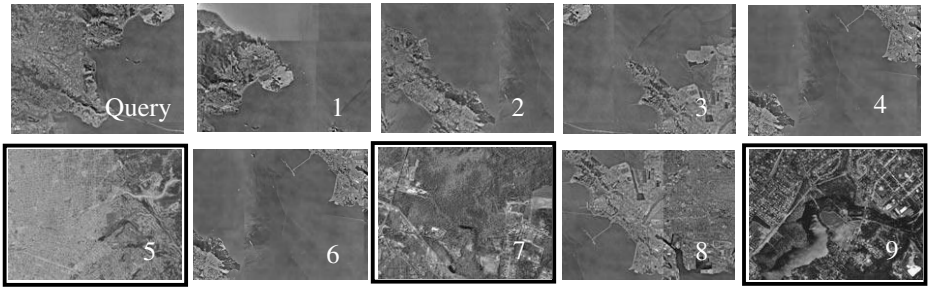


Fig. 6. Example of a group of images (California) used for retrieval test. The top left image is the query and others are displayed according to their retrieval ranking. Pictures 5, 7 and 9 are outliers that do not belong to the group. In this case the machine wins (see Figure 5-b) by 1,5%

Fig. 7. Example of images where human retrieval is far better (25%). From left to right : the original image, the human segmentation, the machine segmentation with 64x64 blocks, same with 32x32 blocks. The improvement in segmentation is enough that the difference gets down to 9%

5 Conclusion

The main contribution of the present paper is to show that, due to genuine properties of the Peano-codes, retrieval is greatly improved in stringing this reasonable set of simple methods : 1) image representation by regions (ARG) featured by their texture obtained by mere fuzzy classification. 2) texture description from the co-occurrence of optimal motifs (primitives of a Peano scan) 3) retrieval based on a distance

between the vectors of the same Co-occurrence matrices per region. In comparing the segmentation results and then retrieval results by expert men and machines over 350 images it was proven that the method can be made fully automatic in the satellite images case, provided the block size for texture characterization be adapted, that was done. Further work deals with applying the same automatic method to images of road landscapes to contribute to vehicle autonomy.

References

1. Rui, Y., Huang, T.S., Chang, S.F.: Image Retrieval: Current Techniques, Promising Directions and Open Issues. Journal of Visual Communication and Image Representation, Vol. 10. (1999) 39-62.
2. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system. IEEE Computer, vol. 28. (1995) 23-32.
3. Bach, J. R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R.: The Virage Image Search Engine: An Open framework for Image Image Management. Proc. Storage and Retrieval for Still Image and Video Databases. SPIE, vol. 2670. (1996) 76-87.
4. Pentland, A. P., Picard R., Sclaroff, S.: Photobook: Content-based manipulation of image databases. Int. Journal of Computer Vision, vol. 18. no. 3. (1996) 233-254.
5. Ma, W.Y., Manjunath, B. S.: A texture thesaurus for browsing large aerial photographs. Journal of the American Society for Information Science, Wiley for ASIS, vol. 49. no.7. (1998) 633-48.
6. Ma, W.Y., Manjunath, B.S.: Netra: a toolbox for navigating large image databases. Multimedia Systems, vol 7. no 3. Springer-Verlag, Berlin, Germany.(1999) 184–198.
7. Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., Malik, J.: Blobworld: a system for region-based image indexing and retrieval. Proceedings of the Third International Conference on Visual Information Systems. (1999) 509–516.
8. Chang, S.K., Shi, Q.Y., Yan, C.W.: Iconic indexing by 2-D strings. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9. no. 3. (1987) 413-428.
9. Gudivada, V.N., Raghavan, V.V.: Design and evaluation of algorithms for image retrieval by spatial similarity. ACM Transactions on Information systems, vol. 13. no. 2. (1995) 115-144.
10. El-Kwae, E.A., Kabuka, M.: A robust framework for content-based retrieval by spatial similarity in image databases. ACM Transaction on Information Systems, vol. 17. no. 2. (1999) 174 -198.
11. Petrakis, Euripides.G.M., Faloutsos, C.: Similarity Searching in Medical Image Databases. IEEE Transactions on Knowledge and Data Engineering, vol.9. no 3. (1997) 435-447.
12. Seetharaman, G., Zavidovique, B.: Image processing in a tree of Peano coded images. IEEE-CAMP 97 Boston. (1997) 229-234.
13. Jhanwar, N., Chaudhuri, S., Seetharaman, G., Zavidovique, B.: Content Based Image Retrieval Using Motif Cooccurence Matrix. Proc. 3rd Indian Conference on Computer Vision Graphics and Image Processing. Ahmedabad, India (2002).
14. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algoritms. Plenum Press, New York (1981).

Significant Perceptual Regions by Active-Nets

David García-Pérez[1], Antonio Mosquera[1],
Marcos Ortega[2], and Manuel G. Penedo[2]

[1] Grupo de Visión Artificial, Departamento de Electrónica y Computación
Universidad de Santiago de Compostela, Spain
davidgp@usc.es, mosquera@dec.usc.es
[2] VARPA Group, Departamento de Computación,
Universidad da Coruña, Spain
infmoh00@ucv.udc.es, cipenedo@dc.fi.udc.es

Abstract. The available visual information is quickly growing now a days, it is the reason of the emerging of a new research field, oriented to the automatic retrieval of this kind of information. These systems usually uses perceptual features of the images (color, shape, texture, ...). There is an important gap between the features used by the CBIR systems and the human perception of the information of an image. This work introduces a technique to extract significant perceptual regions of an image. The developed algorithm uses a bidimensional active model, active nets, these nets are guided by the chromatic components of a perceptual color space of the tested image. The restriction to only chromatic information made the fitting of an active net to the significant perceptual regions more tolerant to illumination problems of the image. The final objective will be to associate significant perceptual regions with semantic descriptors of the objects present in an image.

1 Introduction

With the advent of large image databases with complex images, efficient content-based retrieval of images has become an important issue. In the past years different proposals of Content-Based Image Retrieval (CBIR) systems were presented [1,2,3,4]. The image retrieval process used in a CBIR system usually follows the next steps (fig. 1): First, the user of the system introduces an image as query (image example [5], sketch [6], ...). Automatically, the system calculates a group of image descriptors (modeling) (such as color features [7,8], shape proprieties [9,10], texture [10], ...). Through an indexed algorithm, the system determines which group of images of the database is similar to the query. The system, then, organizes the image group using a ranking algorithm, and shows the results.

The actual CBIR systems uses perceptual features of the images such as color, shape, texture [1,2,3,4]. Although these systems show very interesting results, they have a problem. A human being compares two images based on semantic features, not perceptual ones [3,8]. The human being sees objects in the images and the relation between those objects. It is interesting that a new CBIR system tries to imitate, in some level, the way that a human being compares

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 795–802, 2004.
© Springer-Verlag Berlin Heidelberg 2004

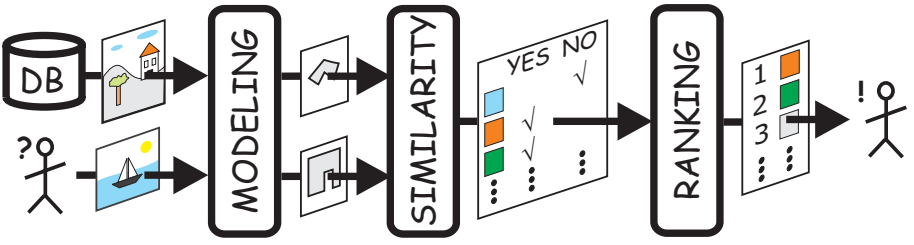


Fig. 1. Diagram of the proposed CBIR implementation.

two images. To try to minimize this gap, in this paper, a module to extract Significant Perceptual Regions (SPR) of digital images is introduced (this module corresponds with the modeling block of the fig . 1). In the future, those SPR could be indexed using a similarity module or be cross-reference with linguistic labels. The extraction of the SPRs is made by the use of a bidimensional active model, active nets, these nets are guided by the information about the color distribution of an image. The results of this SPR extraction module have to be valid, so, the development of the similarity module can be started.

The rest of this paper is organized as follows. In the next section, the active nets model is described. Sect. 3 focus on the external energies used to guide the active nets to describe the relevant zones of an image. Finally, main ideas, results and future work are summarized in Sect. 4.

2 Active Nets

An active net [11] is a discrete implementation of an elastic sheet. Active nets are deformed under the influence of internal and external forces. The internal forces represent inherent features of the physical sheet, such as, shape, stability and contraction of the net. And the external forces make the net change shape to characterize a relevant image feature.

An active net can be described as an active sheet defined by $v(r,s) = (x(r,s), y(r,s))$ where $(r,s) \in ([0,1] \times [0,1])$. The parameter domain, where (r,s) exits, is discretized as a regular grid defined by the internode spacing (k,l). This parameterization is defining a two dimensional net that will be acting in a two-dimensional world (digital image). Then, the energy equation for an active net that is acting in a digital image is

$$E(v) = \sum_{(r,s)} \left(E_{int}(v(r,s)) + E_{ext}(v(r,s)) \right) \tag{1}$$

where E_{int} is the internal energy of the net that controls the shape and structure of the active net and E_{ext} adapts the net to the desire zones in the image.

The main advantage of using an active net over other segmentation image techniques to extract a relevant region, is the ability of an active net to retrieve internal information of the image region.

2.1 Internal Energy

The internal energy term is defined as

$$\begin{aligned} E_{int}\left(v\left(r,s\right)\right) =& \alpha\left(\left|v_r\left(r,s\right)\right|^2+\left|v_s\left(r,s\right)\right|^2\right) \\ &+\beta\left(\left|v_{rr}\left(r,s\right)\right|^2+2\left|v_{rs}\left(r,s\right)\right|^2+\left|v_{ss}\left(r,s\right)\right|^2\right) \end{aligned} \tag{2}$$

where the subscripts indicate the partial derivatives and α and β are coefficients controlling the first and second order smoothness of the net. The internal energy ensures a C^1 continuous net. The first derivatives make the net contract and the second derivatives enforce smoothness and rigidity of the net. The definition of the internal energy in the equation 2 is continuous in the model parameters.

2.2 External Energy

The external energy, E_{ext}, is representing the external forces that act over the active net. These forces are designed to make the net be attracted to some kind of objects or zones present in an image [12,13]. In this work, forces dependent on the image will be used; then, the external energy can be defined as

$$E_{ext}\left(v\left(r,s\right)\right)=\sum_{energies} f\left[I\left(v\left(r,s\right)\right)\right] \tag{3}$$

where f is a general function of the intensity of the image $I\left(v\right)$. In the next section the external energies used in this work will be explained.

3 External Energies

The extraction of SPR is based on color information. The main idea is to detect regions in a digital image with similar color distributions, but, due to the nature of the algorithm used, inside those regions can be subregions with a different color distribution. To detect a color region it is needed to know where the borders of the region are and how its color distribution inside of it is structured.

Two external energies are used to guide the active net to the desire zones of an image. A first external energy makes a color histogram analysis to extract internal information of a SPR. And a second energy detects the borders inside an image, in the zones that have a difference between colors. Those two energies are explained more in detail in the next two subsections.

3.1 Color Distances in Perceptual Color Spaces

The idea of this external energy is to make a study of the color distribution of a digital image, to determine the relevant colors of it. This study is made by the analysis of a bidimensional histogram of the chromatic components of a perceptual color space. The use of perceptual color spaces has several advantages

due to their properties, mainly the feature that the distance measure between two points of the perceptual color space is similar to the human concept of distance between two colors [14]. At the same time, using only the chrominance information, the algorithm is more tolerant to illumination conditions. (The fig. 2.b[1] shows an] example of a bidimensional histogram of 256 bins, it is interesting to point that a $L^*a^*b^*$ was chosen as perceptual color space).

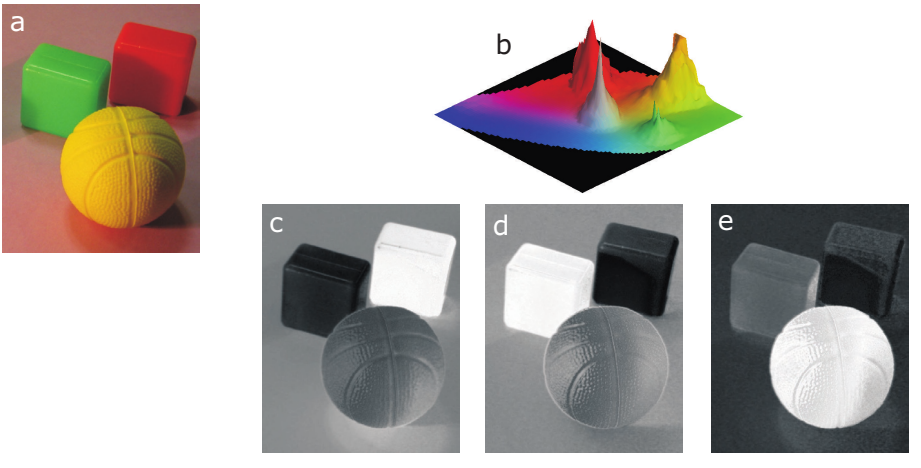


Fig. 2. Creation of the color distance images, a) the original image, b) the bidimensional histograma, each maximum is representing one of the dominant colors, the red tone of the right cube, the green tone of the left cube, the yellow tone of the ball and the background tone, c) the color distance image for the red cube, d) the color image distance for the green cube and e) the color image distance for the yellow cube.

The use of a bidimensional histogram is due to keep a uniform spatial color distribution, so, similar colors will be in the same zone of the histogram. The selection of the representative bins of the $L^*a^*b^*$ color space is given by the use of a K-Means algorithm [15]. After the histogram of an image is created, a study of maximums in the histogram is made to determine the relevant colors of the image: First, the histogram is filtered using a low pass filter to elimite small maximums. Then, all local maximums are selected. To each maximum a surronding zone is selected, this zone is determined by the more near histogram bin where the gradient is positive or when the value of a histogram bin is higher than a 10% of the local maximum. Finally, a color distance image is created to each relevant color zone of the image (To each pixel of the image that is asigned to any of the color in the relavant zone his value is changed to white, and the

[1] A color version of the paper images can be found at the next url: http://www-gva.dec.usc.es/medulio/iciar2004/

rest of pixels will have a grey value proportional to the Euclidean distance of their color value to the maximum color value of the relevant zone) (fig. 2).

3.2 Color Ratios

As it was said in the introduction of this section, it is needed to find the borders of the continuous color regions to get an anchor point for the external nodes of the active net. To do so, a study of color differences in the image is made using a color space that is invariant to illumination, discontinuing shadowing and shading cues. This color space was proposed by Theo Gevers [7].

The color space is defined by the relation of the components R,G,B between two neighbor pixels in an image, the definition is

$$m_i = \left(C_1^{\boldsymbol{x}_1} C_2^{\boldsymbol{x}_2}\right) / \left(C_1^{\boldsymbol{x}_2} C_2^{\boldsymbol{x}_1}\right) \tag{4}$$

where $C_1, C_2 \in \{R, G, B\}$ and $C_1 \neq C_2$, and where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ denote the image locations of two neighboring pixels. Then, the color relations are defined by

$$m_1 = \frac{R^{\boldsymbol{x}_1} G^{\boldsymbol{x}_2}}{R^{\boldsymbol{x}_2} G^{\boldsymbol{x}_1}} \qquad m_2 = \frac{R^{\boldsymbol{x}_1} B^{\boldsymbol{x}_2}}{R^{\boldsymbol{x}_2} B^{\boldsymbol{x}_1}} \qquad m_3 = \frac{G^{\boldsymbol{x}_1} B^{\boldsymbol{x}_2}}{G^{\boldsymbol{x}_2} B^{\boldsymbol{x}_1}} \tag{5}$$

Taking logarithms of both sides of equations 4, the color ratios can be seen as differences at two neighboring locations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (without any loss of generality, all the results derived for m_1 will also hold for m_2 and m_3).

$$d_{m_1}\left(\boldsymbol{x}_1, \boldsymbol{x}_2\right) = \ln m\left(R^{\boldsymbol{x}_1}, R^{\boldsymbol{x}_2}, G^{\boldsymbol{x}_1}, G^{\boldsymbol{x}_2}\right) = \ln R^{\boldsymbol{x}_1} + \ln G^{\boldsymbol{x}_2} - lnR^{\boldsymbol{x}_2} - lnG^{\boldsymbol{x}_1} \tag{6}$$

When these differences are taken between neighboring pixels in a particular direction, they correspond to finite-difference differentiation. To find color ratio edges in images, the edge detection is used, where the component of the gradient vector in the x and y direction is defined by

$$\begin{aligned} M_{m_1}^x(\boldsymbol{x}) = \frac{1}{4} \big(& d_{m_1}\left((x-1, y-1), (x+1, y+1)\right) + \\ & + 2 d_{m_1}\left((x+1, y), (x-1, y)\right) + \\ & + d_{m_1}\left((x-1, y+1), (x+1, y-1)\right) \big) \end{aligned} \tag{7}$$

$$\begin{aligned} M_{m_1}^y(\boldsymbol{x}) = \frac{1}{4} \big(& d_{m_1}\left((x-1, y-1), (x+1, y+1)\right) + \\ & + 2 d_{m_1}\left((x, y+1), (x, y-1)\right) + \\ & + d_{m_1}\left((x-1, y+1), (x+1, y-1)\right) \big) \end{aligned} \tag{8}$$

then, the gradient magnitude is represented by

$$||\nabla M_{m_1}(\boldsymbol{x})|| = \sqrt{M_{m_1}^x(\boldsymbol{x})^2 + M_{m_1}^y(\boldsymbol{x})^2} \tag{9}$$

The fig. 3 shows the results of this algorithm used against the original image of the fig. 2. The figure shows the result of multiplying fig. 2.c, fig. 2.d and fig. 2.e with fig. 3.a. Thanks to this, one border image per relevant zone is obtained, in this particular case, three separate color border images are gotten.

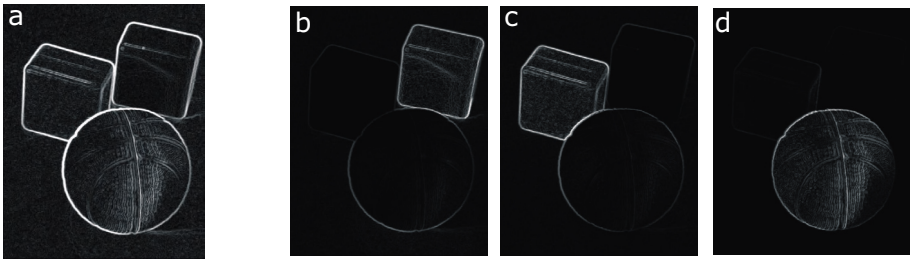


Fig. 3. Results of the color ratios algorithm, a) the external energy calculated to the image fig. 1.a, b) the result of multiplied fig. 1.c with fig. 2.a, with this operation, only the border of the red cube is shown, c) the result of multiplying fig. 1.d with fig. 2.a and d) the result of multiplying fig. 1.e with fig. 2.a

4 Region Extraction

The process of region extraction is the next one: First, the external energies (Sect. 3) for a digital image are calculated. Then, for each relevant color of the image, an active net is created. This net is guided to the region in two steps; First, a net of $24x24$ nodes is used for searching for zones of the image with relevant information (the parameters of this net are: $\alpha = 3$, $\beta = 0.01$, coefficient of external energy of Sect. 3.2 = 1, coeficient of external energy of Sect. 3.3 = 6)[12,13]. Second, a new active net is guided, focused only on the zone of the image selected by the first net. The coefficients of this second active net are the same ones uses for the first active net, but, its number of nodes is proportional to the spacial extension of the delemited object. This region was delimited by the first net. This last net will adapt its shape and internal nodes to the SPR (these result are showed in the fig. 4).

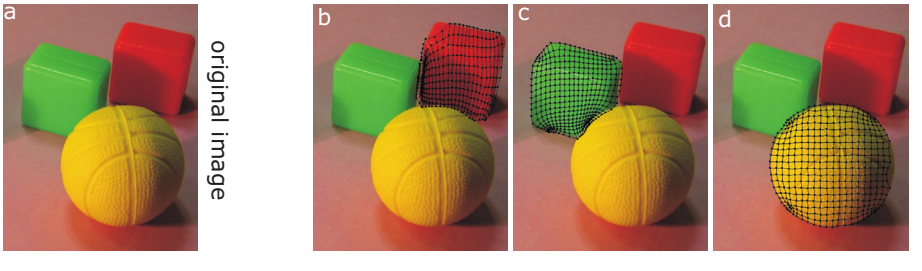


Fig. 4. a) The original image of the first example (fig. 2), b) the first active net focused on the yellow ball, c) the second active net focused on the red cube, d) the third active net focused on the green cube.

Fig. 5. a) A first example showing the results (b,c,d) of the red color tone of the car, e) second example showing the results (f,g,h) of the blue tone of the container, and, i) third example showing the results (j,k,l) of the red tone of the extintor.

5 Results and Conclusions

The fig. 5 shows several examples of results of the algorithm. It's necessary to point out that our technique has differences with simple color segmentation. The main difference is that the presented algorithm retrieves internal information of a region of an image, even, if some zones of this region have not the same color properties that were used to guide the active net. (the fig. 5.d and fig. 5.i show two examples of this behavior). This feature of the algorithm is quite interesting in the information retrieval world, since the algorithm is getting structural information of SRPs of digital images.

The work presented in this paper is one of the subprojects to develop a complete semantic CBIR system. With those promising results, as future work, we are looking forward to staring our next steps of our project. It is interesting to start working in the development of a similarity metric between two active nets, so a distance metric can be obtained to compare two SPR of a digital image. After the similarity metric is done, a cross-reference module can be developed to associate linguistic labels to the SPRs.

References

1. del Bimbo, A.: Visual Information Retrieval. Morgan Kaufmann Plublisers, Inc. (1999).
2. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. **22**, n. 12 (1999) 1349-1380.
3. Santini, S.: Exploratory Image Databases: Content-Based Retrieval. Academic Press (2001).
4. Rui, Y., Huang, T., Mehrotra, S.: Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of Visual Communications and Image Representation, vol. **10** (1999) 39-62.
5. Brunelli, R. and Mich, O.: Image Retrieval by Examples. IEEE Transactions on Multimedia. Vol. **2**, No. 3 (2000) 164-171.
6. del Bimbo, A., Pala, P.: Visual Image Retrieval by Elastic Matching of User Sketches. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. **19**, No. 2 (1997).
7. Gevers, T.: Color Image Invariant Segmentation and Retrieval. Ph. D., Wiskunde, Informatica, Natuurkunde and Sterrenunde (WINS), Amsterdam (1996).
8. Colombo, C., del Bimbo, A.: Color-Induced Image Representation and Retrieval. Pattern Recognition. Vol. **32** (1999) 1685-1695.
9. Berreti,S., del Bimbo, A., Pala, P.: Retrieval by Shape Similarity with Perceptual Distance and Effective Indexing. IEEE Transactions on Multimedia. Vol **2**, No. 4 (2000) 225-239.
10. Pala, P., Santini, S.: Image Retrieval by Shape and Texture. Pattern Recognition, **32** (1999) 517-527.
11. Bro-Nielsen, M.: Active Nets and Cubes. Morten Bro-Nielsen: Active Nets and Cubes, IMM Tech. Rep 94-13 (1994).
12. Ansia, F.M., Penedo, M.G., Mariño, C., López, J., Mosquera, A.: Automatic 3D Shape Reconstruction of Bones using Active Nets Based Segmentation. 15th International Conference on Pattern Recognition, Bacelona (2000).
13. Ansia, F.M., Penedo, M.G., Mariño, C., López, J., Mosquera, A.: Morphological Analysis with Active Nets. 4th Internation Conference on Advances in Pattern Recognition and Digital Techniques, ICAPRDT'99, Calcuta (1999).
14. Sangwine, S., Horne, R.: The Colour Image Procesing Handbook. Chapman & Hall (1998).
15. Weisstein, E. W.: K-Means Clustering Algorithm. MathWorld–A Wolfram Web Resource. (2004) http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html
16. Colombo, C., del Bimbo, A.: Visible Image Retrieval. Castelli, V., Bergman, L.D. (eds). Image Databases: Search and Retrieval of Digital Imagery. John Wiley Sons, Inc. (2002) 11-33.

Improving the Boosted Correlogram

Nicholas R. Howe and Amanda Ricketson

Smith College, Northampton, MA, USA,
nhowe@cs.smith.edu

Abstract. Introduced seven years ago, the correlogram is a simple statistical image descriptor that nevertheless performs strongly on image retrieval tasks. As a result it has found wide use as a component inside larger systems for content-based image and video retrieval. Yet few studies have examined potential variants of the correlogram or compared their performance to the original. This paper presents systematic experiments on the correlogram and several variants under different conditions, showing that the results may vary significantly depending on both the variant chosen and its mode of application. As expected, the experimental setup combining correlogram variants with boosting shows the best results of those tested. Under these prime conditions, a novel variant of the correlogram shows a higher average precision for many image categories than the form commonly used.

1 Introduction

An image rarely reveals anything of interest in its raw pixel color data. For most tasks, pertinent information must be extracted computationally from the raw pixel intensities, yielding new forms of data that describe the image more effectively for the task at hand. Both image retrieval and the related task of image classification depend on effective image descriptors for success. Yet the development of effective descriptors for image and video indexing remains an area of basic research. Although not suitable for all tasks, simple descriptors that represent an image holistically (rather than by parts or regions) have proven remarkably effective in many areas, and are widely used, both outright for indexing and as components in larger systems. Six or seven years ago, the holistic descriptor of choice was the color histogram; today, as judged by recent citations, it is the color correlogram [2,11].

Given the success of the color correlogram as an image descriptor for indexing and classification, it is somewhat surprising how little research explores the details of its implementation and possible variants. In part this may be attributed to a sentiment among researchers that holistic representations lack the sophistication required for "real" image retrieval. Some denigrate the correlogram as too simple to capture the nuances of real semantic categories. Yet in experiments it handily beats other supposedly more nuanced representations [6,8]. More to the point, the fact of its widespread use merits a second look. While the correlogram's holistic approach may not be in tune with current thinking about how

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 803–810, 2004.
© Springer-Verlag Berlin Heidelberg 2004

image retrieval should work, it offers great strengths as a *component* of a larger system. This observation motivates the work in this paper, which seeks ways to improve upon the correlogram in certain applications.

The next section of the paper considers the origins and definition of the standard correlogram, and proposes several variants for investigation. A short summary of recent work in boosting for classification and retrieval follows. Section 3 describes a set of experiments comparing the correlogram variants on a selection of image classification/retrieval tasks. Finally, Section 4 concludes with an analysis of the lessons learned and potential further steps.

2 Correlogram Variants and Boosting

The color correlogram has proven its worth as an image descriptor for both comparison and retrieval. Relatively compact and simple to implement, yet more subtle and powerful than the color histogram, it has become perhaps the most widely used image descriptor today. Previous work has shown that applying boosting techniques to the correlogram representation yields a high quality image classifier, better than many other published boosted image classification/retrieval algorithms [7], and that boosting can function as a feature selector [1,14].

The descriptor that has become known as the correlogram comprises a feature vector computed on an image discretized into n color bins. ($n = 128$ in this paper.) Each component has a succinct probabilistic interpretation: given a pixel p of color c_i, what is the chance that a pixel chosen at random from a specified neighborhood around p also has color c_i? The standard treatment uses concentric ring neighborhoods with square radii of 1, 3, 5, and 7 pixels, allowing for fast computation via dynamic programming. In the equations below, $\Phi(p)$ represents the color of pixel p, and $d(p_1, p_2)$ represents the chessboard distance between pixels p_1 and p_2.

$$C_{c_i,r_1,r_2} = P\left(\Phi(p_2) = c_i \,|\, \Phi(p_1) = c_i \wedge p_2 \in B_{r_1,r_2}(p_1)\right) \tag{1}$$

$$B_{r_1,r_2}(p_1) = \{p_2 \,|\, r_1 < d(p_1, p_2) \leq r_2\} \tag{2}$$

The correlogram as described above first appeared in 1997 [10] and was developed further as part of a family of related descriptors in the Ph.D. dissertation of Jing Huang [9]. Huang referred the commonly used descriptor given above as the *banded autocorrelogram*. In this terminology, *banded* refers to the square ring neighborhoods used to compute the correlogram, and the *auto-* prefix indicates that all the measurements involve frequencies of pixels of the same color. Huang describes but does not further explore a more general set of statistics defined over a set of distance bands and all possible pairs of colors (c_i, c_j). A single component of this descriptor considers all pixels of some color c_i, and measures the fraction of pixels within a particular distance band that are a second color c_j.

$$C^*_{c_i,c_j,r_1,r_2} = P\left(\Phi(p_2) = c_j \,|\, \Phi(p_1) = c_i \wedge p_2 \in B_{r_1,r_2}(p_1)\right) \tag{3}$$

Although the general correlogram requires significantly greater storage than the autocorrelogram, two considerations argue against writing it off immediately. First, recent research on other large image descriptors has shown that they can be effective if applied in combination with effective feature selection algorithms [14]. Second, study of the general correlogram may motivate more compact representations that nevertheless capture the additional information contained in the general correlogram.

This paper introduces a novel image descriptor that represents a compromise in both size and descriptiveness between the autocorrelogram and the general correlogram. Called the *color band correlogram*, it groups colors into color distance bands analogous to the spatial distance bands of the standard correlogram. Each component of the color band correlogram corresponds to a specified initial color c_i, a distance band specified by the bounds r_1 and r_2, and a color band specified by perceptual difference in color space from c_i lying between ρ_1 and ρ_2. The value of the component equals the mean fraction of pixels falling within the specified spatial neighborhood that have colors in the specified color band.

$$C^{CB}_{c_i,r_1,r_2,\rho_1,\rho_2} = P\left(\Phi(p_2) \in \beta_{\rho_1,\rho_2}(c_i) \,|\, \Phi(p_1) = c_i \wedge p_2 \in B_{r_1,r_2}(p_1)\right) \tag{4}$$

$$\beta_{\rho_1,\rho_2}(c_i) = \{c_j \,|\, \rho_1 < \delta(c_i, c_j) \leq \rho_2\} \tag{5}$$

In the equation above, δ represents a perceptual distance function in color space, and ρ_1 and ρ_2 are similarity bounds demarking a set of colors around the central color c_i. In practice correlograms may be computed for two or three color bands, corresponding respectively to an exact color match c_i, a close color match (a handful of colors directly surrounding c_i), and perhaps a more relaxed color match (colors similar to c_i but not in the closely matching category). With three color bands, the color band correlogram requires three times the storage of the autocorrelogram. This reprises the difference in storage between the histogram and the autocorrelogram, which differs by a factor equal to the number of distance bands.

The extra information in the correlogram variants described above may allow higher accuracy in some cases, but may also prove a liability if the inclusion of less relevant features drowns out the more important ones. In other words, the compactness and simplicity of the autocorrelogram may be an advantage under some circumstances. Interestingly, others have studied image descriptors that include large numbers of mostly irrelevant features. Although these descriptors yield poor results when used directly for retrieval, they can become competitive when applied in conjunction with a feature selection algorithm [14]. Boosting has served successfully in this capacity, although it was not originally designed as a feature selector.

The experiments in Section 3 compare the performance of the three correlogram variants in both their original form and using AdaBoost [4] as a feature selector. We hypothesize that the correlogram variants that contain more information will benefit most from boosting, since the boosting process can act as a feature selector. With images where the extra information is relevant to the

query task, the more complex variants should outperform the autocorrelogram; where it is not relevant they should do about the same. The unboosted variants, on the other hand, should suffer somewhat when they include extra features not relevant to the image category being retrieved. One caveat applies: if the amount of training data is not sufficient, boosting may not be able to properly extract features that generalize to unseen images. The experimental results should indicate whether this is a common problem in practice.

This paper breaks no new ground with regard to boosting algorithms themselves; the reader should refer elsewhere for details [5]. Boosting works by repeatedly learning to classify a labeled training set under different weightings of the training instances. The reweighting serves to focus effort on boundaries and special cases, sharpening the definition of the target class. Both theory and practice indicate that the weighted vote of all the classifiers created during training will be more accurate than the original unboosted classifier [12,13].

Note that boosting is typically used not for retrieval but for *classification*, and it requires a training set of both positive and negative instances of the class to be retrieved. Yet it also can perform retrieval. Once trained, a boosted classifier assigns a score to any image that can be used for ranking of unknown images with respect to the trained category. Although some have developed ways to apply boosting within the canonical single-image query model [14], using it most naturally motivates a shift in methodology away from query-by-example toward query-by-category. For example, boosting could be used to train a library of classification models for keyword-based queries, or as input to some larger system. This paper adopts a methodology based upon trained image classifiers throughout, even for the unboosted experiments.

3 Experiments

The experiments divide naturally into two parts: those involving unboosted techniques, and those that involve boosted techniques. The methodologies are similar. All experiments share a 5x2-fold cross validation setup, a common classification testing framework [3]. They differ in the amount of training data used: the unboosted techniques can use all the available data, while the boosted experiments must hold some out (as described below).

For the unboosted descriptors, there are two further divisions into sets of experiments, depending upon the style in which the training data are used. The first style mimics query-by-example: each positive image in the training set forms a single-image query against the which images from the test set are ranked. The average of all these single-image queries gives the overall recall-precision figures for the test fold.

The second style of unboosted experiment builds an unboosted nearest-neighbor classifier. It selects the best exemplars of the class using a greedy additive approach: single images are added from the target class to the exemplar set one by one. The classification rate on the training set forms the criterion for selecting the next exemplar to add; when no new images can improve the train-

ing error, selection stops. The exemplar set then forms the positive examples for the nearest-neighbor classifier. Previous work has shown that this approach works better than simply using all the positive training instance for classification, since some of these may be particularly poor exemplars that can lead the classifier astray [8].

For the boosted experiments, the training data are further split into two equal subsets, one of which is used to train the boosted classifier, while the other (called the *holdout set*) is used to prevent overtraining. (Overtraining refers to situations where a classifier becomes too attuned to the particular set of examples used in training, and cannot generalize to the differences present in new data.) When performance on the holdout set ceases to improve, training stops. Although this method avoids overtraining, overall performance can be lower than if all the data were used for training. Nevertheless, the holdout set method maximizes fairness to the different methods, since they all receive optimal training on the data available.

The image library consists of 20100 images from the Corel photo CD collection, and is described in detail elsewhere [8]. Fifteen image categories chosen to represent a range of difficulty and subject matter make up the target classes. The names of the fifteen categories appear in the tables of results.

Tables 1 and 2 summarize the results of testing on the retrieval performance of the unboosted image descriptors. All numbers given in the tables are average precision. Table 1 shows the results for single-image queries, while Table 2 shows the results for the greedy-exemplar approach. Each row contains results for one image class, while the columns represent the autocorrelogram, two forms of color band correlogram, and general correlogram respectively. (The color band correlograms differ in that the first uses two bands, while the second uses three.) Since the random fold choice over five replications of the experiment leads to substantial variance, the standard deviation of each number shown in the table does not reliably indicate the significance of differences when comparing results between columns. A paired sample t-test accounts for the variance due to the random fold choice and reliably indicates which differences achieve statistical significance. The table uses bold type for performances of the correlogram variants that differ significantly from that of the autocorrelogram, and underlines the cases that represent improvements.

The two tables show that increasing the number of features without boosting tends to decrease the average precision. Although the color band correlograms do better on a few categories, the general correlogram (with the largest number of features by far) does uniformly worse than the autocorrelogram. These results suggest that irrelevant information in the additional features added to the correlogram variants contain is misguiding the retrieval process.

By contrast, boosting changes the results entirely. Table 3 summarizes the the retrieval performance of the boosted image descriptors, in the same format as the tables above. With boosting, the virtues of the correlogram variants become evident: the descriptors with the most features do the best. Although the large variances on some categories limit the number of statistically significant results

Table 1. Average precision for correlogram descriptors on 15 image classes, using unboosted single-image queries. From left to right, columns show the autocorrelogram, color band correlogram with two bands, color band correlogram with three bands, and general correlogram. Numbers that differ significantly from the autocorrelogram are bold, and improvements are underlined. Units are percentages; i.e., perfect retrieval = 100.

Class	Auto.	CB2	CB3	GC
Race Cars	3.4 ± 0.3	**2.6 ± 0.3**	**2.7 ± 0.3**	**1.0 ± 0.2**
Wolves	2.7 ± 0.2	**2.1 ± 0.2**	**2.2 ± 0.3**	**2.1 ± 0.3**
Churches	1.2 ± 0.1	**0.93 ± 0.08**	**0.94 ± 0.07**	**0.87 ± 0.15**
Tigers	10 ± 1	**8.3 ± 1.0**	10 ± 2	**8.4 ± 1.9**
Caves	1.9 ± 0.1	**1.4 ± 0.1**	**1.6 ± 0.1**	**1.3 ± 0.1**
Doors	1.4 ± 0.2	1.4 ± 0.2	**1.5 ± 0.3**	**0.96 ± 0.23**
Stained Glass	29 ± 4	**27 ± 3**	**32 ± 3**	**11 ± 2**
Candy	2.6 ± 0.4	**2.2 ± 0.3**	**2.1 ± 0.4**	**1.6 ± 0.3**
MVs	1.3 ± 0.2	1.3 ± 0.2	**1.2 ± 0.2**	**1.0 ± 0.2**
Bridges	1.2 ± 0.1	**0.99 ± 0.06**	**0.98 ± 0.05**	**1.0 ± 0.1**
Swimmers	4.2 ± 0.4	**5.3 ± 0.4**	**4.7 ± 0.3**	**1.4 ± 0.2**
Divers	12 ± 1	**4.7 ± 0.6**	**4.8 ± 0.7**	**2.9 ± 0.7**
Suns	5.5 ± 0.3	**9.3 ± 0.7**	**7.5 ± 0.4**	**2.5 ± 0.2**
Brown Bears	1.2 ± 0.2	**0.97 ± 0.09**	**0.96 ± 0.15**	**0.82 ± 0.18**
Cheetahs	4.5 ± 0.3	**3.7 ± 0.3**	**3.8 ± 0.3**	**3.7 ± 0.5**

Table 2. Average precision for correlogram descriptors on 15 image classes, using greedily chosen exemplars in a nearest-neighbor classifier. From left to right, columns show the autocorrelogram, color band correlogram with two bands, color band correlogram with three bands. Numbers that differ significantly from the autocorrelogram are bold, and improvements are underlined. Units are percentages; i.e., perfect retrieval = 100.

Class	Auto.	CB2	CB3	GC
Race Cars	6.5 ± 6.2	**0.79 ± 0.25**	**0.73 ± 0.17**	**0.36 ± 0.03**
Wolves	6.5 ± 1.4	6.1 ± 1.9	**5.2 ± 1.9**	**3.0 ± 1.4**
Churches	1.5 ± 0.3	1.1 ± 0.8	1.4 ± 1.1	1.5 ± 1.4
Tigers	26 ± 7	**17 ± 6**	**20 ± 6**	**15 ± 6**
Caves	1.3 ± 0.2	**1.1 ± 0.2**	**1.0 ± 0.1**	**0.59 ± 0.05**
Doors	1.5 ± 0.7	**2.2 ± 1.1**	**2.7 ± 1.5**	**0.95 ± 1.00**
Stained Glass	9.5 ± 7.6	10.0 ± 5.0	**15 ± 5**	**0.32 ± 0.09**
Candy	1.5 ± 0.8	**0.72 ± 0.11**	1.1 ± 0.8	**0.58 ± 0.19**
MVs	2.4 ± 1.1	2.6 ± 1.0	2.5 ± 1.0	**1.4 ± 1.0**
Bridges	1.7 ± 0.8	**1.1 ± 0.2**	**1.1 ± 0.2**	**1.1 ± 0.2**
Swimmers	5.6 ± 5.2	**8.7 ± 4.6**	**8.7 ± 4.7**	**1.3 ± 1.4**
Divers	21 ± 5	**11 ± 4**	**11 ± 4**	**5.2 ± 3.1**
Suns	4.8 ± 1.5	**7.4 ± 2.5**	6.1 ± 2.6	**1.7 ± 0.5**
Brown Bears	2.1 ± 1.7	**0.94 ± 0.47**	**1.2 ± 1.3**	**1.7 ± 1.6**
Cheetahs	6.9 ± 4.4	7.6 ± 3.3	7.6 ± 4.6	**2.8 ± 2.5**

($p < .05$), all the comparisons that achieve significance favor the more complex correlogram versions. This suggests that the boosting process can effectively select the features relevant to the query class, and that giving it more features to work with can enhance this action.

As a practical matter, the fact that $CB3$ can achieve performance near the levels of the general correlogram is encouraging, since it requires only 2% of the storage space. Building a retrieval system based on the general correlogram would be daunting due to its large storage and memory requirements. Thus the future may belong to representations like $CB3$ that combine expressiveness with relative compactness.

Table 3. Average precision for correlogram descriptors on 15 image classes, using boosted classifiers. From left to right, columns show the autocorrelogram, color band correlogram with two bands, color band correlogram with three bands. Numbers that differ significantly from the autocorrelogram are bold, and improvements are underlined. Units are percentages; i.e., perfect retrieval = 100.

Class	Auto.	CB2	CB3	GC
Race Cars	9.4 ± 8.6	19 ± 11	**22 ± 12**	**20 ± 14**
Wolves	2.3 ± 4.2	2.4 ± 3.5	2.6 ± 2.5	1.6 ± 1.4
Churches	0.66 ± 1.25	0.76 ± 1.74	0.57 ± 0.67	0.48 ± 0.84
Tigers	16 ± 10	16 ± 5	15 ± 8	18 ± 8
Caves	0.91 ± 1.18	0.82 ± 1.18	1.7 ± 2.4	5.5 ± 15.9
Doors	2.0 ± 2.9	3.0 ± 4.3	1.5 ± 2.1	1.3 ± 2.1
Stained Glass	44 ± 14	50 ± 12	**55 ± 16**	**64 ± 9**
Candy	12 ± 7	11 ± 9	12 ± 9	8.3 ± 10.5
MVs	1.1 ± 2.0	0.28 ± 0.41	0.23 ± 0.22	0.56 ± 0.52
Bridges	0.084 ± 0.098	1.2 ± 2.7	0.16 ± 0.14	1.9 ± 5.2
Swimmers	13 ± 10	21 ± 5	20 ± 10	**22 ± 8**
Divers	49 ± 13	54 ± 12	50 ± 12	52 ± 12
Suns	29 ± 11	32 ± 8	31 ± 6	30 ± 8
Brown Bears	1.3 ± 3.6	5.7 ± 15.9	0.82 ± 1.23	0.89 ± 1.92
Cheetahs	4.4 ± 4.0	7.2 ± 6.9	11 ± 10	9.7 ± 7.1

4 Conclusion

This paper has systematically examined several variants of the correlogram under a variety of experimental conditions. Boosted classification gives the best average precision over all the experimental frameworks. This result is not unexpected; previous work has shown that boosting improves the retrieval performance of the correlogram [7]. Other work has also shown that boosting can act as a feature selector, choosing features that are correlated with the target class and weeding out those that are not (which might otherwise mislead a classifier by drowning out the significant features) [14]. This paper combines these two

insights by augmenting the standard autocorrelogram with additional features based upon correlations with bands of similar colors. While the new features may not be as relevant for image classification and retrieval as those in the standard autocorrelogram, they can still improve retrieval performance when applied with boosting. This observation, and its experimental confirmation, shows that more remains to be discovered about the humble correlogram.

References

1. V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: A method for efficient approximate similarity rankings. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page (to appear), 2004.
2. I.J. Cox, M.L. Miller, T.P. Minka, T.V. Papathornas, and P.N. Yianilos. The bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments. *IEEE Tran. On Image Processing*, 9(1):20–37, 2000.
3. T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998. Revised December 30, 1997.
4. Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, 1996.
5. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Dept. of Statistics, Stanford University, 1998.
6. N. Howe. Percentile blobs for image similarity. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 78–83, Santa Barbara, CA, June 1998. IEEE Computer Society.
7. N. Howe. A closer look at boosted image retrieval. In *Image and Video Retrieval, Second International Conference*, pages 61–70. Springer, 2003.
8. N. R. Howe. *Analysis and Representations for Automatic Comparison, Classification and Retrieval of Digital Images.* PhD thesis, Cornell University, May 2001.
9. J. Huang. *Color-Spatial Image Indexing and Applications.* PhD thesis, Cornell University, August 1998.
10. J. Huang, S. K. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997.
11. F. Jing, M. Li, H. Zhang, and B. Zhang. Support vector machines for region-based image retrieval. In *Proc. IEEE International Conference on Multimedia & Expo, 2003*, 2003.
12. R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
13. R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.
14. K. Tieu and P. Viola. Boosting image retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume I, pages 228–235, 2000.

Distance Map Retrieval*

László Czúni, Dezső Csordás, and Gergely Császár

Department of Image Processing and Neurocomputing, University of Veszprém,
Egyetem u. 10, 8200 Veszprém, Hungary
czuni@almos.vein.hu

Abstract. The paper describes a new method for indexing and retrieval of photographic images: we propose to involve the distance information of objects for indexing. We applied the "range-from focus" technique for distance estimation combined by statistical segmentation. This technique doesn't require stereo and it is easy to be implemented in conventional digital cameras. Test results illustrate relative and absolute position retrieval methods.

Keywords: Image retrieval, range from focus

1 Introduction

It is an interesting question is how humans remember the 3D structure of real-world scenes: what is the importance of the 3D spatial relations of objects in different types of images and how to exploit it in image retrieval systems? We suppose that the utilization of 3D structure is unreasonably forgotten still recent times, while the representation and storage of depth information of image objects is a relatively simple and explicit way of describing visual sceneries. Since distance is a natural feature of objects in our environment non-specialist application users may easily imagine what distance means in images contrary to other commonly used features, such as edge density, textures, histograms, adjacency maps, etc.

In our paper we discuss a new but simple method how range imaging can help the retrieval of images from collections. While it is clear that the combination of color and depth would lead to superior results, in our first experiments we show different ways of managing pure depth maps for retrieval.

2 Capturing and Preprocessing Depth Information

2.1 Estimating Distance via Measuring Image Sharpness

If thinking about the general application of the idea it is evitable that we should choose a simple method for depth estimation. Most of today's cameras use the "range

* This work is supported by the State Scientific Research Fund of Hungary (OTKA T 037829).

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 811–817, 2004.
© Springer-Verlag Berlin Heidelberg 2004

from focus" method to set the lens position to get a sharp image. It means that they use a so-called "focus function" to estimate image sharpness before shooting an image. This method does not require stereo, fast but can result in noisy depth-maps especially over large distances. In [3] the relative performance of stereo and focus techniques are compared but only for distance below 1 meter, while in our application distance within 10 meters is to be utilized. According to our simulation experiments this leads to decreased reliability of data requiring post-processing.

In our experiments we chose this technique to generate distance maps, although other non-stereo methods may be also adequate from different aspects when talking about possible consumer applications ("range from defocus" [8],"range from zoom", etc.).

The well-know formula describing the relation between the focal length of a lens (f), the object distance (u) and the lens position (v) is:

$$\frac{1}{f} = \frac{1}{u} - \frac{1}{v} \tag{1}$$

That is if we work with fixed f (fixed zoom) and find the best lens position (v) with minimal blur we can estimate v; the only question is how to find the best focused position. Since we want to get a depth-map over the whole image area the focus is measured at every pixel and at every image location the lens' position (v) with the smallest blur is stored. Unfortunately, the image structure can be quite complex and in some cases it would require a long range to sample to find all areas focused. This could be decreased with adaptive focus measurements and with depth from defocus techniques. Although focus measuring does not need sophisticated algorithms and does require to shot several images, as done in today's auto-focus cameras, to have the computations done in the camera requires special camera hardware not available in our case. For this reason and to minimize computational load in our simulations we made 8 shots with different lens positions then downloaded them to the host PC to be processed off-line. (Here we mention that focus computations can already be carried out in programmable CMOS cameras such as [5].)

The optimal method to measure focus in a given configuration depends on the OTF (optical transfer function), the noise behavior, the camera parameters, and even the object that is observed [6]. Unfortunately, these pieces of information are not available in our case and practical considerations (such as the size of the area where focus is measured or the computational complexity of a focus measure function) can also be crucial when implementing the focus measure in a commercial low-price camera. Although physical parameters of the camera (such as iris, focal-length) have a great impact on the estimation process either, our uncalibrated camera gave satisfactory results. This was due to the fact that we didn't need precise depth-maps; rather the relative position of image regions was important. (We applied some zoom to decrease the depth-of-field to get more steep focus measure functions and made estimations on block-based averages). According to our test results the Laplacian operator *(L)* overperformed other focus measure functions (gradient, variance, entropy) similar to the experiments described in [6]. Since we don't need high-resolution depth-maps and want to decrease uncertainty we averaged the focus measure $|L|$ in blocks of size of

app. 30x30. The focal length of the camera was set between 16 and 24mm and the 8 images were taken focusing at 0.7, 1, 2, 3, 5, 7, 10; and at ∞ meters object distance. Fig. 1 illustrates a color image and the relevant gray-scale depth-map.

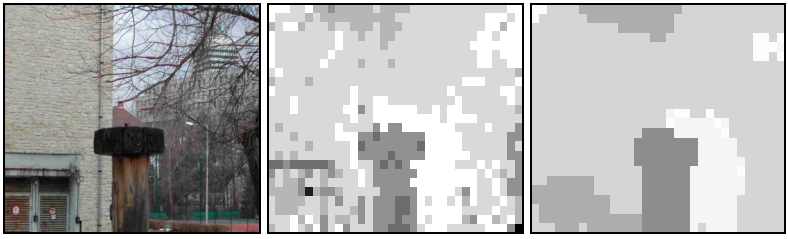


Fig. 1. Input image, related depth-map, and segmented depth information. Closer objects appear darker (doors belong to small building in front of the wall).

2.2 Segmenting Distance Maps

To reduce the noise of estimation uncertainty and to give a better representation for symbolic description, indexing, and storage we segmented the depth-maps with a Markov Random Field (MRF) related technique [7]. Other segmentation methods are also the subject of future experiments and evaluation but not discussed in this paper. Basically MRF methods are complex but can handle any arbitrary shapes and can be applied in the presence of strong noise. Compared to other MRF algorithms our technique is simple, the applied Modified Metropolis Dynamics (MMD) has fast convergence [4] and it works without any a-priori model information. The only parameter to be set ( ) controls homogeneity and was set to 1. The segmentation problem is solved with a MAP (Maximum A Posteriori) estimation of gray-scale pixel value classes (w) based on the initial observation (f and its smoothed version S) and on the rule that neighboring pixels are probably taking the same value on the segmented image. This is implemented in an energy optimization algorithm where the energy at a pixel location (p) to be minimized consists of two terms added:

$$\mathrm{E}_p(\omega) = \frac{(\omega_p - \mu_p)^2}{2\sigma_p^2} + \sum_{\{p,r\} \in C_p} V(\omega_p, \omega_r) \qquad (2)$$

where

$$\mu_p = \frac{f_p + S_p}{2}; \sigma_p = \frac{|f_p - S_p|}{2}; \text{ and } V(\omega_p, \omega_r) = \begin{cases} -\beta, & \text{if } \omega_p = \omega_r \\ +\beta, & \text{if } \omega_p \neq \omega_r \end{cases}. \quad (3);(4)$$

In our implementation w was selected from the 9 possible distance classes (including one for areas where the distance estimation did not work due to the lack of texture). The first term in Eq. 2 is responsible for getting a result that is close to our original observations while the second term gives homogeneity of neighboring regions ($\{p,r\} \in C_p$ denotes that pixels p and r form a neighboring pair called "clique"). The

relaxation algorithm is controlled with MMD. Fig. 1 also shows an MRF segmented depth-map.

3 Image and Depth-Map Indexing and Retrieval

In the following experiments we investigate the usefulness of pure depth-map search then in the next Chapter we propose future research directions how other symbolic solutions can be used to retrieve distance maps. Currently the number of image and depth-map pairs is around 100. In future we are planning to automate the recording process to increase the size of the database significantly. Depth-maps had a resolution of 40x30 blocks. We made several experiments with *query by example* and *query by sketch* methods.

Considering the noisy input data and the retrieval process there are several things to be taken into account:

- Our depth-maps had a range of 10 meters. Over this distance all objects are considered at ∞.
- In many examples the foreground level (usually grass, pavement, floor, etc.) is also visible in the bottom of the images. People defining a query do not take this into consideration still this can significantly modify the results.
- There can be several regions without texture. In these cases the focus function cannot be evaluated leading to unclassified/unmeasured areas. In our test these areas are eliminated and not considered in comparisons (colored with yellow in depth images).

Fig. 2. Results *of absolute query by example* and depth-maps; query is the first element

When people recall a scenario from their memory they may be unsure about the accurate distance of objects. Many cases they remember only the relative position along the Z-axis that makes the situation more difficult. For this reason we implemented two different retrieval methods: one is based on absolute and the other on relative comparisons.

Absolute position comparison: This technique computes the overall difference of the query and candidate depth-maps based on the l_2 norm. Rated results of one *query by example* search are in Fig. 2 and in Table 1 where the first element of the table is the query image itself and all values are normalized between 0 and 10 within each search.

Relative position comparison: Since homogeneous regions don't contain information about the relative position of neighboring regions only those blocks are investigated that have different right or bottom neighbor. As only these two neighbors are investigated and each one can have 3 states ("closer", "at the same distance", "far"), there are 9 types of blocks in the depth map. To describe the depth structure histograms with the 9 corresponding bins are generated and compared as a measure of depth-map similarity. To take into consideration the spatial distribution the image is cut into four quadrants and the NW, NE, SE and SW regions are compared independently then the error is accumulated for the whole depth-map. In several cases the relative position comparison outperforms the other as illustrated in Fig. 3 where the post boxes, shot from different distance with different zoom at different times of the year, are listed in the first positions. At the same time other images retrieved within the first 10 are far from expectations. Probably, the most important disadvantage of this technique is that only the structure of neighboring areas is measured; the relative position of regions that are not neighbors is not represented directly.

Fig. 3. Results *of relative query by example* and depth-maps; query is the first element. Yellow is for unmeasured areas.

Table 1. Numeric results of the query of Fig. 2

0	**3.5**	**3.51**	**3.62**	**3.65**
3.72	**3.8**	**3.97**	**3.98**	**4.0**

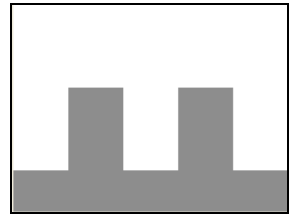


Fig. 4. Symbolic query by sketch. White areas mean „don't care" (not compared).

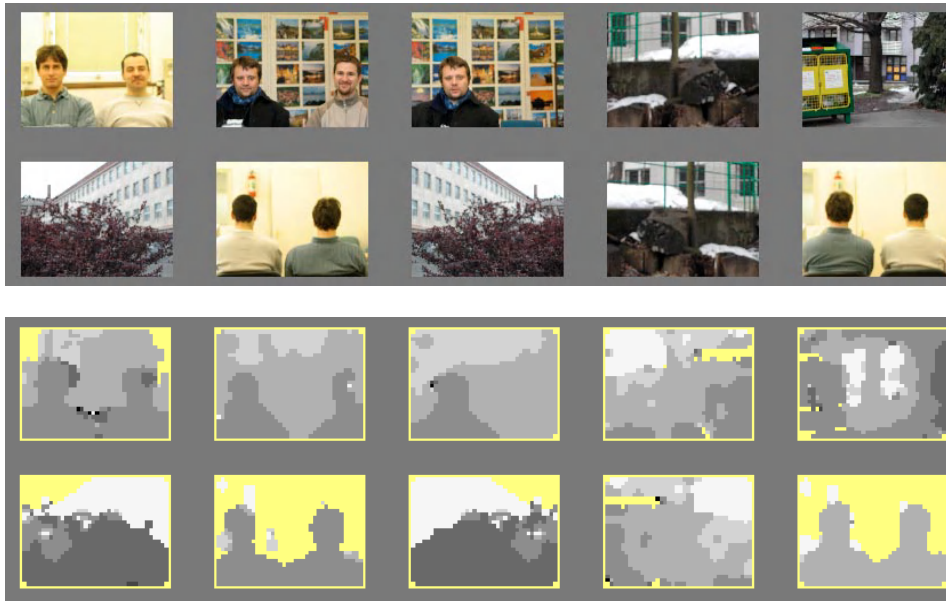


Fig. 5. Top 10 results of the query of Fig. 4. Yellow is for unmeasured areas.

4 Conclusions and Future Work

In this paper we discussed a new idea for indexing and retrieving photographic images. We proposed to use the "range from focus" method for depth-map generation-

since it does not require stereo and could be implemented in commercial cameras. Since depth-maps are affected by many parameters and are noisy over 1 m distance segmentation is applied for post-processing.

Two simple techniques are introduced for pure depth retrieval: absolute and relative distance comparisons are tested and illustrated with some examples. Both techniques have advantages and disadvantages; a possible numeric evaluation of their performance requires a much larger database.

Besides developing techniques to get more reliable depth-maps, currently we are experimenting with different symbolic representations [1] of depth maps and also with the combination of color for multimodal retrieval purposes. Symbolic description would be useful in the Jpeg 2000 and in the MPEG-7 frameworks [9,10] but the noisy and irregular shapes make conventional techniques inefficient. Simple solutions, such as a non-overlapping mosaic grid of depth layers to implement tiling, are suitable for the JPEG 2000 standard but give no answer in case of complex structures as some shown in Fig. 3. One example for such a symbolic query is in Fig. 4 and in Fig. 5. In this case we were looking for two people sitting in front of the camera. Retrieved result (with the relative technique) lists all 4 such images of the database in the first 10 top rated results.

References

[1] Shi-Kuo Chang and Erland Jungert: Symbolic Projection for Image Information Retrieval and Spatial Reasoning, Academic Press, London, 1996.

[2] L. Czúni, A. Licsár: Method of fixing, storing and retrieving images and the associated distance indexes. P0204432 Hungarian Patent Office, Budapest, 2003

[3] S. Das and N. Ahuja: Perfomance Analysis of Stereo, Vergence and Focus as Depth Cues for Active Vision, IEEE Trans. on PAMI, Vol. 17, No.12, pp. 1213-1219, December, 1995

[4] Z. Kato, J. Zerubia, M. Berthod: Satellite image classification using a Modified Metropolis Dynamics, In Proc. ICASSP, San-Francisco, California, USA, Mar. 1992

[5] T. Roska, Á. Zarándy, S. Zöld, P. Földesy and P. Szolgay: The Computational Infrastructure of Analogic CNN Computing - Part I: The CNN-UM Chip Prototyping System, IEEE Trans. on Circuits and Systems I: Special Issue on Bio-Inspired Processors and Cellular Neural Networks for Vision, Vol. 46, pp. 261-268, 1999

[6] M. Subbarao and J.-K. Tyan: Selecting the Optimal Focus Measure for Autofocusing and Depth-From-Focus. IEEE Trans. on PAMI, Vol.20, No.8, August 1998

[7] T. Szirányi, J. Zerubia: Markov Random Field Image Segmentation using Cellular Neural Network. IEEE Trans. on Circuits and Systems I., Vol. 44, pp. 86-89, January 1997

[8] Y. Xiong and S. Shafer: Depth from Focusing and Defocusing. Tech. report CMU-ROI-TR-93-07, Robotics Institute, Carnegie Mellon University, March 1993

[9] Jpeg 2000 Standard, Part I. and II. (ISO/IEC FCD15444-1/2 : 2000)

[10] José M. Martínez: MPEG-7 Overview (ISO/IEC JTC1/SC29/WG11N5525)

Grass Field Segmentation, the First Step Toward Player Tracking, Deep Compression, and Content Based Football Image Retrieval

Kaveh Kangarloo[1,2] and Ehsanollah Kabir[3]

[1]Dept. of Electrical Eng., Azad University, Central Tehran Branch, Tehran, Iran
[2]Dept. of Electrical Eng., Azad University, Science and Research unit, Tehran, Iran
kangarloo@iauctb.org
[3] Dept. of Electrical Eng., Tarbiat Modarres University, Tehran, Iran
kabir@modares.ac.ir

Abstract. In this paper, a method is presented which can be used for the segmentation of grass field in video images taken from football matches. Grass field is a green and nearly soft region. Therefore color and texture are two suitable features which can be used to describe it. As HSI color space is more stable against illumination changes in comparison with other color spaces, the hue is selected as color feature.
Sub-band images containing high frequency information in horizontal, vertical and diagonal directions that are obtained by applying wavelet transform on image intensity have been used for texture description. Classification of grass and non-grass fields is done using an MLP classifier. The results revealed that the proposed method is able to recognize grass and non-grass samples accurately.

Keywords: Football, Grass field, Image segmentation, Color, Texture, Wavelet transform, Classification.

1 Introduction

Grass field segmentation can be used for detection and tracking of players, content based video indexing and retrieval. In a research on player detection, motion and color features were used for grass recognition [1]. In this research, by thresholding the motion histogram, grass field is segmented. In other words, still green pixels, are labeled as grass. In another research, by applying a threshold on color histogram, green pixels are segmented and then considering motion parameters and edge density, players are segmented [2]. Here, moving pixels are detected and then using a moving window, background and foreground regions are separated from each other.
In a similar research, done for scene recognition and analyzing the images taken from baseball and tennis court, at first, the land area is recognized by applying a threshold value on color histogram. Then considering the camera movements, scene shots are segmented [3,4].

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 818–824, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In another research, some frames are randomly selected and dominant colors in them is taken as the grass color. Then by applying a threshold value on the color histogram, the grass field is segmented and based on its shape, far and near field images are classified [5].
In general, for image segmentation, different pixels should be classified based on their similarities. In applications such as surveillance and traffic monitoring, segmentation is based on motion, whereas in applications such as face recognition or image retrieval, it is based on color or shape of connected regions. There are different methods for image segmentation based on similarity of pixels but the two methods, histogram thresholding and feature classification are used more than the others [6].
In the first method, by studying valleys and picks in the feature space histogram and selecting a threshold value, similar regions are segmented. In some cases, user determines the threshold value, whereas in other cases the threshold value is calculated based on an entropy criteria or techniques such as watershed algorithm [7].
In the second method, segmentation is based on feature clustering. General methods such as k-means, mean-shift, self-organized map or graph-cut [8] are the main techniques in this case. It should be stated that in all mentioned methods segmentation could be improved based on the shape of obtained regions. Techniques such as MRF [9], split and merge or studying the boundaries [10] are some of these methods.

Although, motion information or scene perspective in former frames are significant items that improve the segmentation accuracy, but the main purpose of this research is to divide an image to grass and non-grass regions just using within-frame information. In this paper, a method is presented that hue as color feature with 3 texture features obtained from wavelet sub-bands, are used for segmentation. This paper is organized as follow. Section 2 introduces the applied features. In section 3, on the basis of feature-space and clusters related to grass and non-grass samples the classification method is selected. Section 4, provides the experimental results and draws the conclusion.

2 Feature Extraction

2.1 Color Feature

Grass area in video images could be considered as a green and plain area. Color is one of the most important factors that can provide valuable information about the image. Most cameras produce RGB signals. Due to sensitivity to the light changes, color components are transformed to other suitable color spaces [11].
In dynamic scenes, usually based on the application, different color spaces are used. Normalized RGB, HSI, YUV, CIE-Luv and CIE-Lab are some of them. Although, these color spaces have special applications in machine vision, the use of HSI color space, because of its high stability against illumination changes, shadow and Gamma coefficient, is more than the others [12]. In this space H, S and I stand for hue, saturation and intensity respectively. In this research we decided to use the H component as color feature.

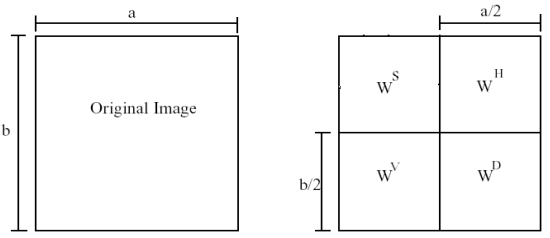


Fig. 1. Image wavelet decomposition.

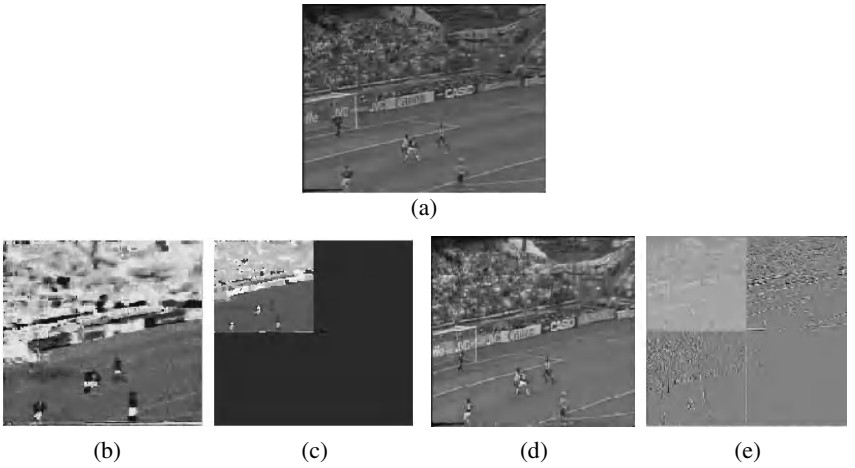
2.2 Texture Features

There are different techniques for texture analysis mainly divided into four groups as, structural, statistical, spectral and techniques based on modeling. In structural methods, texture is described by a series of basic parts named micro textures [13]. Micro textures are determined in a way that, based on the application concerned, texture could be synthesized or recognized easily. In statistical methods, statistical features describe the image. By studying these features or their changes we could classify different regions. Co-occurrence matrix is one of the main techniques in this group [14].

In techniques, based on modeling, the concerned texture is modeled. Markov and fractal modeling are the most common methods in this group [15]. In spectral methods, usually depending on the application, Fourier, wavelet or Gabor transforms are used. Here, by applying one of these transforms, we obtain an image that its changes in intensity or color, is more obvious [16].

In this paper we used the wavelet transform to extract texture features. Generally the wavelet transform is used to decompose an image into a set of four independent spatially oriented frequency channels, named sub-band images [17,18]. One of them, W^S, represents low frequency information *(average)* and the other three components, W^H, W^V, W^D, contain high frequency spectral information *(details)* in horizontal, vertical and diagonal directions respectively (Fig.1). To each image pixel, three corresponding elements of W^H, W^V and W^D is assigned as texture features.

In football matches, images are normally taken from far field. In these images the size of players does not exceed more than 30*50 pixels. On the other hand, in compressed images for decreasing the information rate, smoothing on the color is more than intensity and the resolution in horizontal and vertical directions is descended down to its half [19]. Therefore players are seen as a small section with monotonous color. If wavelet transform applied on the color components, the sub-band images that indicate color changes in horizontal, vertical and diagonal directions would bear no considerable information and nearly all color information shall be collected in the low frequency sub-band image. For this reason, the wavelet transform is applied to image intensity. Fig.2 shows the result of applying Haar wavelet transform on hue and intensity of a sample image. It is clear that, the intensity sub-band images contain more information compared to hue.



(a) (b) (c) (d) (e)

Fig. 2. The result of Haar wavelet decomposition of hue and intensity. (a) Main image, (b) Hue, (c) Hue wavelet decomposition, (d) Intensity, (e) Intensity wavelet decomposition

3 Classification

In order to classify pixels to grass and non-grass, we should collect some samples from different images and find a suitable solution based on their diffusion against each other. For this reason, from 40 video clips of football matches that is taken in different tournaments, 800 pixel samples were randomly selected and the features, hue, W^H, W^V and W^D, for each pixel is calculated. The dispersion of grass and non-grass samples fgor hue and texture features is shown in Fig.3. The graphs from left to right show grass and non-grass clusters based on color and one of the features W^H, W^V and W^D. As shown, grass and non-grass dispersion is in a way that clusters are nearly separated from each other.

Based on dispersion of grass and non-grass samples, we decided to use an MLP classifier [20]. Based on the shape of grass and non-grass clusters, a network with 2 hidden layers will be able to distinct them. For training the network, 400 grass and non-grass samples were randomly selected from collected samples. In learning phase, the features of each sample were considered as input and their belonging to grass and non-grass sets as ± 1 were assumed as output of the network. 99% accuracy in recognition of samples showed that the applied perceptron classifier with 4 input nodes, 3 and 5 nodes in two hidden layers and one node in output, 4-3-5-1, carries out classification in a perfect manner.

In Fig.4 the result of proposed algorithm on two sample images is shown. As it is clear, even in those images with low resolution, the operation of the system is satisfactory and the players, lines and even ball could be recognized from the background. Certainly, information concerning the motion or scene perspective in past frames are features that could be used to improve the recognition rate.

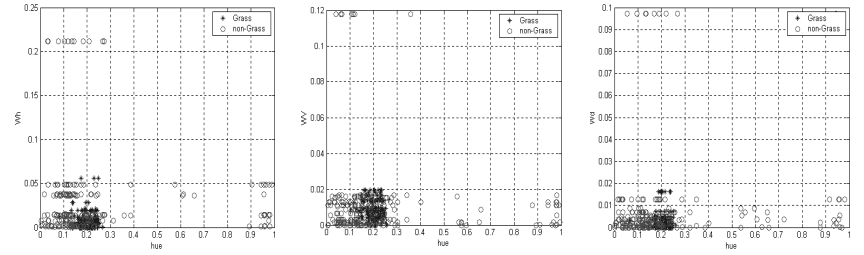


Fig. 3. Grass and none-grass samples dispersion based on color and one texture feature. Horizontal axes in images from left to right, indicate the hue and vertical axes indicate the W^H, W^V and W^D features respectively.

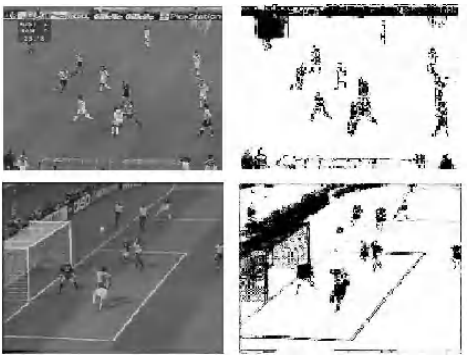


Fig. 4. Result of the proposed algorithm on two sample images of size 180*240 pixels

4 Conclusion

In this paper, a method based on wavelet transform presented that can be used for grass field segmentation in soccer video images. Since wavelet transform is applied to image intensity, sudden illumination changes such as long shadow will cause great errors (Fig.5). In a similar research performed by the authors, the color dispersion is used as a feature to describe the texture. We call it color smoothness method. The idea behind it is to estimate the color dispersion in horizontal and vertical directions on the base of mask operations. In the other word, decision of belonging each pixel to grass or non-grass regions is made based on color information. The major drawback of this method is obstacle recognition of image details such as lines or even some parts of players. (Fig.6).

In order to show the privileges and limitations of these two methods, some tests were done on two different video images. The first set contains images taken from far field. The other one includes players selected from the first set images. In Fig.7 the result of both methods for segmentation of players is shown.

As it is clear, when images are taken from far field or the intensity changes highly, the wavelet-based method labels several pixels erroneously whereas for segmentation of players, the color smoothness method is not effective and the wavelet-based algorithm

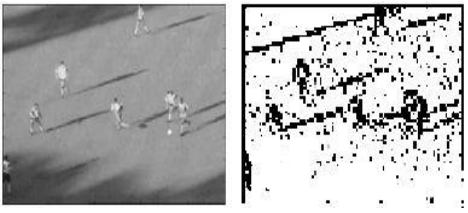


Fig. 5. Some error caused during grass recognition.

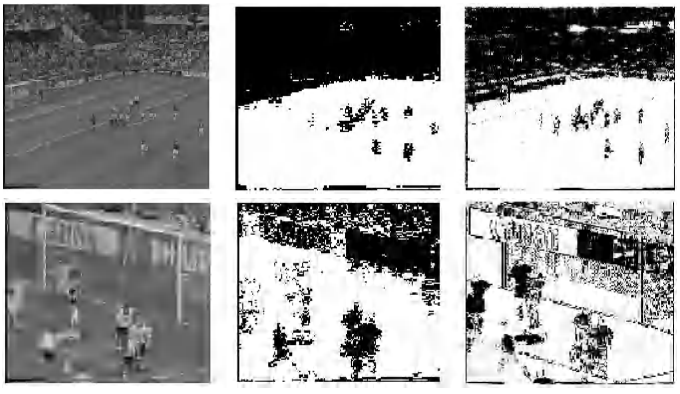


Fig. 6. Results of the two proposed methods on two sample images. Left) Main images, Middle) Grass field segmentation based on color smoothness method, Right) Grass field segmentation based on wavelet transform

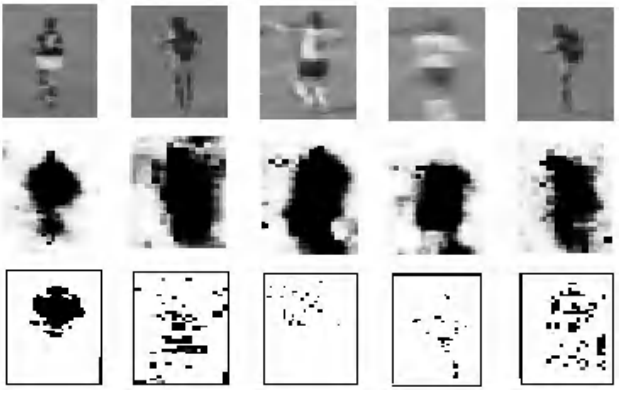


Fig. 7. Results of the proposed algorithms on second image set. Upper row) Main images, Middle row) Grass field segmentation based on wavelet transform, Lower row) Grass field segmentation based on color smoothness method.

submitted an acceptable accuracy. On the basis of aforementioned contents, for recognition the scene, image retrieval or studying the camera movements, the color smoothness method is effective. In case where grass field segmentation is used for game analysis, identifying ball movements or player detection and tracking, the wavelet-based method is suggested. Certainly, applying later processing such as image

merging, adaptive thresholding or using motion information are the items that can be used to improve the accuracy.

References

1. Seo, Y., Choi, S., Kim, H. and Hong, K.S.: Where are the ball and players?: Soccer Game Analysis with Color-based Tracking and Image Mosaic. Proceedings of Int. Conference on Image Analysis and Processing (ICIAP), (1997), 196-203
2. Utsumi, O., Miura, K., Ide, I., Sakai, S. and Tanaka, H.: An Object Detection Method for Describing Soccer Games from Video. Proceedings of IEEE Int. Conference on Multimedia and Expo. (ICME), vol. 1, (2002), 45-48
3. Sudhir, G., Lee, J. C. M., Jain, A. K.: Automatic Classification of Tennis Video for High-level Content-based Retrieval. International Workshop on Content-Based Access of Image and Video Databases (CAIVD), (1998), 81-90
4. Hua, W., Han, M. and Gong, Y.: Baseball Scene Classification Using Multimedia Features. Proceedings of IEEE International Conference on Multimedia and Expo., Vol. 1, (2002), 821-824
5. Xu, P., Xie, L., Chang, S.F., Divakaran, A., Vetro, A. and Sun, H.: Algorithms and System for Segmentation and Structure Analysis in Soccer Video. Proceedings of IEEE Int. Conference on Multimedia and Expo.(ICME), (2001), 928-931
6. Pal, N.R., Pal, S.K.: A Review on Image Segmentation Technique. Pattern Recognition Letters, Vol. 26, (1993), 1277-1294
7. Bleau, A. and Joshua Leon, L.: Watershed-Based Segmentation and Region Merging. Computer Vision and Image Understanding, Vol. 77, (2000), 317-370
8. Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, (2000), 888 - 905
9. Marroquin, J.L., Santana, E.A. and Botello, S.: Hidden Markov Measure Field Models for Image Segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25, (2003), 1380-1387
10. Bhalerao, A. and Wilson, R., Unsupervised Image Segmentation Combining Region and Boundary Estimation, IVC, Vol. 19, pp. 353-368, 2001.
11. Finlayson, G. and Schaefer, G.: Hue That is Invariant to Brightness and Gamma. Proceedings of British Machine Vision Conference, (2000), 303-312
12. Buluswar, S.D. and Draper, B.A.: Color Models for Outdoor Machine Vision. Computer Vision and Image Understanding, Vol. 85, pp. 71-99, 2002.
13. Lin, H.C, Chiu, C.Y. and Yang, S.N.: Finding Textures by Textual Descriptions, Visual Examples and Relevance Feedbacks. Pattern Recognition Letters, Vol. 24, (2003), 2255-2267
14. Zhu, S.C.: Statistical Modeling and Conceptualization of Visual Patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25, (2003), 691-712
15. Pentland, A.: Fractal-Based Description of Natural Scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 6, (1984), 661-674
16. Azencott, R., Wang, J.P. and Younes, L.: Texture Classification Using Windowed Fourier Filters. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, (1997), 148-153
17. Abramovich, F., Bailey, T. and Sapatinas, T.: Wavelet Analysis and its Statistical Applications. Journal of the Royal Statistical Society (JRSSD), Vol. 48, (2000), 1-30
18. Arivazhagan, S. and Ganesan, L.: Texture Segmentation using Wavelet Transform. Pattern Recognition Letters, Vol. 24, (2003), 3197-3203
19. ISO-11172-2: Generic Coding of Moving Pictures and Associated Audio (MPEG-1).
20. Fasulo. D.: An Analysis of Recent Work on Clustering Algorithms. Technical Report 01-03-2, University of Washington, April 1999

Spatio-temporal Primitive Extraction Using Hermite and Laguerre Filters for Early Vision Video Indexing*

Carlos Joel Rivero-Moreno and Stéphane Bres

LIRIS, FRE 2672 CNRS, Lab. d'InfoRmatique en Images et Systèmes d'information,
INSA de Lyon, Bât. Jules Verne, 17 av. Jean Capelle, Villeurbanne Cedex, 69621 FRANCE
rivero@rfv.insa-lyon.fr, stephane.bres@liris.cnrs.fr

Abstract. In this paper we integrate spatial and temporal information, which are extracted separately from a video sequence, for indexing and retrieval purposes. We focus on two filter families that are suitable models of the human visual system for spatial and temporal information encoding. They are special cases of polynomial transforms that perform local decompositions of a signal. Spatial primitives are extracted using Hermite filters, which agree with the Gaussian derivative model of receptive field profiles. Temporal events are characterized by Laguerre filters, which preserve the causality constraint in the temporal domain. Integration of both models gives a spatio-temporal feature extractor based on early vision. They are efficiently implemented as two independent sets of discrete channels, Krawtchouk and Meixner, whose outputs are combined for indexing a video sequence. Results encourage our model for video indexing and retrieval.

1 Introduction

Video indexing and retrieval [2] is an important issue in both multimedia applications and management of huge audiovisual databases. The goal is to retrieve the exact image sequence (monitoring) or similar image sequences with respect to a given query. The latter implies the ill-posed problem of defining a similarity measure for video searching. Most of time, it leads to compare *signatures* based on features extracted from videos. This one can be viewed as a dimensionality reduction process. Without loss of generality, an indexing system requires then two stages. The first one consists of feature extraction of relevant information. The second stage is indexing based on the extracted features vectors in order to achieve dimensionality reduction.

It is well known that the human visual system (HVS) codes efficiently visual stimuli. Both neurophysiology and psychophysics support the notion that early visual processing can be described by a set of channels operating in parallel that transform the input signal to obtain a coded version of the stimulus characteristics [6]. This code can subsequently be used as the basis for all kinds of perceptual attributes. It is thus desirable to have a feature extractor that approximates, in such a way, the channels used to described the HVS.

* This work was supported by the National Council of Science and Technology (CONACyT) of Mexico, grant 111539, and by the SEP of Mexico.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 825–832, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper we present a novel approach to create a video signature for indexing and retrieval purposes. It is based on two polynomial transforms [5] which model the set of channels describing the HVS. In general, a polynomial transform decomposes locally a signal into a set of orthogonal polynomials with respect to the window used for localizing the signal. The Hermite transform [5] is used to extract spatial primitives while the generalized Laguerre transform [1] characterizes temporal events. We only need the analysis parts of such transforms since they encode the required visual information. Furthermore, we treat separately the spatial and temporal dimensions in the feature extraction process since they differ essentially in that the temporal domain must adhere to the causality condition. Last but not least, we present the discrete representations of the two independent sets of channels. They correspond to Krawtchouk and Meixner filters, for Hermite and Laguerre filters, respectively. An efficient implementation is achieved by their normalized recurrence relations.

2 Hermite and Krawtchouk Filters

In order to extract spatial features we use Hermite filters. They correspond to analysis filters of the forward Hermite transform [5] and agree with the Gaussian derivative model of the HVS [10]. We will focus on their cartesian representation which is more oriented to extract spatial primitives such as edges, lines, bars, and corners, into the vertical, horizontal, and oblique directions rather than oriented textures. However, they have similarities to Gabor filters [7], which are more used, essentially for texture, in image processing and feature extraction. Indeed, Hermite and Gabor filters are equivalent models of receptive field profiles (RFPs) of the HVS [9]. Besides these properties, a discrete equivalent representation exists for Hermite filters based on Krawtchouk filters, which allows to have an efficient implementation on discrete data for video indexing purposes.

2.1 Cartesian Hermite Filters

Hermite filters $d_{n-m,m}(x,y)$ decompose a localized signal $l_v(x-p,y-q) = v^2(x-p,y-q)\ l(x,y)$ by a Gaussian window $v(x,y)$ with spread σ and unit energy, which is defined as:

$$v(x,y) = 1/(\sigma\sqrt{\pi})e^{-(x^2+y^2)/(2\sigma^2)} \quad . \tag{1}$$

into a set of Hermite orthogonal polynomials $H_{n-m,m}(x/\sigma, y/\sigma)$. Coefficients $l_{n-m,m}(p,q)$ at lattice positions $(p,q)\in P$ are then derived from the signal $l(x,y)$ by convolving with the Hermite filters. These filters are equal to Gaussian derivatives where $n–m$ and m are respectively the derivative orders, $(n–m,m)$, in x- and y-directions, for $n=0,\ldots,D$ and $m=0,\ldots,n$. Thus, the two parameters of Hermite filters are the maximum derivative order D (or polynomial degree) and the scale σ.

Hermite filters are separable both in spatial and polar coordinates, so they can be implemented very efficiently. Thus, $d_{n-m,m}(x,y) = d_{n-m}(x)\ d_m(y)$, where each 1-D filter is:

$$d_n(x) = \left((-1)^n \Big/ (\sqrt{2^n \cdot n!}\sqrt{\pi}\sigma) \right) H_n(x/\sigma) e^{-x^2/\sigma^2} \quad . \tag{2}$$

where Hermite polynomials $H_n(x)$, which are orthogonal with respect to the weighting function $exp(-x^2)$, are defined by Rodrigues' formula [3] as:

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \quad . \tag{3}$$

In the frequency domain, these filters are Gaussian-like band-pass filters with extreme value for $(\omega\sigma)^2 = 2n$ [9], and hence filters of increasing order analyze successively higher frequencies in the signal.

2.2 Krawtchouk Filters

Krawtchouk filters are the discrete equivalent of Hermite filters. They are equal to Krawtchouk polynomials multiplied by a binomial window $v^2(x) = C_N^x / 2^N$, which is the discrete counterpart of a Gaussian window. These polynomials are orthonormal with respect to this window and they are defined as [3]:

$$K_n(x) = \frac{1}{\sqrt{C_N^n}} \sum_{\tau=0}^{n} (-1)^{n-\tau} C_{N-x}^{n-\tau} C_x^{\tau} \quad . \tag{4}$$

for x=0,…,N and n=0,…,D with $D \le N$.

It can be shown that the Krawtchouk filters of length N approximates the Hermite filters of spread $\sigma = \sqrt{N/2}$. In order to achieve fast computations, we present a normalized recurrence relation to compute these filters:

$$K_{n+1}(x) = \frac{1}{\sqrt{(N-n)(n+1)}} \left[(2x-N) K_n(x) - \sqrt{n(N-n+1)} K_{n-1}(x) \right] \quad . \tag{5}$$

for $n \ge 1$ and with initial conditions $K_0(x) = 1$, $K_1(x) = \frac{2}{\sqrt{N}} \left(x - \frac{N}{2} \right)$.

3 Laguerre and Meixner Filters

Temporal and spatial processing differ essentially in that the temporal domain must adhere to the causality condition. It means that we can only use what has occurred in the past. This one naturally supposes, on the one hand, that events closer time t_0 should have more weight than past events (which tend to vanish), and on the other hand, variations of such events along time might be measured by time derivatives or, which is equivalent, to fit some oscillatory function. These suppositions lead to a kind of event localization from the past up to present time t_0, i.e. a smoothing causal kernel or causal localization window is applied to the signal. As it was argued in [4],

the only primitive scale-space kernels with one side support are the truncated exponential-like functions. We emphasize here the term "exponential-*like*" since functions involving exponentials modulated by a time polynomial is a *generalized* case of such kernels.

The Laguerre transform is another polynomial transform that uses a monomial-modulated exponential function as localization window. There is psychophysical evidence that the early visual processing of temporal stimuli in the HVS is described by this transform and channel responses resemble those of Laguerre filters [1]. Due to these properties, Laguerre filters will be used as a temporal feature extractor. Furthermore, an efficient implementation for video indexing purposes can be achieved by their discrete equivalent representation, i.e. the Meixner filters.

3.1 Generalized Laguerre Filters

Generalized Laguerre filters $d_n(t)$ decompose a localized temporal signal $l_v(t\text{-}t_0) = v^2(t\text{-}t_0)\ l(t)$ by a gamma window (monomial-modulated exponential-like window) $v(t)$, with order of generalization $\alpha \geq 0$ and spread $\sigma > 0$, which is defined as [1]:

$$v(t) = \sqrt{\sigma}(-\sigma t)^{\alpha/2} e^{\sigma t/2} u(-t) \ . \tag{6}$$

where u is the heaviside function ($u(t)=1$ for $t \geq 0$, $u(t)=0$ for $t<0$), into a set of generalized Laguerre orthogonal polynomials $L_n^{(\alpha)}(-\sigma t)$. Coefficients $l_n(t_0)$ at different times $t_0 \in R$ are then derived from the signal $l(t)$ by convolving with the generalized Laguerre filters. These filters are equal to time derivatives of the analysis window modulated by a monomial of order n, where n is the derivative order, for $n=0,\ldots,D$. Thus, the three parameters of generalized Laguerre filters are the maximum derivative order D (or polynomial degree), the scale σ, and the order of generalization α. Time-invariant filters, i.e. exponential decreasing filters, are obtained for $\alpha=0$. For $\alpha>0$, these filters have a non-symmetric bell-shaped envelope, i.e. a gamma-shaped window. It also implies that analyzed events correspond to those in a close past of t_0. In this case, temporal information is more on the basis of past events than on current ones. Besides, for large α the window $v(t)$ increasingly resembles a Gaussian window. These 1-D filters are then defined as:

$$d_n(t) = \left(\sqrt{n!/\Gamma(n+\alpha+1)}\right)\sigma(\sigma t)^{\alpha} e^{-\sigma t} L_n^{(\alpha)}(\sigma t) u(t) \ . \tag{7}$$

where Γ is the gamma function [1]. The generalized Laguerre polynomials $L_n^{(\alpha)}(t)$, which are orthogonal with respect to the weighting function $t^{\alpha} \cdot e^{-t}$, are defined by Rodrigues' formula [3] as:

$$L_n^{(\alpha)}(t) = \frac{t^{-\alpha} e^{t}}{n!} \frac{d^n}{dt^n}\left(t^{n+\alpha} e^{-t}\right) \ . \tag{8}$$

From (8) one can see that generalized Laguerre filters are related to time derivatives of the localizing window $v(t)$ defined in (6). Hence, filters of increasing order analyze successively higher frequencies or temporal variations in the signal.

3.2 Meixner Filters

Meixner filters are the discrete equivalent of generalized Laguerre filters. They are equal to Meixner polynomials multiplied by a square window $v^2(x) = c^x (b)_x / x!$, which is the discrete counterpart of a gamma window and it behaves similarly to a Poisson kernel [4]. $(b)_x$ is the Pochhammer symbol defined by $(b)_0 = 1$ and $(b)_x = b(b+1)(b+2)\ldots(b+x-1)$, $x=1,2,\ldots$. Parameters b and c are equivalent to parameters of the generalized Laguerre filters α and σ, respectively. However, for the discrete case, $b>0$ and $0<c<1$. The equivalent time-invariant case ($\alpha=0$) arises for $b=1$. The Meixner polynomials are orthonormal with respect to this window and they are defined as [3]:

$$M_n(x) = \sqrt{\frac{c^n (1-c)^b}{n!(b)_n}} \sum_{\tau=0}^{n} C_n^\tau (b+\tau)_{n-\tau} (x-\tau+1)_\tau (1-1/c)^\tau \quad . \tag{9}$$

for $x=0,1,2,\ldots,\infty$ and $n=0,\ldots,D$. In general, one fixes the maximum value of x to N.

In order to achieve fast computations, we present a normalized recurrence relation to compute these filters:

$$M_{n+1}(x) = \frac{1}{\sqrt{c(n+1)(n+b)}} \left[\left[(c-1)x + n + (n+b)c \right] M_n - \sqrt{nc(n+b-1)} M_{n-1} \right] \quad . \tag{10}$$

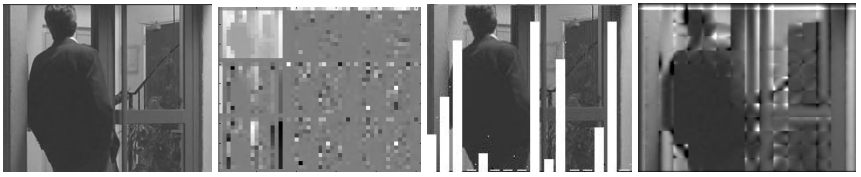
for $n \geq 0$, $M_n = M_n(x)$, and with initial conditions $M_{-1}(x) = 0$, $M_0(x) = (1-c)^{b/2}$.

4 Spatio-temporal Primitive Extraction and Indexing

In this study, features are extracted from either gray-scale or luminance-based pixel values. Let Y be the luminance of an image frame, then it can be obtained from *RGB* components by the transformation $Y = 0.299R + 0.587G + 0.114B$. Images under consideration are previously adapted to match their central part to video format 4:3.

4.1 From Spatial to Temporal Indexing: Hermite-Based Feature Extraction

We apply the 2-D separable Krawtchouk filters in (5) to each frame of a given input image sequence (digital video) in order to extract spatial primitives. Fig. 1 shows a sample frame and its Hermite decomposition. In the most right-hand there is its reconstruction from coefficients up to order 2. This image lets us see that coefficients characterize well enough primitives such as edges, lines, bars, and corners extracted in the preferred vertical and horizontal directions. A perfect image reconstruction requires all coefficient orders up until the analysis window length. However, comparisons between original and reconstructed images show that the general structure and visual information are preserved. Hermite coefficients show that essentially vertical and horizontal patterns are extracted. Patterns in oblique directions stand out too in combinations of x- and y-derivatives.



(a) (b) (c) (d)

Fig. 1. (a) *Original image* in format 4:3, with size of 384x288 (b) *Hermite coefficients* with parameters N=48, T=24, D=2 (c) *Histogram-signature* by horizontal projection (*bars are quantized values of histogram and dots are real values*) (d) *Reconstructed image* up to order 2

We have adaptively computed the analysis window length, N, in order to always have 16 and 12 histogram bins in the horizontal and vertical directions, respectively. We do that for reducing dimensionality. Besides, coefficients are subsampled by a factor $T = N/2$ so histogram bins correspond to central window positions with half overlapping. We chose the maximum order expansion as D=2 since the most relevant patterns are concentrated around these coefficients. The final signature consists of constructing horizontal and vertical histograms based on projecting values of Hermite coefficients into horizontal and vertical directions for each window position. Projection is only over first-order coefficients, (1,0) and (0,1), since they correspond to gradient-like information. Histograms are quantized with 4 bits (16 levels) to reduce signature size. The first three moments $\{\mu_0,\mu_1,\mu_2\}$ of remaining coefficients but the zero-order one (0,0) are included in the signature. Finally, a temporal characterization for video similarity matching for retrieval purposes is achieved by applying a similarity measure suitable for histogram comparisons such as the Kullback-Leibler divergence [8].

4.2 From Temporal to Spatial Indexing: Laguerre-Based Feature Extraction

Temporal events are characterized by applying the 1-D Meixner filters in (10) to time vectors of an image sequence. Each time vector corresponds to a spatial position and its length is the number of frames that are covered by the analysis window. Thus, there are as much vectors as either the number of pixels in a frame or the number of spatial positions each one integrating spatial information within its neighborhood. For this case, we use the zero-order Hermite coefficient, (0,0), presented in the previous section. There are then 16x12=192 spatial neighborhoods. This one represents a dimensionality reduction. The number of frames is equal to the analysis window length in time, N. Important care must be taken when selecting parameters b and c of Meixner filters. In general, b controls the window shape and c the number of temporal samples involved in the analysis process. The larger c the more past events are considered. A trade-off exists in choosing b and c since both of them shift the time t_0, found at window's maximum, where the sequence is estimated. Another parameter is the maximum order expansion, D, which defines the number of primitives (filters) for characterizing temporal events at time t_0. There will then be one coefficient value for each of the primitives along the temporal window. We use the set of coefficients as the set of temporal indexes. Thus, there will be $(D+1)\cdot K$ indexes, where K=192 is the number of spatial positions.

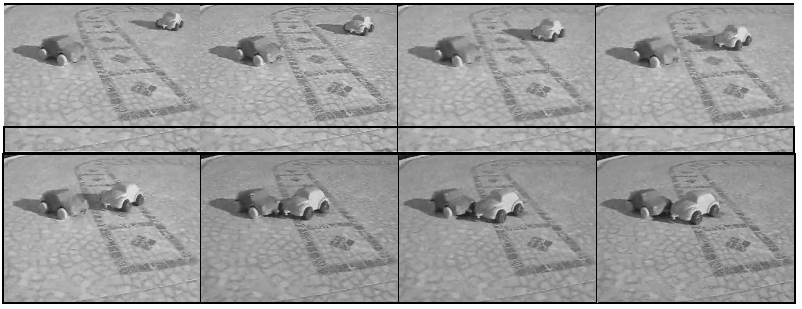


Fig. 2. Sample frames of an input image sequence containing a moving toy car (*the sequence runs line-wise from left to right: each sampled frame is one out of four*)

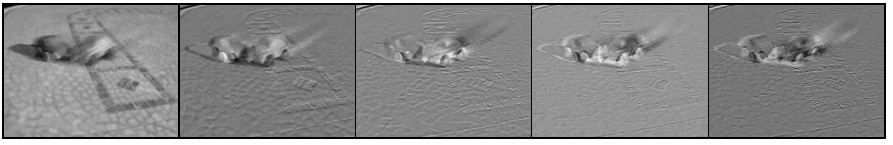


Fig. 3. *Laguerre coefficients* for all pixels of an image sequence of 30 frames (one second length) with parameters N=30 and D=4 (*the order filter goes from 0 to D=4 from left to right*)

Fig. 4. *Reconstructed samples* of sequence shown in figure 2 from their Laguerre coefficients

Experiments have shown that a good sequence interval reconstruction is achieved for N=30, D=4, b=12, and c=0.4. There are thus 960 indexes characterizing one second of temporal events. In this case, t_0=0.8s which corresponds to frame 24 for a buffer of 31 frames. Fig. 2 shows samples of an input image sequence with a moving object (a toy car). Fig. 3 shows their temporal Laguerre coefficients for all pixels. They clearly stand out moving objects and trajectories up until the last frames around t_0 and reveal temporal events from a visual perception viewpoint. Fig. 4 shows the reconstructed sequence for the same samples of fig. 2. One can see that integration over time really takes information around time t_0 (last frames) into account whereas information in the past vanishes. Spatial information in this approach is included by the spatial arrangement of Laguerre coefficients (fig. 3). A complete image sequence is indexed by applying repeatedly the described process (i..e. by convolution) and keeping coefficients. A query could be any sub-sequence lasting one or more seconds. For retrieval purposes, a similarity measure based on a correlation distance, $1-|\rho|$,

where ρ is a normalized correlation coefficient in [-1;1], has been applied between two coefficient sets.

5 Conclusion

We presented two different approaches for video indexing. The first one extracts *spatial* (primitive) features applying the Hermite transform to each frame. Extension to *time* is then achieved by the ordered sequence in time of such features. The second one extracts *temporal* (event) features applying the Laguerre transform to time vectors at pixel or spatial region positions. *Spatial* extension is then achieved by the spatial arrangement of such features. We highlight the difference of extracting spatial and temporal features by the two approaches. Construction of the final spatio-temporal signature based on the previous early vision filters considers the two approaches at the same time. Dimensionality reduction in the spatial domain is achieved by considering histograms of primitive projections into the cartesian axes and moments. In the temporal domain this one is achieved by integrating temporal events over a time support and keeping weights of the first primitives which characterize such events.

References

1. den Brinker, A.C., Roufs, J.A.J.: Evidence for a Generalized Laguerre Transform of Temporal Events by the Visual System. Biological Cybernetics. Vol. 67 (1992) 395-402
2. Flickner, M. et al.: Query by Image and Video Content: The QBIC System. IEEE Computer. Vol. 28 9 (1995) 23-32
3. Koekoek, R., Swarttouw, R.F.: The Askey-scheme of Hypergeometric Orthogonal Polynomials and its q-Analogue. Delft University of Technology, Faculty of Information Technology and Systems, Department of Technical Mathematics and Informatics. Report 98-17 (1998)
4. Lindeberg, T., Fagerström, D.: Scale-Space with Causal Time Direction. In: Proc. European Conference on Computer Vision. Vol. 1064 (1996) 229-240
5. Martens, J.-B.: The Hermite Transform – Theory. IEEE Trans. Acoust. Speech Signal Processing. Vol. 38 9 (1990) 1595-1606
6. Palmer, S.E.: Vision Science. Photons to Phenomenology. The MIT Press, Cambridge, Massachusetts. (1999)
7. Porat, M., Zeevi, Y.Y.: The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision. IEEE Trans. Pattern Analysis Mach. Intell. Vol. 10 (1988) 452-468
8. Puzicha, J., Hofmann, T., Buhmann, J.M.: Non-Parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (1997) 267-272
9. Rivero-Moreno, C.J., Bres, S.: Conditions of Similarity between Hermite and Gabor Filters as Models of the Human Visual System. In: Petkov, N., Westenberg, M.A. (eds.): Computer Analysis of Images and Patterns. Lectures Notes in Computer Science, Vol. 2756. Springer-Verlag, Berlin Heidelberg (2003) 762-769
10. Young, R.A., Lesperance, R.M., Meyer, W.W.: The Gaussian Derivative Model for Spatial-Temporal Vision: I. Cortical Model. Spatial Vision. Vol. 14 3,4 (2001) 261-319

Non-parametric Performance Comparison in Pictorial Query by Content Systems

Sergio Domínguez

Universidad Politecnica de Madrid - DISAM (http://www.disam.upm.es/vision)

Abstract. In this paper the author addresses the problem of performance comparison between CBIR systems. The main objective is the introduction of a new non-parametric method for this task. This method has different advantages that are herein explained, since performance evaluation can be stablished on a query-by-query or an averaged basis; it is straigthforwardly computed; can be easily shown in a graphic, including information for different sizes of the retrieval set; and can be easily interpreted, yielding a fast and clear idea of the comparative performance of the CBIR systems being analyzed. Moreover, it can overcome the drawbacks of the precision vs. recall graphics qualitative comparison.

1 Introduction

Performance evaluation is one of the open issues CBIR community is actively researching, in order to find efficient and meaningful methods to set an assessment on the quality of evaluated systems. Major efforts focus on defining benchmarking sets, using TREC as the leading reference. This is, undoubtedly, one of the most important fields of research on the present stage of CBIR technology, since this community lacks of a common ground truth to set the measures of performance for a given system, leading to some degree of uncertainty in the evaluation of the actual behaviour of the system. Even having a common benchmarking framework, some problems would arise, like the ground truth acquisition. The solution to these problems are clearly guided by TREC methodologies, as addressed in some previous works like [6], [4], [5].

Supposing that the common testbed is defined, the next problem to deal with is the definition of a commonly accepted measure for system performance. The present state of the art in this field is inherited almost exclusively from IR background knowledge. Most commonly accepted methods for performance evaluation are the precision and recall based measures, like precision vs. recall graphs, or precision measures for different sizes of the retrieved set. Nevertheless, all these methods have drawbacks, avoiding CBIR community to assume them (or part of them) as a valid way to set a clear evaluation of performance. This problems have been already addresed in [4], [1], [5], [6].

Once the definition of the performance measure is fixed, there are still some remaining problems, being one of them the comparison between the behaviour of two systems. Once again, the methodologies being presently used are inherited

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 833–839, 2004.
© Springer-Verlag Berlin Heidelberg 2004

from the IR community. A very good review on performance comparison methods can be found in [1].

In Section 2, a brief review on the state of the art of performance comparison techniques is introduced. Section 3 presents the new method for comparing performances of different CBIR systems. Section 4 introduces some examples on how to use this measure. Finally, Section 5 summarizes some conslusions about this work.

2 Performance Comparison Techniques

The most common way to set performance comparisons among different retrieval systems is to apply some matched pair test to the individual performance measures taken from each of them in performing a retrieval test. Typically, these comparison tests have been divided into parametrical and non-parametrical; t-test lays in the first group and Wilcoxon and sign tests in the second [1].

All these methods are based on the same departure information: performance scores of the compared systems under the same retrieval test. These values are substracted, and the null hypothesis is tested, i.e. whether the average of these differences is null or not. The conclusion is the confirmation of this hypothesis, the systems under comparison have not significant differences in performance, or its rejection, meaning that one of them outperforms the other.

Being this a good methodology for setting global assesments on performance comparisons, has several disadvantages that suggest the need for more detailed conclusions in comparisons. Summarizing them:

- The original performance score for measuring the individual performance must be single valued in order allow the definition of differences. However, no single valued measure reliably resembles the real performance of a retrieval system. In fact PR curves is the most common and preferred method for such evaluations [5].
- Even in the case that PR curves could fit into these matched pair tests, different authors have stated that this measure does not depict clearly the actual performance of a CBIR system, adding another factor of uncertainty [4][1].
- It is hard to envisage the actual difference between two system performances given just a single value (the result of the matched pairs test) to justify this decision. Frequently any user would like to have more detailed information, just like PR graphics offers more than a single PR pair for a given test and a given retrieval size. This fact points to the necessity of having comparison graphs, showing how the comparison evolves as retrieval conditions change.

Given all these requirements, the author introduces a new non-parametric technique for evaluating performance comparisons between two CBIR systems, being explained in the following section.

3 Retrieved Set Based Performance Comparison

3.1 Binary Set Operators Definition

Let $X = \{x_1, x_2, \ldots, x_k, \ldots, x_n\}$ be a set such that $x_i \in \{0, 1\}$, i.e. each x_i is a binary variable. Given this set, a cardinality measure can be established as:

$$R_k(X) = \frac{1}{k} \sum_{i=1}^{k} x_i \tag{1}$$

such that $R_k(X)$ computes the proportion of elements of X having a value of 1 among the first k elements of X.

Let $Y = \{y_1, y_2, \ldots, y_k, \ldots, y_n\}$ be a set such that $y_i \in \{0, 1\}$. Given these two sets, the following operators can be defined:

Greater than

$$\{X > Y\} = \{g_1, g_2, \ldots, g_k, \ldots, g_n\} \tag{2}$$
$$g_i = \begin{cases} 1 \text{ if } R_i(X) > R_i(Y) \\ 0 \text{ otherwise} \end{cases}$$

Equals to

$$\{X = Y\} = \{e_1, e_2, \ldots, e_k, \ldots, e_n\} \tag{3}$$
$$e_i = \begin{cases} 1 \text{ if } R_i(X) = R_i(Y) \\ 0 \text{ otherwise} \end{cases}$$

Lower than

$$\{X < Y\} = \{l_1, l_2, \ldots, l_k, \ldots, l_n\} \tag{4}$$
$$l_i = \begin{cases} 1 \text{ if } R_i(X) < R_i(Y) \\ 0 \text{ otherwise} \end{cases}$$

Given these definitions, it can be easily derived that:

- $\{X \geq Y\} = \{X > Y\} + \{X = Y\}$
- $\{X \leq Y\} = \{X < Y\} + \{X = Y\}$

Appliying the cardinality function $R_k(.)$ to these comparison sets, can be easily stablished as well:

$$R_k(\{X > Y\}) + R_k(\{X = Y\}) + R_k(\{X < Y\}) = 1 \tag{5}$$

and therefore:

$$R_k(\{X > Y\}) + R_k(\{X = Y\}) = 1 - R_k(\{X < Y\}) \tag{6}$$

3.2 Precision and Recall

Once this operators and functions have been defined, let Q^A be an ordered binary set representing the relevance of the results of a query Q performed over the retrieval system A, such that $Q^A = \{q_1^A, q_2^A, \ldots, q_k^A, \ldots, q_n^A\}$, where each element q_i^A is defined as:

$$q_i^A = \begin{cases} 1 \text{ if } i\text{-th retrieved element is relevant to the query} \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

such that q_i^A is less similar to the query as i grows up.

Henceforth it is quite clear that precision can be obtained from this set by applying $R_n(Q^A)$:

$$Precision = R_n(Q^A) = \frac{1}{n}\sum_{i=1}^{n} q_i^A = \frac{\text{Number of relevant elements}}{\text{Number of retrieved elements}} \tag{8}$$

As well as recall can be obtained as:

$$Recall = \frac{n}{N} R_n(Q^A) = \frac{1}{N}\sum_{i=1}^{n} q_i^A = \frac{\text{Number of relevant elements}}{\text{Number of relevant elements}} \tag{9}$$

where n is the size of the retrieved set, and N is the number of relevant elements in the database.

3.3 Performance Comparison Measure

Let Q^B be a new binary set, representing the results of the retrieval system B for the same query Q submited to the system A in the former case, $Q^B = \{q_1^B, q_2^B, \ldots, q_k^B, \ldots, q_n^B\}$, where each q_i^B is defined using the same rule as for q_i^A.

Given these conditions, the author proposes as a measure for comparing the performance of retrieval systems A and B the evaluation of the following pair:

$$[R_n(\{Q^A > Q^B\}), R_n(\{Q^A = Q^B\})] \tag{10}$$

Note that these two numbers are completely equivalent to stablish a comparison between precision and recall of the two systems, since the proposed measure is based upon the number of relevant retrieved elements, the same as precision and recall. Given the fact that they comprise the information of two systems working on the same database, the number of relevant elements included in it are the same for both systems. Of course the size of the retrieved sets are the same, and therefore, in terms of performance comparison, the amount of relevant elements in the retrieved sets is the only difference for establishing a valid performance comparison.

This pair of values clearly reflects every posibility for the results of comparison:

– A is outperforming B in the retrieval:

$$R_n(\{Q^A > Q^B\}) > R_n(\{Q^A = Q^B\}) \tag{11}$$

and

$$R_n(\{Q^A > Q^B\}) + R_n(\{Q^A = Q^B\}) > R_n(\{Q^A < Q^B\}) \tag{12}$$

– if they have a similar performance:

$$R_n(\{Q^A > Q^B\}) < R_n(\{Q^A = Q^B\}) \tag{13}$$

and

$$R_n(\{Q^A > Q^B\}) \approx R_n(\{Q^A < Q^B\}) \tag{14}$$

– or if B is better than A

$$R_n(\{Q^A > Q^B\}) < R_n(\{Q^A = Q^B\}) \tag{15}$$

and

$$R_n(\{Q^A > Q^B\}) + R_n(\{Q^A = Q^B\}) < R_n(\{Q^A < Q^B\}) \tag{16}$$

This measure can also be averaged, as precision and recall can be, in order to extend its meaning from a single retrieval experiment to a batch:

$$[\overline{R_n}(\{Q^A > Q^B\}), \overline{R_n}(\{Q^A = Q^B\})] \tag{17}$$

with all the properties remaining as explained above.

Some examples of comparison and their interpretation are:

– $[1, 0]$: A has a better performance than B for any size of the retrieval set
– $[0, 1]$: A and B have the same performance for any size of the retrieved set
– $[0, 0]$: A has a worst performance than B for any size of the retrieved set
– $[a, 1-a]$: A has never a worst performance than B for any size of the retrieved set

Of course this measure of comparison can be represented in a graphical form, by simply drawing two curves containing pairs $(k, R_k(\{Q^A > Q^B\}))$ and $(k, R_k(\{Q^A = Q^B\}))$, $\forall k = 1, \ldots, n$. This graphical representation has the advantage of showing the evolution of this comparison as the retrieved set grows up in the number of elements taken into account, yielding a result comparable in scope to the graphics of precision and recall evolution.

In figures 1 and 2 this graphical representation can be observed for a performance comparison between two retrieval systems. The CBIR systems under comparison are two shape-based retrieval algorithms searching in a trademark database of approximately 8300 images. The first one uses a contour based representation of shapes, while the second uses moment based invariants.

Although results in Figure 1 are clearly depicted, they could be even clearer represented in the way shown in Figure 2, where series $R_k(\{Q^A > Q^B\})$ and $R_k(\{Q^A \geq Q^B\})$ are shown; in this way, for a given k equality result can be found as the overhead of superiority result, and the rest to 1 is the inferiority result.

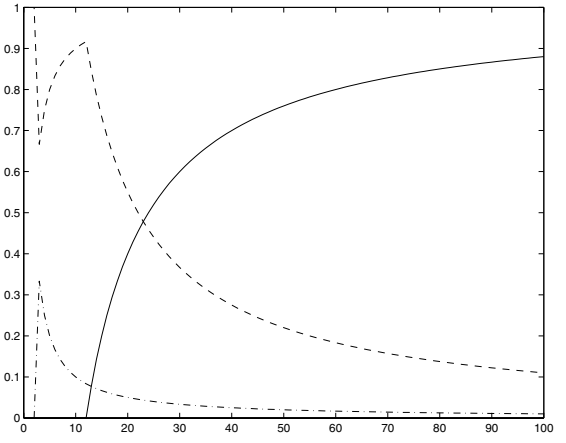


Fig. 1. In this figure, the result of the comparison between two retrieval systems for the same query is depicted. Three curves can be observed: in continuous line, the series of $R_k(\{Q^A > Q^B\}$; in dashed line, the series of $R_k(\{Q^A = Q^B\}$; finally in dot-dash line $R_k(\{Q^A < Q^B\})$, althought it would be not necessary given the relationship among the three series values for each index

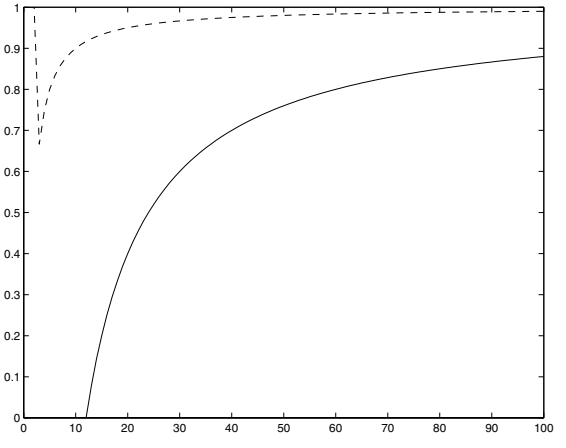


Fig. 2. In this figure, the result of the comparison between two retrieval systems for the same query is depicted. Two curves can be observed: in continuous line, the series of $R_k(\{Q^A > Q^B\}$; in dashed line, the series of $R_k(\{Q^A \geq Q^B\}$; finally, for a given value of k, $R_k(\{Q^A < Q^B\})$, can be found as the difference between this last curve and 1

Results depicted in these figures can be interpreted as follows: for a small retrieved set, more precisely up to twelve elements, system B seems to be slightly better than A, although it can be stated that their performance is almost the same. However, as the retrieved set increases, system A gets clearly better than

A, being clearly more powerful for big retrieved set (up to one hundred elements in the figure).

4 Conclusions

In this paper, a new non-parametric method for establishing performance comparison between two CBIR systems has been introduced. This method has the ability of overcoming some of the drawbacks of traditional matched pairs tests which have been used for this sake; summarizing these advantages:

- The proposed comparison method is not based upon any traditional performance measure or score; since none of them are commonly accepted as complete, the author proposes to use the retrieval set as the basis for setting the comparison, being this a direct method, in the sense that no intermediate score is needed.
- The comparison results can be easily depicted on a graph with a straightforward and clear interpretation results. This graph offers quantitative information, not qualitative, being the evaluation of performance differences on the observer's side, and therefore leaving open the final decision depending on each particular necessities.
- The comparison graph depicts a detailed comparison procedure under different conditions, namely different retrieval set sizes. Therefore, some traditional performance measures, like precision and recall, can be also easily compared by evaluating these graphs, since, as stated in previous section, the comparison metrics is based upon the same information used to compute those values.

References

1. David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 1993.
2. Henning Muller, Wolfgang Muller, and Thierry Pun. Automated Benchmarking in Content-Based Image Retrieval. Technical Report 01.01, Universite de Geneve, Centre Universitaire D'Informatique, Groupe Vision, january 2001.
3. Henning Muller, Wolfgang Muller, and David McG. Squire. Performance Evaluation in Content-Based IMage Retrieval: Overview and Proposals. Technical Report 99.05, Universite de Geneve, Centre Universitaire D'Informatique, Groupe Vision, december 1999.
4. Henning Muller, Wolfgang Muller, David McG. Squire, Stephan Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, (22):593–601, 2001.
5. J.R. Smith. Image retrieval evaluation, 1998.
6. Liu Wenyin, Zhong Su, Stan Li, Yan-Feng Sun, and Hongjiang Zhang. A performance evaluation protocol for content-based image retrieval algorithms/systems. In *IEEE CVPR Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

Hierarchical Watersheds with Inter-pixel Boundaries

Luc Brun[1], Philippe Vautrot[1], and Fernand Meyer[2]

[1] Laboratoire d'Études et de Recherche en Informatique(EA 2618)
Université de Reims - Chemin des Rouliers
51687 Reims Cedex 2 - France
{luc.brun,philippe.vautrot}@univ-reims.fr
and
[2] Centre de Morphologie Mathématique (CMM)
35, rue Saint Honoré
77305 Fontainebleau Cedex - France
Fernand.Meyer@cmm.ensmp.fr

Abstract. Watersheds are the latest segmentation tool developed in mathematical morphology. These algorithms produce a segmentation of an image into a set of basins separated by watershed pixels. The over segmentation produced by these algorithms is reduced by removing all contours with a low saliency. This contour's saliency is generally defined from the minimal height of the watershed pixels along the contour. However, such a definition does not allow to define a contour's saliency in case of thick watersheds. Moreover, the set of basins which corresponds to the intuitive notion of regions does not define an image partition. In this paper we propose a method which allows to aggregate the watershed pixels to the basins while preserving the notion of contour and the associated saliency. The model used to encode the image partition is then decimated according to the contour saliency to obtain a hierarchy of partitions.

1 Introduction

Segmentation and contour extraction are important tasks in image analysis. Among the multitude of methods, the watershed transformation [10] introduced in the late 70's arises as a popular image segmentation algorithm. This method usually based on the gradient of the image presents the main advantage of providing closed curves leading to a proper definition of regions.

In mathematical morphology an image is considered as a topographic relief where each gray level (image intensity or image gradient) is interpreted as an altitude. The traditional watershed algorithm [10] simulates a flooding process. The minima of the image are pierced by holes and the water springing through the holes slowly immerse the whole relief. To prevent streams of water coming from different holes to intermingle, a dam is erected at the meeting locations. The set of all these obstacles represents the watersheds whereas the resulting separated lakes define the so-called basins attached to each regional minima.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 840–847, 2004.
© Springer-Verlag Berlin Heidelberg 2004

Due to the flooding process one drip of water falling from a watershed pixel into an adjacent basin should follow an always descending path until the minimum of the basin. This property may be formalized as follows :

Definition 1 (Crest line property). *For each watershed pixel adjacent to a basin there exists an always descending path from the pixel to the regional minima of the basin.*

Note that not all watershed algorithms guarantee the preservation of this property [8].

A well known drawback of watershed algorithms is the over segmentation often produced by these algorithms. Since the contours appear to be correct the over segmentation problem turns out to be equivalent to a proper valuation of the saliency of each contour. One important feature to determine the contour saliency is its pass value generally defined [8] as the minimal height of the watershed pixels along the contour. This property may also be defined from the watershed pixel's pass values:

Definition 2 (watershed pixel's pass value). *Given a watershed pixel P adjacent to two basins, its pass value is defined as the minimal altitude one has to reach to connect the two basins while passing by P.*

Najman [8] proposed to valuate each contour by the minimal difference between the contour's pass value and the maximal depth of the two basins which merge along it during a flooding process. Such a valuation is called the dynamic of the contour.

However, the minimum of the pass values along a contour is quite sensitive to the noise that may appear along it. Moreover, despite their names, the watershed contours do not always define valid connected paths. Indeed, Najman and Vincent [8,10] have shown that the existence of thick watershed areas is induced by the definition of the watersheds and can't thus be avoided. The determination of the adjacency between the basins and the valuation of the associated contours is then conditioned to a proper thinning [7] of the thick watershed areas.

In this paper we propose a method which allows to aggregate the watershed pixels to the basins while preserving the pixel's pass value information (Section 2). The contours between the basins are encoded by a model based on inter pixel boundary paths, the pixel's pass values being stored within this model (Section 3). The different paths encoding the borders of the partition are then encoded by a graph data structure and a decimation process based on the dynamic of contours is applied on the graph in order to obtain a hierarchy of image partitions (Section 4).

2 Computing the Altitude of Pixel Watershed

As mentioned in Section 1, a watershed algorithm produces a partition of an image into a set of basins $B_1, \ldots, B_p$ and a set of watershed pixels $\mathcal{W}$. In the following of this article we will consider that the 4 neighborhood is used for the

basins and that all watershed pixels adjacent to a basin satisfy the crest line property (Definition 1).

The saliency of a contour between two basins is generally measured from the contour's pass value [8] (Section 1). However, the definition of a watershed pixel's pass value (Definition 2) does not hold for thick watershed areas where many watershed pixels are adjacent to a single basin or surrounded by other watershed pixels and thus not adjacent to any basins.

Thick watersheds induce two different problems for the determination of the contour's pass values. First of all, within the watershed framework a contour corresponds to a border between two basins. In case of a thick watershed area, the basic idea which consists to consider all basins adjacent to the thick area as adjacent may not lead to a valid partition of the image into 4 connected regions. Therefore, the adjacency between the basins and thus the existence and location of the contours is relative to a labeling of the watershed pixels to the different basins. Secondly, in order to be coherent with Definition 2, the definition of the pass value on the resulting contours should correspond to the minimal altitude one has to climb to connect the two basins separated by such a contour.

We perform the labeling of watershed pixels by an aggregation process which merges all watershed pixels into a set of final basins $B'_1, \ldots, B'_p$. Let us define the altitude of a watershed pixel $P \in B'_i$ as the minimal height one as to reach from the minimum of B'_i to reach P. Note that the altitude of a watershed pixel corresponds to its pass value when the pixel is adjacent to at least two basins.

The computation of the watershed pixel's altitude requires to consider all paths joining a watershed pixel $P \in B'_i$ to the connected component m_i of the basin B_i with a minimal height. This set of paths is defined as follows :

$$\Pi_i(P) = \{\pi \subset B'_i | \pi(1) \in m_i \text{ and } \pi(q) = P\} \text{ with } q = |\pi| \tag{1}$$

The altitude of P is then defined as :

$$Alt(P) = min_{\pi \in \Pi_i(P)} max_{j \in \{1,\ldots,|\pi|\}} h(\pi(j)) \tag{2}$$

where h denotes the height of each pixel in the image.

Note that, using the crest line property, the minimum m_i of B_i may be replaced by B_i in equation 1 without changing the value of $Alt(P)$.

Using the above definitions, the altitude of a watershed pixel is equal to its height only if there is an always descending path included in the basin to which it is aggregated and which connects it to the minimum of the basin. Moreover, since a local minima defines a basin, a thick watershed area can not contain any local minima. Therefore, any watershed pixel in a thick watershed area can be connected to a basin by an always descending path. The determination of such paths is insured by Algorithm 2 which performs an immersion process on the watersheds (Fig. 1). More precisely, this algorithm performs the following steps:

1. All watershed pixels adjacent to an initial basin B_i are put into a queue,
2. while the queue is not empty
   a) One pixel with a minimal altitude is removed from the queue and merged with an adjacent basin.

b) All the watershed pixels adjacent to the removed pixel and not already in the queue are added to it.

input : watersheds $\mathcal{W}$, bassins $\{B_1, \ldots, B_p\}$, image h
$\mathcal{B} = \bigcup_{k=1}^{p} B_k$
$\mathcal{F}_0$ **: set of watershed pixels adjacent to a bassin**
$N(P)$**: 4 neighborhood of** P**.**
k=0
while $\mathcal{F}_k \neq \emptyset$
  select P_{k+1} **such that** $h(P_{k+1}) = min\{h(P), P \in \mathcal{F}_k\}$
  $\mathcal{F}_{k+1} = (\mathcal{F}_k - \{P_{k+1}\}) \cup [N(P_{k+1}) \cap (\mathcal{W} - \mathcal{F}_k)]$
  select j **such that** $N(P_{k+1}) \cap B_j^{(k)} \neq \emptyset$
  $B_j^{(k+1)} = B_j^{(k)} \cup \{P_{k+1}\}, \ \forall i \neq j \ B_i^{(k+1)} = B_i^k$
  $\mathcal{W} = \mathcal{W} - \{P_{k+1}\}$
  k=k+1

Algorithm 1: Immersion of watershed pixels

Each watershed pixel is either adjacent to a basin or put into the queue by an adjacent watershed pixel. This adjacent watershed pixel being merged into a basin, each pixel removed from the queue is adjacent to at least one basin. Moreover, the algorithm removes at each step one pixel from the watershed set. This set being finite the algorithm terminates. Finally, once one watershed pixel is processed all its neighbors must be processed since they are put into the queue which is empty when the algorithm terminates. Therefore, if we suppose that one watershed pixel is not processed by the algorithm we must suppose that the whole connected component of watershed pixels including it is not processed. This last assumption contradict the fact that the image is partitioned into the set of watersheds and initial basins. Indeed, all pixels adjacent to a basin are initially put into the queue and thus processed. Therefore the algorithm terminates and proceeds all watershed pixels.

The proof that we can find for each watershed pixel P an always descending path included in the final basin to which P is aggregated and connecting P to the minimum of this basin may be established using a recurrence hypothesis. The basic idea of the proof is as follows: If we suppose that at step k all previously dequeued pixels satisfy the property, the pixel P_{k+1} should have a greater height than these pixels (otherwise it would have been dequeued before them). Therefore, P_{k+1} is either directly adjacent to a basin and the proof is provided by the crest line property (Definition 1) or it is surrounded by watershed pixels and adjacent to a previously dequeued one. In this case we can concatenate P_{k+1} to the path associated to one of its dequeued neighbor. The resulting path is always descending by construction.

The queue may be efficiently implemented by an array of lists. In this case each watershed pixel is considered twice: once to be put into the queue and once to be removed from it. Note that at each step of Algorithm 2, several pixels with a minimal altitude may belong to the queue. The order defined on such pixels

2	3	6	2
3	6	6	6
255	7	6	6
2	255	7	6
1	2	255	5

(a) initial image

2	3	6	2
3	6	6	6
255	7	6	6
2	255	7	6
1	2	255	5

(b) watersheds

A_2	A_3	A_6	A_2
A_3	A_6	A_6	B_6
C_{255}	A_7	B_6	B_6
C_2	C_{255}	D_7	D_6
C_1	C_2	C_{255}	D_5

(c) partition

Fig. 1. A thick watershed(b) produced by the initial image (a) and the resulting image partition produced by Algorithm2(c). The indexes in (c) denote the height of the pixels.

influences the growing speed of the different basins and thus the final partition. A priority may be defined on these pixels based either on a geodesic distance between the basins or an external criteria such as a distance between the feature vectors of the watershed pixels and the ones of the basins. Note however, that the priority that may be established on the queue has no influence on the altitudes of the watershed pixels along the final contours.

3 Transferring Pixel's Altitudes to Lignels

Algorithm 2(Section 2) aggregates each watershed pixel to a basin and provides thus a partition of the image into a set of final basins $B'_1, \ldots, B'_p$. One may think to define the border between two basins B'_1 and B'_2 as the set of watersheds belonging to B'_1 (respectively B'_2) and having one neighbor in B'_2 (respectively B'_1). However, assuming a 4 connectedness for the basins, such a definition of borders does not defines valid 8-connected contours between regions. Indeed, within thick areas two adjacent watershed pixels may belong to different basins. The border between both basins is in this case two pixel large.

This last drawback may be overcome by defining basins's boundaries as *inter-pixel boundary paths* [5,4]. Using such a representation, the boundaries of the basins are defined as 4-connected paths in the $P_{\frac{1}{2}}$ plane (Figure 2(b)):

$$P_{\frac{1}{2}} = \{(i + \tfrac{1}{2}, j + \tfrac{1}{2}), \text{ with } (i, j) \in \mathbb{Z}^2\}.$$

This approach has been first described by Brice and Fennema [1] when introducing grouping segmentation algorithms. Later several discrete topologies have been developed which provide formal tools to study such a representation [5,4].

Two adjacent successive half integer points along an inter-pixel boundary path are said to be joined by a *lignel* [4] (also denoted crack [9]). Each lignel joins two half integer points along a boundary path between two basins and separates two pixels belonging to these two basins (Figure 2(a) and (b)). The set of lignels encoding the borders of the partition is denoted by $\mathcal{L}$.

The definition of the watershed pixel's pass value (Definition 2) may thus be extended to lignels as follows:

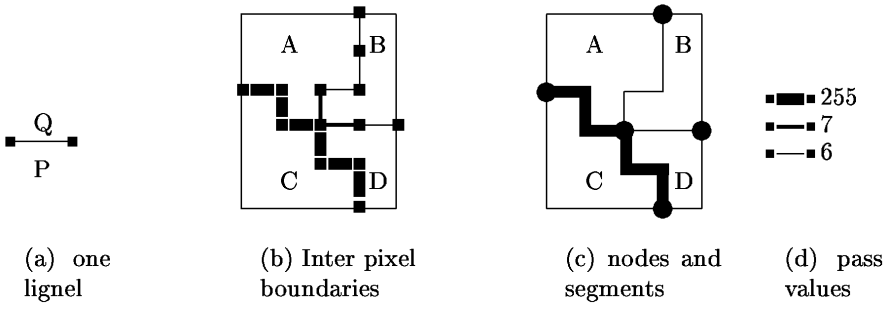


Fig. 2. One lignel element (a) and the encoding of the image partition (Fig. 1(b)) by inter pixel boundary paths(b). The symbols ■ denote the half integer points encoding the borders of the partition. The width of the lines in (d) represent the values attached to lignels(b) and segments(c).

Definition 3 (Lignel's pass value). *Given one lignel $l \in \mathcal{L}$ between two pixels P and Q belonging to two different basins, B'_i, B'_j the pass value of l is defined as the minimal altitude one as to reach to connect B'_i and B'_j while passing by P and Q.*

Given two basins, B'_1 and B'_2 and a lignel l between two watershed pixels P and Q belonging respectively to B'_1 and B'_2, the pass value of l may be formally defined using the set of paths joining B'_1 and B'_2 and passing by P and Q :

$$\begin{aligned} \Pi(P,Q) = \quad & \{\pi \subset B'_1 \cup B'_2 | \pi(0) \in B_1, \pi(q) \in B_2 \text{ and} \\ & \exists j \in \{1, \ldots, q-1\} \text{ such that } (\pi(j), \pi(j+1)) = (P,Q)\} \end{aligned}$$

with $q = |\pi|$.

The pass value of l is then defined as in Section 2 by:

$$pass_value(l) = min_{\pi \in \Pi(P,Q)} max_{j \in \{1,\ldots,|\pi|\}} h(\pi(j))$$

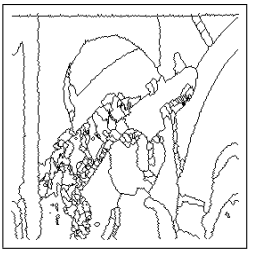
Let us suppose that P and Q are both watershed pixels. Using the image partition provided by Algorithm 2, there is two always descending path π_1 and π_2 respectively from P to B'_1 and from Q to B'_2. If we denote by π_2^{-1} the reverse path of π_2, the path $\pi = \pi_1.\pi_2^{-1}$ belongs to $\Pi(P,Q)$ and has an altitude equal to $max(h(P), h(Q))$. Since all paths in $\Pi(P,Q)$ must pass by P and Q, π has a minimal altitude. If P (resp. Q) is not a watershed pixel we have by the crest line property (Definition 1) $h(Q) > h(P)$ (resp. $h(P) > h(Q)$) and in this case the pass value of l is equal to the height of Q (resp. P). Therefore, in all cases the pass value of $l \in \mathcal{L}$ is equal to the maximal height of P and Q:

$$\forall l \in \mathcal{L} \quad pass_value(l) = max(h(P), h(Q))$$

where l separates the pixels P and Q.

(a) initial image

(b) segment's dynamics



(c) dynamics's thresholding

Fig. 3. The dynamic of the segments (b) deduced from the watersheds of image (a) and one level of the hierarchy where all contours with a dynamic lower than 16 have been removed(c)

4 Computation of Contour's Dynamics

The methods described in Sections 2 and 3 allows us to encode the boundaries of the partition by a set of inter-pixel boundaries points joined by lignel elements. Each lignel is valuated by its pass value (Definition 3).

This set of lignels may be structured as follows: We call a *segment* a maximal path between two basins (Figure 2(c)). The intersection between several segments is called a *node*. Moreover, in order to preserve the topology consistency it is necessary to add an arbitrary node on each boundary reduced to a single loop. It can be shown [3] that a closed single path in the $P_{\frac{1}{2}}$ plane satisfies Jordan's theorem [5]. Therefore, any connected path joining two pixels respectively inside and outside a basin must cross the basin's boundary. This last property allows us to define implicitly each basin by a closed segment or by the concatenation of the segments and nodes which belong to its boundary.

Each segment is thus composed of a sequence of valuated lignels encoding the different pass values along the segment. The pass value of each segment should thus be defined from the ones of its lignels. In our implementation, the segment's pass value is fixed to the median value of its lignel's pass values. Note that this value is generally fixed to the minimal altitude along the contour which roughly corresponds in our model to the minimal pass value of the lignels along the segment. Our choice of a median value prevents the pass value of the contour to be artificially lowered by the presence of noise along the contour.

The set of nodes and segments can then be encoded with a graph by associating nodes and segments respectively with vertices and edges. The pass value of each segment is thus attached to the associated edge. The basins are then defined as the faces of the graph and may be encoded by the vertices of the dual graph. The initial graph and its dual may be encoded using either the dual graph [6] or combinatorial map [2] models. We choose to use the combinatorial map model which allows us to encode only implicitly the structure of the dual graph.

Given the edge's pass values, the dynamic of the edges are computed [8] (Section 1). A hierarchy of segmentations is then built on the edge's dynamics by removing iteratively the edges with the lowest dynamics.

Figure 3 shows the dynamic of contours (b) computed on the Lenna test Image(a) using a Deriche gradient operator as initial image for the watersheds algorithm [10]. The partition obtained by removing all edges with a dynamic lower than 16 is shown in Figure 3(c).

5 Conclusion

We have defined in this article a construction scheme for hierarchical watersheds based on inter-pixel boundaries. This representation allows to remove the ambiguities induced by the presence of thick watershed areas without loss of information. Moreover, our model may be used in conjunction with any watershed algorithm satisfying the crest line property. Our model encodes a partition of the image into 4-connected basin. The topological soundness of this model and its theoretical background allows us to encode it efficiently using combinatorial maps. Our future work will consist to enhance the definition of the segment's pass value based on the values of its lignels. Such results have a direct influence on the value of the dynamics and thus on the final hierarchy.

References

1. R. Brice and C. Fennema. Scene analysis using regions. *Artificial intelligence*, 1:205–226, 1970.
2. L. Brun and W. Kropatsch. Combinatorial pyramids. In Suvisoft, editor, *IEEE International conference on Image Processing (ICIP)*, volume II, pages 33–37, Barcelona, September 2003. IEEE.
3. J. P. Domenger. *Conception et implémentation du noyau graphique d'un environnement 2D$\frac{1}{2}$ d'édition d'images discrètes*. PhD thesis, Labri Université Bordeaux I, 351 cours de la libération 33405 Talence, avril 1992.
4. J. Françon. Topologie de Khalimsky et Kovalevsky et algorithmes graphiques. In *DGCI'91.*, Strasbourg, September 1991.
5. E. Khalimsky, R. Kopperman, and P. Meyer. Boundaries in digital planes. *Journal of applied Math. and Stocastic Analysis*, 3:27–55, 1990.
6. W. G. Kropatsch and H. Macho. Finding the structure of connected components using dual irregular pyramids. In *Cinquième Colloque DGCI*, pages 147–158. LLAIC1, Université d'Auvergne, ISBN 2-87663-040-0, September 1995.
7. J. Marchadier, D. Arquès, and S. Michelin. Thinning grayscale well-composed images. *Pattern Recognition Letters*, 25:581–590, 2004.
8. L. Najman and M. Couprie. Watershed algorithms and contrast preservation. In *DGCI'2003*, volume 2886, pages 62–71. LNCS, Springer Verlag, 2003.
9. A. Rosenfeld. Digital topology. *Amer. math. monthly*, 86:621–630, 1979.
10. L. Vincent and P. Soille. Watersheds in digital spaces : an efficient algorithm based on immersion simulations. *IEEETPAMI*, 13(6):583–598, 1991.

From Min Tree to Watershed Lake Tree: Theory and Implementation

Xiaoqiang Huang, Mark Fisher, and Yanong Zhu

University of East Anglia, Norwich, NR4 7TJ, UK
xq.huang@uea.ac.uk, {mhf,yz}@cmp.uea.ac.uk

Abstract. Segmentation is a classical problem in image processing that has been an active research topic for more than three decades. Classical tools provided by mathematical morphology for segmenting images are the connected set operators and the watershed transformation. Both of these operations can be applied to form hierarchies of nested partitions at increasing scales. This paper studies two image partition hierarchies founded in mathematical morphology, namely the max/min tree and the watershed lake tree. By considering watershed and max/min tree image descriptions we show that a watershed lake tree comprises a subset of min tree vertices.

1 Introduction

Mathematical morphology has provided a powerful set of nonlinear image analysis tools since its advent. Its basic theory was formulated by the early work done by Matheron [12] and Serra [20,21] between 1970s and early 1980s. Mathematical morphology has been applied widely to applications in materials science, microscopic imaging, pattern recognition, medical imaging and computer vision.

The watershed transformation [2] and connected set operators [19] are two important classes of morphological operators. The watershed transformation is usually applied to the gradient image to provide semantically meaningful image partitions bounded by edges. The most widely used implementation of the watershed transformation is based on an immersion simulation proposed by L. Vincent and P. Soille [25] in which the topography of an image is represented as comprising catchment basins separated by watersheds. The algorithm simulates the effect of piercing a hole in the minima of each catchment basin and gradually immersing the surface in a bath of water. The water progressively floods each catchment basin and where water from adjacent catchment basins merge a 'dam' is constructed to separate them. Unfortunately, when the watershed transformation is used directly as a segmentation tool it tends to produce an over-segmentation of the image and it is difficult to identify semantically meaningful image partitions. In order to suppress the over-segmentation problem, a number of techniques have been proposed. Meyer et. al. suggested the use of markers [14] to roughly locate the objects of interest before the watershed transformation is applied. Grimaud [8] came up with the idea of valuating the minima

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 848–857, 2004.
© Springer-Verlag Berlin Heidelberg 2004

on a contrast criterion he referred to as the 'dynamics'. A simple thresholding on the dynamics can then to be applied to select minima that correspond to structures of more interest. Najman and Schmitt [15] consider the problem of watershed saliency by weighting the frontier between each catchment basin by the altitude of the lowest pixel between them. Suppressing all frontiers (dams) below some threshold λ produces a coarser partition. Cichosz and Meyer [5] obtained a hierarchy of nested partitions called a critical lake tree by incrementally increasing a parameter λ (where λ is a function of the 'area', 'depth' and 'volume' of the catchment basin) and Fisher and Aldridge [7] produced a watershed tree directly by tracking the immersion process.

Connected set operators provide another morphological image decomposition based on flat zones [19] (largest connected set of the space where the image is constant). When applied iteratively at increasing scales this simplifies the image and preserves the scale causality. The first pair of connected set operators were called opening and closing by reconstruction. A further development of the connected operator was the introduction of two adaptive morphological operators [4], namely NOP (new opening operator) and NCP (new closing operator). They only remove details consisting of fewer pixels than a given number λ, while preserving the other details. L. Vincent proposed a much faster implementation [23] of NOP and NCP operators later and he coined the concept of area opening and closing to replace NOP and NCP. Since then, the theory of connected operators has attracted a lot of research attention and has been rapidly advancing [19,18,6,3,9,22].

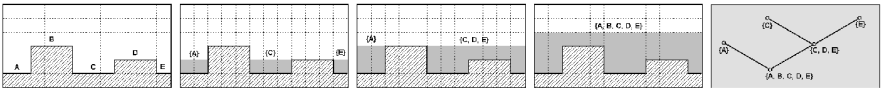
When the watershed and connected set operators are considered under the framework of scale space, either of them generates a pyramid tree structure of image regions. The watershed lake tree [5,7] is generated from the watershed transformation and the max/min tree [18] is developed by applying area opening/closing operators. These two tree structures have been proposed for solving problems of object extraction motivated by the object based image compression standard, MPEG-4. MPEG-4 targets not only large coding gains but also object oriented coding. The advent of the MPEG-4 standard has encouraged many researchers [7,5,17,18,1] to create more powerful tools for image analysis that are capable of performing fast and robust segmentation, with minimal interaction by the user.

Both the watershed lake tree and the max/min tree structures have been studied for image segmentation but surprisingly they have been considered separately by researchers. In this paper we are interested in making a comparison between the watershed lake tree and the max/min tree. This paper presents a clear description of the similarities and differences between these two tree structures. It also shows that a watershed lake tree structure is actually a subset of a min tree structure generated from a same image. Once a min tree of an image is constructed, its corresponding watershed lake tree can be obtained by pruning the vertices. This comparison work will benefit those who are interested in image segmentation and object extraction tools founded in tree based image representations.

This paper is organised as follows: section 2 describes how to generate a watershed lake tree from an image. Section 3 presents definitions of the terms related to the max/min tree; section 4 compares the watershed lake tree structure with the min tree structure. The implementation strategy for transforming a min tree into a watershed lake tree is also presented in this section. Conclusions are drawn in the final section.

2 From Watershed to Watershed Lake Tree

We can use the immersion simulation to generate a watershed tree [5,7] if we imagine that, instead of erecting a 'dam' to separate floods emerging from adjacent catchment basins the waters are allowed to mix and form larger pools. Thus, at an immersion increment before a mixing event we have two pools (child nodes), which are subsumed, an increment later, into a larger pool (father node). The process continues by progressively flooding deeper catchment basins until finally, at the end of the immersion process, a root node is formed to represent the remaining pool containing all of the image pixels.



(a) (b) (c) (d) (e)

Fig. 1. Generating a watershed tree from an example 1-D signal.

A simulation generating a watershed tree from an example 1-D signal is shown in Figure 1. Note letters in the figure represent the support regions of the flat zones of which the signal is made. A node in the watershed lake tree is associated with three parameters considered for the characterisation of the lakes: depth, area, and volume. Based on the labelled watershed lake tree, a family of connected filters can be constructed. Nodes whose parameters fall below a given threshold level are pruned and their corresponding minima are flooded. This reconstruction applied to the original image allows one to filter the regions which do not fulfil a given condition/criterion, i. e. when the depth parameter is considered, lightly contrasted regions are subsumed; when the area is taken into account, small regions are merged with their neighbours. In this way a set of segmentations can be obtained by pruning operations performed on a given watershed lake tree. Experiments performed in [5] show that amongst the three properties proposed, 'volume' best describes salient regions in terms of psychovisual perception.

3 From Area Opening/Closing to Max/Min Tree

An image or function f is considered as a mapping from a finite rectangular subset E, the underlying grid, of the discrete plane Z^2 into a discrete set {0, 1, ..., NG} of gray-levels, where NG represents the maximum possible gray value (usually NG = 255). A binary image is often regarded as the set $X \subseteq E$ and can only take values 0 or 1.

Definition: Level Sets. Let E_h and E^h denote the upper level set and lower level set, respectively, resulting from thresholding the function f at gray value h. Upper level sets and lower level sets of gray scale image are defined below:

$$E_h = \{x \in E \mid f(x) \geq h\}, \quad E^h = \{x \in E \mid f(x) < h\} \tag{1}$$

Definition: Connected Components. Let p a point of the grid E. The value of the function f at point p is denote by $f(p)$. Let $N_E(p)$ denote the set of the neighbours of pixel p for the set E. Useful neighbourhood relationships are normally defined by 4 or 8 connectivity. Let x and y be two pixels on E, there exists a path connectivity x and y if, and only if, there exits an n-tuple of pixels $(p_0, p_1, \ldots, p_n)$ such that $p_0 = x$ and $p_n = y$ and $p_i \in N_E(p_{i-1}), \quad \forall i = 1, ..., n$. Then the connected component of E that contains p can be defined as the union of all the pathes included in E with origin in p. Let $E_h^k, k \geq 1$, denote the k'th connected component of the upper level set E_h and E_k^h the k'th connected component of the lower level set E^h. Given the definition of connected component, the level sets can also be defined as:

$$E_h = \bigcup_k E_h^k, \quad E^h = \bigcup_k E_k^h \tag{2}$$

The difference between a connected component and a flat zone is that a connected component is a subset of the grid E while a flat zone is a subset of the function f.

Definition: Max and Min Tree. The max tree is a structured representation of the connected components of the upper level sets of an image. This representation is developed to deal with area opening. Both Salember [18] and Meijster [13] claimed that once a max tree of an image has been constructed, computing an area opening at scale λ reduces to removing all nodes representing connected components which have an area smaller than λ from the tree. The image can then be reconstructed at the new scale directly from the tree.

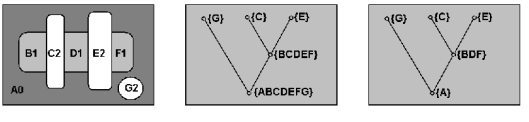
Each node in the max tree is associated with a connected component. The leaves of the tree correspond to the regional maxima of the image and the links between the nodes describe inclusion (father-child) relationships between connected components. Let N_k^i and N_{k+1}^j denote two nodes associated with with connected components E_k^i and E_{k+1}^j from two successive levels sets E_k and E_{k+1} respectively. The node N_{k+1}^i is a child of the node N_k^j if E_k^i and E_{k+1}^j satisfy:

$$\text{father}(N_{k+1}^i) = N_k^j \quad \text{if} \quad E_{k+1}^i \subseteq E_k^j$$

Additionally, each node in the tree has associated attributes representing 1) the number of total pixels belonging to its associated connected component, namely, area; 2) the gray level h of the level set that its associated connected component belongs to. However, a tree generated from an image directly according to above definition is not the final max tree. Two tasks need to be executed to remove redundant data from the tree.

Firstly, two connected components E_k^i and E_m^j from two upper level sets at different gray value may contain identical subset of E. Therefore, for those nodes associated with the same subset of E, the node of highest gray level h will be retained and the other nodes will be removed. Secondly, due to the inclusion relationship encoded in the tree structure, a connected component associated with a max tree node is a subset of the connected component associated with its father's node. In order to remove the redundancy, a node in the final max tree actually stores only the difference between its associated connected component and all the connected components belonging to its child nodes for memory efficiency. However, a node in the max tree still logically represents its original associated connected component, which can be reconstructed by taking the union of the sets associated to its descendants. Only after performing above two tasks is the final tree constructed. An example of generating a max tree from an image is shown in Figure 2 (reproduced from [16]).

The min tree is a structural representation of connected components of lower level sets of an image. Each node in a min tree is associated connected components from a lower level set. The inclusion relationship among the min tree nodes is established and redundant data are removed in a similar way to the max tree.



(a) (b) (c)

Fig. 2. (a): A simple image only with seven flat zones, (b): the logical representation form of the max tree created from the image on the left (a), (c): the physical storage form of the same max tree

The max and min tree can also be regarded as representation structures of flat zones and granules of image f and $-f$, respectively. Actually, the max/min tree logically encodes the granules of the image $f/-f$ and physically forms a partition of flat zones of the image $f/-f$. However, this is out of the scope of this paper; the reader is referred to [16] for more details. In addition, there are a number of image trees similar to the max/min tree, i.e. the opening/closing tree [24] from L. Vincent, the sieve opening/closing tree [1] from Bangham, and

the component tree [11] from Jones. The difference is that the opening/closing tree and the sieve tree arose to encode the image granules resulting from applying granule functions to the image. The component tree is developed to perform attribute opening/closing. However, the max/min tree can be regarded as a special instance of the opening/closing tree or the component tree and as a sibling to the sieve tree. We proceed to establish the relationship between the watershed lake tree and the min tree.

4 From Min Tree to Watershed Lake Tree

4.1 Theory

We have described before how the immersion simulation is used to implement the watershed transformation and watershed lake tree. Given the definition of level sets and connected components, we can say that the procedure of the immersion simulation is also the procedure of generating the lower level sets of an image. During a flooding at gray value h, the regions covered by the water forms the lower level sets at level h. Let L_k^h denotes the k'th lake during the immersion simulation at gray value h, the support region (the surface) of the lake L_k^h is a connected component E_k^h of the lower level set E^h. When two lakes L_m^h and L_n^h merge into a bigger lake L_k^{h+1} at gray level of $h+1$, there exists a relationship among their related connected components:

$$E_m^h \neq E_n^h,\ E_m^h \subset E_k^{h+1},\ E_n^h \subset E_k^{h+1}$$

A node of a watershed lake tree is associated with a lake created during the immersion simulation and is created only when a mixture event happens, a node N_m^h is said to be a child of another node N_k^{h+1} if E_m^h and E_k^{h+1} satisfy:

1: father$(N_m^h) = N_k^{h+1}$ if $E_m^h \in E_k^{h+1}$.
2: $\exists E_n^h,\quad m \neq n,\quad E_n^h \in E_k^{h+1}$.

If we take the second criterion away, the above definition becomes the definition of the inclusion relationship for the min tree. Clearly, if above criteria holds, $father(E_n^h) = N_k^{h+1}$. Thus, a non leaf node in the watershed lake tree has at least two children nodes. The min tree can be also generated by the immersion simulation. The difference is that in the case of generating a watershed lake tree, new nodes are created only when at least two pools are mixed into one big pool while in the case generating a min tree, a new node is created whenever a pool surface become larger due to the raised water. Obviously, the mixture of two small pools is a special case of when the pool surface become larger. Therefore, we could conclude a watershed lake tree structure can be regarded as a subset of its corresponding min tree.

The similarity between the watershed lake tree structure and the min tree structures are illustrated in Figure 3. When a mixture event happens, the surface of a small pool becomes a proper connected component for a new created node in the lake tree structure. Note that the surface of the new big pool is not a

proper connected component for the new created father node at this stage until the new pool mixes with other pools in a later immersion stage. For example, in Figure 3(c), the pool $\{C_1\}$ and $\{C_3\}$ merged together and they generated two new nodes in the lake tree (see Figure 3(g)). However, the surface of the new pool $\{C\}$ is not the final connected component associated with the new created father node. It is when the water carried on raising one more level, the pool $\{C\}$ became $\{B_3, C\}$ and is about to merge with another pool $\{D_2, D_3, E\}$ that it becomes the connected component for the new node in the lake tree. When the immersion reaches the highest place in the topological structure (the global maximum), the corresponding pool covers the whole region and it becomes the root of the lake tree.

When the immersion simulation is applied to generate the min tree, new nodes are created in a similar way. Given a signal, when new nodes need to be created in the lake tree, new nodes will also be created in the min tree. In addition, when a pool surface becomes bigger due to the immersion effect and there is no mixture event, the pool just before the increment of the immersion also becomes a new node in the min tree. This incident only happens to the min tree and not to the lake tree. For example, the pool surface of $\{D_3, E\}$ in Figure 3(c) becomes bigger in the next increment of immersion, thus there is node in the min tree to correspond this pool. However, there is no such a node in the lake tree. In the case the enlargements of all the pool surfaces are caused by the mixture events, the watershed lake tree structure and the min tree structure will be exactly the same. For example, the watershed lake tree structure obtained in Figure 1 is also a min tree for the same signal. In summary, the similarity between these two tree structures is that they have the same number of branches and these branches form the same tree structure. The difference is that the numbers of nodes in their branches are generally different.

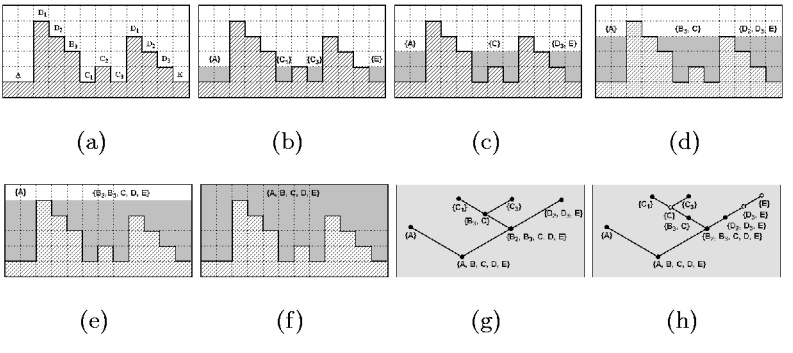


Fig. 3. A lake tree (g) and a min tree (h) construction simulations from a same signal. Please note in this figure $\{B\} = \{B1, B2, B3\}, \{C\} = \{C1, C2, C3\}, \{D\} = \{D1, D2, D3\}$. Please note that nodes in black in the min tree are those nodes which form the watershed lake tree

4.2 Implementation Details

As mentioned before, given a min tree constructed from an image, the watershed lake tree of the same image could be constructed by pruning the nodes from the min tree. The pruning criteria is described as follows: those min tree nodes having siblings should be retained and those that have no siblings should be removed (obviously, this rule does not apply to the root node). The root node of a min tree will become the root node of the watershed lake tree.

The min tree structure is normally built through its dual tree structure, namely the max tree. Given an image f, the min tree structure of f is equivalent to the max tree structure of $-f$, where $-f = NG - f$ and NG represents the maximum possible gray value (usually $NG = 255$). Please refer to [18,10] for more details of how a min tree can be built through a max tree. Once a min tree of an image is created, a set of attributes associated with a min tree node usually include: 1) a node identification number(ID); 2) its father's ID; 3) a list of children IDs; 4) a list of pixel coordinates; 5) number of pixels belonging to a node (area); 6) a gray value. Please note when pixels coordinates information are saved, the physical storage form of the tree is considered for memory efficiency.

In our implementation of the max/min tree structure [10], a two dimensional linked list is used as the data structure to accommodate the tree. In order to access an element in the two dimensional linked list directly, a hash table is created when the min tree construction is finished. The node index (m, n) is used as the key for the hash table. This hash table turns the complexity of visiting an element in a linked list from $O(N)$ into $O(1)$. Please refer to [10] for more details. During the transformation of a min tree into a watershed tree, this hash table helps to reduce the processing time.

Apart from the branch pruning process, we also need to consider the attributes associated with the watershed lake tree and the min tree. Fortunately, all the attributes associated with the min tree node are required by the watershed lake tree. However, the purpose of the attribute 'gray value' has changed. In the watershed lake tree, the attribute 'gray value' represents the level at the point (pixel) where lakes merge. In addition, the min tree only have one measure 'area', while the watershed lake tree has three measures 'area', 'height' and 'volume' that need to be computed in the image reconstruction. Therefore, we need calculate the values of 'height' and 'volume' for the watershed lake tree nodes.

Clearly, there is a need to consider updating the father-child relationships in the min tree after a node is deleted from the tree. There are two situations we need to consider. First, the deleted node only has one child node. This is a easy situation to deal with. According to tree pruning criteria, the child node of the deleted node will also be deleted from the tree. For example, when the node $\{D_3, E\}$ is deleted from the min tree in Figure 3(h), its child node $\{E\}$ will also deleted. However, if the deleted node has more than one child node, these children nodes need to be kept in the min tree after the tree pruning process. There is a need to assign new father nodes for these children nodes. Clearly, one of the deleted node's ancestors will become the new father node. The closest ancestor node that has at least one sibling will become the new father node. For

example, when the node $\{C\}$ is deleted from the min tree, the node $\{B_3, C\}$ become the new father of the two child node $\{C_1\}$ and $\{C_3\}$.

In summary, transforming a min tree to a watershed lake tree involves following three steps:

1. Check each node except the root node in the min tree if it has at least one sibling. If the answer to a node is YES, the node will be labelled as a valid watershed lake tree node. Otherwise, it will labelled as a invalid watershed lake tree node.
2. Calculate two new attributes 'height' and 'volume' for the lake tree nodes. The attribute 'area' from a min tree node will be used as attribute 'area' for a lake tree node.
3. Restore the father-child relationship for those nodes whose father nodes are not labelled as valid watershed lake tree nodes. For those nodes whose father nodes are labelled as valid watershed lake tree nodes, there is no need to update their father-child relationship.

5 Conclusion

In this paper, we discussed the similarities between two well know tree based image representation structures, namely the watershed lake tree and the min tree. Both trees have achieved some success in image segmentation and object extraction. However, these two trees are often treated separately by researchers. In this paper, a framework comparing the two trees is established.

We also feel that creating a watershed lake tree from its corresponding min tree is more straightforward than the existing watershed lake tree construction strategy [7], which is based on the the immersion simulation algorithm [25] used to generate the watershed transformation.

References

1. J. A. Bangham, J. R. Hidalgo, R. Harvey, and G. Cawley. The segmentation of images via scale-space trees. In *9th BMVC*, pages :33–43, 1998.
2. S. Beucher and C. Lantuéjoul. Use of watershed in contour detection. In *Proc. Int'l Workshop Image Processing, Real-Time Edge and Motion Detection/Estimation*, Rennes, France, 1979.
3. Edmond J. Breen and Ronald Jones. Attribute openings, thinnings, and granulometries. *Computer Vision and Image Understanding*, 64(3):377–389, 1996.
4. F. Cheng and A. N. Venetsanopoulos. An adaptive morphological filter for image processing. *IEEE Transactions on Image Processing*, 1(4):533–539, 1992.
5. J. Cichosz and F. Meyer. Morphological multiscale image segmentation. In *WIAMIS'97*, pages 161–166, 1997.
6. J. Crespo, J. Serra, and R.W. Schafer. Theoretical aspects of morphological filters by reconstruction. *Signal Processing*, 47(2):201–225, 1995.
7. Mark Fisher and Richard Aldrige. Hierarchical segmentation of images using a watershed scale-space trees. In *IEE Int. Conf. Image Processing and its Applications*, pages 522–526, 1999.

8. M. Grimaud. A new measure of contrast: the dynamics. In *Proceedings of SPIE Conference on Image Algebra and Morphological Image Processing*, volume 1769, pages 292–304, 1992.
9. Henk J. A. M. Heijmans. Connected morphological operators for binary images. *Computer Vision and Image Understanding: CVIU*, 73(1):99–120, 1999.
10. Xiaoqiang Huang, Mark Fisher, and Dan Smith. An efficient implementation of max tree with linked list and hash table. In *Proceedings of International Conference on Digital Image Computing-Techniques and Applications*, pages 299–308, Macquarie University, Sydney, Australia, December 2003.
11. R. Jones. Connected filtering and segmentation using component trees. *Computer Vision and Image Understanding*, 75(3):215–228, 1999.
12. G. Matheron. *Random Sets and Integral Geometry*. John Wiley and Sons, New York, 1975.
13. A. Meijster and M. Wilkinson. A comparison of algorithms for connected set openings and closings. *IEEE PAMI*, 24(4):484–494, 2002.
14. F. Meyer and S. Beucher. Morphological segmentation. *J. Visual Communication and Image Representation*, 1:21–46, 1990.
15. L. Najman and M. Schmitt. Geodesic saliency of watershed contours and hierarchical segmentation. *IEEE PAMI*, 18(12):1163–1173, 1996.
16. Luis Garrido Ostermann. *Hierarchical Region Based Processing of Image and Video Sequences: Application to Filtering, Segmentation and Information Retrieval*. PhD thesis, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, Spain, April 2002.
17. P. Salembier and L. Garrido. Binary partition tree as an efficent representation for image processing, segmentation, and information retrieval. *IEEE Transactions on Image Processing*, 9(4):561–576, 2000.
18. P. Salembier, A. Oliveras, and L. Garrido. Anti-extensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing*, 7(4):555–570, 1998.
19. P. Salembier and J. Serra. Flat zones filtering, connected operators and filters by reconstruction. *IEEE Transactions on Image Processing*, 3(8):1153–1160, 1995.
20. J. Serra. *Image Analysis and Mathematical Morphology*, volume I. Academic Press, London, 1982.
21. J. Serra. *Image Analysis and Mathematical Morphology: Theorectical Advances*, volume II. Academic Press, London, 1988.
22. J. Serra and P. Salembier. Connected operators and pyramids. In *Proceedings of SPIE Conference on Image Algebra and Mathematical Morphology*, volume 2030, pages 65–76, 1993.
23. L. Vincent. Grayscale area openings and closings: Their efficent implementation and applications. In *WMMASP'93*, pages 22–27, 1993.
24. L. Vincent. Fast grayscale granulometry algorithms. In *SMMAIP'94*, pages 265–272, 1994.
25. L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE PAMI*, 13(6):583–598, 1991.

From Min Tree to Watershed Lake Tree: Evaluation

Xiaoqiang Huang and Mark Fisher

University of East Anglia, Norwich, NR4 7TJ, UK
xq.huang@uea.ac.uk, mhf@cmp.uea.ac.uk

Abstract. Recently, several tree based hierarchical image descriptions have been proposed for image segmentation and analysis. This paper considers the problem of evaluating such algorithms. Recently we proposed a new algorithm for constructing the watershed lake tree by transforming the min tree structure as these two image trees share some similarities. We use this algorithm to illustrate the evaluation approach. The algorithm is evaluated by considering its computational complexity, memory usage and the cost of manipulating the resulting tree structure. Our results show that considerable care is needed when evaluating algorithms of this kind. In particular, comparisons cannot be made simply on the basis of computational complexity alone and other parameters such as image/tree 'complexity' also need to be considered.

1 Introduction

Tree based hierarchical image descriptions have become more important in the field of image analysis and segmentation over last decade. A hierarchy (from coarse to fine) of partitions of an image can be described by using a tree structure. Compared to traditional pixel based image representations, the tree structure is more flexible and powerful in presenting image content at different scales. Recently, image trees have achieved some success in image filtering, segmentation, indexing and retrieval. Those proposed include: the critical lake tree [2] (and equivalently the watershed lake tree [3]), The Max/Min tree [9], the Component Tree [6], the Sieve Tree [1], the Inclusion Tree [8] and the Binary partition Tree [9].

However, standard evaluation methods for image tree construction algorithms do not exist. Most image tree construction algorithms are evaluated by their computing cost and memory usage [7,9]. In addition, the evaluation of computing cost is mainly conducted by measuring the time required to construct image trees from images of different size. We believe that more factors need to be taken into account when evaluating an algorithm for building image trees. For instance, according to our experiments (reported in section 3) not only image size, but also image complexity and tree complexity needs to be considered. In addition, because the power of image tree descriptions lie in their ability to encode image regions, we should also consider how easily an image tree can be manipulated (i. e. nodes merged/deleted).

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 858–865, 2004.
© Springer-Verlag Berlin Heidelberg 2004

In this paper we propose a fairer and more comprehensive method of evaluating algorithms for building image trees. We illustrate the approach by taking a new method for building watershed lake trees [5] as an example. This paper is organised as follows: section 2 describes the proposed evaluation method and section 3 gives experimental results. The conclusion and suggestions for further work are presented in the final section.

2 Evaluation Method

There are two aspects to consider when evaluating image tree data structures. Firstly the computational complexity due to building the tree and secondly the costs associated with manipulating the tree structure once it is built.

2.1 Cost of Building the Image Tree

As described before, given an algorithm for building image trees, there are three factors that affect the computing cost of constructing an image tree: the image size, the image complexity and the tree complexity. The term 'image complexity' and 'tree complexity' can be defined using the concept of a 'flat zone'. Flat zones in a grey image are the largest connected sets of the space where the image grey-level is constant.

$$image\ complexity = \frac{total\ number\ of\ flat\ zones}{image\ size} \tag{1}$$

$$tree\ complexity = \frac{total\ number\ of\ tree\ nodes}{total\ number\ of\ flat\ zones} \tag{2}$$

With this in mind, one needs to be very careful in choosing test images, particularly if one wants to test an algorithm's performance with respect to image size (keeping the image and tree complexity constant). It is difficult to find a group of real images that meet this criteria. One of the solutions is to use synthetic images, where we can control the image and tree complexity. However, in this paper we use a more straightforward approach. Given an grey image, we first change the value of pixels on the border line into zero. Then we make copies of this modified image and 'stitch' them together using a 'step-and-repeat' process to get an image of larger size. The more copies we make, the bigger image we can get (an example is given in Figure 1). Images created in this way have almost constant image and tree complexity. Additionally, it is also necessary to evaluate the performance over a group of images of same size but different content. This test enables us to determine sensitive the algorithm is to image content (i.e. how the computational cost varies with image and tree complexity).

(a) (b)

Fig. 1. Two images of different size but of same image complexity and tree complexity

2.2 Cost of Manipulating the Tree

Image tree descriptions are important for image analysis tasks and therefore the ease with which the final tree structure can be manipulated is an important consideration. For example, recently the max/min tree has been proposed for the efficient implementation of area opening and closing operators. Thus, when one comes to evaluate an algorithm for building a max/min tree, one needs to additionally evaluate its performance in respect to implementing the area opening/closing operators by manipulating the resulting max/min tree structure.

In this paper we will illustrate the approach by evaluating an algorithm for building a watershed lake tree. In this case we will consider how efficiently the resulting structure can be manipulated to implement a critical connected filter [2].

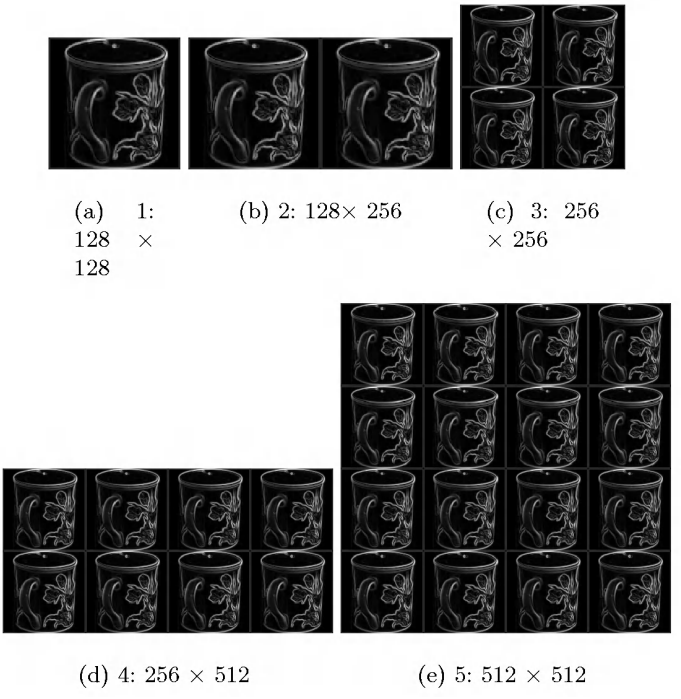
2.3 Memory Usage

There might be a number of different methods constructing a same image tree. However, given clear definitions of the tree structure and attributes associated with tree nodes, it seems that the memory used to store the final image tree shall be fixed. However, different image tree construction algorithms may use different data structures (array, linked list, vector, etc...) and this causes a variation of memory usage. Thus, if one wants to compare two different methods of constructing the same image tree, one needs to compare the memory usage. In addition, one also needs to determine if the data types used are appropriate.

3 Evaluating the Watershed Lake Tree Construction Algorithm

3.1 Test Images

We have chosen three different groups of images for the following evaluation. The first group (Figure 2) comprises images of different size but the same image and tree complexity. The second group (Figure 3) comprises images of the same size but different image and tree complexity. The third group (Figure 4) is used to test the cost of manipulating the tree. All the test images used in this paper are gradient images as the boundaries of the watershed regions correspond to



(a) 1: 128 × 128 (b) 2: 128× 256 (c) 3: 256 × 256

(d) 4: 256 × 512 (e) 5: 512 × 512

Fig. 2. Our first group of test images.

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5 (f) 6

Fig. 3. Our second group of test images. Images are all in size of 128 × 128

the ridges of the surface. The specification of the computer in which all experiments were carried out is as follows: OS: Windows 2000, CPU: 1.3GHZ, Memory: 512M. Please note all the experimental data are averages of twenty individual experiments.

3.2 Computing Cost and Memory Usage

The time cost and memory usage of constructing watershed lake trees using the first group of test images (same image/tree complexity but of different size) are shown in Table 1. As we construct the min tree before the watershed lake tree

(a) 1 (b) 2

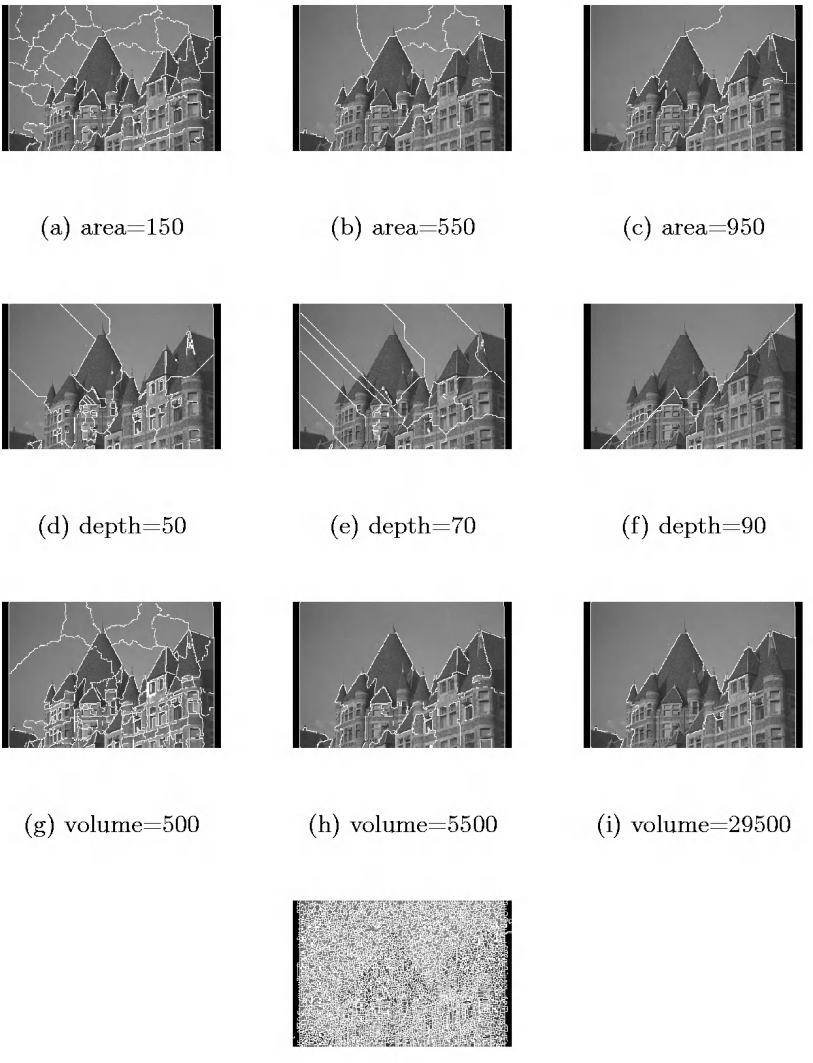
Fig. 4. Our third group of test images. Images are all in size of 384 ×256

is created (from the min tree), we list both the time cost and memory usage of the two trees in the tables. The 'time: Total' and 'memory: Total' rows show the cost of constructing a watershed lake tree from an image. From this table, we can conclude that both the min tree and the watershed lake construction algorithms generally perform linearly with respect to image size.

Table 1. Time cost (milli-seconds) and memory usage (bytes) for building min tree and watershed lake tree from images shown in Figure 2.

Trees	Image Number				
	1	2	3	4	5
time: Min Tree	54.5	108.5	222.5	453	931.5
time: Watershed Lake Tree	3.5	10	18	39.5	83.5
time: Total	58	118	240.5	492.5	1015
memory: Min Tree	298360	582328	1146552	2278584	4535224
memory: Watershed Lake Tree	51459	101372	199385	397187	789165
memory: Total	349819	683700	1345937	2675771	5324389

The implementation of the min tree algorithm is evaluated and discussed in [4]. Hence we only analyse the computing complexity of transforming a min tree into a watershed lake tree here. As described in [5], there are three steps needed to fulfill this task. In the first step 'Labelling valid watershed lake tree nodes' we visit all the nodes in a min tree and decide what nodes are valid watershed lake tree nodes. A min tree node is only labelled as a valid watershed lake tree node if it has a sibling. Min tree nodes without any siblings will only be visited once and those nodes with at least one siblings will be visited twice. Thus, in the worst case that all min tree nodes (apart from the root node) have at least one siblings, all the min tree nodes will be visited twice. Thus, the computing complexity of this step is $O(N)$, where N is the total number of min tree nodes. In the second step 'Calculating new attributes for the watershed lake tree nodes', all min tree nodes will be visited no more than twice to calculate the 'depth' and 'volume' value for min tree nodes. In order to calculate these two attributes for the watershed lake tree nodes, all min tree nodes will be visited another three



(a) area=150 (b) area=550 (c) area=950

(d) depth=50 (e) depth=70 (f) depth=90

(g) volume=500 (h) volume=5500 (i) volume=29500

(j) the original watershed image

Fig. 5. Results of first applying the watershed transform to reconstructed images based on different criterion type and threshold values and then adding watershed lines to the original image.

times (the worst situation). Thus, the computing complexity of this step is also $O(N)$. In the last step 'Updating father-child relationships', all min tree nodes are checked to determine if their father nodes are also valid watershed lake tree

Table 2. Time cost (milli-seconds) and memory usage (bytes) of constructing a watershed lake tree from the images shown in Fig 3

Image	Flat zones	Nodes (Min)	Nodes (WLT)	Time (Min)	Time (WLT)	Total Memory
No. 1	5019	2331	1315	50.5	3	333604
No. 2	8158	1845	1237	51.5	1	298782
No. 3	6729	2597	1506	51.5	3.5	352831
No. 4	8212	2131	1152	50.5	3	304725
No. 5	7218	2331	1390	51	3.5	324667
No. 6	8730	2748	1685	53	4.5	374686

Table 3. Test on performance of implementing (milli-seconds) critical connected filters through watershed lake tree

Image Number	Min Tree	WLT Tree	Total	area	depth	volume
1	367	45.5	412.5	17	21	19
2	397	66.5	463.5	20	25	21

nodes. If not, a new father node needs to be allocated. All min tree nodes will be visited three times in the worst case. Thus, the computing complexity of this step is also $O(N)$. In summary, the computing complexity of transforming a min tree into a watershed lake tree is $O(N)$. This is supported by the experimental results shown in Table 1.

The time cost and memory usage of constructing watershed lake trees using the second group of images are shown in Table 2. This table shows that creating watershed lake trees and min trees from images with high image/tree complexities needs more time.

3.3 Manipulating Cost

Based on the watershed lake tree, a family of critical connected filters [2] can be constructed. The node whose value is below a given threshold level is pruned and its corresponding minimum is flooded. This reconstruction applied to the original image allows us to filter the regions which do not fulfil a given condition for a given criterion: lake area, depth and volume. This family of connected filters helps to reduce the problem of over segmentation with the watershed transform. Figure 5 shows the results of first applying the watershed transform to the reconstructed images based on different criterion type and threshold value and then adding watershed lines to the original image. Note that the watershed lake trees are built from the gradients of the original image. Table 3 shows the time cost of constructing watershed lake trees and performing different types of connected filters from the third group of test images shown in Figure 4. Note that the time cost of applying a connected filter to an image is independent of the threshold value. The results show that compared to the tree construction cost, performing critical connected filter operations is reasonably fast.

4 Conclusion and Further Work

This paper evaluates an algorithm for building watershed lake trees using a method that is fairer and more comprehensive than that used previously. This paper also demonstrated that it is possible to create a watershed lake tree from its corresponding min tree. The experimental results show that the proposed algorithm works very efficiently ($O(N)$). Further work will focus on evaluating and comparing other approaches for building watershed lake trees.

References

1. J. A. Bangham, J. R. Hidalgo, R. Harvey, and G. Cawley. The segmentation of images via scale-space trees. In *9th BMVC*, pages 33–43, 1998.
2. J. Cichosz and F. Meyer. Morphological multiscale image segmentation. In *WIAMIS'97*, pages 161–166, 1997.
3. Mark Fisher and Richard Aldrige. Hierarchical segmentation of images using a watershed scale-space trees. In *IEE Int. Conf. Image Processing and its Applications*, pages 522–526, 1999.
4. Xiaoqiang Huang, Mark Fisher, and Dan Smith. An efficient implementation of max tree with linked list and hash table. In *Proceedings of International Conference on Digital Image Computing-Techniques and Applications*, pages 299–308, Macquarie University, Sydney, Australia, December 2003.
5. Xiaoqiang Huang, Mark Fisher, and Yanong Zhu. From min tree to watershed lake tree:theory and implementation. In *Proceedings of Int'l Conf. on Image Analysis and Recognition*, Porto, Portugal, 2004.
6. R. Jones. Connected filtering and segmentation using component trees. *Computer Vision and Image Understanding*, 75(3):215–228, 1999.
7. A. Meijster and M. Wilkinson. A comparison of algorithms for connected set openings and closings. *IEEE PAMI*, 24(4):484–494, 2002.
8. Pascal Monasse and Frédéric Guichard. Scale-space from a level lines tree. *Journal of Visual Communication and Image Representation*, 11(2):224–236, 2000.
9. Luis Garrido Ostermann. *Hierarchical Region Based Processing of Image and Video Sequences: Application to Filtering, Segmentation and Information Retrieval*. PhD thesis, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, Spain, April 2002.

Optimizing Texture Primitives Description Based on Variography and Mathematical Morphology

Assia Kourgli, Aichouche Belhadj-aissa, and Lynda Bouchemakh

Image Processing Laboratory, Electronic Institute, U.S.T.H.B., B.P. 32, El Alia, 16111 Bab Ezzouar, Algeria
{a_kourgli, h.Belhadj@lycos.com, lbouchemakh@yahoo.fr}

Abstract. This paper proposes a novel method of optimising texture primitives detection based on the mathematical morphology. Indeed, successful textural analysis relies on the careful selection of the adapted window size. We use variography to optimise the shape of structuring elements to fit the shape of the unit patterns that form a texture. The variogram is essentially a "variance of differences" in the values as a function of the separation distance. This variance therefore changes as the separation distance increases where repetitive structures are described as hole-effects. We used the local minima (hole-effects) to find size, shape an orientation of unit pattern of image textures and thus to determine the optimal structuring element which will be used in mathematical morphological texture analysis. Some of Brodatz's natural texture images have been used for evaluating the performance of the structuring elements found in the characterisation and discrimination of the texture aspects of images. Promising results are obtained and presented.

1 Introduction

A number of techniques for texture analysis and texture discrimination have been proposed and have achieved considerable success, although generally under well-defined and rather limited operating conditions. These techniques can be classified as either structural or statistical [4]. The structural approach is based on the description of unit patterns and their placement rules. In the statistical approach, the aim is to characterise the stochastic properties of the spatial distribution of gray levels in an image by estimating first and higher order statistics from a local neighbourhood [11], [6].

Mathematical morphology is a structural method of processing images according to the images' topological properties. It has been successfully used in many applications including object recognition, image enhancement, texture analysis and industrial inspection [14]. However, structuring elements can vary greatly in their weights, sizes and shapes, depending on the specific applications. Several adaptive techniques have been used for finding optimal morphological structuring elements. Some technique use neural networks and fuzzy system [10], [16].

We shown [7] that the variogram measure commonly used in geostatistics can be used to characterise the placement rules and the unit patterns of texture. Indeed, local

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 866–873, 2004.
© Springer-Verlag Berlin Heidelberg 2004

minima of the variogram measure can be used to characterise size, shape, orientation and placement rules of unit patterns of a texture. The purpose of this paper is to demonstrate the use of variograms to customise window sizes (structuring element shape) for use in mathematical morphological texture analysis.

2 Variogram

Semi-variograms are essentially graphs measuring the difference between grade values relative to the distance separating them, in a particular orientation [9]. They provide a description of the similarity or dissimilarity between pairs of values as a function of their separation vector 'd' [1].

For certain applications such as remote sensing or image processing, whose data sets contain huge amounts of closely spaced and regularly gridded information, summarising the pattern of spatial continuity is, by itself an important goal [12]. The variogram characterises the spatial variability of the random variable (in this case, the gray levels of the texture image), it quantifies how different the measured values are likely to be as the separation distance between the sampling points increases [3].

Numerically, the variogram function is calculated in units of distances. It is typically estimated by the "sample" or experimental variogram [15]:

$$2\gamma(d)=\frac{1}{N(d)}\sum_{i}\left[Z(x_i)-Z(x_i+d)\right]^2 \quad (1)$$

Where

d = inter-sample spacing distance (in pixels),

N = number of gray level pairs within an image,

$Z(x_i)$, $Z(x_i+d)$ = Gray level pairs.

The result is a variogram value, which is plotted against distance. Fig. 2. shows the variogram values (various distances over the four angles 0°, 45°, 90° and 135°) computed for the raffia texture shown in Fig. 1.

A typical semi-variogram has three characteristics, which are of assistance in enhancing the understanding of the phenomenon behaviour. These are as follows:

- Range of influence : The distance between the samples at which the semi-variogram appears to level off.
- Nesting structures : repetitive structures are described as hole-effects.
- Sill : The level of variability that indicates that there is no further correlation between pairs of samples.

A single semi-variogram summarises the spatial variability between the samples along a given relative orientation. Since the nesting structure on the semi-variogram defines the separation distance up to which the samples can be considered repetitive, it can also be used as a tool for identifying the principal directions and size of texture structures and thus the optimal structuring element. In all likelihood, this will be the distance that exhibits a hole-effect.

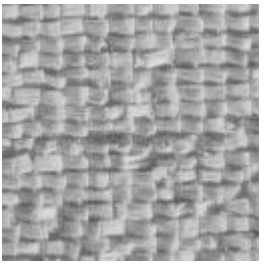


Fig. 1. Raffia texture

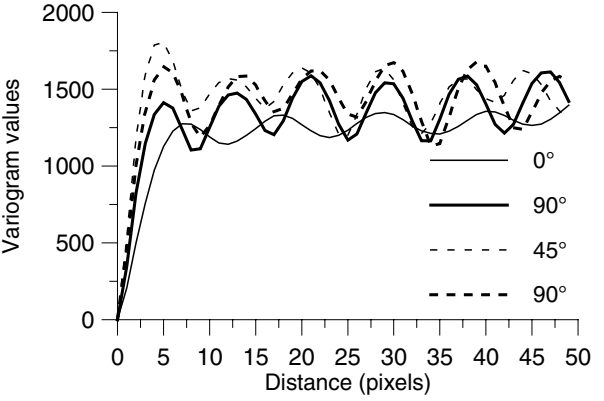


Fig. 2. Graphs of $\gamma(h)$ measures versus distance at the four orientations for Raffia texture.

For natural texture, the variogram measure is never equal to zero but local minima represent candidates' points from which the structure size can be computed. So, the minimum variogram values being significantly different from zero indicates that the replication process and/or the unit patterns are not always the same [6]. Fig. 2. shows that the points where variogram values are minimum (hole-effect) correspond to inter-sample spacing distance values 'd' which are integer multiple of 12 pixels for the horizontal direction and of 8 pixels for the vertical one.

3 Mathematical Morphology

Both linear convolution and morphological methods are widely used in texture image processing. One of the characteristics among them is that they both require applying a template to a given image, pixel by pixel, to yield a new image [13]. In the case of convolution the template is linear and is usually called convolution window or mask, while in mathematical morphology, it is referred to a structuring element.

Mathematical morphology involves the study of the different ways in which a structuring element interacts with a given set, modifies it shapes, and extracts the resultant set [10]. Structuring elements can vary greatly in their weights, sizes and shapes, depending on the specific applications.

The structuring element is used as a tool to manipulate the image using various operations. The basic operations are erosion and dilatation [5]. We use the notation of [8] to introduce morphology on functions.

Suppose that a set A in the Euclidean N-space (E^N) is given. Let F and K be $\{x \in E^{N-1} \mid$ for some $y \in E$, $(x,y) \in A\}$, and let the domains of the gray-scale image be f and the structuring element (kernel/template) be k, respectively.

The dilatation of f by k, which is denoted by $f \oplus k$, is defined as:

$$(f \oplus k)(x, y) = \max \{f(x + m, y + n) + k(m, n)\} \quad (2)$$

for all $(m, n) \in K$ and $(x + m, y + n) \in F$.

The erosion of f by k, which is denoted by $f \theta k$, is defined as:

$$(f \theta k) (x, y) = \min \{f (x + m, y + n) - k (m, n)\} \quad (3)$$

for all $(m, n) \in K$ and $(x + m, y + n) \in F$.

Based on these operations, closing and opening are defined. The closing operation is a dilatation followed by erosion while the opening operation is an erosion followed by a dilatation .

The selection of structuring element k used by the dilatation and erosion functions is very important to the system as this determines the manner in which the individual objects are supposed to be connected.

As the variogram provides a measure of the spatial dependence of data, we use variography to optimise the shape of structuring elements to fit the shape of the unit patterns that form a texture.

4 Experimental Results

In texture analysis the first and most important task is to extract texture features which most completely embody information about the spatial distribution of intensity variations in an image.

In order to evaluate the performance of the variogram measure for texture characterization, we use the following Brodatz's [2] textures: raffia, herringbone weave and woollen cloth (Fig .1. and Fig. 3.). The texture images were taken from *USC-SIPI Image Database* [18] and are of size 128 × 128 with 256 gray levels and the variogram values (Fig. 2., Fig. 4. and Fig. 5.) were computed for these textures.

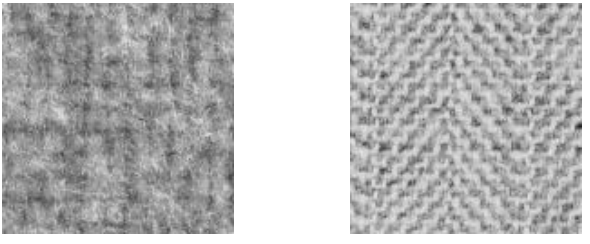


Fig. 3. Woollen cloth and herringbone weave textures.

The plots shown in Fig. 4. indicate that we have a structure repetition each 6 pixels in the both diagonal directions (45° et 135°) for herringbone weave texture, hence, we choose a structuring element k_1 whose size is half the nesting structure (3 × 3 pixels) and shape in respect with the main directions (both diagonal directions) of the texture.

We can also, estimate the size of the unit pattern of woollen cloth texture using the graphs of variogram measures. The plots shown in Fig. 5. indicate that the unit pattern is a square of 12 pixels by 12 pixels, hence, we choose a square structuring element k_2 whose size is 6 × 6 pixels. As before, for raffia texture, Fig. 2. shows that the

size of the unit pattern is 12 × 8, where 12 is the horizontal dimension and 8 is the vertical one, thus, the corresponding structuring element k_3 is a rectangle of 6 × 4 pixels.

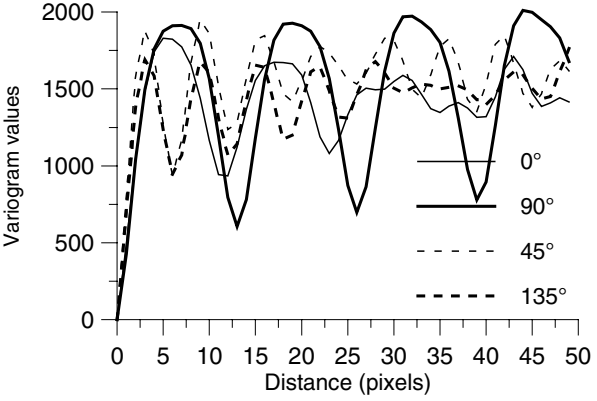


Fig. 4. Graphs of $\gamma(h)$ measures versus distance at the four orientations for herringbone weave texture.

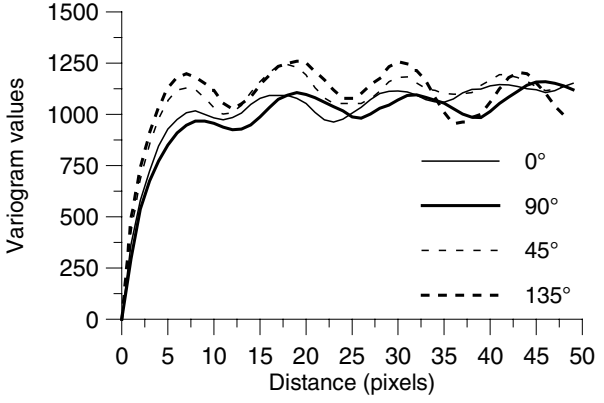


Fig. 5. Graphs of $\gamma(h)$ measures versus distance at the four orientations for woollen cloth texture.

We obtain the following structuring elements k_1, k_2 and k_3:

$$k_1 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix}, k_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, k3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad (4)$$

Erosion, dilatation, opening and closing, with the above structuring elements were applied to the three texture images shown in Fig. 1. And Fig. 3. The target images for these operations are shown in Fig. 6., Fig.7. and Fig. 8.

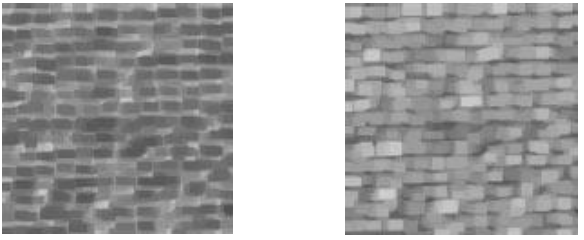


Fig. 6. Results of erosion and opening of raffia texture of Fig.1.

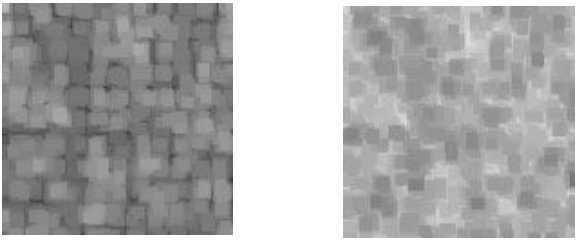


Fig. 7. Results of opening and closing for woollen cloth texture of Fig. 3.

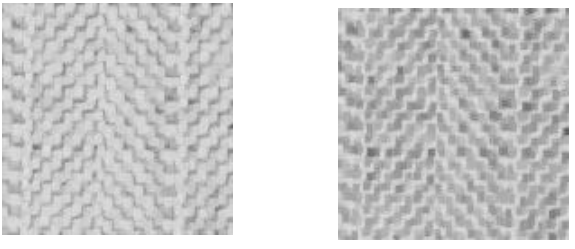


Fig. 8. Results of dilatation and closing for herringbone texture of Fig. 3.

Fig.6. shows that the results of erosion and opening of raffia texture (Fig. 1.) by the structuring element k_1 allow to enhance unit patterns which looks homogenous, while contours are well preserved. Similar results (Fig. 7. And Fig. 8.) are obtained for woollen cloth and herringbone weave textures (Fig. 3.), we note that the shape and the boundaries of the unit patterns are enhanced, hence, the structuring elements found are well adapted.

To assert these observations, we applied a contour enhancement algorithm [17] (Sobel filter) to the original texture images and the morphological transforms ones. The results obtained are illustrated by Fig. 9., the upper images illustrate the results of contour enhancement applied to the original textures while, lower ones illustrate the results of contour enhancement applied to the eroded texture using the optimal structuring elements found.

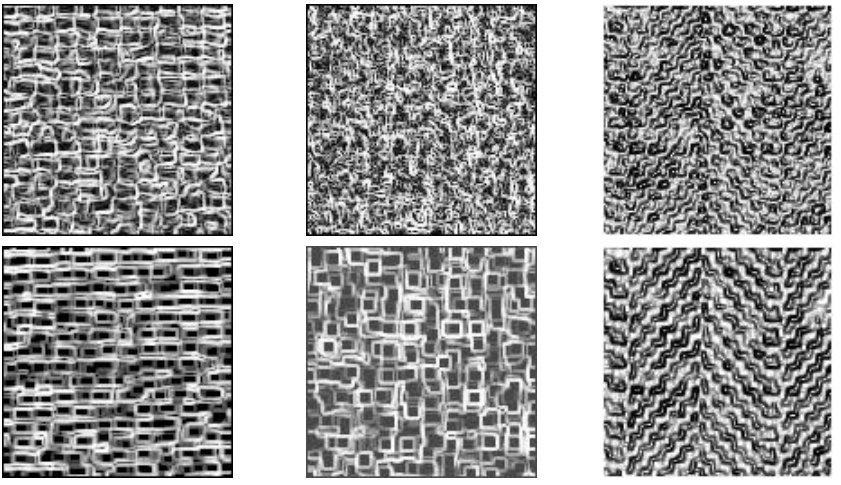


Fig. 9. Results of contour enhancement applied to raffia, woollen cloth and herringbone weave textures (up) and applied to the results of theirs erosion (down)

These three examples demonstrate that good results can be obtained when the variogram measure is used to compute the size and shape of structuring elements. Indeed, texture primitives are well characterised. Other textures from Brodatz's album have also been analysed. The textures examined include water, beach sand, wood grain, grass, beach sand, pigskin and pressed calf leather.

4 Conclusion

We have investigated a morphological approach to texture analysis which attempts to optimize the structuring element size and shape using the variogram representation for primitive textures descriptions. Some characteristics (nesting structures) of the variogram representation have been computed which was introduced for efficient implementation of mathematical morphological functions. Indeed, we demonstrate the ability of the variogram to identify the features of the texture and its underlying structure. Experimental results show that interesting results can be obtained.

Of course, it is far from constituting a complete independent analysis method. Some further study is required, including the adaptive structuring elements. Whereas the structuring elements used in most applications remain constant as they probe an image, there are situations where structuring elements must change their size, shape, orientation during probing this change can be made on the basis of the variogram to obtain an adaptive structuring element.

References

1. R. A. Brazier, R.A., Boomer, K.: Enhancing the Sampling Procedure through a Geostatistical Analysis. available via http://www.essc.psu.edu/~brazier/ geo.html.
2. Brodatz, P.: Textures: A Photographic Album for Artists and Designers. Dover Publications, New York (1966)
3. Cressie, N.A.: Statistics for Spatial Data. Wiley-Interscience, New York, Revised edition (1993)
4. Haralick, R.M.: Statistical Image Texture Analysis. Chapter 11 of Handbook of Pattern Recognition and Image Processing. Academic Press Edition (1986) 247-279
5. Jackway, P.T., Deriche, M.: Scale-Space Properties of the Multiscale Morphological Dilatation-Erosion. IEEE Trans. Pat. Anal. and Mach. Intel. vol.18, no. 1 (1996) 38-51
6. Kourgli, A., Belhadj-aissa, A.: Approche structurale de génération d'images de texture. International Journal of Remote Sensing vol. 18, no. 17 (1997) 3611-3627
7. Kourgli,A., Belhadj-aissa, A.: Characterising Textural Primitives using Variography. Proc. IMVIP 2000, Belfast, Ireland (2000) 165-175
8. Krishnamurthy, S., Iyengar, S.S., Hoyler, R.J., Lybanon, M.: Histogram-Based Morphological Edge Detector. IEEE Trans. on Geos. and Rem. Sens., vol. 32, no. 4 (1994) 759-767
9. Lacaze, B., Rambal, S., Winkel, T.: Identifying spatial patterns of Mediterranean landscapes from geostatistical analysis of remotely-sensed data. Int. of Rem. Sens. vol. 15, no. 12 (1994) 2437-2450
10. Lee, K., -H., Morale, A., Ko, S., -J.: Adaptive Basis Matrix for the morphological Function Processing Opening and Closing. IEEE Trans. on Image Proc., vol. 6, no. 5 (1997) 769-774
11. Reed, T., R., Du Buf, J.,M.,H.: A Review of Recent Texture Segmentation and Feature Extraction Techniques. CVGIP: Image Understanding vol. 57, no. 3 (1993) 359-372
12. Srivastava, M., R., Parker, H.M.: Robust Measures of Spatial Continuity. Geostatistics Proceedings of the Third Int. Geostatistics Congress, Avigon, France, vol.1 (1988) 295-308
13. Sussner, P., Ritter, G., X.: Decompostion of Gray-Scale Morphological Templates Using the Rank Method. IEEE Trans. Pattern Anal. and Mach. Intel. vol.19, no. 6 (1997) 649-658
14. Sutherland, K., Ironside, J. W.: Automatic Texture Segmentation Using Morphological Filtering on Images of the Human Cerebellum. available via http://citeseer.nj.nec.com
15. Thomas, G., S.: Interactive Analysis and Modelling of Semi-Variograms. Proceeding, 1st International Conference on Information Technologies in the Minerals Industry (via the Internet), December 2-13, Paper GT67, A A Balkema, (1997) available via: http://www.snowdenau. com/techno/visor/papers/ GT67T1
16. Verly, J. G., Delanoy, R L.: Adaptive Mathematical Morphology for Range Imagery. IEEE Transactions on Image Processing, vol. 2, no. 2 (1993) 272-275
17. Cocquerz, J-P., and al: Analyse d'images: Filtrage et Segmentation,Masson, Paris, 1995.
18. USC-SIPI Image Database: http://sipi.usc.edu/database.cgi

Author Index

Abad, Francisco I-688
Abdel-Dayem, Amr R. II-191
Adán, Antonio II-33
Aguiar, Rui II-158
Ahmed, Maher I-368, I-400
Ahn, Sang Chul I-261
Al Shaher, Abdullah I-335
Al-Mazeed, Ahmad II-363
Alajlan, Naif I-139, I-745
Alba-Castro, José Luis II-323, 660
Alegre, Enrique II-589
Alemán-Flores, Miguel II-339
Alexander, Simon K. I-236
Álvarez-León, Luis II-339
Alves, E. Ivo II-489
Ampornaramveth, V. I-530
Angel, L. I-705
Antequera, T. II-150
Ascenso, João I-588
Atine, Jean-Charles I-769
Atkinson, Gary I-621
Austin, Jim II-684
Ávila, Bruno Tenório II-234, II-249
Ávila, J.A. II-150
Azhar, Hannan Bin I-556
Azimifar, Zohreh II-331

Baek, Sunkyoung I-471
Bailly, G. II-100
Bak, EunSang I-49
Bandeira, Lourenço II-226
Banerjee, A. II-421
Banerjee, N. II-421
Barata, Teresa II-489
Barreira, N. II-43
Batista, Jorge P. II-552
Batouche, Mohamed I-147
Bedini, Luigi II-241
Beirão, Céu L. II-841
Belhadj-aissa, Aichouche I-866
Berar, M. II-100
Bernardino, Alexandre I-538, II-454
Bevilacqua, Alessandro II-481
Bhuiyan, M.A. I-530
Borgosz, Jan I-721
Bouchemakh, Lynda I-866
Bozma, H. Işıl I-285
Brahma, S. II-421
Brassart, Eric II-471
Breckon, Toby P. I-680
Bres, Stéphane I-825
Brun, Luc I-840
Bruni, V. I-179
Bueno, Gloria II-33
Bui, Tien D. I-82

Caderno, I.G. II-132
Caldas Pinto, João R. I-253, II-226, II-802
Calpe-Maravilla, J. II-429
Camahort, Emilio I-688
Campilho, Ana II-166
Campilho, Aurélio II-59, II-108, II-158, II-166, II-372
Camps-Valls, G. II-429
Carmona-Poyato, A. I-424
Caro, A. II-150
Carreira, M.J. I-212, II-132
Castelán, Mario I-613
Castrillon, M. II-725
Cazes, T.B. II-389
Chanda, Bhabatosh II-217
Chen, Jia-Xin II-581
Chen, Mei I-220
Chen, Xinjian I-360
Chen, Yan II-200
Chen, Ying I-269
Chen, Zezhi I-638
Cherifi, Hocine I-580, II-289
Chi, Yanling I-761
Cho, Miyoung I-471
Cho, Sang-Hyun II-597
Choe, J. II-446
Choe, Jihwan I-597
Choi, E. II-446
Chowdhury, S.P. II-217
Chung, Yongwha II-770
Civanlar, Reha I-285
Clérentin, Arnaud II-471
Cloppet, F. II-84

Conte, D. II-614
Cooray, Saman II-741
Cordeiro, Viviane I-187
Cordova, M.S. II-834
Corkidi, G. II-834
Correia Miguel V. II-372, II-397
Cosío, Fernando Arámbula II-76
Császár, Gergely I-811
Csordás, Dezső I-811
Cyganek, Boguslaw I-721
Czúni, László I-811

Dang, Anrong I-195, I-269
Das, A.K. II-217
Dawood, Mohammad II-544
De Backer, Steve II-497
de Mello, Carlos A.B. II-209
De Santo, M. I-564
de With, Peter H.N. II-651
Debruyn, Walter II-497
Dejnozkova, Eva I-416
Delahoche, Laurent II-471
Denis, Nicolas I-318
Deniz, O. II-725
Desvignes, M. II-100
Dikici, Çağatay I-285
Dimond, Keith I-556
Dios, J.R. Martinez-de I-90, I-376
Ditrich, Frank I-629
Di Stefano, Luigi I-408, II-437, II-481
Doguscu, Sema I-432
Dokladal, Petr I-416
Domínguez, Sergio I-318, I-833
Dopico, Antonio G. II-397
Dosil, Raquel I-655
Doulaverakis, C. I-310
Draa, Amer I-147
du Buf, J.M. Hans I-664
Durán, M.L. II-150

El Hassouni, Mohammed I-580
El Rube', Ibrahim I-368
El-Sakka, Mahmoud R. II-191, II-759
Elarbi Boudihir, M. II-563

Falcon, A. II-725
Fang, Jianzhong I-503
Fathy, Mahmood II-623
Faure, A. II-84
Fdez-Vidal, Xosé R. I-655
Feitosa, R.Q. II-389
Feng, Xiangchu I-479
Feng, Xiaoyi II-668
Fernández, Cesar I-547
Fernández, J.J. II-141
Fernández-García, N.L. I-424
Ferreiro-Armán, M. II-323
Fieguth, Paul I-9, I-114, I-163, I-236, I-572, I-745, II-314, II-331
Figueiredo, Mário A. T. II-841
Filip, Jiří II-298
Fisher, Mark I-848, I-858
Fisher, Robert B. I-680
Flusser, Jan I-122
Foggia, P. II-614

Galindo, E. II-834
Galinski, Grzegorz I-729
Gao, Song I-82
Gao, Wen II-520, II-778
Gao, Xinbo I-74, II-381
Garcia, Bernardo II-166
Garcia, Christophe II-717
García, D. I-705
García, I. II-141
García-Pérez, David I-795
García-Sevilla, Pedro I-25
Ghaffar, Rizwan II-512
Glory, E. II-84
Gómez-Chova, L. II-429
Gomez-Ulla, F. II-132
Gonçalves, Paulo J. Sequeira I-253
González, F. II-132
González, J.M. I-705
González-Jiménez, Daniel II-660
Gou, Shuiping I-41
Gregson, Peter H. I-130
Gu, Junxia II-381
Guidobaldi, C. II-614
Guimarães, Leticia I-187
Gunn, Steve II-363

Hadid, Abdenour II-668
Hafiane, Adel I-787
Haindl, Michal II-298, II-306
Hamou, Ali K. II-191
Han, Dongil I-384
Hancock, Edwin R. I-327, I-335, I-352, I-613, I-621, II-733
Hanson, Allen I-519

Hao, Pengwei I-195, I-269
Hasanuzzaman, M. I-530
Havasi, Laszlo II-347
Hernández, Sergio II-826
Heseltine, Thomas II-684
Hotta, Kazuhiro II-405
Howe, Nicholas R. I-803
Huang, Xiaoqiang I-848, I-858

Ideses, Ianir II-273
Iivarinen, Jukka I-753
Izri, Sonia II-471

Jafri, Noman II-512
Jalba, Andrei C. I-1
Jamzad, Mansour II-794
Jeong, Pangyu I-228
Jeong, T. II-446
Jernigan, Ed I-139, I-163, II-331
Ji, Hongbing I-74
Jia, Ying II-572
Jiang, Xiaoyi II-544
Jiao, Licheng I-41, I-455, I-479, I-487, II-504
Jin, Fu I-572
Jin, Guoying I-605
Jung, Kwanho I-471

Kabir, Ehsanollah I-818
Kamel, Mohamed I-244, I-368, I-400, I-745, II-25, II-51
Kang, Hang-Bong II-597
Kangarloo, Kaveh I-818
Kartikeyan, B. II-421
Kempeneers, Pieter II-497
Khan, Shoab Ahmed II-512
Khelifi, S.F. II-563
Kim, Hyoung-Gon I-261
Kim, Ig-Jae I-261
Kim, Kichul II-770
Kim, Min II-770
Kim, Pankoo I-471
Kim, Tae-Yong II-528, II-536
Kobatake, Hidefumi I-697
Kong, Hyunjang I-471
Kopricky, Miroslav I-400
Kourgli, Assia I-866
Kucharski, Krzysztof I-511
Kumazawa, Itsuo II-9
Kutics, Andrea I-737
Kwon, Yong-Moo I-261
Kwon, Young-Bin I-392

Lam, Kin-Man I-65
Landabaso, Jose-Luis II-463
Lanza, Alessandro II-481
Laurent, Christophe II-717
Lee, Chulhee I-597, II-446
Lee, Seong-Whan II-536
Lee, Tae-Seong I-261
Lefèvre, Sébastien II-606
Leung, Maylor K.H. I-761
Le Troter, Arnaud II-265
Li, Gang I-171
Li, Jie II-381
Li, Minglu II-116
Li, Xin II-572
Li, Yang II-733
Li, Yanxia II-200
Liang, Bojian I-638
Lieutaud, Simon I-778
Limongiello, A. II-614
Lins, Rafael Dueire II-175, II-234, II-249
Lipikorn, Rajalida I-697
Liu, Kang I-487
Liu, Shaohui II-520, II-778
Liu, Yazhou II-520, II-778
Lotfizad, A. Mojtaba II-623
Lukac, Rastislav I-155, II-1, II-124, II-281
Luo, Bin I-327

Ma, Xiuli I-455
Madeira, Joaquim II-68
Madrid-Cuevas, F.J. I-424
Majumdar, A.K. I-33
Majumder, K.L. II-421
Mandal, S. II-217
Manuel, João II-92
Marengoni, Mauricio I-519
Marhic, Bruno II-471
Mariño, C. II-132
Marques, Jorge S. I-204
Martín-Guerrero, J.D. II-429
Martín-Herrero, J. II-323
Martínez-Albalá, Antonio II-33
Martínez-Usó, Adolfo I-25
Mattoccia, Stefano I-408, II-437
Mavromatis, Sebastien II-265

McDermid, John I-638
McGeorge, Peter I-295
Meas-Yedid, V. II-84
Medina, Olaya II-818
Medina-Carnicer, R. I-424
Melo, José II-454
Mendez, J. II-725
Mendonça, Ana Maria II-108, II-158
Mendonça, L.F. I-253
Mery, Domingo I-647, II-818, II-826
Meyer, Fernand I-840
Meynet, Julien II-709
Micó, Luisa I-440
Mikeš, Stanislav II-306
Mirmehdi, Majid I-212, II-810
Mochi, Matteo II-241
Mohamed, S.S. II-51
Mohan, M. I-33
Mola, Martino II-437
Moon, Daesung II-770
Moreira, Rui II-108
Moreno, J. II-429
Moreno, Plinio I-538
Mosquera, Antonio I-795
Mota, G.L.A. II-389

Naftel, Andrew II-454
Nair, P. II-421
Nakagawa, Akihiko I-737
Nedevschi, Sergiu I-228
Neves, António J.R. I-277
Nezamoddini-Kachouie, Nezamoddin I-163
Nicponski, Henry II-633
Nixon, Mark II-363
Nourine, R. II-563
Nunes, Luis M. II-397

O'Connor, Noel II-741
O'Leary, Paul II-849
Ochoa, Felipe I-647
Oh, Sang-Rok I-384
Oliver, Gabriel I-672
Olivo-Marin, J-Ch. II-84
Ollero, A. I-90, I-376
Ortega, Marcos I-795
Ortigosa, P.M. II-141
Ortiz, Alberto I-672
Ouda, Abdelkader H. II-759

Paiva, António R.C. I-302
Palacios, R. II-150
Palma, Duarte I-588
Pan, Sung Bum II-770
Pardas, Montse II-463
Pardo, Xosé M. I-655
Park, Hanhoon II-700
Park, Jaehwa I-392
Park, Jihun II-528, II-536
Park, Jong-Il II-700
Park, Sunghun II-528
Pastor, Moisés II-183
Pavan, Massimiliano I-17
Payá, Luis I-547
Payan, Y. II-100
Pears, Nick E. I-638, II-684
Pelillo, Marcello I-17
Penas, M. I-212
Penedo, Manuel G. I-212, I-795, II-43, II-132
Peng, Ning-Song II-581
Percannella, G. I-564
Pereira, Fernando I-588
Petrakis, E. I-310
Pezoa, Jorge E. II-413
Pietikäinen, Matti II-668
Pimentel, Luís II-226
Pina, Pedro II-226, II-489
Pinho, Armando J. I-277, I-302
Pinho, Raquel Ramos II-92
Pinset, Ch. II-84
Pla, Filiberto I-25
Plataniotis, Konstantinos N. II-1, II-281
Podenok, Leonid P. I-447
Popovici, Vlad II-709

Qin, Li II-17
Qiu, Guoping I-65, I-503

Ramalho, Mário II-226
Ramel, J.Y. II-786
Ramella, Giuliana I-57
Rautkorpi, Rami I-753
Redondo, J.L. II-141
Reindl, Ingo II-849
Reinoso, Oscar I-547
Richardson, Iain I-295
Ricketson, Amanda I-803
Rico-Juan, Juan Ramón I-440

Riseman, Edward I-519
Rital, Soufiane II-289
Rivero-Moreno, Carlos Joel I-825
Rizkalla, K. II-51
Robles, Vanessa II-589
Rodrigues, João I-664
Rodríguez, P.G. II-150
Roerdink, Jos B.T.M. I-1
Rueda, Luis II-17
Ryoo, Seung Taek I-98

Sabri, Mahdi II-314
Sadykhov, Rauf K. I-447
Sáez, Doris II-826
Sahin, Turker I-495, II-355
Sahraie, Arash I-295
Salama, M.M.A. I-244, II-25, II-51
Salerno, Emanuele II-241
Samokhval, Vladimir A. I-447
Sanches, João M. I-204
Sánchez, F. I-705
San Pedro, José I-318
Sanniti di Baja, Gabriella I-57
Sansone, C. I-564
Santos, Beatriz Sousa II-68
Santos, Jorge A. II-397
Santos-Victor, José I-538, II-454
Saraiva, José II-489
Sarkar, A. II-421
Schaefer, Gerald I-778, II-257
Schäfers, Klaus P. II-544
Scheres, Ben II-166
Scheunders, Paul II-497
Seabra Lopes, Luís I-463
Sebastián, J.M. I-705, II-589
Sener, Sait I-344
Sequeira, Jean II-265
Serrano-López, A.J. II-429
Shan, Tan I-479, II-504
Shimizu, Akinobu I-697
Shirai, Yoshiaki I-530
Silva, Augusto II-68
Silva, José Silvestre II-68
Skarbek, Władysław I-511, I-729
Smolka, Bogdan I-155, II-1, II-124, II-281
Soares, André I-187
Song, Binheng I-171
Sousa, António V. II-158
Sousa, João M.C. I-253, II-802
Sroubek, Filip I-122
Stamon, G. II-84
Suesse, Herbert I-629
Sun, Luo II-572
Sun, Qiang I-41
Sun, Yufei II-200
Sural, Shamik I-33
Susin, Altamiro I-187
Sziranyi, Tamas II-347
Szlavik, Zoltan II-347

Taboada, B. II-834
Takahashi, Haruhisa II-405
Talbi, Hichem I-147
Tao, Linmi I-605, II-572
Tavares, R.S. II-92
Tax, David M.J. I-463
Thiran, Jean-Philippe II-709
Thomas, Barry T. I-212, II-810
Tian, Jie I-360
Tombari, Federico I-408
Tonazzini, Anna II-241
Torres, Sergio N. II-413
Toselli, Alejandro II-183
Traver, V. Javier I-538
Tsui, Hung Tat I-713
Tsunekawa, Takuya II-405
Twardowski, Tomasz I-721

Ueno, H. I-530
Unel, Mustafa I-344, I-432, I-495, II-355
Uvarov, Andrey A. I-447

Vadivel, A. I-33
Vagionitis, S. I-310
Vautrot, Philippe I-840
Vega-Alvarado, L. II-834
Venetsanopoulos, Anastasios N. II-1
Vento, M. I-564, II-614
Vicente, M. Asunción I-547
Vidal, Enrique II-183
Vidal, René I-647
Vincent, Nicole II-606, II-786
Vinhais, Carlos II-59
Visani, Muriel II-717
Vitulano, D. I-179
Vivó, Roberto I-688
Voss, Klaus I-629
Vrscay, Edward R. I-236

Wang, Lei I-74
Wang, Lijun II-520, II-778
Wang, QingHua I-463
Wang, Yuzhong I-106
Wesolkowski, Slawo I-9
Wilkinson, Michael H.F. I-1
Wilson, Richard C. I-327
Winger, Lowell I-572
Wirotius, M. II-786
Wnukowicz, Karol I-729
Wong, Kwan-Yee Kenneth II-676
Wong, Shu-Fai II-676

Xiao, Bai I-352
Xie, Jun I-713
Xie, Xianghua II-810
Xu, Guangyou I-605, II-572
Xu, Li-Qun II-463
Xu, Qianren I-244, II-25

Yaghmaee, Farzin II-794
Yang, Jie I-106, II-581
Yang, Xin I-360, II-643, II-692
Yano, Koji II-9
Yao, Hongxun II-520, II-778
Yaroslavsky, Leonid II-273
Yazdi, Hadi Sadoghi II-623
Yi, Hongwen I-130
Yin, Jianping II-750
You, Bum-Jae I-384
Yu, Hang I-352

Zavidovique, Bertrand I-787
Zervakis, M. I-310
Zhang, Chao I-195
Zhang, Guomin II-750
Zhang, Tao I-530
Zhang, Xiangrong II-504
Zhang, Yuzhi II-200
Zhao, Yongqiang II-116
Zhong, Ying I-295
Zhou, Dake II-643, II-692
Zhou, Yue I-106
Zhu, En II-750
Zhu, Yanong I-848
Zilberstein, Shlomo I-519
Zuo, Fei II-651