Lecture Notes in Artificial Intelligence 3802

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Yue Hao Jiming Liu Yuping Wang
Yiu-ming Cheung Hujun Yin
Licheng Jiao Jianfeng Ma
Yong-Chang Jiao (Eds.)

Computational Intelligence and Security

International Conference, CIS 2005
Xi'an, China, December 15-19, 2005
Proceedings, Part II

Springer

Volume Editors

Yue Hao
E-mail: yhao@xidian.edu.cn

Jiming Liu
E-mail: jiming@comp.hkbu.edu.hk

Yuping Wang
E-mail: ywang@xidian.edu.cn

Yiu-ming Cheung
E-mail: ymc@comp.hkbu.edu.hk

Hujun Yin
E-mail: hujun.yin@manchester.ac.uk

Licheng Jiao
E-mail: lchjiao@mail.xidian.edu.cn

Jianfeng Ma
E-mail: jfma@mail.xidian.edu.cn

Yong-Chang Jiao
E-mail: ychjiao@xidian.edu.cn

Library of Congress Control Number: 2005937071

CR Subject Classification (1998): I.2, H.3, H.4, H.5, F.2.2, I.4

ISSN 0302-9743
ISBN-10 3-540-30819-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-30819-5 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 2005
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11596981 06/3142 5 4 3 2 1 0

Preface

The International Conference on Computational Intelligence and Security (CIS) is an annual international conference that brings together researchers, engineers, developers and practitioners from both academia and industry to share experience and exchange and cross-fertilize ideas on all areas of computational intelligence and information security. The conference serves as a forum for the dissemination of state-of-the-art research and the development, and implementations of systems, technologies and applications in these two broad, interrelated fields. This year CIS 2005 was co-organized by the IEEE (Hong Kong) Computational Intelligence Chapter and Xidian University, and co-sponsored by Hong Kong Baptist University, National Natural Science Foundation of China, Key Laboratory of Computer Networks and Information Security of the Ministry of Education of China, and Guangdong University of Technology. CIS 2005 received in total 1802 submissions from 41 countries and regions all over the world. All of them were strictly peer reviewed by the Program Committee and experts in the field. Finally, 337 high-quality papers were accepted yielding an acceptance rate of 18.7%. Among them, 84 papers are the extended papers and 253 are the regular papers. The conference was greatly enriched by a wide range of topics covering all areas of computational intelligence and information security. Furthermore, tutorials and workshops were held for discussions of the proposed ideas. Such practice is extremely important for the effective development of the two fields and computer science in general.

We would like to thank the organizers: the IEEE (Hong Kong) Computational Intelligence Chapter and Xidian University for their great contributions and efforts in this big event. Thanks also go to the sponsors, the Institute of Electrical and Electronics Engineers (IEEE), Hong Kong Baptist University (HKBU), National Natural Science Foundation of China, Key Laboratory of Computer Networks and Information Security of the Ministry of Education of China, Guangdong University of Technology (GDUT), and the publisher, Springer, for their unremitting support and collaboration to make CIS 2005 possible and successful. Furthermore, we would like to sincerely thank the Program Committee members and additional reviewers for their professional, efficient input to the review process. Last but not the least, the Organizing Committee is much appreciated for their enormous efforts and marvelous work.

October 2005

Yue Hao and Jiming Liu
General Co-chairs of CIS 2005
Yuping Wang, Yiu-ming Cheung
Hujun Yin and Licheng Jiao
Program Committee Co-chairs of CIS 2005
Jianfeng Ma and Yong-Chang Jiao
Organizing Committee Co-chairs of CIS 2005

Organization

CIS 2005 was organized by IEEE (Hong Kong) Computational Intelligence Chapter and Xidian University.

General Co-chairs

Yue Hao	Xidian University, China
Jiming Liu	Hong Kong Baptist University, Hong Kong, China

Steering Committee Chair

Yiu-ming Cheung	Hong Kong Baptist University, Hong Kong, China

Organizing Committee

Jianfeng Ma	Xidian University, China (Co-chair)
Yong-Chang Jiao	Xidian University, China (Co-chair)
Hailin Liu	Guangdong University of Technology, China (Tutorial and Workshop Chair)
Hecheng Li	Xidian University, China (Treasurer)
Lixia Han	Xidian University, China (Publicity Chair)
Yuen-tan Hou	Hong Kong Baptist University, Hong Kong, China (Publicity Chair)
Liang Ming	Xidian University, China (Registration Chair)
Yuanyuan Zuo	Xidian University, China (Local Arrangement Chair)
Shuguang Zhao	Xidian University, China (Publication Chair)
Kapluk Chan	Nanyang Technological University, Singapore (Asia Liaison)
Yan Wu	Xidian University, China (Secretary)
Jingxuan Wei	Xidian University, China (Secretary)
Rongzu Yu	Xidian University, China (Web Master)

Program Committee

Yiu-ming Cheung (Co-chair) (Hong Kong, China)
Licheng Jiao (Co-chair) (China)

Yuping Wang (Co-chair) (China)
Hujun Yin (Co-chair) (UK) Michel Abdalla (France)
Khurshid Ahmad (UK)
Francesco Amigoni (Italy)
Sherlock Au (Hong Kong, China)
Dunin-Keplicz Barbara (Poland)
Mike Barley (New Zealand)
Chaib-draa Brahim (Canada)
Tony Browne (UK)
Scott Buffett (Canada)
Matthew Casey (UK)
Dario Catalano (France)
Kapluk Chan (Singapore)
Keith Chun-chung Chan (Hong Kong, China)
Michael Chau (Hong Kong, China)
Sheng Chen (UK)
Songcan Chen (China)
Zheng Chen (China)
Xiaochun Cheng (UK)
William Cheung (Hong Kong, China)
Sungzoon Cho (Korea)
Paolo Ciancarini (Italy)
Stelvio Cimato (Italy)
Helder Coelho (Portugal)
Carlos Coello (USA)
Emilio Corchado (Spain)
Wei Dai (Australia)
Joerg Denzinger (Canada)
Tharam Dillon (Australia)
Tom Downs (Australia)
Richard Everson (UK)
Bin Fang (China)
Marcus Gallagher (Australia)
Matjaz Gams (Slovenia)
John Qiang Gan (UK)
Joseph A. Giampapa (USA)
Maria Gini (USA)
Eric Gregoire (France)
Heikki Helin (Finland)
Tony Holden (UK)
Vasant Honavar (USA)
Mike Howard (USA)
Huosheng Hu (UK)
Yupu Hu (China)
Marc van Hulle (Belgium)
Michael N. Huhns (USA)
Samuel Kaski (Finland)
Sokratis Katsikas (Greece)
Hiroyuki Kawano (Japan)
John Keane (UK)
Alvin Kwan (Hong Kong, China)
Kwok-Yan Lam (Singapore)
Loo Hay Lee (Singapore)
Bicheng Li (China)
Guoping Liu (UK)
Huan Liu (USA)
Zhe-Ming Lu (Germany)
Magdon-Ismail Malik (USA)
Xiamu Niu (China)
Wenjiang Pei (China)
Hartmut Pohl (Germany)
Javier Lopez (Spain)
V. J. Rayward-Smith (UK)
Henry H.Q. Rong (Hong Kong, China)
Guenter Rudolph (Germany)
Patrizia Scandurra (Italy)
Bernhard Sendhoff (Germany)
Michael Small (Hong Kong, China)
Vic Rayward Smith (UK)
Fischer-Huebner Simone (Sweden)
Stephanie Teufel (Switzerland)
Peter Tino (UK)
Christos Tjortjis (UK)
Vicenc Torra (Spain)
Kwok-ching Tsui (Hong Kong, China)
Bogdan Vrusias (UK)
Bing Wang (UK)
Ke Wang (Canada)
Haotian Wu (Hong Kong, China)
Gaoxi Xiao (Singapore)
Hongji Yang (UK)
Shuang-Hua Yang (UK)
Zheng Rong Yang (UK)
Xinfeng Ye (New Zealand)
Benjamin Yen (Hong Kong, China)
Dingli Yu (UK)
Jeffrey Yu (Hong Kong, China)
Qingfu Zhang (UK)

Additional Reviewers

Ailing Chen
Asim Karim
bian Yang
Bin Song
Binsheng Liu
Bo An
Bo Chen
Bo Cheng
Bo Liu
Bo Zhang
Bobby D. Gerardo
Byunggil Lee
Changan Yuan
Changhan Kim
Changji Wang
Changjie Wang
Changsheng Yi
Chanyun Yang
Chiachen Lin
Chienchih Yu
Chinhung Liu
Chong Liu
Chong Wang
Chong Wu
Chongzhao Han
Chundong Wang
Chunhong Cao
Chunxiang Gu
Chunxiang Xu
Chunyan Liang
Congfu Xu
Cuiran Li
Daoyi Dong
Darong Huang
Dayong Deng
Dechang Pi
Deji Wang
Dong Zhang
Dongmei Fu
Enhong Chen
Eunjun Yoon
Fan Zhang
Fang Liu
Fei Liu
Fei Yu
Feng Liu
Fengzhan Tian
Fuyan Liu
Fuzheng Yang
Guangming Shi
Guicheng Wang
Guiguang Ding
Guimin Chen
Gumin Jeong
Guogang Tian
Guojiang Fu
Guowei Yang
Haewon Choi
Haiguang Chen
Haiyan Jin
Hansuh Koo
Heejo Lee
Heejun Song
Hong Liu
Hong Zhang
Hongbin Shen
Hongbing Liu
Hongfang Zhou
Hongsheng Su
Hongtao Wu
Hongwei Gao
Hongwei Huo
Hongxia Cai
Hongzhen Zheng
Horong Henry
Hsinhung Wu
Hua Xu
Hua Zhong
Huafeng Deng
Huanjun Liu
Huantong Geng
Huaqiu Wang
Hui Wang
Huiliang Zhang
Huiying Li
Hyun Sun Kang
Hyungkyu Yang
Hyungwoo Kang
Jeanshyan Wang
Jian Liao
Jian Wang
Jian Zhao
Jianming Fu
Jianping Zeng
Jianyu Xiao
Jiawei Zhang
Jieyu Meng
Jiguo Li
Jing Han
Jing Liu
Jingmei liu
Jinlong Wang
Jinqian Liang
Jin-Seon Lee
Jinwei Wang
Jongwan Kim
Juan Zhang
Jun Kong
Jun Ma
Jun Zhang
Junbo Gao
Juncheol Park
Junying Zhang
K. W. Chau
Kang-Soo You
Kanta Matsuura
Ke Liao
Keisuke Takemori
Kihyeon Kwon
Kongfa Hu
Kyoungmi Lee
Lei Sun
Li Wang
Liang Zhang
Liangli Ma
Liangxiao Jiang
Lianying Zhou
Liaojun Pang
Libiao Zhang

Licheng Wang
Liefeng Bo
Lijun Wu
Limin Wang
Lin Lin
Ling Chen
Ling Zhang
Lingfang Zeng
Linhuan Wang
Liping Yan
Meng Zhang
Mengjie Yu
Ming Dong
Ming Li
Ming Li
Minxia Luo
Murat Ekinci
Ni Zhang
Noria Foukia
Omran Mahamed
Pengfei Liu
Ping Guo
Purui Su
Qi Liu
Qi Xia
Qian Chen
Qiang Guo
Qiang Wang
Qiang Wei
Qiang Zhang
Qin Wang
Qinghao Meng
Qinghua Hu
Qiuhua Zheng
Qizhi Zhang
Ravi Prakash
Ronghua Shang
Rongjun Li
Rongqing Yi
Rongxing Lu
Ruo Hu
Sangho Park
Sangyoung Lee
Seonghoon Lee
Seunggwan Lee
Shangmin Luan
Shaowei Wang
Shaozhen Chen
Shengfeng Tian
Shenghui Su
Shi Min
Shifeng Rui
Shiguo Lian
Shuping Yao
Sinjae Kang
Songwook Lee
Sooyeon Shin
Sunghae Jun
Sungjune Hong
Suresh Sundaram
Tangfei Tao
Tao Guan
Tao Peng
Teemupekka Virtanen
Terry House
Tianding Chen
Tianyang Dong
Tieli Sun
Tingzhu Huangy
W. X. Wang
Wei Chen
Wei Hu
Wei Yan
Wei Zhang
Weiguo Han
Weijun Chen
Weimin Xue
Weixing Wang
Weizhen Yan
Wencang Zhao
Wenjie Li
Wen Quan
Wensen An
Wenyuan Wang
Xia Xu
Xiangchong Liu
Xiangchu Feng
Xiangyong Li
Xianhua Dai
Xiaofeng Chen
Xiaofeng Liu
Xiaofeng Rong
Xiaohong Hao
Xiaohui Yuan
Xiaoliang He
Xiaoyan Tao
Xiaozhu Lin
Xijian Ping
Xinbo Gao
Xinchen Zhang
Xingzhou Zhang
Xinling Shi
Xinman Zhang
XiuPing Guo
Xiuqin Chu
Xuebing Zhou
Xuelong Chen
Yajun Guo
Yan Wang
Yanchun Liang
Yao Wang
Yeonseung Ryu
Yi Li
Yibo Zhang
Yichun Liu
Yifei Pu
Yijia Zhang
Yijuan Shi
Yijun Mo
Yilin Lin
Yiming Zhou
Yinliang Zhao
Yintian Liu
Yong Xu
Yong Zhang
Yongjie Wang
Yongjun Li
Yuan Shu
Yuan Yuan
Yubao Liu
Yufeng Zhang
Yufu Ning
Yukun Cao
Yunfeng Li
Yuqian Zhao

Zaobin Gan
Zhansheng Liu
Zhaofeng Ma
Zhenchuan Chai
Zhengtao Jiang
Zhenlong Du
Zhi Liu
Zhian Cheng
Zhicheng Chen
Zhigang Ma
Zhihong Deng
Zhihua Cai
Zhiqing Meng
Zhisong Pan
Zhiwei Ni
Zhiwu Liao
Zhong Liu
Ziyi Chen
Zongben Xu

Sponsoring Institutions

IEEE (Hong Kong) Computational Intelligence Chapter
Xidian University
Hong Kong Baptist University
National Natural Science Foundation of China
Guangdong University of Technology

Table of Contents – Part II

Cryptography and Coding

A Fast Inversion Algorithm and Low-Complexity Architecture over $GF(2^m)$
Sosun Kim, Nam Su Chang, Chang Han Kim, Young-Ho Park, Jongin Lim 1

An ID-Based Optimistic Fair Signature Exchange Protocol from Pairings
Chunxiang Gu, Yuefei Zhu, Yajuan Zhang 9

FMS Attack-Resistant WEP Implementation Is Still Broken
Toshihiro Ohigashi, Yoshiaki Shiraishi, Masakatu Morii 17

Design of a New Kind of Encryption Kernel Based on RSA Algorithm
Ping Dong, Xiangdong Shi, Jiehui Yang 27

On the Security of Condorcet Electronic Voting Scheme
Yoon Cheol Lee, Hiroshi Doi 33

Special Distribution of the Shortest Linear Recurring Sequences in $Z/(p)$ Field
Qian Yin, Yunlun Luo, Ping Guo 43

Cryptanalysis of a Cellular Automata Cryptosystem
Jingmei Liu, Xiangguo Cheng, Xinmei Wang 49

A New Conceptual Framework Within Information Privacy: Meta Privacy
Geoff Skinner, Song Han, Elizabeth Chang 55

Error Oracle Attacks on Several Modes of Operation
Fengtong Wen, Wenling Wu, Qiaoyan Wen 62

Stability of the Linear Complexity of the Generalized Self-shrinking Sequences
Lihua Dong, Yong Zeng, Yupu Hu 68

On the Construction of Some Optimal Polynomial Codes
Yajing Li, Weihong Chen 74

Perceptual Hashing of Video Content Based on Differential Block Similarity
Xuebing Zhou, Martin Schmucker, Christopher L. Brown 80

Cryptographic Protocols

Secure Software Smartcard Resilient to Capture
Seung Wook Jung, Christoph Ruland 86

Revised Fischlin's (Blind) Signature Schemes
Kewei Lv 96

Certificateless Threshold Signature Schemes
Licheng Wang, Zhenfu Cao, Xiangxue Li, Haifeng Qian 104

An Efficient Certificateless Signature Scheme
M. Choudary Gorantla, Ashutosh Saxena 110

ID-Based Restrictive Partially Blind Signatures
Xiaofeng Chen, Fangguo Zhang, Shengli Liu 117

Batch Verification with DSA-Type Digital Signatures for Ubiquitous Computing
Seungwon Lee, Seongje Cho, Jongmoo Choi, Yookun Cho 125

On Anonymity of Group Signatures
Sujing Zhou, Dongdai Lin 131

The Running-Mode Analysis of Two-Party Optimistic Fair Exchange Protocols
Yuqing Zhang, Zhiling Wang, Bo Yang 137

Password-Based Group Key Exchange Secure Against Insider Guessing Attacks
Jin Wook Byun, Dong Hoon Lee, Jongin Lim 143

On the Security of Some Password-Based Key Agreement Schemes
Qiang Tang, Chris J. Mitchell 149

A New Group Rekeying Method in Secure Multicast
Yong Xu, Yuxiang Sun 155

Pairing-Based Provable Blind Signature Scheme Without Random Oracles
Jian Liao, Yinghao Qi, Peiwei Huang, Mentian Rong 161

Efficient ID-Based Proxy Signature and Proxy Signcryption Form Bilinear Pairings
Qin Wang, Zhenfu Cao 167

An Identity-Based Threshold Signcryption Scheme with Semantic Security
Changgen Peng, Xiang Li 173

A Token-Based Single Sign-On Protocol
Li Hui, Shen Ting 180

Simple Threshold RSA Signature Scheme Based on Simple Secret Sharing
Shaohua Tang 186

Efficient Compilers for Authenticated Group Key Exchange
Qiang Tang, Chris J. Mitchell 192

Insider Impersonation-MIM Attack to Tripartite Key Agreement Scheme and an Efficient Protocol for Multiple Keys
Lihua Wang, Takeshi Okamoto, Tsuyoshi Takagi, Eiji Okamoto 198

Intrusion Detection

An Immune System Inspired Approach of Collaborative Intrusion Detection System Using Mobile Agents in Wireless Ad Hoc Networks
Ki-Won Yeom, Ji-Hyung Park 204

A New User-Habit Based Approach for Early Warning of Worms
Ping Wang, Binxing Fang, Xiaochun Yun 212

A Multi-gigabit Virus Detection Algorithm Using Ternary CAM
Il-Seop Song, Youngseok Lee, Taeck-Geun Kwon 220

Sampling Distance Analysis of Gigantic Data Mining for Intrusion Detection Systems
Yong Zeng, Jianfeng Ma 228

Hardware-Software Hybrid Packet Processing for Intrusion Detection Systems
Saraswathi Sachidananda, Srividya Gopalan, Sridhar Varadarajan . . . 236

D-S Evidence Theory and Its Data Fusion Application in Intrusion Detection
Junfeng Tian, Weidong Zhao, Ruizhong Du 244

A New Network Anomaly Detection Technique Based on Per-Flow and Per-Service Statistics
Yuji Waizumi, Daisuke Kudo, Nei Kato, Yoshiaki Nemoto 252

SoIDPS: Sensor Objects-Based Intrusion Detection and Prevention System and Its Implementation
SeongJe Cho, Hye-Young Chang, HongGeun Kim, WoongChul Choi 260

A Statistical Model for Detecting Abnormality in Static-Priority Scheduling Networks with Differentiated Services
Ming Li, Wei Zhao 267

Tamper Detection for Ubiquitous RFID-Enabled Supply Chain
Vidyasagar Potdar, Chen Wu, Elizabeth Chang 273

Measuring the Histogram Feature Vector for Anomaly Network Traffic
Wei Yan 279

Efficient Small Face Detection in Surveillance Images Using Major Color Component and LDA Scheme
Kyunghwan Baek, Heejun Jang, Youngjun Han, Hernsoo Hahn 285

Fast Motion Detection Based on Accumulative Optical Flow and Double Background Model
Jin Zheng, Bo Li, Bing Zhou, Wei Li 291

Reducing Worm Detection Time and False Alarm in Virus Throttling
Jangbok Kim, Jaehong Shim, Gihyun Jung, Kyunghee Choi 297

Protection Against Format String Attacks by Binary Rewriting
Jin Ho You, Seong Chae Seo, Young Dae Kim, Jun Yong Choi, Sang Jun Lee, Byung Ki Kim 303

Masquerade Detection System Based on Principal Component Analysis and Radial Basics Function
Zhanchun Li, Zhitang Li, Yao Li, Bin Liu 309

Anomaly Detection Method Based on HMMs Using System Call and Call Stack Information
Cheng Zhang, Qinke Peng 315

Parallel Optimization Technology for Backbone Network Intrusion Detection System
Xiaojuan Sun, Xinliang Zhou, Ninghui Sun, Mingyu Chen 322

Attack Scenario Construction Based on Rule and Fuzzy Clustering
Linru Ma, Lin Yang, Jianxin Wang 328

A CBR Engine Adapting to IDS
Lingjuan Li, Wenyu Tang, Ruchuan Wang 334

Application of Fuzzy Logic for Distributed Intrusion Detection
Hee Suk Seo, Tae Ho Cho 340

Security Models and Architecture

Dynamic Access Control for Pervasive Grid Applications
Syed Naqvi, Michel Riguidel 348

On the Security of the Canetti-Krawczyk Model
Xinghua Li, Jianfeng Ma, SangJae Moon 356

A Novel Architecture for Detecting and Defending Against Flooding-Based DDoS Attacks
Yi Shi, Xinyu Yang 364

A Variant of Poly1305 MAC and Its Security Proof
Dayin Wang, Dongdai Lin, Wenling Wu 375

Covert Channel Identification Founded on Information Flow Analysis
Jianjun Shen, Sihan Qing, Qingni Shen, Liping Li 381

Real-Time Risk Assessment with Network Sensors and Intrusion Detection Systems
André Årnes, Karin Sallhammar, Kjetil Haslum, Tønnes Brekne, Marie Elisabeth Gaup Moe, Svein Johan Knapskog 388

Design and Implementation of a Parallel Crypto Server
Xiaofeng Rong, Xiaojuan Gao, Ruidan Su, Lihua Zhou 398

Survivability Computation of Networked Information Systems
Xuegang Lin, Rongsheng Xu, Miaoliang Zhu 407

Assessment of Windows System Security Using Vulnerability Relationship Graph
Yongzheng Zhang, Binxing Fang, Yue Chi, Xiaochun Yun 415

A New (t, n)-Threshold Multi-secret Sharing Scheme
HuiXian Li, ChunTian Cheng, LiaoJun Pang 421

An Efficient Message Broadcast Authentication Scheme for Sensor Networks
Sang-ho Park, Taekyoung Kwon 427

Digital Image Authentication Based on Error-Correction Codes
Fan Zhang, Xinhong Zhang, Zhiguo Chen 433

Design and Implementation of Efficient Cipher Engine for IEEE 802.11i Compatible with IEEE 802.11n and IEEE 802.11e
Duhyun Bae, Gwanyeon Kim, Jiho Kim, Sehyun Park, Ohyoung Song 439

Secure Delegation-by-Warrant ID-Based Proxy Signcryption Scheme
Shanshan Duan, Zhenfu Cao, Yuan Zhou 445

Building Security Requirements Using State Transition Diagram at Security Threat Location
Seong Chae Seo, Jin Ho You, Young Dae Kim, Jun Yong Choi, Sang Jun Lee, Byung Ki Kim 451

Study on Security iSCSI Based on SSH
Weiping Liu, Wandong Cai 457

A Scheduling Algorithm Based on a Trust Mechanism in Grid
Kenli Li, Yan He, Renfa Li, Tao Yang 463

Enhanced Security and Privacy Mechanism of RFID Service for Pervasive Mobile Device
Byungil Lee, Howon Kim 469

Worm Propagation Modeling and Analysis on Network
Yunkai Zhang, Fangwei Wang, Changguang Wang, Jianfeng Ma 476

An Extensible AAA Infrastructure for IPv6
Hong Zhang, Haixin Duan, Wu Liu, Jianping Wu 482

The Security Proof of a 4-Way Handshake Protocol in IEEE 802.11i
Fan Zhang, Jianfeng Ma, SangJae Moon 488

A Noble Key Pre-distribution Scheme with LU Matrix for Secure Wireless Sensor Networks
Chang Won Park, Sung Jin Choi, Hee Yong Youn 494

Security Management

A Virtual Bridge Certificate Authority Model
Haibo Tian, Xi Sun, Yumin Wang 500

Weak Signals in Information Security Management
Jorma Kajava, Reijo Savola, Rauno Varonen 508

PDTM: A Policy-Driven Trust Management Framework in Distributed Systems
Wu Liu, Haixin Duan, Jianping Wu, Xing Li 518

Methodology of Quantitative Risk Assessment for Information System Security
Mengquan Lin, Qiangmin Wang, Jianhua Li 526

A Secure and Efficient (t, n) Threshold Verifiable Multi-secret Sharing Scheme
Mei-juan Huang, Jian-zhong Zhang, Shu-cui Xie 532

Improvement on an Optimized Protocol for Mobile Network Authentication and Security
ChinChen Chang, JungSan Lee 538

Neural Network Based Flow Forecast and Diagnosis
Qianmu Li, Manwu Xu, Hong Zhang, Fengyu Liu 542

Protecting Personal Data with Various Granularities: A Logic-Based Access Control Approach
Bat-Odon Purevjii, Masayoshi Aritsugi, Sayaka Imai, Yoshinari Kanamori, Cherri M. Pancake 548

Enhancement of an Authenticated Multiple-Key Agreement Protocol Without Using Conventional One-Way Function
Huifeng Huang, Chinchen Chang 554

Topology-Based Macroscopical Response and Control Technology for Network Security Event
Hui He, Mingzeng Hu, Weizhe Zhang, Hongli Zhang, Zhi Yang 560

Watermarking and Information Hiding

Adaptive Hiding Scheme Based on VQ-Indices Using Commutable Codewords
Chinchen Chang, Chiachen Lin, Junbin Yeh 567

Reversible Data Hiding for Image Based on Histogram Modification of Wavelet Coefficients
Xiaoping Liang, Xiaoyun Wu, Jiwu Huang 573

An Image Steganography Using Pixel Characteristics
Young-Ran Park, Hyun-Ho Kang, Sang-Uk Shin, Ki-Ryong Kwon . . . 581

Alternatives for Multimedia Messaging System Steganography
Konstantinos Papapanagiotou, Emmanouel Kellinis, Giannis F. Marias, Panagiotis Georgiadis 589

Error Concealment for Video Transmission Based on Watermarking
Shuai Wan, Yilin Chang, Fuzheng Yang 597

Applying the AES and Its Extended Versions in a General Framework for Hiding Information in Digital Images
Tran Minh Triet, Duong Anh Duc 605

An Image Hiding Algorithm Based on Bit Plane
Bin Liu, Zhitang Li, Zhanchun Li 611

A Blind Audio Watermarking Algorithm Robust Against Synchronization Attack
Xiangyang Wang, Hong Zhao 617

Semi-fragile Watermarking Algorithm for Detection and Localization of Temper Using Hybrid Watermarking Method in MPEG-2 Video
Hyun-Mi Kim, Ik-Hwan Cho, A-Young Cho, Dong-Seok Jeong 623

Public Watermarking Scheme Based on Multiresolution Representation and Double Hilbert Scanning
Zhiqiang Yao, Liping Chen, Rihong Pan, Boxian Zou, Licong Chen 629

Performance Evaluation of Watermarking Techniques for Secure Multimodal Biometric Systems
Daesung Moon, Taehae Kim, SeungHwan Jung, Yongwha Chung, Kiyoung Moon, Dosung Ahn, Sang-Kyoon Kim 635

An Improvement of Auto-correlation Based Video Watermarking Scheme Using Independent Component Analysis
Seong-Whan Kim, Hyun-Sung Sung 643

A Digital Watermarking Technique Based on Wavelet Packages
Chen Xu, Weiqiang Zhang, Francis R. Austin 649

A Spectral Images Digital Watermarking Algorithm
Long Ma, Changjun Li, Shuni Song 655

Restoration in Secure Text Document Image Authentication Using Erasable Watermarks
Niladri B. Puhan, Anthony T.S. Ho 661

Web and Network Applications

The Study of RED Algorithm Used Multicast Router Based Buffer Management
Won-Hyuck Choi, Doo-Hyun Kim, Kwnag-Jae Lee, Jung-Sun Kim 669

Genetic Algorithm Utilized in Cost-Reduction Driven Web Service Selection
Lei Cao, Jian Cao, Minglu Li 679

MacroOS: A Pervasive Computing Platform Supporting Context Awareness and Context Management
Xiaohua Luo, Kougen Zheng, Zhaohui Wu, Yunhe Pan 687

A Frame for Selecting Replicated Multicast Servers Using Genetic Algorithm
Qin Liu, Chanle Wu 695

On a Novel Methodology for Estimating Available Bandwidth Along Network Paths
Shaohe Lv, Jianping Yin, Zhiping Cai, Chi Liu 703

A New AQM Algorithm for Enhancing Internet Capability Against Unresponsive Flows
Liyuan Zhao, Keqin Liu, Jun Zheng 711

Client Server Access: Wired vs. Wireless LEO Satellite-ATM Connectivity; A (MS-Ro-BAC) Experiment
Terry C. House 719

An Algorithm for Automatic Inference of Referential Integrities During Translation from Relational Database to XML Schema
Jinhyung Kim, Dongwon Jeong, Doo-Kwon Baik 725

A Fuzzy Integral Method to Merge Search Engine Results on Web
Shuning Cui, Boqin Feng 731

The Next Generation PARLAY X with QoS/QoE
Sungjune Hong, Sunyoung Han 737

A Design of Platform for QoS-Guaranteed Multimedia Services Provisioning on IP-Based Convergence Network
Seong-Woo Kim, Young-Chul Jung, Young-Tak Kim 743

Introduction of Knowledge Management System for Technical Support in Construction Industries
Tai Sik Lee, Dong Wook Lee, Jeong Hyun Kim 749

An Event Correlation Approach Based on the Combination of IHU and Codebook
Qiuhua Zheng, Yuntao Qian 757

Image and Signal Processing

Face Recognition Based on Support Vector Machine Fusion and Wavelet Transform
Bicheng Li, Hujun Yin 764

A Dynamic Face and Fingerprint Fusion System for Identity Authentication
Jun Zhou, Guangda Su, Yafeng Deng, Kai Meng, Congcong Li 772

Image Recognition for Security Verification Using Real-Time Joint Transform Correlation with Scanning Technique
Kyu B. Doh, Jungho Ohn, Ting-C Poon 780

Binarized Revocable Biometrics in Face Recognition
Ying-Han Pang, Andrew Teoh Beng Jin, David Ngo Chek Ling 788

Short Critical Area Computational Method Using Mathematical Morphology
Junping Wang, Yue Hao 796

A Robust Lane Detection Approach Based on MAP Estimate and Particle Swarm Optimization
Yong Zhou, Xiaofeng Hu, Qingtai Ye 804

MFCC and SVM Based Recognition of Chinese Vowels
Fuhai Li, Jinwen Ma, Dezhi Huang 812

A Spatial/Frequency Hybrid Vector Quantizer Based on a Classification in the DCT Domain
Zhe-Ming Lu, Hui Pei, Hans Burkhardt 820

Removing of Metal Highlight Spots Based on Total Variation Inpainting with Multi-sources-flashing
Ji Bai, Lizhuang Ma, Li Yao, Tingting Yao, Ying Zhang 826

Component-Based Online Learning for Face Detection and Verification
Kyoung-Mi Lee 832

SPIHT Algorithm Based on Fast Lifting Wavelet Transform in Image Compression
Wenbing Fan, Jing Chen, Jina Zhen 838

Modified EZW Coding for Stereo Residual
Han-Suh Koo, Chang-Sung Jeong 845

Optimal Prototype Filters for Near-Perfect-Reconstruction Cosine-Modulated Filter Banks
Xuemei Xie, Guangming Shi, Xuyang Chen 851

Fast Motion Estimation Scheme for Real Time Multimedia Streaming with H.264
Chan Lim, Hyun-Soo Kang, Tae-Yong Kim 857

Motion-Compensated 3D Wavelet Video Coding Based on Adaptive Temporal Lifting Filter Implementation
Guiguang Ding, Qionghai Dai, Wenli Xu 863

Accurate Contouring Technique for Object Boundary Extraction in Stereoscopic Imageries
Shin Hyoung Kim, Jong Whan Jang, Seung Phil Lee, Jae Ho Choi 869

Robust Object Tracking Based on Uncertainty Factorization Subspace Constraints Optical Flow
Yunshu Hou, Yanning Zhang, Rongchun Zhao 875

Bearings-Only Target Tracking Using Node Selection Based on an Accelerated Ant Colony Optimization
Benlian Xu, Zhiquan Wang 881

Image Classification and Delineation of Fragments
Weixing Wang 887

A Novel Wavelet Image Coding Based on Non-uniform Scalar Quantization
Guoyuo Wang, Wentao Wang 893

A General Image Based Nematode Identification System Design
Bai-Tao Zhou, Won Nah, Kang-Woong Lee, Joong-Hwan Baek 899

A Novel SVD-Based RLS Blind Adaptive Multiuser Detector for CDMA Systems
Ling Zhang, Xian-Da Zhang 905

New Electronic Digital Image Stabilization Algorithm in Wavelet Transform Domain
Jung-Youp Suk, Gun-Woo Lee, Kuhn-Il Lee 911

Line Segments and Dominate Points Detection Based on Hough Transform
Z.W. Liao, S.X. Hu, T.Z. Huang 917

The Study of the Auto Color Image Segmentation
Jian Zhuang, Haifeng Du, Jinhua Zhang, Sun'an Wang 923

Regularized Image Restoration by Means of Fusion for Digital Auto Focusing
Vivek Maik, Jeongho Shin, Joonki Paik 929

Fast Ray-Space Interpolation Based on Occlusion Analysis and Feature Points Detection
Gangyi Jiang, Liangzhong Fan, Mei Yu, Rangding Wang, Xien Ye, Yong-Deak Kim 935

Non-parametric ICA Algorithm for Hybrid Sources Based on GKNN Estimation
Fasong Wang, Hongwei Li, Rui Li, Shaoquan Yu 941

SUSAN Window Based Cost Calculation for Fast Stereo Matching
Kyu-Yeol Chae, Won-Pyo Dong, Chang-Sung Jeong 947

An Efficient Adaptive De-blocking Algorithm
Zhiliang Xu, Shengli Xie, Youjun Xiang 953

Facial Features Location by Analytic Boosted Cascade Detector
Lei Wang, Beiji Zou, Jiaguang Sun 959

New Approach for Segmentation and Pattern Recognition of Jacquard Images
Zhilin Feng, Jianwei Yin, Zhaoyang He, Wuheng Zuo, Jinxiang Dong 965

Nonstationarity of Network Traffic Within Multi-scale Burstiness Constraint
Jinwu Wei, Jiangxing Wu 971

Principle of Image Encrypting Algorithm Based on Magic Cube Transformation
Li Zhang, Shiming Ji, Yi Xie, Qiaoling Yuan, Yuehua Wan, Guanjun Bao 977

A Study on Motion Prediction and Coding for In-Band Motion Compensated Temporal Filtering
Dongdong Zhang, Wenjun Zhang, Li Song, Hongkai Xiong 983

Adaptive Sampling for Monte Carlo Global Illumination Using Tsallis Entropy
Qing Xu, Shiqiang Bao, Rui Zhang, Ruijuan Hu, Mateu Sbert 989

Applications

Incremental Fuzzy Decision Tree-Based Network Forensic System
Zaiqiang Liu, Dengguo Feng 995

Robust Reliable Control for a Class of Fuzzy Dynamic Systems with Time-Varying Delay
Youqing Wang, Donghua Zhou 1003

Using Concept Taxonomies for Effective Tree Induction
Hong Yan Yi, B. de la Iglesia, V.J. Rayward-Smith 1011

A Similarity-Based Recommendation Filtering Algorithm for Establishing Reputation-Based Trust in Peer-to-Peer Electronic Communities
Jingtao Li, Yinan Jing, Peng Fu, Gendu Zhang, Yongqiang Chen . . . 1017

Automatic Classification of Korean Traditional Music Using Robust Multi-feature Clustering
Kyu-Sik Park, Youn-Ho Cho, Sang-Hun Oh 1025

A Private and Efficient Mobile Payment Protocol
Changjie Wang, Ho-fung Leung 1030

Universal Designated-Verifier Proxy Blind Signatures for E-Commerce
Tianjie Cao, Dongdai Lin, Rui Xue 1036

An Efficient Control Method for Elevator Group Control System
Ulvi Dagdelen, Aytekin Bagis, Dervis Karaboga 1042

Next Generation Military Communication Systems Architecture
Qijian Xu, Naitong Zhang, Jie Zhang, Yu Sun 1048

Early Warning for Network Worms
Antti Tikkanen, Teemupekka Virtanen 1054

Skeleton Representation of Character Based on Multiscale Approach
Xinhua You, Bin Fang, Xinge You, Zhenyu He, Dan Zhang, Yuan Yan Tang 1060

Channel Equalization Based on Two Weights Neural Network
Wenming Cao, Wanfang Chai, Shoujue Wang 1068

Assessment of Uncertainty in Mineral Prospectivity Prediction Using Interval Neutrosophic Set
Pawalai Kraipeerapun, Chun Che Fung, Warick Brown 1074

Ring-Based Anonymous Fingerprinting Scheme
Qiang Lei, Zhengtao Jiang, Yumin Wang 1080

Scalable and Robust Fingerprinting Scheme Using Statistically Secure Extension of Anti-collusion Code
Jae-Min Seol, Seong-Whan Kim 1086

Broadcast Encryption Using Identity-Based Public-Key Cryptosystem
Lv Xixiang, Bo Yang 1092

Multimedia Digital Right Management Using Selective Scrambling for Mobile Handset
Goo-Rak Kwon, Tea-Young Lee, Kyoung-Ho Kim, Jae-Do Jin, Sung-Jea Ko 1098

Design and Implementation of Crypto Co-processor and Its Application to Security Systems
HoWon Kim, Mun-Kyu Lee, Dong-Kyue Kim, Sang-Kyoon Chung, Kyoil Chung 1104

Continuous Speech Research Based on HyperSausage Neuron
Wenming Cao, Jianqing Li, Shoujue Wang 1110

Variable-Rate Channel Coding for Space-Time Coded MIMO System
Changcai Han, Dongfeng Yuan 1116

A New Watermarking Method Based on DWT
Xiang-chu Feng, Yongdong Yang 1122

Efficient Point Rendering Method Using Sequential Level-of-Detail
Daniel Kang, Byeong-Seok Shin 1127

Construction of a Class of Compactly Supported Biorthogonal Multiple Vector-Valued Wavelets
Tongqi Zhang, Qingjiang Chen 1134

Metabolic Visualization and Intelligent Shape Analysis of the Hippocampus
Yoo-Joo Choi, Jeong-Sik Kim, Min-Jeong Kim, Soo-Mi Choi, Myoung-Hee Kim 1140

Characteristic Classification and Correlation Analysis of Source-Level Vulnerabilities in the Linux Kernel
Kwangsun Ko, Insook Jang, Yong-hyeog Kang, Jinseok Lee, Young Ik Eom 1149

Author Index 1157

Table of Contents – Part I

Learning and Fuzzy Systems

Empirical Analysis of Database Privacy Using Twofold Integrals
Jordi Nin, Vicenç Torra 1

Intrusion Detection Alert Verification Based on Multi-level Fuzzy Comprehensive Evaluation
Chengpo Mu, Houkuan Huang, Shengfeng Tian 9

Improving the Scalability of Automatic Programming
Henrik Berg, Roland Olsson 17

Texture Segmentation by Unsupervised Learning and Histogram Analysis Using Boundary Tracing
Woobeom Lee, Wookhyun Kim 25

An Improved Bayesian Network Learning Algorithm Based on Dependency Analysis
Fengzhan Tian, Shengfeng Tian, Jian Yu, Houkuan Huang 33

Mining Common Patterns on Graphs
Ivan Olmos, Jesus A. Gonzalez, Mauricio Osorio 41

Moderated Innovations in Self-poised Ensemble Learning
Ricardo Ñanculef, Carlos Valle, Héctor Allende, Claudio Moraga 49

An Adaptive Framework for Solving Multiple Hard Problems Under Time Constraints
Sandip Aine, Rajeev Kumar, P.P. Chakrabarti 57

An RLS-Based Natural Actor-Critic Algorithm for Locomotion of a Two-Linked Robot Arm
Jooyoung Park, Jongho Kim, Daesung Kang 65

Dynamic Clustering Using Multi-objective Evolutionary Algorithm
Enhong Chen, Feng Wang 73

Multimodal FeedForward Self-organizing Maps
Andrew P. Papliński, Lennart Gustafsson 81

Decision Fusion Based Unsupervised Texture Image Segmentation
Hua Zhong, Licheng Jiao 89

Speaker Adaptation Techniques for Speech Recognition with a Speaker-Independent Phonetic Recognizer
Weon-Goo Kim, MinSeok Jang 95

Fuzzy QoS Controllers in Diff-Serv Scheduler Using Genetic Algorithms
Baolin Sun, Qiu Yang, Jun Ma, Hua Chen 101

Neural Network Based Algorithms for Diagnosis and Classification of Breast Cancer Tumor
In-Sung Jung, Devinder Thapa, Gi-Nam Wang 107

New Learning Algorithm for Hierarchical Structure Learning Automata Operating in P-Model Stationary Random Environment
Yoshio Mogami 115

A TFN-Based AHP Model for Solving Group Decision-Making Problems
Jian Cao, Gengui Zhou, Feng Ye 121

A Tactics for Robot Soccer with Fuzzy Logic Mediator
Jeongjun Lee, Dongmin Ji, Wonchang Lee, Geuntaek Kang, Moon G. Joo 127

Gait Control for Biped Robot Using Fuzzy Wavelet Neural Network
Pengfei Liu, Jiuqiang Han 133

A New Approach for Regression: Visual Regression Approach
Deyu Meng, Chen Xu, Wenfeng Jing 139

Orthogonally Rotational Transformation for Naive Bayes Learning
Limin Wang, Chunhong Cao, Haijun Li, Haixia Chen, Liyan Dong 145

Efficient Learning Bayesian Networks Using PSO
Tao Du, S.S. Zhang, Zongjiang Wang 151

Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode
Zengyou He, Shengchun Deng, Xiaofei Xu 157

Distance Protection of Compensated Transmission Line Using Computational Intelligence
S.R. Samantaray, P.K. Dash, G. Panda, B.K. Panigrahi 163

Computational Intelligence for Network Intrusion Detection: Recent Contributions
Asim Karim 170

Evolutionary Computation

Design of a Switching PID Controller Using Advanced Genetic Algorithm for a Nonlinear System
Jung-Shik Kong, Bo-Hee Lee, Jin-Geol Kim 176

Preference Bi-objective Evolutionary Algorithm for Constrained Optimization
Yuping Wang, Dalian Liu, Yiu-Ming Cheung 184

Self-adaptive Differential Evolution
Mahamed G.H. Omran, Ayed Salman, Andries P. Engelbrecht 192

Steady-State Evolutionary Algorithm for Multimodal Function Global Optimization
Ziyi Chen, Lishan Kang 200

Finding Optimal Addition Chains Using a Genetic Algorithm Approach
Nareli Cruz-Cortés, Francisco Rodríguez-Henríquez, Raúl Juárez-Morales, Carlos A. Coello Coello 208

Using Reconfigurable Architecture-Based Intrinsic Incremental Evolution to Evolve a Character Classification System
Jin Wang, Je Kyo Jung, Yong-min Lee, Chong Ho Lee 216

On the Relevance of Using Gene Expression Programming in Destination-Based Traffic Engineering
Antoine B. Bagula, Hong F. Wang 224

Model and Convergence for the Combination of Genetic Algorithm and Ant Algorithm
Jianli Ding, Wansheng Tang, Yufu Ning 230

Moving Block Sequence and Organizational Evolutionary Algorithm for General Floorplanning
Jing Liu, Weicai Zhong, Licheng Jiao 238

Integrating the Simplified Interpolation into the Genetic Algorithm for Constrained Optimization Problems
Hong Li, Yong-Chang Jiao, Yuping Wang 247

Using Ensemble Method to Improve the Performance of Genetic Algorithm
Shude Zhou, Zengqi Sun 255

Parallel Mining for Classification Rules with Ant Colony Algorithm
Ling Chen, Li Tu 261

A Genetic Algorithm Approach on Reverse Logistics Optimization for Product Return Distribution Network
Gengui Zhou, Zhenyu Cao, Jian Cao, Zhiqing Meng 267

Multi-objective Evolutionary Design and Knowledge Discovery of Logic Circuits with an Improved Genetic Algorithm
Shuguang Zhao, Licheng Jiao, Jun Zhao 273

Robust Mobile Robot Localization Using an Evolutionary Particle Filter
Bo Yin, Zhiqiang Wei, Xiaodong Zhuang 279

Hybrid Genetic Algorithm for Solving the Degree-Constrained Minimal Bandwidth Multicast Routing Problem
Yun Pan, Zhenwei Yu, Licheng Wang 285

Using Fuzzy Possibilistic Mean and Variance in Portfolio Selection Model
Weiguo Zhang, Yingluo Wang 291

A Novel Genetic Algorithm for Multi-criteria Minimum Spanning Tree Problem
Lixia Han, Yuping Wang 297

Intelligent Agents and Systems

A Software Architecture for Multi-agent Systems
Vasu S. Alagar, Mao Zheng 303

User-Oriented Multimedia Service Using Smart Sensor Agent Module in the Intelligent Home
Jong-Hyuk Park, Jun Choi, Sang-Jin Lee, Hye-Ung Park, Deok-Gyu Lee 313

Meta-learning Experiences with the MINDFUL System
Ciro Castiello, Anna Maria Fanelli 321

Learning Cooperation from Classifier Systems
Trung Hau Tran, Cédric Sanza, Yves Duthen 329

Location Management Using Hierarchical Structured Agents for Distributed Databases
Romeo Mark A. Mateo, Bobby D. Gerardo, Jaewan Lee 337

On_line Measurement System of Virtual Dielectric Loss Based on Wavelets and LabVIEW and Correlation Technics
BaoBao Wang, Ye Wang 343

Model Checking Temporal Logics of Knowledge and Its Application in Security Verification
Lijun Wu, Kaile Su, Qingliang Chen 349

A Computational Approach for Belief Change
Shangmin Luan, Guozhong Dai 355

Feature Selection by Fuzzy Inference and Its Application to Spam-Mail Filtering
Jong-Wan Kim, Sin-Jae Kang 361

Design of Multiagent Framework for Cellular Networks
A.K. Sharma, Dimple Juneja 367

Transitive Dependence Based Formation of Virtual Organizations
Bo An, Chunyan Miao, Zhiqi Shen, Yuan Miao, Daijie Cheng 375

An Agent Based Education Resource Purvey System
Xiaochun Cheng, Xin He, Xiaoqi Ma, Dongdai Zhou, Peijun Duan, Shaochun Zhong 381

Model of Game Agent for Auction-Based Negotiation
Jun Hu, Chun Guan 387

An Autonomous Mobile Robot Based on Quantum Algorithm
Daoyi Dong, Chunlin Chen, Chenbin Zhang, Zonghai Chen 393

A MPC and Genetic Algorithm Based Approach for Multiple UAVs Cooperative Search
Jing Tian, Yanxing Zheng, Huayong Zhu, Lincheng Shen 399

Self-organization Evolution of Supply Networks: System Modeling and Simulation Based on Multi-agent
Gang Li, Linyan Sun, Ping Ji, Haiquan Li 405

Modeling and Analysis of Multi-agent Systems Based on π-Calculus
Fenglei Liu, Zhenhua Yu, Yuanli Cai 410

A Cooperation Mechanism in Agent-Based Autonomic Storage Systems
Jingli Zhou, Gang Liu, Shengsheng Yu, Yong Su 416

A Mobile Agent Based Spam Filter System
Xiaochun Cheng, Xiaoqi Ma, Long Wang, Shaochun Zhong 422

Hexagon-Based Q-Learning to Find a Hidden Target Object
Han-Ul Yoon, Kwee-Bo Sim 428

Intelligent Information Retrieval

A Naive Statistics Method for Electronic Program Guide Recommendation System
Jin An Xu, Kenji Araki 434

A Hybrid Text Classification System Using Sentential Frequent Itemsets
Shizhu Liu, Heping Hu 442

An Approach of Information Extraction from Web Documents for Automatic Ontology Generation
Ki-Won Yeom, Ji-Hyung Park 450

Improving Text Categorization Using the Importance of Words in Different Categories
Zhihong Deng, Ming Zhang 458

Image Copy Detection with Rotating Tolerance
Mingni Wu, Chiachen Lin, Chinchen Chang 464

Interactive and Adaptive Search Context for the User with the Exploration of Personalized View Reformulation
Supratip Ghose, Geun-Sik Jo 470

Integrating Collaborate and Content-Based Filtering for Personalized Information Recommendation
Zhiyun Xin, Jizhong Zhao, Ming Gu, Jiaguang Sun 476

Interest Region-Based Image Retrieval System Based on Graph-Cut Segmentation and Feature Vectors
Dongfeng Han, Wenhui Li, Xiaomo Wang, Yanjie She 483

A Method for Automating the Extraction of Specialized Information from the Web
Ling Lin, Antonio Liotta, Andrew Hippisley 489

An Incremental Updating Method for Clustering-Based High-Dimensional Data Indexing
Ben Wang, John Q. Gan 495

Support Vector Machine

Typhoon Track Prediction by a Support Vector Machine Using Data Reduction Methods
Hee-Jun Song, Sung-Hoe Huh, Joo-Hong Kim, Chang-Hoi Ho, Seon-Ki Park 503

Forecasting Tourism Demand Using a Multifactor Support Vector Machine Model
Ping-Feng Pai, Wei-Chiang Hong, Chih-Sheng Lin 512

A Study of Modelling Non-stationary Time Series Using Support Vector Machines with Fuzzy Segmentation Information
Shaomin Zhang, Lijia Zhi, Shukuan Lin 520

Support Vector Machine Based Trajectory Metamodel for Conceptual Design of Multi-stage Space Launch Vehicle
Saqlain Akhtar, He Linshu 528

Transductive Support Vector Machines Using Simulated Annealing
Fan Sun, Maosong Sun 536

Input Selection for Support Vector Machines Using Genetic Algorithms
Hee-Jun Song, Seon-Gu Lee, Sung-Hoe Huh 544

Associating *k*NN and SVM for Higher Classification Accuracy
Che-Chang Hsu, Chan-Yun Yang, Jr-Syu Yang 550

Multi-class SVMs Based on SOM Decoding Algorithm and Its Application in Pattern Recognition
Xiaoyan Tao, Hongbing Ji 556

Selective Dissemination of XML Documents Using GAs and SVM
K.G. Srinivasa, S. Sharath, K.R. Venugopal, Lalit M. Patnaik 562

A Smoothing Support Vector Machine Based on Exact Penalty Function
Zhiqing Meng, Gengui Zhou, Yihua Zhu, Lifang Peng 568

Speech Acts Tagging System for Korean Using Support Vector Machines
Songwook Lee, Jongmin Eun, Jungyun Seo 574

A New Support Vector Machine for Multi-class Classification
Zhiquan Qi, Yingjie Tian, Naiyang Deng 580

Support Vector Classification with Nominal Attributes
Yingjie Tian, Naiyang Deng 586

A New Smooth Support Vector Machine
Yubo Yuan, Chunzhong Li 592

The Application of Support Vector Machine in the Potentiality Evaluation for Revegetation of Abandoned Lands from Coal Mining Activities
Chuanli Zhuang, Zetian Fu, Ping Yang, Xiaoshuan Zhang 598

Prediction of T-cell Epitopes Using Support Vector Machine and Similarity Kernel
Feng Shi, Jing Huang 604

Radial Basis Function Support Vector Machine Based Soft-Magnetic Ring Core Inspection
Liangjiang Liu, Yaonan Wang 609

Direct Adaptive NN Control of a Class of Feedforward Systems
Wang Dongliang, Liu Bin, Zhang Zengke 616

Swarm Intelligence

Performance of an Ant Colony Optimization (ACO) Algorithm on the Dynamic Load-Balanced Clustering Problem in Ad Hoc Networks
Chin Kuan Ho, Hong Tat Ewe 622

Hybrid Particle Swarm Optimization for Flow Shop Scheduling with Stochastic Processing Time
Bo Liu, Ling Wang, Yi-hui Jin 630

Particle Swarm Optimizer with C-*Pg* Mutation
Guojiang Fu, Shaomei Wang, Mingjun Chen, Ning Li 638

Algal Bloom Prediction with Particle Swarm Optimization Algorithm
K.W. Chau 645

Synthesis of the Antenna Array Using a Modified Particle Swarm Optimization Algorithm
Tengbo Chen, Yong-Chang Jiao, Fushun Zhang 651

An Ant Colony Optimization Approach to the Degree-Constrained Minimum Spanning Tree Problem
Y.T. Bau, C.K. Ho, H.T. Ewe 657

Crowd Avoidance Strategy in Particle Swarm Algorithm
Guimin Chen, Qi Han, Jianyuan Jia, Wenchao Song 663

Particle Swarm Optimization with Multiscale Searching Method
Xiaohui Yuan, Jing Peng, Yasumasa Nishiura 669

Outcome-Space Branch and Bound Algorithm for Solving Linear Multiplicative Programming
Yuelin Gao, Chengxian Xu, Yueting Yang 675

A Binary Ant Colony Optimization for the Unconstrained Function Optimization Problem
Min Kong, Peng Tian 682

Data Mining

Mining Dynamic Association Rules in Databases
Jinfeng Liu, Gang Rong 688

A Novel Typical-Sample-Weighted Clustering Algorithm for Large Data Sets
Jie Li, Xinbo Gao, Licheng Jiao 696

Mining Weighted Generalized Fuzzy Association Rules with Fuzzy Taxonomies
Shen Bin, Yao Min, Yuan Bo 704

Concept Chain Based Text Clustering
Shaoxu Song, Jian Zhang, Chunping Li 713

An Efficient Range Query Under the Time Warping Distance
Chuyu Li, Long Jin, Sungbo Seo, Keun Ho Ryu 721

Robust Scene Boundary Detection Based on Audiovisual Information
Soon-tak Lee, Joon-sik Baek, Joong-hwan Baek 729

An FP-Tree Based Approach for Mining All Strongly Correlated Item Pairs
Zengyou He, Shengchun Deng, Xiaofei Xu 735

An Improved kNN Algorithm – Fuzzy kNN
Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Zhihai Wang, Youli Qu 741

A New Integrated Personalized Recommendation Algorithm
Hongfang Zhou, Boqin Feng, Lintao Lv, Zhurong Wang 747

An Improved EMASK Algorithm for Privacy-Preserving Frequent Pattern Mining
Congfu Xu, Jinlong Wang, Hongwei Dan, Yunhe Pan 752

CR*-Tree: An Improved R-Tree Using Cost Model
Haibo Chen, Zhanquan Wang 758

Grid-ODF: Detecting Outliers Effectively and Efficiently in Large Multi-dimensional Databases
Wei Wang, Ji Zhang, Hai Wang 765

Clustering XML Documents by Structure Based on Common Neighbor
Xizhe Zhang, Tianyang Lv, Zhengxuan Wang, Wanli Zuo 771

A Generalized Global Convergence Theory of Projection-Type Neural Networks for Optimization
Rui Zhang, Zongben Xu 777

Hierarchical Recognition of English Calling Card by Using Multiresolution Images and Enhanced Neural Network
Kwang-Baek Kim, Sungshin Kim 785

An Novel Artificial Immune Systems Multi-objective Optimization Algorithm for 0/1 Knapsack Problems
Wenping Ma, Licheng Jiao, Maoguo Gong, Fang Liu 793

RD-Based Seeded Region Growing for Extraction of Breast Tumor in an Ultrasound Volume
Jong In Kwak, Sang Hyun Kim, Nam Chul Kim 799

Improving Classification for Microarray Data Sets by Constructing Synthetic Data
Shun Bian, Wenjia Wang 809

A Method to Locate the Position of Mobile Robot Using Extended Kalman Filter
Ping Wei, Chengxian Xu, Fengji Zhao 815

Simulated Annealing with Injecting Star-Alignment for Multiple Sequence Alignments
Hongwei Huo, Hua Ming 821

A Noise-Insensitive Hierarchical Min-Max Octree for Visualization of Ultrasound Datasets
Sukhyun Lim, Kang-hee Seo, Byeong-Seok Shin 827

A Novel Fusing Algorithm for Retinal Fundus Images
Bin Fang, Xinge You, Yuan Yan Tang 833

Improving PSO-Based Multiobjective Optimization Using Competition and Immunity Clonal
Xiaohua Zhang, Hongyun Meng, Licheng Jiao 839

Clonal Selection Algorithm for Dynamic Multiobjective Optimization
Ronghua Shang, Licheng Jiao, Maoguo Gong, Bin Lu 846

Key Frame Extraction Based on Evolutionary Artificial Immune Network
Fang Liu, Xiaoying Pan 852

Clonal Selection Algorithm with Immunologic Regulation for Function Optimization
Hang Yu, Maoguo Gong, Licheng Jiao, Bin Zhang 858

A Fault-Tolerant and Minimum-Energy Path-Preserving Topology Control Algorithm for Wireless Multi-hop Networks
Zhong Shen, Yilin Chang, Can Cui, Xin Zhang 864

Computational Biomechanics and Experimental Verification of Vascular Stent
Yuexuan Wang, Hong Yi, Zhonghua Ni 870

Numerical Computing of Brain Electric Field in Electroencephalogram
Dexin Zhao, Zhiyong Feng, Wenjie Li, Shugang Tang 878

A Novel Multi-stage 3D Medical Image Segmentation: Methodology and Validation
Jianfeng Xu, Lixu Gu, Xiahai Zhuang, Terry Peters 884

Medical Image Alignment by Normal Vector Information
Xiahai Zhuang, Lixu Gu, Jianfeng Xu 890

Global Exponential Stability of Non-autonomous Delayed Neural Networks
Qiang Zhang, Dongsheng Zhou, Xiaopeng Wei 896

A Prediction Method for Time Series Based on Wavelet Neural Networks
Xiaobing Gan, Ying Liu, Francis R. Austin 902

Training Multi-layer Perceptrons Using MiniMin Approach
Liefeng Bo, Ling Wang, Licheng Jiao 909

Two Adaptive Matching Learning Algorithms for Independent Component Analysis
Jinwen Ma, Fei Ge, Dengpan Gao 915

Bioprocess Modeling Using Genetic Programming Based on a Double Penalty Strategy
Yanling Wu, Jiangang Lu, Youxian Sun, Peifei Yu 921

An Improved Gibbs Sampling Algorithm for Finding TFBS
Caisheng He, Xianhua Dai 927

Pattern Recognition

A Novel Fisher Criterion Based S_t-Subspace Linear Discriminant Method for Face Recognition
Wensheng Chen, Pong C. Yuen, Jian Huang, Jianhuang Lai 933

EmoEars: An Emotion Recognition System for Mandarin Speech
Bo Xie, Ling Chen, Gen-Cai Chen, Chun Chen 941

User Identification Using User's Walking Pattern over the ubiFloorII
Jaeseok Yun, Woontack Woo, Jeha Ryu 949

Evolving RBF Neural Networks for Pattern Classification
Zheng Qin, Junying Chen, Yu Liu, Jiang Lu 957

Discrimination of Patchoulis of Different Geographical Origins with Two-Dimensional IR Correlation Spectroscopy and Wavelet Transform
Daqi Zhan, Suqin Sun, Yiu-ming Cheung 965

Gait Recognition Using View Distance Vectors
Murat Ekinci, Eyup Gedikli 973

HMM Parameter Adaptation Using the Truncated First-Order VTS and EM Algorithm for Robust Speech Recognition
Haifeng Shen, Qunxia Li, Jun Guo, Gang Liu 979

Model Type Recognition Using De-interlacing and Block Code Generation
Cheol-Ki Kim, Sang-Gul Lee, Kwang-Baek Kim 985

R-functions Based Classification for Abnormal Software Process Detection
Anton Bougaev, Aleksey Urmanov 991

A Unified Framework for Shot Boundary Detection
Bing Han, Xinbo Gao, Hongbing Ji 997

Image Recognition with LPP Mixtures
SiBao Chen, Min Kong, Bin Luo 1003

Line-Based Camera Calibration
Xiuqin Chu, Fangming Hu, Yushan Li 1009

Shot Boundary Detection Based on SVM and TMRA
Wei Fang, Sen Liu, Huamin Feng, Yong Fang 1015

Robust Pattern Recognition Scheme for Devanagari Script
Amit Dhurandhar, Kartik Shankarnarayanan, Rakesh Jawale 1021

Credit Evaluation Model and Applications Based on Probabilistic Neural Network
Sulin Pang 1027

Fingerprint Ridge Line Extraction Based on Tracing and Directional Feedback
Rui Ma, Yaxuan Qi, Changshui Zhang, Jiaxin Wang 1033

A New Method for Human Gait Recognition Using Temporal Analysis
Han Su, Fenggang Huang 1039

Microcalcification Patterns Recognition Based Combination of Autoassociator and Classifier
Wencang Zhao, Xinbo Yu, Fengxiang Li 1045

Improved Method for Gradient-Threshold Edge Detector Based on HVS
Fuzheng Yang, Shuai Wan, Yilin Chang 1051

MUSC: Multigrid Shape Codes and Their Applications to Image Retrieval
Arindam Biswas, Partha Bhowmick, Bhargab B. Bhattacharya 1057

Applications

Adaptation of Intelligent Characters to Changes of Game Environments
Byeong Heon Cho, Sung Hoon Jung, Kwang-Hyun Shim, Yeong Rak Seong, Ha Ryoung Oh 1064

An Knowledge Model for Self-regenerative Service Activations Adaptation Across Standards
Mengjie Yu, David Llewellyn Jones, A. Taleb-Bendiab 1074

An Agent for the HCARD Model in the Distributed Environment
Bobby D. Gerardo, Jae-Wan Lee, Jae-jeong Hwang, Jung-Eun Kim 1082

A New Class of Filled Functions for Global Minimization
Xiaoliang He, Chengxian Xu, Chuanchao Zhu 1088

Modified PSO Algorithm for Deadlock Control in FMS
Hesuan Hu, Zhiwu Li, Weidong Wang 1094

Optimization Design of Controller Periods Using Evolution Strategy
Hong Jin, Hui Wang, Hongan Wang, Guozhong Dai 1100

Application of Multi-objective Evolutionary Algorithm in Coordinated Design of PSS and SVC Controllers
Zhenyu Zou, Quanyuan Jiang, Pengxiang Zhang, Yijia Cao 1106

Author Index 1113

A Fast Inversion Algorithm and Low-Complexity Architecture over $GF(2^m)^\star$

Sosun Kim[1], Nam Su Chang[2], Chang Han Kim[3], Young-Ho Park[4], and Jongin Lim[2]

[1] Platform Security R&D, Softforum Co., Seoul, Korea
ssosuny@softforum.com
[2] Center for Information and Security Technologies(CIST), Korea Univ., Seoul, Korea
{ns-chang, jilim}@korea.ac.kr
[3] Dept. of Information and Security, Semyung Univ., Jecheon, Korea
chkim@semyung.ac.kr
[4] Dept. of Information Security, Sejong Cyber Univ., Seoul, Korea
youngho@cybersejong.ac.kr

Abstract. The performance of public-key cryptosystems is mainly appointed by the underlying finite field arithmetic. Among the basic arithmetic operations over finite field, the multiplicative inversion is the most time consuming operation. In this paper, a fast inversion algorithm over $GF(2^m)$ with the polynomial basis representation is proposed. The proposed algorithm executes in about 27.5% or 45.6% less iterations than the extended binary gcd algorithm (EBGA) or the montgomery inverse algorithm (MIA) over $GF(2^{163})$, respectively. In addition, we propose a new hardware architecture to apply for low-complexity systems. The proposed architecture takes approximately 48.3% or 24.9% less the number of reduction operations than [4] or [8] over $GF(2^{239})$, respectively. Furthermore, it executes in about 21.8% less the number of addition operations than [8] over $GF(2^{163})$.

1 Introduction

Finite field arithmetic has many applications in coding theory and cryptography. The performance of the public-key cryptosystems is especially appointed by the underlying the efficiency of it. Among the basic arithmetic operations over $GF(2^m)$, the multiplicative inversion is the highest computational complexity.

Several VLSI algorithms for multiplicative inversion over $GF(2^m)$ have been proposed [2], [3], [8], [9] and [10]. They use a polynomial basis for the representation of field elements, and carry out inversion $2m$ steps through one directional shifts and additions in $GF(2^m)$. In [9] and [10], Wu et al. proposed the systolic

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1–8, 2005.
© Springer-Verlag Berlin Heidelberg 2005

array structure to compute the inverses over $GF(2^m)$ based on EBGA. In the practical applications where the dimension of the field may vary, it becomes area-complexity or time-complexity even impractical. It is not suitable for low-weight and low-power systems, i.e., smartcard, the mobile phone SIM card.

In this paper, we propose a fast inversion algorithm based on EBGA over $GF(2^m)$. It also uses the polynomial basis for the representation of the field element. The proposed algorithm performs the multiplicative inversion $2m$ steps less than the previous algorithms. The proposed algorithm executes in about 27.5% or 45.6% less steps than EBGA or MIA over $GF(2^{163})$, respectively. This paper takes it as focal point hardware architecture to be suitable for low-complexity systems. It executes in about 48.3% or 24.9% less the number of reduction operations than [4] or [8] over $GF(2^{239})$, respectively. Furthermore, it takes approximately 21.9% or 21.8% less the number of addition operations than [4] or [8] over $GF(2^{163})$, respectively.

This paper is organized as follows: In Section 2, we introduce EBGA for inversion over $GF(2^m)$. In Section 3, we propose a fast inversion algorithm over $GF(2^m)$ based on EBGA in Section 2. In Section 4, we propose a new hardware architecture of the proposed algorithm. We present simulation analysis of the proposed algorithm over flexible size fields and conclude in Section 5.

2 Previous Works

Finite field $GF(2^m)$ is an extension field of $GF(2)$. Let

$$G(x) = x^m + \sum_{i=0}^{m-1} g_i x^i,\ g_i \in GF(2)$$

be an irreducible polynomial of degree m over $GF(2)$. Any element $A(x)$ in $GF(2^m)$ can be represented as

$$A(x) = a_{m-1}x^{m-1} + a_{m-2}x^{m-2} + \ldots + a_1 x + a_0,$$

where $a_i \in GF(2)$. We call $A(x)$ the polynomial representation of A, and $A = (a_{m-1}, a_{m-2}, \ldots, a_1, a_0)$ is the polynomial basis vector representation of $A(x)$.

Addition and subtraction of two elements in $GF(2^m)$ are operated by bitwise exclusive-OR operation. Let $A(x)$ and $B(x)$ be polynomial representation of elements in $GF(2^m)$. Multiplication over $GF(2^m)$ is defined as follows:

$$A(x) \cdot B(x) = A(x) \times B(x)\ mod\ G(x),$$

where $\times$ denotes multiplication operator in the polynomial ring [8]. Division over $GF(2^m)$ is defined as follows:

$$A(x) \div B(x) = A(x) \times B(x)^{-1}\ mod\ G(x).$$

If $A(x) \times A(x)^{-1} = 1\ mod\ G(x)$, then $A(x)^{-1}$ is the multiplicative inversion of $A(x)$ in $GF(2^m)$.

In general, the inversion over the polynomial basis is performed by variants of the extended Euclidean algorithm (EEA). Euclidean-based schemes contain EEA, EBGA and MIA. Among these schemes, EBGA is suitable for binary arithmetic operations, and employs simple division-free operations.

Let $U(x)$ and $V(x)$ be two polynomials in $GF(2^m)$. EBGA holds the following two equations:

$$R(x) \times A(x) = U(x) \mod G(x), \tag{1}$$

$$S(x) \times A(x) = V(x) \mod G(x). \tag{2}$$

The variables are initialized as follows: $U(x) = A(x)$, $V(x) = G(x)$, $R(x) = 1$, $S(x) = 0$. It can be seen that Eq.(1) and Eq.(2) always hold each iteration of EBGA in [8]. At the end of EBGA, finally $U(x) = 0$ and $S(x) \times A(x) = 1 \mod G(x)$ holds in Eq.(2). The polynomial $S(x)$ is $A(x)^{-1} \mod G(x)$ in $GF(2^m)$. For such an initial condition, the degree of polynomial $U(x)$ and $V(x)$ may not exceed m each iteration. Thus, EBGA needs a total of $2m$ iterations by one directional shifts in a worst case of [8].

EBGA is constructed by simple division-free operations such as fixed shift operation, addition and comparison. For that reason, EBGA is more efficient than EEA to design an inversion hardware architecture reducing complexity. However, EBGA requires the comparision $deg(U(x))$ and $deg(V(x))$ to determine the next operations.

3 Fast Algorithm for Inversion over $GF(2^m)$

EBGA does not involve in a fixed number of steps to compute inversion over $GF(2^m)$. It is difficult to implement a hardware architecture. In [9] and [10], Wu et al. modified EBGA that it has a fixed number of $2m$ iterations to compute the multiplicative inversion. In [8], Watanabe et al. proposed the comparison operation replaced by observing the rightmost bit from shift registers and variables.

We propose a fast inversion algorithm, that is, Algorithm I over $GF(2^m)$ which is suitable for the low-complexity systems. Algorithm I is based on EBGA, and uses the polynomial basis for the representation of the field element.

We can easily show that Algorithm I holds in Eq.(1) and Eq.(2). At the beginning of Algorithm I, the degree of the irreducible polynomial $G(x)$ is always m and the degree of the polynomial $A(x)$ is at most $(m-1)$. Algorithm I for calculating the inversion over $GF(2^m)$ generated by $G(x)$ is as follows. The polynomials $A(x)$ and $G(x)$ are given by an m-bit and an $(m+1)$-bit binary vector A and G, respectively. In Algorithm I, U, R, and S are m-bit binary vectors and V is an $(m+1)$-bit binary vector.

The degree of the polynomial $U(x)$ is reduced by 1 for each iteration of EBGA in [8]. As a result, it executes $2m$ steps for computing the inversion over $GF(2^m)$ in the worst case. In Algorithm I, the degree of $U(x)$ is reduced by at most 2 for each iteration. When u_0 is 0, Algorithm I is similar to that of EBGA in [8], and the rightmost bit of U is removed. If (u_1, u_0) is $(0, 0)$, it is removed as step 4 in Algorithm I, and the vector U is reduced by 2-bit. Otherwise, Algorithm I

Algorithm I. Fast Inversion Algorithm over $GF(2^m)$

INPUT: $A = A(x)$ and $kG = kG(x)$, $k = 0, 1, x, (1+x)$.
OUTPUT: A^{-1} in $GF(2^m)$.

1. $U = A,\ V = G,\ R = 1,\ S = 0,\ D = (0, ..., 0, 1, 0),\ f = 1,\ P = (1, 0, ..., 0)$.
2. While $p_0 = 0$ do the following:
3. $(kG,\ q)$ = Algorithm II$((u_1, u_0), v_1, (r_1, r_0), (s_1, s_0), g_1)$.
4. If $[(u_1, u_0) = (0, 0)]$ then $U = U/x^2,\ R = (R + kG)/x^2$.
5. If $f = 0$ then
6. If $[d_1 = 1]$ then $f = 1$ Else $D = D/x^2$.
7. If $[d_0 = 1]$ then $f = 1$.
8. Else /*$f = 1$*/
9. If $[d_0 = 1]$ then $P = P/x$ Else $P = P/x^2$.
10. $D = x \cdot D$.
11. Else If $[u_0 = 0]$ then $U = U/x,\ R = (R + kG)/x$.
12. If $f = 0$ then $D = D/x$.
13. If $[d_0 = 1]$ then $f = 1$.
14. Else /*$f = 1$*/
15. If $[d_0 = 0]$ then $P = P/x$.
16. $D = x \cdot D$.
17. Else /*$u_0 = 1$*/
18. If $[d_0 = 1$ or $f = 0]$ then
19. If $[(q = 0)$ and $(r_1, r_0) = (s_1, s_0)]$ then
20. $U = (U + V)/x^2,\ R = (R + S)/x^2$.
21. Else
22. $R = R + kG,\ V' = q \cdot xV,\ S' = S + q \cdot xS$.
23. $U = (U + V')/x^2,\ R = (R + S')/x^2$.
24. If $[d_0 = 1]$ then $D = x \cdot D$ Else /*$f = 0$*/ $D = D/x,\ P = P/x$.
25. If $[d_0 = 1]$ then $f = 1$.
26. Else /*$d_0 = 0$ and $f = 1$*/
27. If $[(q = 0)$ and $(r_1, r_0) = (s_1, s_0)]$ then
28. $U = (U + V)/x^2,\ V = U,\ R = (R + S)/x^2,\ S = R$.
29. Else
30. $R = R + kG,\ V' = q \cdot xV,\ S' = S + q \cdot xS$.
31. $U = (U + V')/x^2,\ V = U,\ R = (R + S')/x^2,\ S = R$.
32. $f = 0$.
33. If $[d_1 = 1]$ then $f = 1$ Else $D = D/x^2,\ P = P/x$.
34. If $[d_0 = 1]$ then $f = 1$.
35. Return (S).

Algorithm II. Determining the vector kG in Algorithm I

INPUT: The bits $BU = (u_1, u_0), BV = v_1, BR = (r_1, r_0), BS = (s_1, s_0)$ and g_1.
OUTPUT: The binary vector kG and a flag q.

1. If $[u_0 = 0]$ then $(r_1', r_0') = (r_1, r_0)$.
2. Else
3. If $[u_1 = v_1]$ then $q = 0,\ (r_1', r_0') = (r_1' + s_1, r_0' + s_0)$.
4. Else $q = 1,\ (r_1', r_0') = (r_1' + s_1 + s_0, r_0' + s_0)$.
5. If $[(r_1', r_0') = (0, 0)]$ then $kG = 0$.
6. Else if $[(r_1', r_0') = (1, 0)]$ then $kG = xG$.
7. Else if $[r_1' = g_1]$ then $kG = G$.
8. Else $kG = (1 + x)G$.
9. Return$((kG,\ q))$.

Table 1. Addition operations in Algorithm I

(u_1, u_0)	(v_1, v_0)	Addition $k = 0, 1, x, (1+x)$	
$(0,1)$	$(0,1)$	$U = U + V$	$R = R + kG + S$
$(0,1)$	$(1,1)$	$U = U + V + xV$	$R = R + kG + S + xS$
$(1,1)$	$(0,1)$	$U = U + V + xV$	$R = R + kG + S + xS$
$(1,1)$	$(1,1)$	$U = U + V$	$R = R + kG + S$

executes addition operations to make $(u_1, u_0) = (0, 0)$ at all times, where (u_1, u_0) is the result of addition. Table 1 shows addition operations that the result (u_1, u_0) of addition is made up $(0, 0)$.

The polynomial $R(x)$ contains the reduction operation which adds the polynomials $R(x)$ and $G(x)$ in EBGA. When $u_0 = 1$, Algorithm I performs operation as described in Table 1 whether $(R+kG+S)/x^2$ or $(R+kG+S+xS)/x^2$ according to u_1 and v_1. The binary vector R requires adding kG which the rightmost bits (r_1, r_0) of the result addition holds $(0, 0)$ at each step. The vector R' denotes the result of R, $R+S$ or $R+S+xS$ in Table 1. We consider the rightmost bits (r'_1, r'_0) of the binary vector R'. The rightmost bits (r'_1, r'_0) hold (r_1, r_0), (r_1+s_1, r_0+s_0) or $(r_1+s_1+s_0, r_0+s_0)$ in Algorithm I. We can know the vector kG which makes $(r_1, r_0) = (0, 0)$ at each step. Algorithm II presents operation which determines the binary vector kG and a flag q where $k = 0, 1, x, (1 + x)$ before the main operative processes execute in the loop. The binary vector kG is selected by the rightmost bits (r'_1, r'_0) of the result of the bitwise operations in Algorithm II. Therefore, Algorithm II precomputes the vector kG and a flag q at each iteration of Algorithm I. The operations between brackets are performed in parallel.

EBGA requires the way which checks whether $U(x) \neq 0$ and $deg(U(x)) < deg(V(x))$. We describe the comparison method introduced in [8]. In order to avoid time-consuming the comparison of $deg(U(x))$ and $deg(V(x))$, we consider upper bounds of them, α and β. Instead of keeping α and β, we hold the difference $d = \alpha - \beta$ and $p = \min(\alpha, \beta)$. We introduce a $(m + 1)$-bit 1-hot counter D for denoting the absolute value of d and a flag f for indicating the sign of d so that $f = 0$ if d is positive and $f = 1$ otherwise. We also introduce a $(m+1)$-bit 1-hot counter P for holding p. Note that in 1-hot counter, only one bit is 1 and the others are 0. We can know whether $deg(U(x)) < deg(V(x))$ by checking d_0 and f where d_0 is the rightmost bit of D, i.e., if $deg(U(x)) < deg(V(x))$, then $d_0 = 0$ and $f = 1$ hold. When $U(x) = 0$, then $V(x) = 1$ holds in Eq.(2). We can check whether $V(x) \neq 1$ instead of $U(x) \neq 0$. Therefore, we can know $V(x) \neq 1$ by checking whether $p_0 = 0$ where p_0 is the rightmost bit of P.

4 New Hardware Architecture for Inversion over $GF(2^m)$

In [9] and [10], Wu et al. proposed the systolic array structure to compute the multiplicative inverses in $GF(2^m)$. However, if the dimension of the field m is large, it requires more areas and resources. Accordingly, it is not suitable for systems to provide little power, space and time. The previous architectures [4], [8]

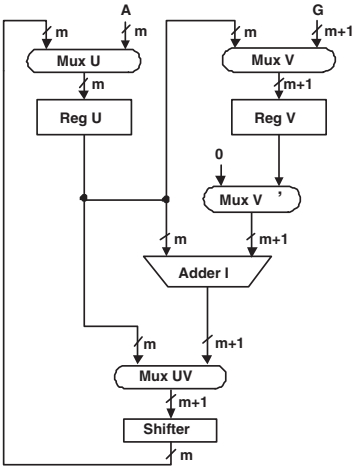


Fig. 1. UV-Block of the proposed architecture

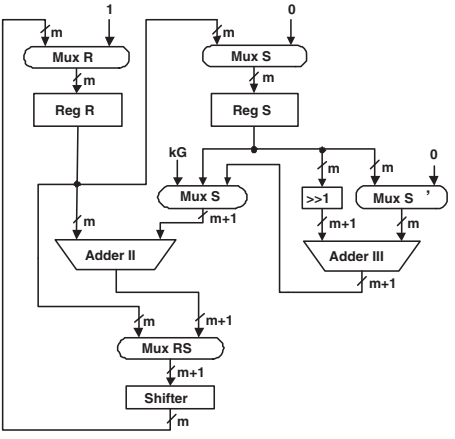


Fig. 2. RS-Block of the proposed architecture

carry out inversion $2m$ steps through iteration of shifts. In addition, they contain the delay of reduction operation, and the critical path of them is increased.

We propose a new hardware architecture focusing under low-complexity systems. The area complexity of the proposed architecture is similar to [8]. The proposed architecture takes approximately 27.5% less step than [8] over $GF(2^{163})$. It performs with the reduction operation and the main operative processes in parallel. Therefore, it is not generated by the delay of executing the reduction operation.

The proposed hardware architecture presents the basic parts: the control logic block, the UV-Block, the RS-Block and 1-hot counters, i.e., D and P. In the main loop of Algorithm I, all the flags depend on parity bits of variables and on 1-hot counters. The control logic block performs that all the flags can be efficiently precomputed. It determines the vector kG through the bitwise operations regarding the rightmost bits of the binary vectors. It consists of an AND gate and 4 XOR gates in the control logic block. The delay of added gates in the control logic block is negligible with respect to the executing time. In addition, we can reduce the complexity of the control logic block than [6]. Since shift operation is faster and simpler than others operations, the proposed architecture is more efficient than the architecture using the counters in [6].

The UV-Block and RS-Block operative parts are shown in Fig.1 and Fig.2, respectively. The shifter performs 1-bit right shift or 2-bit right shift according as the rightmost bits (u_1, u_0) in the circuit. When the rightmost bit $u_0 = 1$, Algorithm I performs processes which present two parts of addition as the following expression:

$$U = (U + V + q \cdot xV)/x^2 = \{(U + V)/x + qV\}/x, \tag{3}$$

$$R = (R + kG + S + q \cdot xS)/x^2 = \{(R + kG)/x + (S/x + qS)\}/x. \tag{4}$$

UV-Block performs the addition of the vectors U and V in Eq.(3). Additionally, RS-Block performs both the reduction operation, i.e., $R + kG$, and the addition

operation, i.e., $S/x + qS$ in parallel. They execute 1-bit right shift by shifter in Fig.1 and Fig.2. Secondly, UV-Block performs the remaining operations in Eq.(3). RS-Block executes the addition of the precomputed vectors $R + kG$ and $S/x + qS$, and performs 1-bit right shift.

5 Comparison and Conclusion

Simulation and quantitative analysis of the number of additions, reductions, and loops were performed for all existing algorithms. It was performed for the recommended elliptic curves and the irreducible polynomials in [1]. A total of $1,000,000$ random element $A(x) \in GF(2^m)$ was computed by each method. Simulation results are presented in Table 2. The tests include all "less than" comparisons except '$U(x) \neq 0$' and '$V(x) \neq 1$' in the main loop, which does not require a operation.

Table 2. The average computational cost in inversions

Field	Algo.	Add.	Red.	Loop	Loop*	Field	Algo.	Add.	Red.	Loop	Loop*
$GF(2^{163})$	Algo. I	**210.9**	**101.4**	**293.2**	**1.80**	$GF(2^{283})$	Algo. I	**369.0**	**176.8**	**513.6**	**1.81**
	MIA** [4]	269.9	135.1	539.1	3.31		MIA[4]	469.3	234.8	938.8	3.32
	EBGA[8]	269.7	135.0	404.6	2.48		EBGA[8]	468.7	234.2	703.2	2.48
	[9]	322.0	161.2	320.5	1.97		[9]	561.8	280.6	565.4	2.00
$GF(2^{193})$	Algo. I	**248.9**	**117.0**	**348.8**	**1.81**	$GF(2^{409})$	Algo. I	**523.8**	**231.4**	**746.8**	**1.83**
	MIA[4]	319.9	160.1	639.1	3.31		MIA[4]	679.9	339.8	1360.2	3.33
	EBGA[8]	314.9	155.0	474.6	2.46		EBGA[8]	638.5	298.9	977.0	2.39
	[9]	376.8	185.6	383.0	1.98		[9]	772.3	365.6	818.1	2.00
$GF(2^{239})$	Algo. I	**289.7**	**102.4**	**433.0**	**1.81**	$GF(2^{571})$	Algo. I	**750.7**	**358.8**	**1039.5**	**1.82**
	MIA[4]	395.8	198.1	790.3	3.31		MIA[4]	950.2	474.8	1903.4	3.33
	EBGA[8]	319.6	121.8	517.3	2.16		EBGA[8]	950.4	475.0	1420.8	2.49
	[9]	397.2	160.1	477.0	2.00		[9]	1139.7	569.6	1146.8	2.01

Add.: The average value of Additions. Red.: The average value of Reductions.
Loop: The average value of Loops. Loop*: Loop/Bit.
**MIA [4] contains the number of operations of Phase I and Phase II.

Table 3. Comparison with previously proposed inverters

	Proposed Inverter	**[4]**	**[8]**
Field	$GF(2^m)$	$GF(p)$ and $GF(2^m)$	$GF(2^m)$
Maximum Cell Delay	$T_{RShift} + T_{Add}$ $+3T_{Mux}$	$T_{RShift} + T_{Add/Sub}$ $+4T_{Mux} + T_{Demux}$	$T_{RShift} + 2T_{Add}$ $+T_{Mux}$
Basic Components and Their Numbers	Right Shifter: 3	Right Shifter: 3	Right Shifter: 2
	$(m+1)$-bit Register: 1 m-bit Register: 3	m-bit Register: 4	$(m+1)$-bit Register: 1 m-bit Register: 3
	2-input Adder: 3	2-input Subtractor: 2 2-input Adder: 1	2-input Adder: 3
	Multiplexer: 9	Multiplexer: 13 De-multiplexer: 1	Multiplexer: 6
Controller and Their Numbers	1-hot counter: 2 flag Register: 1	$(log_2 m)$-bit counter: 1	1-hot counter: 2 flag Register: 1

T_{Mux} : The propagation delay through one $Mutiplexer$ gate.
T_{Add} : The propagation delay through one $Adder$ gate.
$T_{Add/Sub}$: The propagation delay through one $Adder/Subtractor$ gate.
T_{RShift} : The propagation delay through $RightShifter$ gate.

In Algorithm I, the degree of the polynomial $U(x)$ is reduced by at most 2 for each iteration. We can see that Algorithm I executes in about 45.6% or 27.5% less steps than MIA [4] or EBGA [8] over $GF(2^{163})$, respectively. As described in the column "Add.", although the proposed architecture has an extra adder, the number of addition operations is reduced by about 26.8% than [4] over $GF(2^{239})$, and is reduced by about 21.8% than [8] over $GF(2^{163})$. Additionally, it takes approximately 34.5% less the number of addition operations than [9] over $GF(2^{163})$. We can get that Algorithm I takes approximately 24.9% less the number of reduction operations than [8] over $GF(2^{163})$. It executes in about 48.3% less the number of reduction operations than [4] over $GF(2^{239})$. Furthermore, Algorithm I requires a minimum of 1.80 iterations/bit and on average 1.81 iterations/bit to compute the inverse over $GF(2^m)$.

Table 3 shows the comparison of the hardware architectures for inversion over $GF(2^m)$. The area complexity of the proposed architecture is similar to [8]. In addition, the depth of the proposed inverter is a constant independent of m. Since the propagation delay through multiplexer is smaller than that through adder, the time complexity of the proposed inverter is smaller than that of [4] and [8]. This suggests the use of our design in low-complexity devices, i.e., smartcard, the mobile phone SIM card.

References

1. Certicom Research, *SEC 2: Recommended Elliptic Curve Domain Parameters*, version 1.0, September 2000.
2. J. H. Guo, C. L. Wang, *Hardware-efficient systolic architecture for inversion and division in* $GF(2^m)$, IEE Proc. Comput. Digital Tech. vol. 145, no. 4, 1998, pp.272-278.
3. J. H. Guo, C. L. Wang, *Systolic Array Implementation of Euclid's Algorithm for Inversion and Division in* $GF(2^m)$, IEEE Transactions on Computers, vol. 47, no. 10, October 1998, pp.1161-1167.
4. A. Gutub, A. F. Tenca, E. Savas, C. K. Koc, *Scalable and unified hardware to compute Montgomery inverse in* $GF(p)$ *and* $GF(2^m)$, CHES 2002, LNCS 2523, August 2002, pp.484-499.
5. D. Hankerson, J. L. Hernandez, A. Menezes, *Software Implementation of Elliptic Curve Cryptography Over Binary Fields*, Cryptographic Hardware and Embedded Systems, CHES'00, 2000, pp.1-24.
6. R. Lorenzo, *New Algorithm for Classical Modular Inverse*, Cryptographic Hardware and Embedded Systems, CHES'02, LNCS 2523, 2002, pp.57-70.
7. N. Takagi, *A VLSI Algorithm for Modular Division Based on the Binary GCD Algorithm*, IEICE Trans. Fundamentals, vol. E81-A, May 1998, pp. 724-728.
8. Y. Watanabe, N. Takagi, and K. Takagi, *A VLSI Algorithm for Division in* $GF(2^m)$ *Based on Extended Binary GCD Algorithm*, IEICE Trans. Fundamentals, vol. E85-A, May 2002, pp. 994-999.
9. C. H. Wu, C. M. Wu, M. D. Shieh, and Y. T. Hwang, *Systolic VLSI Realization of a Novel Iterative Division Algorithm over* $GF(2^m)$*: a High-Speed, Low-Complexity Design*, 2001 IEEE International Symposium on Circuits and Systems, May 2001, pp.33-36.
10. C. H. Wu, C. M. Wu, M. D. Shieh, and Y. T. Hwang, *An Area-Efficient Systolic Division Circuit over* $GF(2^m)$ *for Secure Communication*, 2002 IEEE International Symposium on Circuits and Systems, August 2002, pp.733-736.

An ID-Based Optimistic Fair Signature Exchange Protocol from Pairings*

Chunxiang Gu, Yuefei Zhu, and Yajuan Zhang

Network Engineering Department, Information Engineering University,
P.O. Box 1001-770, Zhengzhou 450002, P.R. China
gcxiang5209@yahoo.com.cn

Abstract. ID-based public key cryptosystem can be a good alternative for certificate-based public key setting. The protocol for fair exchange of signatures can be widely used in signing digital contracts, e-payment and other electronic commerce. This paper proposes an efficient ID-based verifiably encrypted signature scheme from pairings. Using this new scheme as kernel, we provide an efficient ID-based optimistic fair signature exchange protocol. We offer arguments for the fairness, efficiency and security proof of our new protocol. Our new protocol provides an efficient and secure solution for the problem of fair exchange of signatures in ID-based cryptosystem.

1 Introduction

1.1 ID-Based Public Key Cryptography

In 1984, Shamir[1] first proposed the idea of ID-based public key cryptography (ID-PKC) to simplify key management procedure of traditional certificate-based PKI. In ID-PKC, an entity's public key is directly derived from certain aspects of its identity. Private keys are generated for entities by a trusted third party called a private key generator (PKG). The direct derivation of public keys in ID-PKC eliminates the need for certificates and some of the problems associated with them. The first entire practical and secure ID-based public key encryption scheme was presented in[2]. Since then, a rapid development of ID-PKC has taken place. ID-based public key cryptography has become a good alternative for certificate-based public key setting, especially when efficient key management and moderate security are required.

1.2 Protocols for Fair Exchange of Signatures

As more and more electronic commerce, such as signing digital contracts, e-payment, etc, are being conducted on insecure networks, protocols for fair signature exchange attract much attention in the cryptographic community. Fairness

* Research supported by Found 973 (No. G1999035804), NSFC (No. 90204015, 60473021) and Elitist Youth Foundation of Henan in China (No. 021201400).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 9–16, 2005.
© Springer-Verlag Berlin Heidelberg 2005

means that the two parties to exchange signatures in such a fair way that either party gets the other's signature, or neither party does.

Until recently, there have been two main approaches for achieving fair exchange. The first approach is to ensure that the exchange occurs simultaneously. One way of providing simultaneous exchange is to have the participants exchange information bit by bit in an interleaving manner.

The second approach is to ensure that the exchange will be completed even though one of the entities participating in the exchange refuses to continue. Fair exchange protocols which employ this approach often use a trusted third party (TTP) to store the details of the transaction. These details are released if one of the entities refuse to complete the protocol. The use of the online TTP greatly reduces the efficiency of the protocol. With the assumption that the participators are honest in most situations, more preferable solutions, called ***optimistic fair exchange protocols*** based on off-line TTP, are proposed in [6, 7]. In these protocols, the off-line TTP does not participate in the actual exchange protocol in normal cases, and is invoked only in abnormal cases to dispute the arguments.

In the design of optimistic fair exchange protocols, ***verifiably encrypted signature schemes*** (VESSs) are usually being used as the kernel components. VESS is a special extension of general signature primitive, which enables user Alice to give user Bob a signature on a message M encrypted with an *adjudicator*'s public key, and enables Bob to verify that the encrypted signature indeed contains such a signature. The adjudicator is an off-line TTP, who can reveal the signature when needed.

In the recent years, researches on VESSs and protocols for fair exchange of signatures have got fruit achievements. Several new constructions of VESSs and fair signature exchange protocols [3, 6, 8, 9] have been proposed. However, all these works are in traditional certificate-based public key cryptosystem.

1.3 Contributions and Organization

This paper provides an efficient and secure solution for the fair signature exchange problem in ID-based public key setting. We propose an efficient ID-based VESS based on the ID-based signature scheme due to Cheon.et.al[4] (we call their scheme CKY scheme). Using our new scheme as kernel, we provide an efficient optimistic fair signature exchange protocol in ID-based setting.

The rest of this paper is organized as follows: In Section 2, we construct a new ID-based VESS. In Section 3, we present an ID-based optimistic fair signature exchange protocol. We provide protocol analysis in Section 4. Finally, we conclude in Section 5.

2 A New ID-Based VESS Based on CKY Scheme

Let $(G_1, +)$ and $(G_2, \cdot)$ be two cyclic groups of order q. $\hat{e} : G_1 \times G_1 \rightarrow G_2$ be a map which satisfies the following properties.

1. Bilinear: $\forall P, Q \in G_1, \forall \alpha, \beta \in Z_q, \hat{e}(\alpha P, \beta Q) = \hat{e}(P, Q)^{\alpha\beta}$;
2. Non-degenerate: If P is a generator of G_1, then $\hat{e}(P, P)$ is a generator of G_2;
3. Computable: There is an efficient algorithm to compute $\hat{e}(P, Q)$ for any $P, Q \in G_1$.

Such an bilinear map is called an *admissible bilinear pairing*. The Weil pairings and the Tate pairings of elliptic curves can be used to construct efficient admissible bilinear pairings. The **computational Diffie-Hellman problem (CDHP)** is to compute abP for given $P, aP, bP \in G_1$. We assume through this paper that CDHP is intractable.

Based on the CKY scheme[4], we proposed the following ID-based VESS, which is consists of seven polynomial-time algorithms:

- **Setup**: Given $G_1, G_2, q, \hat{e}, P$, return the system parameters $\Omega = (G_1, G_2, q, \hat{e}, P, P_{pub}, P_a, H_1, H_2)$, the PKG's private key $s \in Z_q^*$ and the adjudicator's private key $s_a \in Z_q^*$, where $P_{pub} = sP$, $P_a = s_aP$, $H_1 : \{0,1\}^* \rightarrow G_1^*$ and $H_2 : \{0,1\}^* \times G_1 \rightarrow Z_q$ are hash functions.
- **Extract**: Given an identity $ID \in \{0,1\}^*$, compute $Q_{ID} = H_1(ID) \in G_1^*$, $D_{ID} = sQ_{ID}$. PKG uses this algorithm to extract the user secret key D_{ID}, and gives D_{ID} to the user by a secure channel.
- **Sign**: Given a private key D_{ID} and a message m, pick $r \in Z_q^*$ at random, compute $U = rP$, $h = H_2(m, U)$, $V = rQ_{ID} + hD_{ID}$, and output a signature (U, V).
- **Verify**: Given a signature (U, V) of an identity ID for a message m, compute $h = H_2(m, U)$, and accept the signature and return 1 if and only if $\hat{e}(P, V) = \hat{e}(U + hP_{pub}, H_1(ID))$.
- **VE_Sign**: Given a secret key D_{ID} and a message m, choose $r_1, r_2 \in Z_q^*$ at random, compute $U_1 = r_1P$, $h = H_2(m, U_1)$, $U_2 = r_2P$, $V = r_1H_1(ID) + hD_{ID} + r_2P_a$, and output a verifiably encrypted signature (U_1, U_2, V).
- **VE_Verify**: Given a verifiably encrypted signature (U_1, U_2, V) of an identity ID for a message m, compute $h = H_2(m, U_1)$, and accept the signature and return 1 if and only if $\hat{e}(P, V) = \hat{e}(U_1 + hP_{pub}, H_1(ID)) \cdot \hat{e}(U_2, P_a)$.
- **Adjudication**: Given the adjudicator's secret key s_a and a valid verifiably encrypted signature (U_1, U_2, V) of an identity ID for a message m, compute $V_1 = V - s_aU_2$, and output the original signature (U_1, V_1).

Validity requires that verifiably encrypted signatures verify, and that adjudicated verifiably encrypted signatures verify as ordinary signatures, i.e., for $\forall m \in \{0,1\}^*$, $ID \in \{0,1\}^*$, $D_{ID} \leftarrow Extract(ID)$, satisfying:

1. $VE_Verify(ID, m, VE_Sign(D_{ID}, m)) = 1$
2. $Verify(ID, m, Adjudication(s_a, VE_Sign(D_{ID}, m))) = 1$

The correctness is easily proved as follows: For a verifiably encrypted signature (U_1, U_2, V) of an identity ID for a message m.

$$\begin{aligned}\hat{e}(P, V) &= \hat{e}(P, r_1H_1(ID) + hD_{ID} + r_2P_a) \\ &= \hat{e}((r_1 + hs)P, H_1(ID)) \cdot \hat{e}(r_2P, P_a) \\ &= \hat{e}(U_1 + hP_{pub}, H_1(ID)) \cdot \hat{e}(U_2, P_a)\end{aligned}$$

That is, $VE_Verify(ID, m, VE_Sign(D_{ID}, m)) = 1$.

On the other hand, $V_1 = V - s_a U_2 = V - s_a r_2 P = V - r_2 P_a = r_1 Q_{ID} + hD_{ID}$. $\hat{e}(P, V_1) = \hat{e}(P, r_1 Q_{ID} + hD_{ID}) = \hat{e}((r_1 + hs)P, H_1(ID)) = \hat{e}((U_1 + hP_{pub}, H_1(ID))$

So we have, $Verify(ID, m, Adjudication(s_a, VE_Sign(D_{ID}, m))) = 1$.

3 An ID-Based Optimistic Fair Signature Exchange Protocol

Based on the ID-based VESS described in Section 2, we present an ID-based optimistic fair signature exchange protocol. Our new protocol consists of four procedures: **Initialization**, **Registration**, **Exchange** and **Dispute**.

- **Initialization**: TTP runs $Setup(1^k)$ to generate system parameters $\Omega = (G_1, G_2,\ q, \hat{e},\ P, P_{pub}, P_a\ \ H_1, H_2)$, the master secret key $s \in Z_q^*$ and the adjudication key $s_a \in Z_q^*$. TTP publishes Ω, while keeps s and s_a secretly.
- **Registration**: A user with identity $ID \in \{0,1\}^*$ registers at the TTP with ID. TTP extracts the user secret key $D_{ID} = Extract(ID)$ and sends D_{ID} to the user by a secure channel.
- **Exchange**: Let Alice be the sponsor with identity ID_A and secret key D_{ID_A}. Alice exchanges a signature of a message m with Bob, whose identity is ID_B and secret key is D_{ID_B}. The exchange procedure is as following:
  - Step1: Alice computes verifiably encrypted signature $(U_1, U_2, V) = VE_Sign(D_{ID_A}, m)$, and sends $(m, (U_1, U_2, V))$ to Bob.
  - Step2: Bob checks the validity of $(m, (U_1, U_2, V))$. If $VE_Verify(ID_A, m, (U_1, U_2, V)) \neq 1$, then aborts. Otherwise, Bob computes ordinary signature $(U_B, V_B) = Sign(D_{ID_B}, m)$, and sends (U_B, V_B) to Alice.
  - Step3: Alice checks the validity of $(m, (U_B, V_B))$. If $Verify(ID_B, m, (U_B, V_B)) \neq 1$, then aborts. Otherwise, Alice computes ordinary signature $(U_A, V_A) = Sign(D_{ID_A}, m)$, and sends (U_A, V_A) to Bob.
  - Step4: If Bob receives (U_A, V_A) and $Verify(ID_A, m, (U_A, V_A)) = 1$, the protocol ends with success. Otherwise, Bob can request to TTP for arbitrament.
- **Dispute**:
  - Step1: Bob sends $(m, (U_1, U_2, V))$ and (U_B, V_B) to TTP.
  - Step2: TTP verifies the validity of (U_1, U_2, V) and (U_B, V_B). If $VE_Verify(ID_A, m, (U_1, U_2, V)) \neq 1$ or $Verify(ID_B, m, (U_B, V_B)) \neq 1$, then aborts. Otherwise, TTP computes $(U_A, V_A) = Adjudication(s_a, (U_1, U_2, V))$
  - Step3: TTP sends (U_A, V_A) to Bob and sends (U_B, V_B) to Alice.

In the protocol, TTP works in an *optimistic* way. That is, TTP does not participate in the actual **Exchange** protocol in normal cases (no argument appears), and is invoked only in abnormal cases to dispute the arguments for fairness. If no dispute occurs, only Bob and Alice need to participate in the exchange.

Note: In the description of the protocol, TTP acts both as PKG and the arbiter. In fact, this two roles can be executed by two different trusted entities.

4 Protocol Analysis

4.1 Fairness and Efficiency

At the end of Step1 of **Exchange**, if Bob aborts after receiving $VE_Sign(D_{ID_A}, m)$, Bob can't get the ordinary signature $Sign(D_{ID_A}, m)$ by himself. If Bob requests to TTP for dispute with valid $VE_Sign(D_{ID_A}, m)$ and $Sign(D_{ID_B}, m)$, TTP computes $Sign(D_{ID_A}, m) = Adjudication(s_a, VE_Sign(D_{ID_A}, m))$, sends $Sign(D_{ID_B}, m)$ to Bob and sends $Sign(D_{ID_B}, m)$ to Alice. That is, either party gets the other's signature, or neither party does. At the end of Step2 or Step3 of **Exchange**, if Bob has send Alice $Sign(D_{ID_B}, m)$ while hasn't received valid $Sign(D_{ID_A}, m)$, Bob can request to TTP for dispute with valid $VE_Sign(D_{ID_A}, m)$ and $Sign(D_{ID_B}, m)$. As a result, either party gets the other's signature.

The efficiency of the protocol can be evaluated by the performance of the algorithms of ID-based VESS. Denote by M a scalar multiplication in $(G_1, +)$ and by $\hat{e}$ a computation of the pairing. Do not take other operations into account.

	Sign	Verify	VE_Sign	VE_Verify	Adjudication	digital certificate
Proposed	$3M$	$2\hat{e} + 1M$	$5M$	$3\hat{e} + 1M$	$1M$	not need

4.2 Security Proof

Security proof is a sticking point for the design of new protocols. The security of our new protocol can be reduced to the security of the basic ID-based VESS. As mentioned in [3], a secure VESS should satisfies the following properties:

- **Signature Unforgeability**: It is difficult to forge a valid ordinary signature.
- **Verifiably Encrypted Signature Unforgeability**: It is difficult to forge a valid verifiably encrypted signature.
- **Opacity**: It is difficult, given a verifiably encrypted signature, to get an ordinary signature on the same message.

In this section, we will extend this security notation to ID-based setting.

The ordinary signature algorithm of our ID-based VESS is the same as that of CKY scheme. The signature unforgeability has been shown in the random oracle model under the hardness assumption of CDHP in [4].

We consider an adversary $\mathcal{F}$ which is assumed to be a polynomial time probabilistic Turing machine which takes as input the global scheme parameters and a random tape. To aid the adversary we allow it to query the following oracles:

- Extract oracle $E(.)$: For input an identity ID, this oracle computes $D_{ID} = Extract(ID)$, and outputs the secret key D_{ID}
- VE_Sign oracle $VS(.)$: For input (ID, m), this oracle computes and outputs a verifiably encrypted signature $\pi = VE_Sign(D_{ID}, m)$, where $D_{ID} = Extract(ID)$.
- Adjudication oracle $A(.)$: For input ID, m and a valid verifiable encrypted signature π of ID for m, this oracle computes and outputs the corresponding ordinary signature δ.

Note: An ordinary signing oracle is not provided, because it can be simulated by a call to $VS(.)$ followed by a call to $A(.)$. In the random oracle model, $\mathcal{F}$ also has the ability to issue queries to the hash function oracles $H_1(.), H_2(.)$ adaptively.

Definition 1. *The advantage in existentially forging a verifiably encrypted signature of an algorithm $\mathcal{F}$ is defined as*

$$Adv_{\mathcal{F}}^{EUF}(k) = Pr\left[\begin{array}{c}\Omega \leftarrow Setup(1^k), (ID, m, \pi) \leftarrow \mathcal{F}^{E(.),VS(.),A(.),H_1(.),H_2(.)}(\Omega) : \\ VE_Verify(ID, m, \pi) = 1, \\ (ID, .) \notin E_l, (ID, m, .) \notin V_l, (ID, m, .) \notin A_l\end{array}\right]$$

*where E_l, V_l and A_l are the query and answer lists coming from $E(.)$, $VS(.)$ and $A(.)$ respectively during the attack. The probability is taken over the coin tosses of the algorithms, of the oracles, and of the forger. An ID-based VESS is said to be **existential unforgeable**, if for any adversary $\mathcal{F}$, $Adv_{\mathcal{F}}^{EUF}(k)$ is negligible.*

Definition 2. *The advantage in opacity attack of an algorithm $\mathcal{F}$ is defined as*

$$Adv_{\mathcal{F}}^{OPA}(k) = Pr\left[\begin{array}{c}\Omega \leftarrow Setup(1^k), (ID, m, \pi, \delta) \leftarrow \mathcal{F}^{E(.),VS(.),A(.),H_1(.),H_2(.)}(\Omega) : \\ VE_Verify(ID, m, \pi) = 1, Verify(ID, m, \delta) = 1, \\ A(ID, m, \pi) = \delta, (ID, .) \notin E_l, (ID, m, .) \notin A_l\end{array}\right]$$

*The probability is taken over the coin tosses of the algorithms, of the oracles, and of the forger. An ID-based VESS is said to be **opaque**, if for any adversary $\mathcal{F}$, $Adv_{\mathcal{F}}^{OPA}(k)$ is negligible.*

Theorem 1. *In the random oracle model, if there is an adversary $\mathcal{F}_0$ which performs, within a time bound T, an existential forgery against our ID-based VESS with probability ε, then there is an adversary $\mathcal{F}_1$ which performs an existential forgery against CKY scheme with probability no less than ε, within a time bound $T + (2n_{VS} + n_A)M$, where n_{VS} and n_A are the number of queries that $\mathcal{F}_0$ can ask to $VS(.)$ and $A(.)$ respectively, M denotes a scalar multiplication in G_1.*

Proof. From $\mathcal{F}_0$, we can construct an adversary $\mathcal{F}_1$ of CKY scheme. The detail proof is similar to the proof of Theorem 4.4 in [3]. Due to the limit of paper length, we omit the details of the proof in this paper.

Theorem 2. *In the random oracle mode, let $\mathcal{F}_0$ be an adversary which has running time T and success probability ε in opaque attack. We denote by n_{h_1}, n_E, n_A and n_{VS} the number of queries that $\mathcal{F}_0$ can ask to the oracles $H_1(.)$, $Extract(.)$, $A(.)$ and $VS(.)$ respectively. Then there is a polynomial-time Turing machine $\mathcal{F}_1$ who can solve the computational Diffie-Hellman problem within expected time $T + (5n_{VS} + n_E + n_A + n_{h_1})M$ with probability $\varepsilon/(e \cdot n_{h_1} \cdot n_{VS})$.*

Proof. Without any loss of generality, we may assume that for any ID, $\mathcal{F}_0$ queries $H_1(.)$ with ID before ID is used as (part of) an input of any query to $E(.)$, $VS(.)$, or $A(.)$. From the adversary $\mathcal{F}_0$, we construct a Turing machine $\mathcal{F}_1$ which outputs abP on input of any given $P, aP, bP \in G_1^*$ as follows:

1. $\mathcal{F}_1$ runs $Setup$) to generate the PKG's private key $s \in Z_q^*$, the adjudicator's private key $s_a \in Z_q^*$ and other system parameters $G_2, \hat{e}, P_{pub}, P_a, H_1, H_2$, where $P_{pub} = sP$, $P_a = s_aP$,

2. $\mathcal{F}_1$ sets $P_a = bP$, $v = 1, j = 1$ and $V_l = \Phi$.
3. $\mathcal{F}_1$ picks randomly t and ι satisfying $1 \leq t \leq n_{h_1}$, $1 \leq \iota \leq n_{VS}$, and picks randomly $x_i \in Z_q, i = 1, 2, ...n_{h_1}$.
4. $\mathcal{F}_1$ gives $\Omega = (G_1, G_2, q, \hat{e}, P, P_{pub}, P_a, H_1, H_2)$ to $\mathcal{F}_0$ as input and lets $\mathcal{F}_0$ run on. During the execution, $\mathcal{F}_1$ emulates $\mathcal{F}_0$'s oracles as follows:
   - $H_1(.)$: For input ID, $\mathcal{F}_1$ checks if $H_1(ID)$ is defined. If not, he defines $H_1(ID) = \begin{cases} k_1 bP & v = t \\ x_v P & v \neq t \end{cases}$, where $k_1 \in_R Z_q^*$, and sets $ID_v = ID$, $v = v + 1$. $\mathcal{F}_1$ returns $H_1(ID)$ to $\mathcal{F}_0$.
   - $H_2(.)$: For input (m, U), $\mathcal{F}_1$ checks if $H_2(m, U)$ is defined. If not, it picks a random $h \in Z_q$, and sets $H_2(m, U) = h$. $\mathcal{F}_1$ returns $H_2(m, U)$ to $\mathcal{F}_0$.
   - $Extract(.)$: For input ID_i, if $i = t$, $\mathcal{F}_1$ returns with $\perp$. Otherwise, $\mathcal{F}_1$ lets $d_i = x_i \cdot P_{pub}$ be the reply to $\mathcal{F}_0$.
   - $VS(.)$: For input ID_i and message m, $\mathcal{F}_1$ emulates the oracle as follows:
     - Case 0: If $j = \iota$, and $i = t$,
       a. Pick randomly $k_2, h \in z_q$;
       b. $U_1 = aP - hP_{pub}$, $U_2 = k_1(k_2 P - aP)$, $V = k_1 k_2 P$;
       c. If $H_2(m, U_1)$ has been defined, $\mathcal{F}_1$ aborts (a collision appears). Otherwise, set $H_2(m, U_1) = h$.
       d. Add $(i, j, ., U_1, U_2, V)$ to VS_{list}.
     - Case 1: If $j \neq \iota$, or $i \neq t$,
       a. Pick randomly $r_2, z_j, h \in z_q$;
       b. If $j \neq \iota$, then compute $U_1 = z_j P - hP_{pub}$, $U_2 = r_2 P$, $V = z_j(H_1(ID_i)) + r_2 P_a$. Otherwise, compute $U_1 = aP - hP_{pub}$, $U_2 = r_2 P$, $V = x_i(aP) + r_2 P_a$;
       c. If $H_2(m, U_1)$ has been defined, $\mathcal{F}_1$ aborts (a collision appears). Otherwise, set $H_2(m, U_1) = h$.
       d. Add (i, j, r_2, U_1, U_2, V) to VS_{list}.

     Set $j = j + 1$ and let (U_1, U_2, V) be the reply to $\mathcal{F}_0$.
   - $A(.)$: For input ID_i, m and valid verifiably encrypted signature (U_1, U_2, V), $\mathcal{F}_1$ obtains the corresponding item (i, j, r_2, U_1, U_2, V) (or $(i, j, ., U_1, U_2, V)$) from the VS_{list}. If $i = t$ and $j = \iota$, $\mathcal{F}_1$ declares failure and aborts. Otherwise, $\mathcal{F}_1$ computes $V_1 = V - r_2 P_a$, and replies to $\mathcal{F}_0$ with (U_1, V_1).
5. If $\mathcal{F}_0$'s output is $(ID_i, m^*, (U_1^*, U_2^*, V^*, V_1^*))$, then $\mathcal{F}_1$ obtains the corresponding item on the VS_{list}. If $i = t$ and $j = \iota$, $\mathcal{F}_1$ computes and successfully outputs $abP = k_1^{-1} V_1^*$. Otherwise, $\mathcal{F}_1$ declares failure and aborts.

This completes the description of $\mathcal{F}_1$. It is easy to see that the probability that a collision appears in the definition of $H_2(m, U_1)$ in step4 is negligible and the probability that $\mathcal{F}_1$ does not abort as a result of $\mathcal{F}_0$'s adjudication queries in step4 is at lest $1/e$ (see Claim 4.6 in [3]). The input Ω given to $\mathcal{F}_0$ is from the same distribution as that produced by $Setup$. If no collision appears and $\mathcal{F}_1$ does not abort, the responses of $\mathcal{F}_1$'s emulations are indistinguishable from $\mathcal{F}_0$'s real oracles. Therefore $\mathcal{F}_0$ will produce a valid and nontrivial $(ID_i, m, U_1^*, U_2^*, V^*, V_1^*)$ with probability at least ε/e. That is, $\mathcal{F}_1$ computes and successfully outputs abP

with probability at least $\varepsilon/(e \cdot n_{h_1} \cdot n_{VS})$ as required. $\mathcal{F}_1$'s running time is approximately the same as $\mathcal{F}_0$'s running time plus the time taken to respond to $\mathcal{F}_0$'s oracle queries. Neglect operations other than $\hat{e}(*,*)$ and scalar multiplication in $(G_1, +)$, the total running time is roughly $T + (5n_{VS} + n_E + n_A + n_{h_1})M$.

5 Conclusion

Protocols for fair exchange of signatures have found numerous practical applications in electronic commerce. In this paper, we propose an efficient and secure ID-based optimistic fair signature exchange protocols based on bilinear pairings. The users in the signature exchange protocol use ID-based setting and need no digital certificates. Our new protocol provides an efficient and secure solution for the problem of fair exchange of signatures in ID-based public key cryptosystem.

References

1. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Advances in Cryptology - CRYPTO'84. Lecture Notes in Computer Science, Vol. 196. Springer-Verlag, Berlin Heidelberg New York (1984) 47-53.
2. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In: Advances in Cryptology- CRYPTO 2001. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York (2001) 213-229.
3. Boneh, D., Gentry, C., Lynn, B., Shacham, H.: Aggregate and Verifiably Encrypted Signature from Bilinear Maps. In: Eurocrypt 2003. Lecture Notes in Computer Science, Vol. 2248. Springer-Verlag, Berlin Heidelberg New York (2003) 514-532.
4. Cheon, J.H., Kim, Y., Yoon, H.: Batch Verifications with ID-based Signatures. In: Information Security and Cryptology - ICISC 2004. Lecture Notes in Computer Science, Vol. 3506. Springer-Verlag, Berlin Heidelberg New York (2005) 233-248.
5. Hess,F.: Efficient identity based signature schemes based on pairings. In: Selected Areas in Cryptography 9th Annual International Workshop, SAC 2002. Lecture Notes in Computer Science, Vol. 2595. Springer-Verlag, Berlin Heidelberg New York (2003) 310-324.
6. Asokan, N., Shoup, V., Waidner, M.: Optimistic fair exchange of signatures. In: Advances in Cryptology - EUROCRYPT 1998. Lecture Notes in Computer Science vol. 1403. Springer-Verlag, Berlin Heidelberg New York (1998) 591-606.
7. Dodis, Y., Reyzin, L., Breaking and reparing optimistic fair exchange from PODC 2003. In: Proc. of the 2003 ACM Workshop on Digital Rights Management. ACM Press, New York (2003) 47-54.
8. Poupard, G., Stern, J.: Fair encryption of RSA keys. In: Proc. of Eurocrypt 2000, Lecture Notes in Computer Science vol. 1807. Springer-Verlag, Berlin Heidelberg New York (2000) 172-189.
9. Camenisch, J., Shoup, V.: Practical verifiable encryption and decryption of discrete logarithms. In: Advances in Cryptology - CRYPTO 2003, Lecture Notes in Computer Science vol. 2729. Springer-Verlag, Berlin Heidelberg New York (2003) 195-211.

FMS Attack-Resistant WEP Implementation Is Still Broken

— Most IVs Leak a Part of Key Information —

Toshihiro Ohigashi[1], Yoshiaki Shiraishi[2], and Masakatu Morii[3]

[1] The University of Tokushima, 2-1 Minamijosanjima, Tokushima 770-8506, Japan
ohigashi@is.tokushima-u.ac.jp
[2] Kinki University, 3-4-1 Kowakae, Higashi-Osaka 577-8502, Japan
zenmei@info.kindai.ac.jp
[3] Kobe University, 1-1 Rokkodai, Kobe 657-8501, Japan
mmorii@kobe-u.ac.jp

Abstract. In this paper, we present an attack to break WEP that avoids weak IVs used in the FMS attack. Our attack is a known IV attack that doesn't need the specific pattern of the IVs. This attack transforms most IVs of WEP into weak IVs. If we attempt to avoid all weak IVs used in our attack, the rate at which IVs are avoided is too large to use practical. When using a 128-bit session key, the efficiency of our attack is $2^{72.1}$ in the most effective case. This implies that our attack can recover a 128-bit session key within realistically possible computational times.

1 Introduction

Wired Equivalent Privacy (WEP) protocol is a part of the IEEE 802.11 standard [1] and is a security protocol for wireless LAN communication. A key recovery attack against WEP, referred to as the FMS attack, was proposed in 2001 [2]. The FMS attack is a chosen Initialization Vector (IV) attack. The IV of the specific pattern used in the FMS attack is referred to as the weak IV, and it leaks the secret key information. When a 128-bit session key is used, $13 \cdot 2^8$ 24-bit IVs are transformed into weak IVs by the FMS attack. The FMS attack can recover the 104-bit secret key using the weak IVs. In order to avoid all the weak IVs used in the FMS attack, we must remove $13/2^{16}$ (about 0.02%) of the IVs. The proposers of the FMS attack recommended two methods of improvement to address the increasing need of securing WEP protocol against the FMS attack [2]. The first involves generating the session key by a secure hash function [3] from the IV and the secret key. The second involves discarding the first 2^n outputs of RC4 [4], where n is the word size of RC4. RC4 is used as the packet encryption algorithm of WEP. Some security protocols, e.g. Wi-Fi Protected Access (WPA) [5], are designed based on these improvements. Another improvement to safeguard against the FMS attack is to avoid weak IVs. The advantage of this improvement is that WEP can be modified to avoid the weak IVs by implementing a minimal number of changes. We refer to this improved

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 17–26, 2005.
© Springer-Verlag Berlin Heidelberg 2005

WEP as the FMS attack-resistant WEP implementation. Some wireless LAN chip makers have incorporated this improvement in their designs.

In this paper, we demonstrate the weakness of the FMS attack-resistant WEP implementation by showing that most IVs are transformed into weak IVs by our attack. Our attack is a known IV attack, that is, the attacker can obtain a part of the session key information from any 24-bit IV. In case of a 128-bit session key, $13/16 \cdot 2^{24}$ 24-bit IVs are transformed into weak IVs by our attack. In order to avoid all the weak IVs used in this attack, we must remove 13/16 (about 81%) of the IVs. The rate at which IVs are avoided is too large to use practical. Our attack can reduce the computational times for recovering the secret key compared with the exhaustive key search. When a 128-bit session key is used, the efficiency of our attack fir recovering a 104-bit secret key is $2^{72.1}$ in the most effective case. This shows that our attack can recover a 104-bit secret key within realistically possible computational times. Finally, the FMS attack-resistant WEP implementation is broken by our attack.

2 WEP Protocol

2.1 Generation of the Session Key

In WEP, a secret key K' is preshared between an access point and a mobile node. The length of K' is either 40-bit or 104-bit. The session key K is generated according to $K = IV||K'$, where IV is a 24-bit IV and $||$ is concatenation. The IV is transmitted in plaintext and changed for every packet. We refrain from discussing WEP that utilises a 40-bit secret key because it is easily broken by exhaustive key search.

2.2 RC4

We follow the description of RC4 as given in [4]. RC4 comprises the Key-Scheduling Algorithm (KSA) and the Pseudo-Random Generation Algorithm (PRGA). The initial state is made by the session key K in the KSA. A keystream is generated from the initial state in the PRGA. The plaintext is XOR-ed with the keystream to obtain the ciphertext. If the attacker can know the first word of the plaintext, it is possible to obtain the first output word of RC4. For WEP, the method to determine the first word of the plaintext was explained in [6].

Key-Scheduling Algorithm: The internal state of RC4 at time t in the KSA consists of a permutation table $S_t^* = (S_t^*[x])_{x=0}^{2^n-1}$ of 2^n n-bit words (in practice, $n = 8$). All the values of S_t^* differ from each other, that is, if $x \neq x'$ then $S_t^*[x] \neq S_t^*[x']$. S_t^* is initialized to $S_0^*[x] = x$. Two n-bit word pointers i_t^* and j_t^* at time t are used; they are initialized to $i_0^* = j_0^* = 0$. The l-bit session key is split into $\lceil l/n \rceil$ n-bit words $K = (K[x])_{x=0}^{\lceil l/n \rceil - 1}$. In WEP, $K[0]$, $K[1]$ and $K[2]$ are the IV entries. The initial state is made by Eqs. (1)–(3) in time $t = 1, 2, \ldots, 2^n$:

$$j_t^* = (j_{t-1}^* + S_{t-1}^*[i_{t-1}^*] + K[i_{t-1}^* \bmod \lceil l/n \rceil]) \bmod 2^n, \tag{1}$$

$$S_t^*[i^*] = \begin{cases} S_{t-1}^*[j_t^*], & i^* = i_{t-1}^* \\ S_{t-1}^*[i_{t-1}^*], & i^* = j_t^* \\ S_{t-1}^*[i^*], & i^* \neq i_{t-1}^*, j_t^*, \end{cases} \tag{2}$$

$$i_t^* = (i_{t-1}^* + 1) \bmod 2^n. \tag{3}$$

Pseudo-Random Generation Algorithm: The internal state of RC4 at time t in the PRGA consists of a permutation table $S_t = (S_t[x])_{x=0}^{2^n-1}$ of 2^n n-bit words. S_0 is initialized to $S_0[x] = S_{2^n}^*[x]$. Two n-bit word pointers i_t and j_t at time t are used; these are initialized to $i_0 = j_0 = 0$. Let Z_t denote the output n-bit word of RC4 at time t. Then, the next-state and output functions of RC4 for every $t \geq 1$ are defined as

$$i_t = (i_{t-1} + 1) \bmod 2^n \ , \ j_t = (j_{t-1} + S_{t-1}[i_t]) \bmod 2^n, \tag{4}$$

$$S_t[i] = \begin{cases} S_{t-1}[j_t], & i = i_t \\ S_{t-1}[i_t], & i = j_t \\ S_{t-1}[i], & i \neq i_t, j_t, \end{cases} \tag{5}$$

$$Z_t = S_t[(S_t[i_t] + S_t[j_t]) \bmod 2^n]. \tag{6}$$

3 Known IV Attack Against WEP

We propose a known IV attack against WEP in order to recover the session key. First, we present the algorithm of our attack. Second, we describe an attack equation and discuss the bias caused by the attack equation. Third, we present the efficiency of our attack by experimental results and statistical analysis. Finally, we discuss the rate of weak IVs used in our attack.

3.1 Algorithm

The goal of our attack is to recover all the secret key entries $K[3], K[4], \ldots, K[\lceil l/n \rceil - 1]$ by using the known entries $K[0]$, $K[1]$ and $K[2]$. Our attack comprises the key recovery and exhaustive key search processes.

Key Recovery Process: In this process, the session key entries are recovered by using weak IVs and an attack equation with high probability. We observe that the IVs are split into $\lceil l/n \rceil$ groups – each of which leaks a session key information $K[a \bmod \lceil l/n \rceil]$. a is defined as follows:

$$a = (K[0] + K[1] + 2) \bmod 2^n, \ a \in \{0, 1, \ldots, 2^n - 1\}. \tag{7}$$

We assume that the attacker can obtain any pair of IV and Z_1. Then, an IV in this group would leak information regarding $K[a \bmod \lceil l/n \rceil]$ by the following attack equation;

$$K[a \bmod \lceil l/n \rceil] = (Z_1 - (a-1) - a) \bmod 2^n. \tag{8}$$

The right-hand side of Eq. (8) can recover $K[a \bmod \lceil l/n \rceil]$ with a higher probability than that by a random search. If the attacker can obtain several IV and Z_1 pairs, the success probability of recovering session key entries increases with a statistical attack using Eq. (8). The following is the procedure for a statistical attack:

Step 1. Obtain a pair of IV and Z_1.
Step 2. Determine a using Eq. (7).
Step 3. Consider the value calculated by Eq. (8) as a candidate for $K[a \bmod \lceil l/n \rceil]$.
Step 4. Repeat Step 1–3 by using different IV and Z_1 pairs. Finally, the candidate of $K[a \bmod \lceil l/n \rceil]$ corresponding to the highest frequency is guessed as the correct value in each $a \bmod \lceil l/n \rceil$.

The success probability of recovering session key entries increases if the number of the obtained IV and Z_1 pairs increases. Due to this reason, the bias of the attack equation exerts a strong influence. All of the secret key entries guessed by the key recovery process are not necessarily correct. Additionally, the number of secret key entries that are recovered correctly is unknown. Therefore, the attacker selects the threshold value k and executes the exhaustive key search process – the procedure to recover the remaining $\lceil l/n \rceil - 3 - k$ secret key entries when at least k bytes of secret key entries is recovered. The success probability and time complexity of our attack depend on the selected k.

Exhaustive Key Search Process: In this process, the attacker assumes that k of the secret key entries guessed by the key recovery process are correct. Then, the remaining $\lceil l/n \rceil - 3 - k$ secret key entries are recovered by the exhaustive key search. The attacker cannot determine the positions of the recovered secret key entries; the ${}_{\lceil l/n \rceil - 3}C_k$ times complexity is required for this process. Therefore, the number of all patterns that should be searched in the exhaustive key search process is written as ${}_{\lceil l/n \rceil - 3}C_k \cdot 2^{(\lceil l/n \rceil - 3 - k)n}$.

3.2 Attack Equation

Definition 1. *In the fortunate case, $K[a \bmod \lceil l/n \rceil]$ is recovered by Eq. (8) with probability 1 under the condition that the IV holds Eq. (7).*

Theorem 1. *When the fortunate case occurs with probability $\alpha(a)$, the probability that $K[a \bmod \lceil l/n \rceil]$ is recovered by Eq. (8) is equal to $\alpha(a) \cdot (2^n - 1)/2^n + 1/2^n$.*

Proof. In the *fortunate case*, Eq. (8) recovers $K[a \bmod \lceil l/n \rceil]$ with probability 1. For the other case that occurs with probability $1 - \alpha(a)$, the probability that $K[a \bmod \lceil l/n \rceil]$ is recovered by Eq. (8) is equal to $1/2^n$ (it is equal to the probability that $K[a \bmod \lceil l/n \rceil]$ is recovered by a truly random value). As a result, the total probability that $K[a \bmod \lceil l/n \rceil]$ is recovered by Eq. (8) is written as follows:

$$\alpha(a) \cdot 1 + (1 - \alpha(a)) \cdot \frac{1}{2^n} = \alpha(a) \cdot \frac{2^n - 1}{2^n} + \frac{1}{2^n}. \tag{9}$$

□

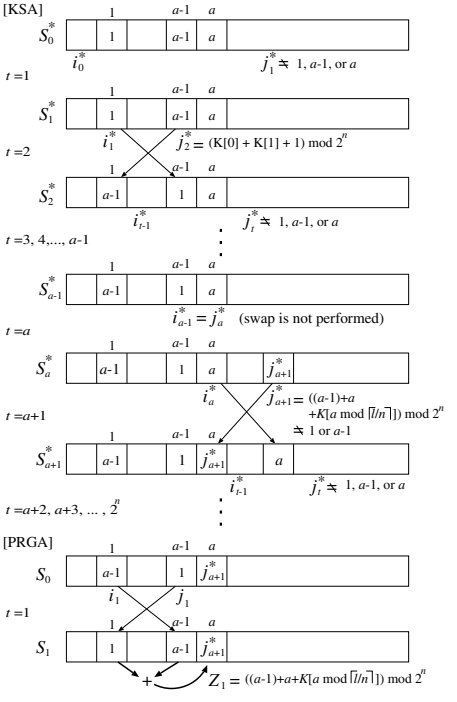


Fig. 1. The *fortunate case* scenario

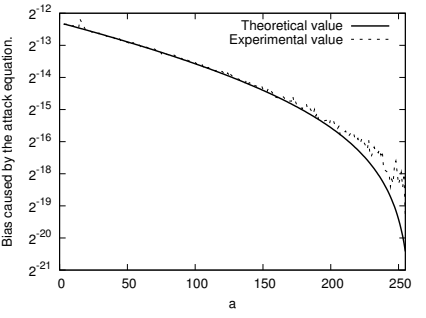


Fig. 2. The theoretical and experimental values of the bias caused by the attack equation

We discuss the bias of Eq. (9) from the probability that $K[a \bmod \lceil l/n \rceil]$ is recovered by a truly random value. We present the condition for the *fortunate case* in order to calculate the theoretical value of $\alpha(a)$. Figure 1 shows the *fortunate case* scenario. We define **KSA**(t) and **PRGA**(t) as the processing in the KSA and PRGA, respectively, at time t. In the following, we explain the condition for the *fortunate case* and calculate the probability that this condition holds.

KSA$(t = 1)$ **[Condition 1]:** We can calculate $i_t^* = t$ by Eq. (3) and $i_0^* = 0$. Then, $i_0^* = 0$ holds. It is assumed that $j_1^* \neq 1$, $a-1$, or a holds. Then, $S_1^*[1] = 1$, $S_1^*[a-1] = a-1$ and $S_1^*[a] = a$ hold because their entries are not swapped by Eq. (2).

(The probability of satisfying Condition 1)

It is assumed that j_1^* is an independent and random value. Then, the probability that $j_1^* \neq 1$, $a-1$, or a holds is $(2^n - 3)/2^n$.

KSA$(t = 2)$ **[Condition 2]:** $i_1^* = 1$ holds as well as Condition 1. $j_2^* = (K[0] + K[1] + S_0^*[0] + S_1^*[1]) \bmod 2^n$ holds by Eqs. (1) and (3). If Condition 1 holds, $j_2^* = (a-1) \bmod 2^n$ holds because $S_0^*[0] = 0$ and $S_1^*[1] = 1$ hold and $(K[0] + K[1] + 2) \bmod 2^n = a$ holds as the condition of the IV. Then, $S_2^*[1] = a-1$, $S_2^*[a-1] = 1$ and $S_2^*[a] = a$ are satisfied by Eq. (2).

(The probability of satisfying Condition 2)

If Condition 1 holds, the probability that Condition 2 holds is 1.

KSA$(t = 3, 4, \ldots, a-1)$ **[Condition 3]:** $i^*_{t-1} = t-1$ holds as well as Condition 1. It is assumed that $j^*_t \neq 1$, $a-1$, or a holds. If Condition 1–2 hold, $S^*_t[1] = a-1$, $S^*_t[a-1] = 1$ and $S^*_t[a] = a$ also hold because their entries are not swapped by Eq. (2).

(The probability of satisfying Condition 3)

It is assumed that j^*_t is an independent and random value. Then, the probability that $j^*_t \neq 1$, $a-1$, or a holds is written as

$$\prod_{t=3}^{a-1} \frac{2^n - 3}{2^n}. \tag{10}$$

KSA$(t = a)$ **[Condition 4]:** $i^*_{a-1} = a-1$ holds as well as Condition 1. It is assumed that $j^*_a = i^*_{a-1}$ holds. Then, the internal-state entries are not swapped by Eq. (2). If Conditions 1–3 hold, $S^*_a[1] = a-1$, $S^*_a[a-1] = 1$ and $S^*_a[a] = a$ also hold.

(The probability of satisfying Condition 4)

It is assumed that j^*_a is an independent and random value. Then, the probability that $j^*_a = i^*_{a-1}$ holds is $1/2^n$.

KSA$(t = a+1)$ **[Condition 5]:** $i^*_a = a$ holds as well as Condition 1. It is assumed that $j^*_{a+1} \neq 1$ or $a-1$ and $S^*_a[j^*_{a+1}] = j^*_{a+1}$ hold. If Conditions 1–4 hold, $j^*_{a+1} = ((a-1) + a + K[a \bmod \lceil l/n \rceil]) \bmod 2^n$ holds by Eq. (1) because $j^*_a = i^*_{a-1} = a-1$ and $S^*_a[i^*_a] = S^*_a[a] = a$ hold. Then, $S^*_{a+1}[1] = a-1$, $S^*_{a+1}[a-1] = 1$ and $S^*_{a+1}[a] = j^*_{a+1} = ((a-1) + a + K[a \bmod \lceil l/n \rceil]) \bmod 2^n$ also hold by Eq. (2) because $S^*_a[j^*_{a+1}] = j^*_{a+1}$ holds.

(The probability of satisfying Condition 5)

We define $f(t)$ as the number of entries that satisfy $S^*_t[x] = x$, for $x \neq 1$, $a-1$, or a, and $x \in \{0, 1, \ldots, 2^n - 1\}$. It is assumed that j^*_{a+1} is an independent and random value. If Conditions 1–4 hold, the probability that $j^*_{a+1} \neq 1$ or $a-1$ and $S^*_a[j^*_{a+1}] = j^*_{a+1}$ hold is written as

$$\frac{1}{2^n} \cdot 1 + \frac{2^n - 3}{2^n} \cdot \frac{f(a)}{2^n - 3}. \tag{11}$$

We describe the manner in which Eq. (11) can be calculated. We separately consider the cases where $j^*_{a+1} = a = i^*_a$ and $j^*_{a+1} \neq 1$, $a-1$, or a hold. The probability that $j^*_{a+1} = a = i^*_a$ holds is $1/2^n$. If $j^*_{a+1} = a = i^*_a$ holds, the probability that $S^*_a[j^*_{a+1}] = j^*_{a+1}$ holds is 1 because $S^*_a[a] = a$ holds. The probability that $j^*_{a+1} \neq 1$, $a-1$ or a holds is $(2^n - 3)/2^n$. If $j^*_{a+1} \neq 1$, $a-1$, or a holds, the probability that $S^*_a[j^*_{a+1}] = j^*_{a+1}$ holds is written as $f(a)/(2^n - 3)$. The details of $f(a)$ are given in Appendix A.

KSA$(t = a+2, a+3, \ldots, 2^n)$ **[Condition 6]:** $i^*_{t-1} = t-1$ holds as well as Condition 1. It is assumed that $j^*_t \neq 1$, $a-1$, or a holds. If Conditions 1–5 hold, $S^*_t[1] = a-1$, $S^*_t[a-1] = 1$ and $S^*_t[a] = ((a-1) + a + K[a \bmod \lceil l/n \rceil]) \bmod 2^n$ also hold because their entries are not swapped by Eq. (2).

(The probability of satisfying Condition 6)

It is assumed that j^*_t is an independent and random value. Then, the probability that $j^*_t \neq 1$, $a-1$, or a holds is written as

$$\prod_{t=a+2}^{2^n} \frac{2^n-3}{2^n}. \tag{12}$$

PRGA$(t=1)$ **[Condition 7]:** If Conditions 1–6 hold, $S_0[1] = a-1$, $S_0[a-1] = 1$ and $S_0[a] = ((a-1)+a+K[a \bmod \lceil l/n \rceil]) \bmod 2^n$ also hold. As the definition of the PRGA indicates, $i_0 = j_0 = 0$ holds. Z_1 is calculated by using Eqs. (4)–(6) as follows:

$$Z_1 = S_1[(S_0[1] + S_0[S_0[1]]) \bmod 2^n] = S_1[a]. \tag{13}$$

$Z_1 = S_1[a] = S_0[a] = ((a-1)+a+K[a \bmod \lceil l/n \rceil]) \bmod 2^n$ holds because $a \neq i_1$, or j_1 holds. Then, $K[a \bmod \lceil l/n \rceil]$ is calculated by using Eq. (8).

(The probability of satisfying Condition 7)

If Conditions 1–6 hold, the probability that Condition 7 holds is 1.

$\alpha(a)$ is given as follows by the product of the probabilities that satisfy Conditions 1–7

$$\alpha(a) = \frac{1}{2^n} \cdot \frac{f(a)+1}{2^n} \cdot \left(\frac{2^n-3}{2^n}\right)^{2^n-3}. \tag{14}$$

We can calculate the bias caused by the attack equation $\alpha(a) \cdot (2^n-1)/2^n$ by using Eq. (14). Figure 2 shows the theoretical and experimental values of the bias caused by the attack equation. The experimental values of the bias caused by the attack equation are obtained by testing the attack equation 2^{32} times. As the figure indicates, the theoretical values can approximate the experimental values.

3.3 Efficiency of the Attack

Experimental Results: We conduct an experiment to verify the efficiency of our attack. The targets are $1,000,000$ secret keys. The length of the session key is 128-bit (secret key length is 104-bit). The experiment was carried out under two conditions. Firstly, the attacker can obtain all the IVs except the weak IVs used in the FMS attack, and secondly, the first outputs of RC4 are known. Next, we observe the rate at which the secret keys with at least k bytes of secret key entries are recovered by the key recovery process of our attack. We list the experimental results in Table 1. The table shows the rate at which the secret keys are recovered, the time complexity involved in the recovery of all the secret key entries and the efficiency of our attack for each k. When at least k bytes of secret key entries are recovered, the time complexity is calculated as follows:

$$2^{3n} \cdot \frac{\lceil l/n \rceil - 3}{\lceil l/n \rceil} + {}_{\lceil l/n \rceil - 3}C_k \cdot 2^{(\lceil l/n \rceil - 3 - k)n}. \tag{15}$$

The number of IVs used in our attack is $2^{3n} \cdot (\lceil l/n \rceil - 3)/\lceil l/n \rceil$; this is indicated as the time complexity of the key recovery process. The number of IVs used

Table 1. The efficiency of our attack

k	complexity	experimental		statistical	
		rate	efficiency	rate	efficiency
1	$2^{99.7}$	$3.53018 \cdot 10^{-1}$	$2^{101.2}$	$3.54392 \cdot 10^{-1}$	$2^{101.2}$
2	$2^{94.3}$	$6.79340 \cdot 10^{-2}$	$2^{98.2}$	$6.70899 \cdot 10^{-2}$	$2^{98.2}$
3	$2^{88.2}$	$8.53100 \cdot 10^{-3}$	$2^{95.0}$	$8.08125 \cdot 10^{-3}$	$2^{95.1}$
4	$2^{81.5}$	$7.44000 \cdot 10^{-4}$	$2^{91.9}$	$6.74740 \cdot 10^{-4}$	$2^{92.0}$
5	$2^{74.3}$	$5.10000 \cdot 10^{-5}$	$2^{88.6}$	$4.09003 \cdot 10^{-5}$	$2^{88.9}$
6	$2^{66.7}$	$3.00000 \cdot 10^{-6}$	$2^{85.1}$	$1.84517 \cdot 10^{-6}$	$2^{85.8}$
7	$2^{58.7}$	—	—	$6.26163 \cdot 10^{-8}$	$2^{82.7}$
8	$2^{50.3}$	—	—	$1.59680 \cdot 10^{-9}$	$2^{79.6}$
9	$2^{41.5}$	—	—	$3.02045 \cdot 10^{-11}$	$2^{76.4}$
10	$2^{32.2}$	—	—	$4.11770 \cdot 10^{-13}$	$2^{73.3}$
11	$2^{24.2}$	—	—	$3.83026 \cdot 10^{-15}$	$2^{72.1}$
12	$2^{23.7}$	—	—	$2.17855 \cdot 10^{-17}$	$2^{79.1}$
13	$2^{23.7}$	—	—	$5.72148 \cdot 10^{-20}$	$2^{87.6}$

in our attack is discussed in Sect. 3.4. ${}_{\lceil l/n \rceil - 3}C_k \cdot 2^{(\lceil l/n \rceil - 3 - k) \cdot n}$ is indicated as the time complexity of the exhaustive key search process. The efficiency is the time complexity per the rate. If the efficiency of our attack is lower than the complexity of the exhaustive key search, WEP that avoid weak IVs used in the FMS attack is broken cryptographically. When $k = 6$, the efficiency of our attack is $2^{85.1}$. The efficiency for $k \geq 7$ is not obtained in this experiment because the targets are too few to obtain the rate at which secret keys are recovered for $k \geq 7$. For $k \geq 7$, the complexity of the simulation to obtain the efficiency is too high for the available computing power.

Statistical Analysis: We statistically obtain the efficiencies of the entire of k by using the experimental result. We can calculate the averages of the biases caused by weak IVs in each $K[a \bmod \lceil l/n \rceil]$ by using Eq. (14). When a 128-bit session key is used, the averages of the biases caused by the weak IVs are in the range of $2^{-13.66}$ to $2^{-13.84}$. The differences in the averages of the biases is sufficiently small as compared with the averages of the biases. Therefore, we assume that the success probability that a secret key is recovered by the key recovery process is the same for all secret key entries. From the experimental results, we obtain $p = 3.30985 \cdot 10^{-2}$ when a 128-bit session key is used, where p is the average of the success probability that a secret key entry is recovered. Then, the rate of the secret keys in which at least k secret key entries of $\lceil l/n \rceil - 3$ secret key entries are recovered by the key recovery process is as follows:

$$\sum_{x=k}^{\lceil l/n \rceil - 3} \left({}_{\lceil l/n \rceil - 3}C_x \cdot p^x \cdot (1-p)^{\lceil l/n \rceil - 3 - x} \right). \tag{16}$$

We calculate the efficiencies of the entire range of k by Eq. (16), as shown in Table 1. As Table 1 indicates, the efficiency of $k = 11$ of our attack is $2^{72.1}$.

This suggests that a 104-bit secret key can be recovered within realistically possible computational times by using our attack. Therefore, our attack poses a new threat to the FMS attack-resistant WEP implementation that use a 128-bit session key.

3.4 Rate of the Weak IVs

We discuss the rate of the weak IVs in our attack. The number of all IVs of WEP is equal to 2^{3n}; these are split equally by Eq. (7) into $\lceil l/n \rceil$ groups. However, three of these groups leak the information of IV entries $K[0]$, $K[1]$ and $K[2]$ by Eq. (8) instead of the information regarding secret key entries $K[3], K[4], \ldots, K[\lceil l/n \rceil - 1]$. It is not necessary to recover the IV entries. Therefore, the rate of weak IVs in our attack is $(\lceil l/n \rceil - 3)/\lceil l/n \rceil$. Further, the number of all weak IVs in our attack is equal to $2^{3n} \cdot (\lceil l/n \rceil - 3)/\lceil l/n \rceil$. When $l = 128$ and $n = 8$, the rate of the weak IVs is about 81%. Therefore, most of the IVs of WEP transform into weak IVs by our attack.

4 Conclusion

In this paper, we demonstrated an attack to break the FMS attack-resistant WEP implementation. Most IVs transform into weak IVs by our attack because our attack is a known IV attack that doesn't need the specific pattern of the IVs. If we attempt to avoid all the weak IVs used in our attack, the rate at which IVs are avoided is too large to use practical. When a 128-bit session key is used, the efficiency of our attack to recover a 104-bit secret key is $2^{72.1}$ in the most effective case. Therefore, our attack can recover the 104-bit secret key of the FMS attack-resistant WEP implementation within realistically possible computational times, that is, the FMS attack-resistant WEP implementation is insecure.

References

1. IEEE Computer Society, "Wireless Lan Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE Std 802.11, 1999 Edition.
2. S. Fluhrer, I. Mantin, and A. Shamir, "Weaknesses in the key scheduling algorithm of RC4," Proc. SAC2001, LNCS, vol.2259, pp.1–24, Springer-Verlag, 2001.
3. NIST, FIPS PUB 180-2 "Secure Hash Standard," Aug. 2002.
4. B. Schneier, Applied Cryptography, Wiley, New York, 1996.
5. Wi-Fi Alliance, "Wi-Fi Protected Access," available at `http://www.weca.net/opensection/protected_access.asp`.
6. A. Stubblefield, J. Loannidis and A. D. Rubin, "A key recovery attack on the 802.11b wired equivalent privacy protocol (WEP)," ACM Trans. Inf. Syst. Secur., vol.7, no.2, pp.319–332, May 2004.

A Details of $f(a)$

We introduce $g(t)$ in order to calculate $f(t)$. We define $g(t)$ as the number of entries that satisfy $S_t^*[y] = y$, for $y \neq 1, a-1$, or a, and $y \in \{i_t^*, i_t^*+1, \ldots, 2^n-1\}$. We carefully calculated the expected values of $f(t)$ and $g(t)$ for time t as shown in Fig. 1.

KSA$(t = 0, 1, 2, 3)$**:**

$$f(0) = 2^n - 3,\ g(0) = 2^n - 3, \tag{17}$$

$$f(1) = (f(0) - 2) \cdot \left(1 - \frac{1}{2^n - 3}\right) + f(0) \cdot \frac{1}{2^n - 3}, \tag{18}$$

$$g(1) = (g(0) - 2) \cdot \left(1 - \frac{1}{2^n - 3}\right) + (g(0) - 1) \cdot \frac{1}{2^n - 3}, \tag{19}$$

$$f(2) = f(1),\ g(2) = g(1). \tag{20}$$

KSA$(t = 3, 4, \ldots, a-1)$**:**

$$\begin{aligned} f(t) = {} & (f(t-1) - 2) \cdot X_1 \cdot X_2 \cdot \left(1 - \frac{1}{f(t-1)}\right) + \\ & (f(t-1) - 1) \cdot \Big(X_1 \cdot (1 - X_2) + (1 - X_1) \cdot X_2\Big) + \\ & f(t-1) \cdot \Big(X_1 \cdot X_2 \cdot \frac{1}{f(t-1)} + (1 - X_1) \cdot (1 - X_2) \cdot (1 - X_3)^2\Big) + \\ & (f(t-1) + 1) \cdot (1 - X_1) \cdot (1 - X_2) \cdot 2 \cdot X_3 \cdot (1 - X_3) + \\ & (f(t-1) + 2) \cdot (1 - X_1) \cdot (1 - X_2) \cdot {X_3}^2, \end{aligned} \tag{21}$$

$$\begin{aligned} g(t) = {} & (g(t-1) - 2) \cdot X_1 \cdot X_2 \cdot \left(1 - \frac{1}{f(t-1)}\right) \cdot \frac{g(t-1) - 1}{f(t-1) - 1} + \\ & (g(t-1) - 1) \cdot \Bigg(X_1 \cdot (1 - X_2) + (1 - X_1) \cdot X_2 \cdot \frac{g(t-1)}{f(t-1)} \\ & + X_1 \cdot X_2 \cdot \left(\frac{1}{f(t-1)} + \left(1 - \frac{1}{f(t-1)}\right) \cdot \left(1 - \frac{g(t-1) - 1}{f(t-1) - 1}\right)\right)\Bigg) + \\ & g(t-1) \cdot \Bigg((1 - X_1) \cdot X_2 \cdot \left(1 - \frac{g(t-1)}{f(t-1)}\right) + (1 - X_1) \cdot (1 - X_2) \\ & \cdot \Big({X_3}^2 \cdot (1 - X_4) + X_3 \cdot (1 - X_3) + X_3 \cdot (1 - X_3) \cdot (1 - X_4) + (1 - X_3)^2\Big)\Bigg) + \\ & (g(t-1) + 1) \cdot (1 - X_1) \cdot (1 - X_2) \cdot \Big({X_3}^2 \cdot X_4 + X_3 \cdot (1 - X_3) \cdot X_4\Big), \end{aligned} \tag{22}$$

where $X_1 = g(t-1)/(2^n - (t-1) - 2)$, $X_2 = f(t-1)/(2^n - 3)$, $X_3 = 1/(2^n - 3 - f(t-1))$ and $X_4 = (2^n - (t-1) - 2 - g(t-1))/(2^n - 3 - f(t-1))$.

KSA$(t = a)$**:**

$$f(a) = f(a-1),\ g(a) = g(a-1) - 1. \tag{23}$$

Design of a New Kind of Encryption Kernel Based on RSA Algorithm

Ping Dong[1], Xiangdong Shi[2], and Jiehui Yang[2]

[1] Information Engineer School, University of Science and Technology Beijing, Beijing 100083, China
dp55@sina.com
[2] Information Engineer School, University of Science and Technology Beijing, Beijing 100083, China
laurelyj@163.com

Abstract. Fast realization of RSA algorithm by hardware is a significant and challenging task. In this paper an ameliorative Montgomery algorithm that makes for hardware realization to actualize the RSA algorithm is proposed. This ameliorative algorithm avoids multiplication operation, which is easier for hardware realization. In the decryption and digital signature process, a combination of this ameliorative Montgomery algorithm and the Chinese remainder theorem is applied, which could quadruple the speed of the decryption and digital signature compared to the encryption. Furthermore, a new hardware model of the encryption kernel based on the ameliorative Montgomery is founded whose correctness and feasibility is validated by Verilog HDL in practice.

1 Introduction

RSA as a public key algorithm is recognized as a relative safer one and is also one of the most popular algorithms at present. In the RSA algorithm, different keys are used respectively for encryption and decryption. Public key (PK) is the information for encryption while correspondingly secret key (SK) is the information for decryption that can't be calculated directly from public key. The encryption algorithm (E) and the decryption algorithm (D) are also public, and the public key and the secret key are corresponding to each other. Senders use public key to encrypt the message and legal receivers use secret key to decrypt it.

2 RSA Algorithm

RSA algorithm is the most widely used public key system that can be applied on secrecy and digital signature. Both of the public key and secret key are the functions of two large primes (more than 100 decimal bits), and it is guessed that the difficulty of deducing the message from the key and secret message is equivalent to factoring the multiplication of two large primes.

In RSA algorithm, each encryption and decryption is corresponding to a public key (n, e) and secret key (n, d) where n is the modulus composed by two large primes p and q which are chosen randomly:

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 27 – 32, 2005.
© Springer-Verlag Berlin Heidelberg 2005

$$N = pd \tag{1}$$

and e is an exponent which satisfies:

$$GCD((p-1)(q-1), e) = 1 and 1 < e < (p-1)(q-1) \tag{2}$$

Secret key d is chosen like that:

$$d = e^{-1} \bmod (p-1)(q-1) \tag{3}$$

Basic mathematical operation used by RSA to encrypt a message M is modular exponentiation:

$$C = M^e \bmod n \tag{4}$$

That binary or general m-nary method can break into a series of modular multiplications.

Decryption is done by calculating:

$$M = C^d \bmod n \tag{5}$$

We can see that the kernel of RSA algorithm is modular exponentiation operation of large number.

3 Modular Exponentiation Algorithm Realizations

As the modular exponentiation operation of large number is the difficulty of RSA algorithm, therefore, predigesting the modular exponentiation operation can speed up the RSA algorithm obviously.

3.1 Binary Square Multiplication Rapid Algorithm

This rapid algorithm scans from the lowest bin to the highest bin and uses a square exponent for assistant space storage of M. The algorithm to take the operation of $C = M^e \bmod n$ is as follows:

First express e into binary form that is:

$$E = \sum_{J=0}^{N-1} e_j 2^j \qquad (e_j = 1 \text{ or } 0) \tag{6}$$

Then make $R_0 = 1$ and $d_0 = M$, and take the following circulation operation:

```
For (i = 0; i < n; i ++)
  {d_{i+1} = d_i^2 mod N;
    if (e_i = = 1)  R_{i+1} R_i * d_i mod N;
    else R_{i+1} = R_i}.
```

Where R_n is the value of C.

The complexity of this algorithm is the sum of square of the nth modulus that is the (n-1)/2 –th modular exponentiation on average. However, because executive state-

ments have no data correlation so they can be processed parallel. If adding a modular multiplication unit, the algorithm speed can be doubled whereas the hardware scale will also be doubled.

3.2 No Final Subtractions Radix-2 Montgomery Algorithm

The original Montgomery algorithm only gives the rapid algorithm for calculating MR-1 mod N when $0 < M < RN$. However, in RSA algorithm modular multiplications in the form of $Z = AB \bmod N$ are often needed and in the hardware realization the value of R is generally 2 which may not satisfy $0 < M < RN$. Furthermore, the original Montgomery algorithm has to judge whether t is larger than n or not in the end, and if $t > n$ the final subtraction operation has to implemented, which needs more chip area cost. So the original Montgomery algorithm can't be used directly and has to been modified.

Actually, the final subtractions can be avoided by adding a loop in the modular multiplication circulation. On the assumption that $N<R^S$ and GCD (N , R) = 1 (because of R = 2, N can meet the condition just when it is an odd number) and $0 < A, B < N$ (where both A and B are integer), we make $A = (a_{n-1}, a_{n-2}\ldots a_1, a_0)$, $B = (b_{n-1}, b_{n-2}\ldots b_1, b_0)$, $N = (n_{n-1}, n_{n-2}\ldots n_1, n_0)$. Then the function of $Z = M (A, B) = ABR^{-(s+1)} \bmod N$ is as follows:

```
Function M (A, B)
      Z0 = 0
      For (i = 0, i < s + 1, i ++)
        {Pi = (Zi + ai * b0) mod R;
          Zi+1 = Zi + ai * B + Pi * N;
          Zi+1 = Zi+1 / R}.
```

Where the last circulation is an empty loop $a_S = 0$ and $R^{-(s+1)}$ can be calculated by extended Euclidean algorithm. However, $R^{-(s+1)}$ has not to be calculated in RSA algorithm, because its infection can be eliminated through the whole algorithm construction.

3.3 Constructing Modular Exponentiation Algorithm Based on No Final Subtraction Montgomery Algorithm and the Square Multiplication

According to the results discussed above, the RSA algorithm to calculate $C = M^e \bmod N$ (where $N<2s$) is given below based on no final subtraction Montgomery algorithm and the square multiplication:

```
e = [es-1 es-2...e1 e0]2
  Z = 2s+1 * 2s+1 mod N; R0 = 1, d0 = m;
   For (i = 0; i < n; i ++)
        {Pi = M (di, Z);
          di+1 = M (di, Pi);
          if (ei == 1)
           Ri+1 = M (Ri, Pi);
          else Ri+1 = Ri}.
```

Where R is the value of C.

3.4 Brief Introduction of Chinese Remainder Theorem

Chinese Remainder theorem can be used to accelerate RSA decryption and digital signature. The receiver who us RSA to decrypt calculates $M=C^d \bmod N$ where $N=PQ$ (P and Q are two large binary primes whose lengths are close). Because the receiver knows the P and Q, so the operation can be separated as two parts:

$$M \equiv M_1 \bmod P$$
$$M \equiv M_2 \bmod Q \tag{7}$$

Where $M_1 = C_1^{d1} \bmod P$, $M_2 = C_2^{d2} \bmod Q$ and $C_1 = C \bmod P$, $C_2 = C \bmod Q$, $d_1 = d \bmod (P - 1)$ and $d_2 = d \bmod (Q- 1)$.

4 Design of Encryption Kernel

In RSA algorithm, the kernel operation in encryption and decryption is modular multiplication of large numbers, which is also the most computational one. The design of encryption kernel bases on radix-2 Montgomery to construct the modular multiplication and uses square multiplication to form the modular exponentiation. This method uses two Montgomery operation modules and other basic control modules to realize the RSA encryption and decryption. Because of small hardware area it takes, the speed can be improved obviously, especially for decrypting with Chinese Remainder Theorem, which makes the speed of RSA signature decryption faster.

4.1 Design of Montgomery Multiplier Module

Montgomery multiplier module (brief in MM module) is designed based on the modification Montgomery algorithm for calculating $ABR^{-(s+1)} \bmod N$ and its structure is illustrated in Fig 1:

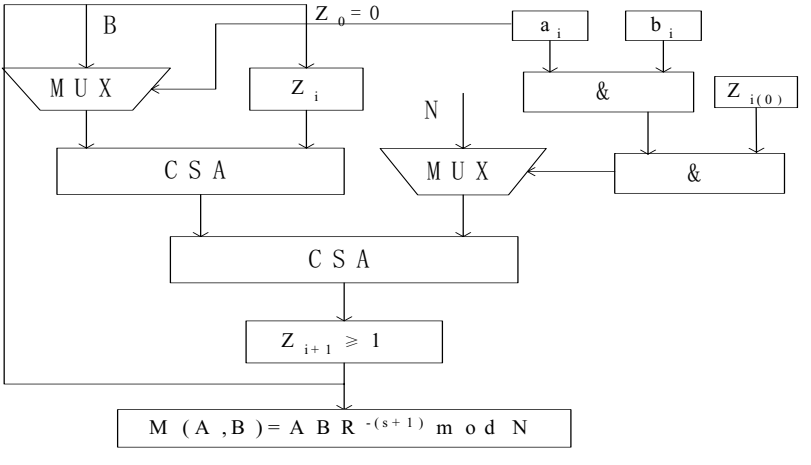


Fig. 1. The structure of Montgomery multiplier module

4.2 Design of Modular Exponentiation Module

The method that uses MM module to construct square multiplication modular exponentiation is showed in Fig 2:

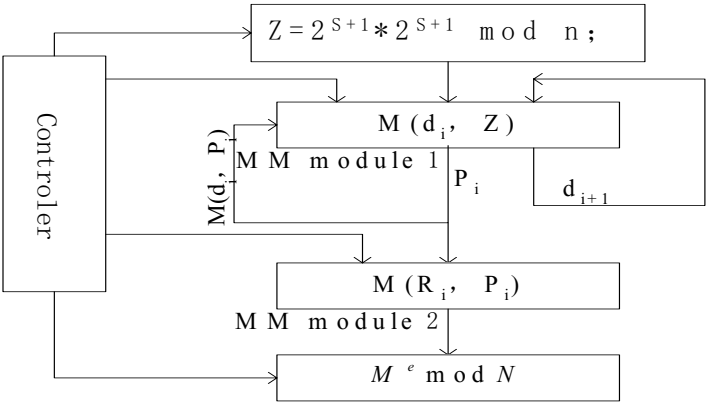


Fig. 2. Design of modular exponentiation module

Where Z is pre-computed. Two MM modules are used for parallel calculating the two-step modular exponentiation, which can save lots of computing time. MM module I takes charge of twice modular multiplication operation, and because $P_i = M(d_i, Z)$ and $d_{i+1} = M(d_i, P_i)$ are correlative to each other, which cannot be calculated at the same time so MM module I is reused. The circulation times and inputs are determined by the controller, and the result of modular exponentiation will be outputted at the end of the last circulation. The same modular exponentiation module is also used in decryption and signature with Chinese Remainder Theorem, however, whose bit number of exponent is much smaller, which can reduce circulation loops. Design the control circuit reasonably could improve the efficiency of decryption and signature greatly.

5 Conclusions

The paper introduces RSA algorithm systematically and analyzes the improvement of Montgomery algorithm. Combined with square multiplication approach, the improved Montgomery algorithm increases the speed of modular multiplication algorithm, simplifies the structures of hardware and saves the area of chips. At the same time, we analyze the use of Chinese Remainder Theorem in decryption and digital signature of RSA, which can improve the speed better. Finally, we put a design approach of 1024-bit RSA encryption kernel. The encryption kernel takes both operation speed and chip area into consideration and is a good way to process long-bit RSA algorithm.

References

1. Alexandre, F., Tenca.: A Scalable Architecture for Modular Multiplication Based on Montgomery's Algorithm. IEEE Trans. Computers, Vol. 52, No. 9 (2003) 1215–1220.
2. Stephen E. Eldridge and Colin D. Walter.: Hardware Implementation of Montgomery's Modular Multiplication Algorithm. IEEE Trans. Computers, Vol. 42, No. 6 (1993) 693 – 699.
3. Montgomery, P. L.: Modular multiplication without trial division. Math. Computation, Vol. 44, n 170, (1985) 519-521
4. Kaihara, Marcelo E., Takagi, Naofumi: A hardware algorithm for modular multiplication/division. IEEE Trans. Computers, Vol. 54, No. 1 (2005) 12-21

On the Security of Condorcet Electronic Voting Scheme

Yoon Cheol Lee[1] and Hiroshi Doi[2]

[1] Department of Information and System Engineering,
Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, Japan
lee@chao.ise.chuo-u.ac.jp
[2] Institute of Information Security,
2-14-1 Tsuruya-cho Kanagawa-ku, Yokohama, Japan
doi@iisec.ac.jp

Abstract. In this paper, we focus on the Condorcet voting scheme in which each voter votes with the full order of the candidates according to preference, and the result of the election is determined by one-on-one comparisons between each candidate. We propose the Condorcet electronic voting scheme that is secure, universally verifiable and satisfying one-on-one comparison privacy. Furthermore the result of the election can be determined without revealing the order of the candidates which each voter specified. We use a matrix to represent the order of all the candidates according to preference, and satisfy one-on-one comparison privacy using homomorphic property.

1 Introduction

An electronic voting scheme is one of the most significant application of cryptographic protocol. This has been studied for over the last twenty years, and many voting schemes have been proposed. There are many different methods of voting and tallying votes. The efficient voting protocols can be categorized by their approaches into three types using blind signatures [11, 16], using homomorphic encryption [2, 3, 13], and using mix-net [4, 10, 12]. The suitability of each of these types varies with the conditions under which it is to be applied. In national presidential elections, each voter votes only for one candidate among the candidates and the top candidate is elected. This is called a plurality voting scheme. Instead of the plurality voting scheme, it is convenient to specify full order of all the candidates according to voter's preference [17]. Such an expansion of voting rights would allow voters to specify their preference.

In this paper, we focus on the Condorcet voting scheme in which each voter votes with the full order of the candidates according to preference, and the result of the election is determined by one-on-one comparisons between each candidate [7, 9, 14]. The idea behind the Condorcet voting scheme is for each voter to specify their preference among the candidates, and then for each pair of candidates to consider how many times one candidate in the pair is preferred to the other. The Condorcet voting scheme is one of the fairest scheme which would come

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 33–42, 2005.
© Springer-Verlag Berlin Heidelberg 2005

closer to an accurate representation of the voters' wishes than any other scheme does [6]. Based on this idea, we propose the Condorcet electronic voting scheme that is secure, universally verifiable and satisfy one-on-one comparison privacy. Furthermore the result of the election can be determined without revealing the order of the candidates which each voter specified. Instead of applying mix-net based scheme simply, we use the pairwise matrix to represent the order of all the candidates according to preference as the ballots of the voters collectively, and satisfy one-on-one comparison privacy using homomorphic property.

2 Overview of the Condorcet Voting

In this section, we describe the Condorcet voting scheme and the security requirements.

2.1 Condorcet Voting

In most of the current plurality voting schemes, it is not always easy to reflect on the preference of the voters exactly. That is far better than no choice at all, of course, but it is nowhere near as good as also being allowed to specify a second and third choice, or beyond. The Condorcet voting scheme was proposed by a French philosopher, mathematician, and early political scientist, Marie Jean Antoine Nicolas Caritat Marquis de Condorcet (1743 $\sim$ 1794) as an improvement to the plurality method for elections. The voters specify the order of the candidates according to preference. The one-on-one comparison then takes into account preference of each voter for one candidate over another. If one of the candidates is preferred to all the candidates in their one-on-one comparisons, then that candidate is elected in the election.

Example 1. *Suppose there are four candidates C_1, C_2, C_3 and C_4 in the election. We assume that the each voter in total 18 voters votes by specifying the full order of the all candidates according to preference as Table 1.* (2) *in Table 1 means that 8 ballots are cast for the order of the candidates according to preference s.t. first C_2, in succession C_3, C_1 and C_4. C_2 is preferred with 14 votes than C_1 and 2 votes than C_3 and C_4. Since C_2 is preferred to all other candidates, C_2 is elected in this election.*

Table 1. Order of Preference

Sequence Number	Ballot	Number of Votes
(1)	C_1 , C_2 , C_3 , C_4	2
(2)	C_2 , C_3 , C_1 , C_4	8
(3)	C_3 , C_4 , C_2 , C_1	7
(4)	C_4 , C_3 , C_2 , C_1	1

2.2 Security Requirements

In this subsection, we describe some requirements for the secure Condorcet electronic voting as bellows. We say that the Condorcet electronic voting scheme is secure if the scheme satisfies the following requirements.

$[R_1]$ **Completeness:** All valid votes are counted correctly.
$[R_2]$ **Soundness:** No dishonest voter can disrupt the voting.
$[R_3]$ **Privacy:** Nobody can figure out who votes for whom.
$[R_4]$ **Unreusability:** No voter can vote more than twice.
$[R_5]$ **Eligibility:** Only legitimated voter can vote.
$[R_6]$ **Robustness:** Voting system can be performed successfully regardless of partial failure.
$[R_7]$ **Universal verifiability:** Correctness of the result is verifiable for any verifiers.
$[R_8]$ **One-on-one comparison privacy:** Any information of the order of all the candidates according to the voter's preference except the result of the election according to one-on-one comparisons are concealed.

3 Building Blocks

Robust threshold ElGamal. We use the robust threshold ElGamal cryptosystem as the basis for our constructions. It is easy to take robustness into account from ElGamal cryptosystem. The object of a threshold scheme is to share a secret key among a set of receivers such that plaintexts can only be decrypted when a substantial set of receivers cooperate. Let g be a generator of cyclic group G of prime order q, and the secret key s is distributed among authorities. The public key $h = g^s$ is published to the participants in the system. An encryption of a vote m is given by $(x, y) = (g^\alpha, g^m h^\alpha)$, where $\alpha \in {}_R Z_q$ is a random number. An ElGamal encryption scheme is homomorphic. Given an ElGamal encryption (x_1, y_1) of m_1, (x_2, y_2) of m_2, $(x_1 x_2, y_1 y_2)$ is an ElGamal encryption of $m_1 + m_2$ by homomorphic property. To decrypt a ciphertext $(x_1 x_2, y_1 y_2)$, the authorities recovers $g^{m_1+m_2}$ jointly without explicitly reconstructing the secret key along with a proof of correct exponentiation [3, 5].

Mix and Match. By using the mix and match techniques [15] constructed with the ElGamal cryptosystem [8], one can determine whether $T \geq 0$, where T is an encrypted value and $-N \leq T \leq N$. When one examines whether $T \geq 0$, there exists one plaintext 1 in the decrypted $D[E(g^T)/E(g^\lambda)]$, $0 \leq \lambda \leq N$, and $E(g^T)/E(g^\lambda) = E(g^{T-\lambda})$ by homomorphic property.

1-out-of-L re-encryption proof. Given encryption E and a batch of re-encryption, $\hat{E}_1, \hat{E}_2, \ldots, \hat{E}_L$, and a witness that $\hat{E}_i$ is a re-encryption of E, one can prove $\hat{E}_i$ is indeed re-encryption of E without revealing i. This proof is called 1-out-of-L re-encryption proof [13]. This can be constructed with the ElGamal cryptosystem.

4 Proposed Voting Scheme

In this section, we construct our secure and universally verifiable Condorcet electronic voting scheme. We describe the model, the representation of votes and the proposed Condorcet voting scheme.

4.1 Model

Our voting scheme is composed of N voters $V_1, \ldots, V_N$, L candidates $C_1, \ldots, C_L$, a bulletin board, a tallying authority T, and the multiple judging authorities J_k. To simplify, N is assumed to be odd. The communication model required for our voting scheme is best viewed as a public broadcast channel with memory. This is called a bulletin board. No party can erase any information from the bulletin board, but each active participant can append messages to its own designated section. The tallying authority T is responsible for collecting ballots from the authorized voters, multiplying, permutating and re-encrypting the ballots. The judging authorities J_k are responsible for keeping their secret keys private, decrypting and figuring out the result of the election satisfying one-on-one comparison privacy using the techniques of [15].

4.2 Representation of Votes

The method of pairwise comparisons revolves around one-on-one comparisons. We need a (L, L) matrix to represent all possible preference, and $m_{ij} = 1$ means a candidate C_i is preferred to a candidate C_j and $m_{ji} = -1$. The one-on-one comparison matrix for each vote of Figure 1 are summed to get the tallied one-on-one comparison matrix of Figure 2. In Figure 1 and 2, the matrix consists of the difference of that times each row was preferred to each column and opposite, and "$\perp$" means the diagonal components that does not reflect on tallying. The tallied one-on-one comparison matrix of Figure 2 is used to determine the result of the election. Now we look for a row which is not negative signs in it. That row is for C_2. What this means is that in one-on-one comparison between C_2 and each other candidates, C_2 prefers that comparisons. So, C_2 is elected in the election.

$$M = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \begin{pmatrix} \perp & -1 & -1 & 1 \\ 1 & \perp & 1 & 1 \\ 1 & -1 & \perp & 1 \\ -1 & -1 & -1 & \perp \end{pmatrix} \end{array}$$

Fig. 1. One-on-one comparison matrix for (2) in Table 1

$$T = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \begin{pmatrix} \perp & -14 & -14 & 2 \\ 14 & \perp & 2 & 2 \\ 14 & -2 & \perp & 16 \\ -2 & -2 & -16 & \perp \end{pmatrix} \end{array}$$

Fig. 2. Tallied one-on-one comparison matrix

Definition 1. (*One-on-one comparison matrix*) *With the Condorcet voting scheme, each voter can specify the order of the candidates according to pref-*

erence. It can be represented as a (L, L) *matrix for all the order of preference when the ballots are calculated. It is defined by one-on-one comparison matrix with special symbol* $\perp$ *in diagonal. The* m_{ij} *is filled up* 1 *when the candidate* C_i *is preferred to* C_j *(*$C_i > C_j$ *in short), and* m_{ji} *is filled up* -1 *at the same time. That is,*

$$m_{ij} = 1 \Leftrightarrow C_i > C_j \text{ and } m_{ij} = -1 \Leftrightarrow C_i < C_j.$$

We will denote the one-on-one comparison matrix by Ω.

As described above, the matrix in Figure 1 is a one-on-one comparison matrix for (2) of Table 1. Furthermore, the matrix of Figure 2 is a tallied one-on-one comparison matrix for Figure 1. Note that $\perp \in_R Z_q$, $| \perp | > N$, if it is assumed for us to put in the concrete number.

Definition 2. (*Basic one-on-one comparison matrix*) *We construct a* (L, L) *matrix composed* m_{ij} $(i < j) = 1$ *and* m_{ij} $(i > j) = -1$ *with* $\perp$ *in diagonal. We call this matrix as a basic one-on-one comparison matrix* B.

$$B = \begin{pmatrix} \perp & 1 & 1 & \cdots & 1 \\ -1 & \perp & 1 & \cdots & 1 \\ \vdots & \ldots & \ddots & \ldots & \vdots \\ -1 & \cdots & -1 & \perp & 1 \\ -1 & \cdots & -1 & -1 & \perp \end{pmatrix}$$

Fig. 3. Basic one-on-one comparison matrix

Definition 3. (*Condorcet voting matrix*) *Using permutation matrix* P, *each voter permutates to represent his votes s.t.* $P^T BP$. *We call this constructed matrix as a Condorcet voting matrix* Θ.

Definition 4. (*Preference matrix*) *It is defined by the Condorcet voting matrix* Θ *s.t.* $m_{ij} = 1 \Leftrightarrow C_i > C_j$ *and* $m_{ij} = -1 \Leftrightarrow C_i < C_j$, *and represent the set of all order of the* L *candidates* C_i, $1 \leq i \leq L$ *according to preference. We denote this preference matrix by* Ψ.

Note that any Condorcet voting matrix $\Theta \in \{\Theta_\upsilon\}$ is one to one corresponding to the preference matrix $\Psi \in \{\Psi_\ell\}$, $1 \leq \upsilon, \ell \leq L!$.

Theorem 1. *For any* $A \in \{\Omega_k\}$, $1 \leq k \leq L!$, *and for any permutation matrix* P, $\tilde{A} = P^T AP$ *satisfies* $\tilde{a}_{ii} = \perp$ *for all* i.

More preciously, $\tilde{a}_{ij} = a_{\sigma^{-1}(i)\sigma^{-1}(j)}$, *(i.e.* $a_{ij} = \tilde{a}_{\sigma(i)\sigma(j)}$*),* σ *is the permutation,* $P = (P_{\sigma(i)})$ *is the permutation matrix, where* $P_{\sigma(i)}$ *means* $\sigma(i)$*-th elment is* 1 *and the other elements are* 0.

(*Proof*).

$$P^T A = ({P^T}_{\sigma(1)}\ {P^T}_{\sigma(2)} \cdots\ {P^T}_{\sigma(L)}) \begin{pmatrix} A_1 \\ \vdots \\ A_L \end{pmatrix} = {P^T}_{\sigma(1)} A_1 + \ldots + {P^T}_{\sigma(L)} A_L$$

$$= \begin{pmatrix} A_{\sigma^{-1}(1)} \\ A_{\sigma^{-1}(2)} \\ \vdots \\ A_{\sigma^{-1}(L)} \end{pmatrix} = \begin{pmatrix} a_{\sigma^{-1}(1)1} & a_{\sigma^{-1}(1)2} & \cdots & a_{\sigma^{-1}(1)L} \\ a_{\sigma^{-1}(2)1} & a_{\sigma^{-1}(2)2} & \cdots & a_{\sigma^{-1}(2)L} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\sigma^{-1}(L)1} & a_{\sigma^{-1}(L)2} & \cdots & a_{\sigma^{-1}(L)L} \end{pmatrix} = \acute{A} = (\acute{A}_1 \ \acute{A}_2 \cdots \acute{A}_L)$$

$$\tilde{A} = \acute{A}P = (\acute{A}_1 \ \acute{A}_2 \cdots \acute{A}_L) \begin{pmatrix} P_{\sigma(1)} \\ \vdots \\ P_{\sigma(L)} \end{pmatrix} = \acute{A}_1 P_{\sigma(1)} + \ldots + \acute{A}_1 P_{\sigma(1)}$$

$$= \begin{pmatrix} \acute{A}_{\sigma^{-1}(1)} \\ \acute{A}_{\sigma^{-1}(2)} \\ \vdots \\ \acute{A}_{\sigma^{-1}(L)} \end{pmatrix} = \begin{pmatrix} a_{\sigma^{-1}(1)\sigma^{-1}(1)} & a_{\sigma^{-1}(1)\sigma^{-1}(2)} & \cdots & a_{\sigma^{-1}(1)\sigma^{-1}(L)} \\ a_{\sigma^{-1}(2)\sigma^{-1}(1)} & a_{\sigma^{-1}(2)\sigma^{-1}(2)} & \cdots & a_{\sigma^{-1}(2)\sigma^{-1}(L)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{\sigma^{-1}(L)\sigma^{-1}(1)} & a_{\sigma^{-1}(L)\sigma^{-1}(2)} & \cdots & a_{\sigma^{-1}(L)\sigma^{-1}(L)} \end{pmatrix}$$

Theorem 2. *For any $F \in \{\Psi_\ell\}$, there exists a permutation matrix P s.t., $F = P^T B P$.*

(*Proof*). We know that F is one to one corresponding to Θ. So it is clear from Theorem 1.

Theorem 3. *For any P, $P^T B P \in \{\Psi_\ell\}$.*

(*Proof*). It is clear from Theorem 1.

Theorem 4. *For any $P, \acute{P}$, $M = P^T B \acute{P}$ and $m_{ii} = \perp$ for all i if and only if $P = \acute{P}$.*

(*Proof*). ($\Leftarrow$) is trivial.
($\Rightarrow$) Let $P \neq \acute{P}$. We assume that the permutation σ and τ are corresponding to P and $\acute{P}$ respectively. By Theorem 1, $m_{ii} = \tilde{a}_{\sigma^{-1}(i)\tau^{-1}(i)}$. Since $\sigma \neq \tau$, there exists i s.t., $\sigma^{-1}(i) \neq \tau^1(i)$ and $m_{ii} \neq \perp$. This is a contradiction. Therefore, $P = \acute{P}$.

4.3 Overview of Our Voting Scheme

We present an overview of our voting scheme as follows. We do not consider the votes that end in a tie, but only preferring or not preferring in this paper.

Each voter specifies the order of the candidates according to preference, and encrypts the basic one-on-one comparison matrix B using an ElGamal cryptosystem [8] and permutates. Each voter casts the calculated ciphertexts as the ballots. The tallying authority calculates the product and computes N ciphertexts $C_{ij\lambda} = (C_{ij1}, \ldots, C_{ijN})$. Then, the tallying authority permutates and re-encrypts $C_{ij\lambda}$, and achieves $\hat{C}_{ij\lambda} = (\hat{C}_{ij1}, \ldots, \hat{C}_{ijN})$. Each judging authority decrypts the ciphertexts $\hat{C}_{ij\lambda}$ jointly without explicitly reconstructing the secret key, then examines whether there exists one plaintext 1 in the decrypted $D(\hat{C}_{ij\lambda})$ to determine the result of the election.

4.4 Protocol

Set Up. The number p and q are prime numbers satisfying $p = \ell q + 1$ for some integer ℓ. The judging authorities J_k will randomly choose s_k as his secret key and corresponding public key $h_k = g^{s_k}$. Each J_k need to prove the knowledge of the secret key. The common public key is $h(= h^s \bmod p)$ [3]. The T publishes a basic one-on-one comparison matrix B.

Voting. Each voter encrypts all the components of the published B using an ElGamal cryptosystem [8] specifying the order of the candidates according to preference. The encrypted cipertexts are following as bellows.

$(x_{ij}, y_{ij}) = (g^{\alpha_{ij}}, g^{b_{ij}} h^{\alpha_{ij}})$, where α_{ij} is chosen randomly by the voter, $b_{ij(j \neq i)}$ is selected from B and $1 \leq i, j \leq L$. Each voter transposes the encrypted ballots $P^T E(B)$, where P is permutation matrix and E is an ElGamal encryption function. Then, each voter computes his ballots by permutating the whole row and the whole column s.t. $E(P^T E(B))P$. The achieved ciphertexts are the components of the one-on-one comparison matrix Ω. Then each voter casts the ciphertexts $E(P^T E(B))P$ as the ballots along with the proof of 1-out-of-L re-encryption of vector for the whole row and the whole column, and the proof for which the ballots in diagonal are composed with $E(\beta)$, where β is the number in diagonal s.t. $\beta \notin \left\{g^1, \cdots, g^N\right\}$ which does not reflect on the result of the election.

Each prover must prove that for an encrypted vote (x_{ij}, y_{ij}), there is a re-encryption in the $L \times L$ encrypted votes with suitable permutation. Assume that $(\hat{x}_{t\ell}, \hat{y}_{t\ell}) = (g^{\xi_{t\ell}} x_{i\ell}, h^{\xi_{t\ell}} y_{i\ell})$ is a re-encryption of $(x_{i\ell}, y_{i\ell})$ for $1 \leq \ell \leq L$. The prover can execute the proof of 1-out-of-L re-encryption of vector described in Figure 4. If the prover executes the proof of 1-out-of-L re-encryption of vector of Figure 4 in L times, the prover can prove for the permutation and re-encryption of the whole row. Furthermore, the prover must execute the same proofs for the whole column. Finally, the prover need prove that it is composed with β in diagonal. Assume that the ciphertext is $(x, y) = (g^r, h^r \beta) = E(\beta)$, where $r \in Z_q$ is a random number. It suffices to prove the knowledge of r satisfying $(\frac{y}{x}\beta^{-1}) = (\frac{h}{g})^r$.

Tallying. T calculates the product of each component in the matrix for the cast ballots from the voters. To examine whether $T_{ij} \geq 0$, where T_{ij} is the total number of votes of the encrypted m_{ij} in the upper (or lower) triangular half, T computes N ciphertexts $C_{ij\lambda}$ as follows.

$C_{ij\lambda} = m_{ij}/E(g^\lambda) = E(g^{T_{ij}})/E(g^\lambda)$,
where $1 \leq \lambda \leq N$, $1 \leq i, j \leq L$, $-N \leq T_{ij} \leq N$.

Then, T permutates and re-encrypts the vector $(C_{ij1}, \ldots, C_{ijN})$ of the ciphertext $C_{ij\lambda}$ to get the vector $(\hat{C}_{ij1}, \ldots, \hat{C}_{ijN})$ of the ciphertext $\hat{C}_{ij\lambda}$. T posts the encrypted vector $\hat{C}_{ij\lambda} = (\hat{C}_{ij1}, \ldots, \hat{C}_{ijN})$ as the tallied ballots. It is enough only to examine the ciphertexts of the upper (or lower) triangular half for determining the result of this election.

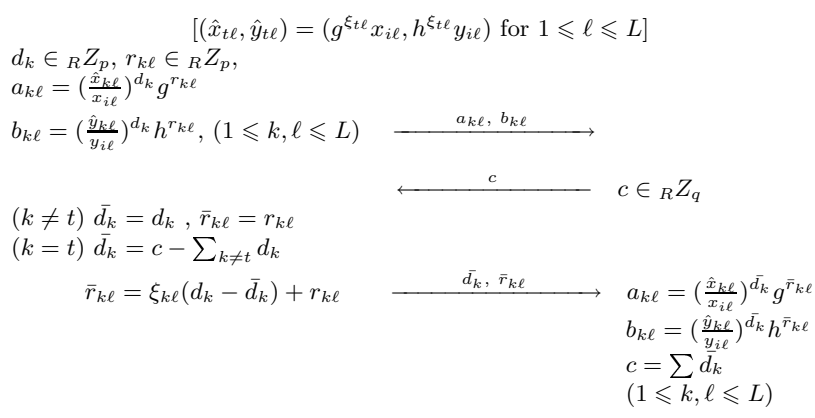


Fig. 4. Proof of 1-out-of L re-encryption of vector

Judging. J_k first decrypts partially the ciphertexts $\hat{C}_{ij\lambda} = (\hat{C}_{ij1}, \ldots, \hat{C}_{ijN})$ to determine whether $T_{ij} \geq 0$ jointly without explicitly reconstructing the secret key. By examining whether $T_{ij} \geq 0$ or not, the judging authorities can only decide whether candidate i is preferred to candidate j or not. The decrypted values tells nothing about the total number of votes T_{ij}, look random and their relations to the T_{ij} are hidden. Then, they examine the decrypted ciphertexts to determine the result of this election using the techniques of [15] as follows.

The total number of votes T_{ij} of the encrypted component m_{ij} have to satisfy $-N \leq T_{ij} \leq N$. Therefore, the problem of examining whether $T_{ij} \geq 0$ is reduced to examine whether $D(E(g^{T_{ij}})) \in \{g^1, \ldots, g^N\}$. To determine whether $T_{ij} \geq 0$, the judging authorities have to check whether there exists one plaintext 1 in the decrypted $D[E(g^{T_{ij}})/E(g^{\lambda})]$. If there exists one plaintext 1, it is concluded that $T_{ij} \geq 0$ that is, candidate i is preferred to candidate j; if there is no plaintext 1, it is concluded that $T_{ij} < 0$ that is, candidate j is preferred to candidate i.

5 Evaluation of Security and Efficiency

The proposed Condorcet electronic voting scheme satisfies in the sense of the security requirements in subsection 2.2. $[R_1]$ is guaranteed since all votes are published on the bulletin board. This assures that all valid votes are counted correctly. $[R_2]$ is satisfied, since the disruption which is for the voter to keep casting invalid ballots can be detected. $[R_3]$ is preserved, since it is infeasible to figure out between the encrypted votes and the permutated and re-encrypted votes. $[R_4]$ is satisfied, since each voter can vote only once on the bulletin board. $[R_5]$ is satisfied, since only the legitimate voters registered on the bulletin board can participate in the election. $[R_6]$ is guaranteed since the threshold ElGamal cryptosystem is used. Since each encrypted ballot cast on the bulletin board must be accompanied with the proofs, its validity can be verified with its proofs. $[R_7]$ is satisfied due to the use of a homomorphic property of the ballots and

the proofs. The tallied ciphertexts can be obtained and verified by any observer against the product for the cast ballots in the upper and lower triangular. Using the techniques of [15], any information of the order of preference except the result of the election according to one-on-one comparisons are concealed. This ensures $[R_8]$. The size of vote is $O(L^2)$, and the size of proof is $O(L^3)$. The computational cost of voting, verifying and mix and match is $O(L^3)$, $O(L^3)$ and $O(NL^2)$ respectively.

Although the Condorcet electronic voting scheme that is secure, universally verifiable and satisfying one-on-one privacy is proposed for the first time in this paper, it can also be realized by using mix-net or blind signature. In that case, the computational cost of voter and the size of data become small. However, $[R_8]$ of security requirements is not satisfied.

6 Conclusion

In this paper, we proposed the Condorcet electronic voting scheme that is secure, universally verifiable and satisfying one-on-one comparison privacy. Furthermore only the result of the election can be determined without revealing the order of the candidates which the voters specified. In mix-net based scheme, the final tally is computed and how many votes the candidate in row received over the candidate in column is revealed. While in the proposed scheme we used the pairwise matrix to represent the order of all the candidates according to preference as the ballots of the voters collectively, and satisfied one-on-one comparison privacy using homomorphic property.

References

1. J. Benaloh, "Verifiable Secret-Ballot Elections," PhD thesis, Yale University, Department of Computer Science Department, New Haven, CT, September, 1987.
2. R. Cramer, M. Franklin, B. Schoenmakers, and M.Yung, "Multi-authority secret ballot elections with linear work," EUROCRYPTO'96, LNCS 1070, pp.72-83, Springer-Verlag, 1996.
3. R. Cramer, R. Gennaro, and B. Schoenmakers, "A secure and optimally efficient multi-authority election scheme," EUROCRYPTO'97, LNCS 1233, pp.103-118, Springer-Verlag, 1997.
4. D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," Communications of the ACM, 24(2):84-88, 1981.
5. Y. Desmedt, "Threshold cryptography," European Transactions on Telecommunications, 5(4), pp.449-457, 1994.
6. P. Dasgupta, E. Maskin, "The Fairest Vote of all," SCIENTIFIC AMERICAN, vol.290, pp.64-69, March, 2004.
7. B. Duncan, "The Theory of Committees and Election," Cambridge University Press, JF1001, B49, ISBN 0-89838-189-4, 1986.
8. T. ElGamal, "A public-key cryptosystem and a signature scheme based on discrete logarithms," IEEE Transactions on Information Theory, IT-31(4):469-472, 1985.
9. P. Fishburn, "Condorcet Social Choice Functions," SIAM Journal of Applied Mathematics, vol.33, pp.469-489, 1977.

10. J. Furukawa, H. Miyauchi, K. Mori, S. Obana and K. Sako, "An Implementation of a Universally Verifiable Electronic Voting Scheme based on Shuffling," Financial Cryptography, LNCS 2357, pp.16-30, Springer-Verlag, 2002.
11. A. Fujioka, T. Okamoto and K. Ohta, "A Practical secret Voting Scheme for Large Scale Elections," AUSCRYPTO'92, pp.244-251, Springer-Verlag, 1992.
12. J. Furukawa and K. Sako, "An Efficient Scheme for Proving an Shuffle," CRYPTO'01, LNCS 2139, pp.368-387, Springer-Verlag, 2001.
13. M. Hirt and K. Sako, "Efficient receipt-free voting based on homomorphic encryption," EUROCRYPTO'00, LNCS 1807, pp.539-556, Springer-Verlag, 2000.
14. M. Iain and H. Fiona, "Condorcet:Foundations of Social Choice and Political Theory," Nuffield College, Edward Elgar Publishing Limited, 1994.
15. M. Jakobsson and A. Juels, "Mix and Match:Secure Function Evaluation via Ciphertexts," ASIACRYPTO'00, LNCS 1976, pp.162-177, Springer-Verlag, 1998.
16. T. Okamoto, "Receipt-free electronic voting schemes for large scale elections," Workshop on Security Protocols, LNCS 1361, pp.25-35, Springer-Verlag, 1997.
17. A. Taylor, "Mathematics and politics: strategy, voting, power and proof, " Springer-Verlag New York, Inc., ISBN 0-387-94500-8, 1995.
18. C. J. Wang and H. F. Leung, "A Secure and Fully Private Borda Voting Protocol with Universal Verifiability," 28th COMPSAC, vol.1, pp.224-229, 2004.

Special Distribution of the Shortest Linear Recurring Sequences in $Z/(p)$ Field

Qian Yin, Yunlun Luo, and Ping Guo

Department of Computer Science, College of Information Science and Technology, Beijing Normal University, Beijing 100875, China
{yinqian, pguo}@bnu.edu.cn

Abstract. In this paper, the distribution of the shortest linear recurring sequences in $Z/(p)$ is studied. It is found that the shortest linear recurrent length is always equal to $n/2$ when n is even and is always equal to $n/2+1$ when n is odd for any sequence whose length is n. In other words, the shortest linear recurring length is always equal to the half of the length of the given sequence. The probability of finding the distribution of the shortest linear recurring length of two sequences in $Z/(p)$ field is also given.

1 Introduction

Cryptology, which studies the mathematical techniques of designing, analyzing and attacking information security services, is attracting many researchers in nowadays. The classic information security goal is obtained by using a primitive called a cipher. While stream cipher, using pseudorandom sequence [1] encrypts one character at a time with a time-varying transformation, is an important cryptographic system in cryptology field. At present, shift register sequence [2] is used most frequently in pseudorandom sequence, now studying shift register sequence becomes an important part of stream cipher system.

Berlekamp-Massey (BM) algorithm was used as the first algorithm to solve the integrated problems of linear feedback shift register (LFSR). As a result, problems on linear complexity become an important norm of intensity in stream cipher [3]. So-called integrated problems about LFSR are related to given sequence, requiring the integrated solution made up of the length of its shortest linear recurrence and the smallest polynomial. According to this method, the results we can get from this algorithm, which can be done in $Z/(p)$ field, are the length of shortest linear shift register in the given sequence and its smallest polynomial.

The relationships between two kinds of algorithm, which is on the integrated problems of shortest linear shift register on ring for single sequence, are given in the reference [4]. The algorithm for two-sequences on the integrated problems of shortest linear shift register is given in the reference [5]. The algorithm for multi-sequences on the integrated problems of shortest linear shift register is given in the reference [6], which is an expansion of the number of sequences in BM algorithm. Other several algorithms for multi-sequences on the integrated problems of shortest linear shift register is given in the reference [7] and [8].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 43–48, 2005.
© Springer-Verlag Berlin Heidelberg 2005

However, there is no specific description on the distribution of the length of shortest linear recurrence in these research works. In this paper, we obtain the distribution of the shortest linear recurrent length (SLRL) completely when we apply BM algorithm to search the shortest linear recurring length of single sequence in $Z/(p)$ field. In the meanwhile, we also give the probability to find the distribution of the SLRL of two sequences in $Z/(p)$ field.

2 Lemmas

Before lemmas are given, we define some notations first. Let

$$A_{l-1} = \begin{pmatrix} a_0 & a_1 & \cdots & a_{l-1} \\ a_1 & a_2 & \cdots & a_l \\ \vdots & \ddots & \cdots & \vdots \\ a_{l-1} & a_l & \cdots & a_{2l-2} \end{pmatrix}; \qquad M_{k,l-1} = \begin{pmatrix} a_0 & \cdots & a_{l-2} & a_{l-1} \\ \vdots & \cdots & \ddots & \vdots \\ a_k & \cdots & a_{k+l-2} & a_{k+l-1} \end{pmatrix};$$

$$S_{l-1} = \{A_{l-1} | a_0, a_1, \cdots, a_{2l-2} \in F(q)\}; \qquad T_n = \{a_0, a_1, \cdots, a_n | a_i \in F(q)\}.$$

Here $F(q)$ stands for a finite field, which has q elements and q is the power of prime number.

Lemma 1. *If the length of the shortest linear recurrence of one certain sequence $\{a_0, a_1, \ldots, a_n\}$ is l and $2l \leq n+1$, then A_{l-1} is reversible and the coefficient of the linear recurrence is unique.*

Proof. We will prove it by induction method.

Case 1 ($l = 1$): The sequence whose elements are all equal to 0 has a smallest linear recurrence whose length is 0. Therefore, if $l = 1$, then $a_0 \neq 0$. As a result, A_{l-1} is reversible and the coefficient of the linear recurrence is unique.

Case 2 ($l > 1$ and $2l \leq n + 1$): We will prove it to be contradictive. If a sequence $\{a_0, a_1, \ldots, a_n\}$ has a shortest linear recurrence l and A_{l-1} is not reversible, then there is a smallest $m(m \leq l-1)$ which will decrease the rank of the matrix $M_{l-1,m}$(the rank is less than $m+1$). As m is the smallest, it is clear that the sequence $\{a_0, a_1, \ldots, a_{m+l-1}\}$ has the shortest linear recurrence whose length is m. At the same time, $m+l > m+l-1 \geq 2m$ is available. Therefore, we know that A_{m-1} is reversible according to the assumption of the induction. We assume k is the biggest and the sequence $\{a_0, a_1, \ldots, a_k\}$ has the shortest linear recurrence whose length is m, then $m + l - 1 \leq k < n$ is available. Therefore, it must be true that the length of the shortest linear recurrence of the sequence $\{a_0, a_1, \ldots, a_{k+1}\}$ is not longer than that of the sequence $\{a_0, a_1, \ldots, a_n\}$. Now let's consider the matrix equations as follows.

$$M_{m,k-m+1} X_{k-m} = 0 \tag{1}$$

where

$$M_{m,k-m+1} = \begin{pmatrix} a_0 & \cdots & a_{k-m} & a_{k-m+1} \\ \vdots & \ddots & \ldots & \vdots \\ a_m & \cdots & a_k & a_{k+1} \end{pmatrix}, \qquad X_{k-m} = \begin{pmatrix} x_0 \\ \vdots \\ x_{k-m} \\ -1 \end{pmatrix}.$$

The rank of the coefficient matrix is smaller than the rank of the augmented matrix, this results in the matrix equations have no solutions. Therefore it deduces that the shortest linear recurrent length of the sequence $\{a_0, a_1, \ldots, a_{k+1}\}$ will be greater than $k-m+1$, which is $\geq l$. It is conflictive. □

Lemma 2. *The number of reversible matrix in S_{l-1} is $(q-1)q^{2l-2}$.*

Proof. We will prove it by induction method:

Case 1 ($l=1$): $A_{l-1}=(a_0)$ is obvious. A_{l-1} is reversible $\Longleftrightarrow a_0 \neq 0$. Therefore, there are $q-1$ reversible matrices in S_0. The result illustrates that the conclusion is right.

Case 2 ($l>1$): We assume k is the smallest one, which can lead to $rank\{M_{k,l-2}\}=k$ be available, then $k \leq l-1$, that is, $2k \leq k+l-1$. If $k>0$, that can know A_{k-1} is reversible through lemma 1 as the length of $\{a_0, a_1, \ldots, a_{k+l-2}\}$'s shortest linear recurrence is k. We then know the number of this kind of A_{k-1} is $(q-1)q^{2k-2}$ by induction. We choose one of them A_{k-1} and let a_{2k-1} be any value. As A_{l-1} is reversible, we know that $rank\{M_{k,l-1}\}=k+1$. Therefore, the value of a_{k+l-1} will have $q-1$ choices. We choose one of them and denote that

$$c_0 a_{k-i} + \cdots + c_{k-1} a_{i-1} + a_i = 0,\ i = k, \cdots, k+l-2. \tag{2}$$

Then A_{l-1} will be changed to the following matrix by primary operation.

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 1 & 0 & \cdots & \cdots & 0 \\ c_0 & \cdots & c_{k-1} & 1 & 0 & \cdots & 0 \\ 0 & c_0 & \cdots & c_{k-1} & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_0 & \cdots & c_{k-1} & 1 \end{pmatrix} A_{l-1} = \begin{pmatrix} a_0 & \cdots & \cdots & a_{l-2} & a_{l-1} \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ a_{k-1} & \cdots & \cdots & a_{k+l-3} & a_{k+l-2} \\ 0 & \cdots & \cdots & 0 & c \\ \vdots & \cdots & \cdots & \cdots & * \\ 0 & \cdots & c & * & * \end{pmatrix}$$

Where $c \neq 0$, "$*$" can be any value we have computed, that is to say $a_{k+1}, \cdots, a_{2l-2}$ can be arbitrary values. Then for these kinds of A_{l-1} and k, the number of reversible matrices is

$$(q-1)q^{2k-2}q(q-1)q^{l-k-1} = (q-1)^2 q^{l+k-2}, k = 1, \cdots, l-2. \tag{3}$$

If $k=0$, then $a_{l-1} \neq 0$. Therefore, $a_l, \cdots, a_{2l-2}$ can be any value. As a result, the number of these kinds of reversible matrices is $(q-1)q^{l-1}$. Therefore, the number of reversible matrices in S_{l-1} is

$$(q-1)q^{l-1} + (\sum_{k=1}^{l-1} q^{k-1})q^{l-1}(q-1)^2 = (q-1)q^{2l-2} \tag{4}$$

□

3 The Distribution of the SLRL of Single Sequence

Theorem 1. *The number of sequence, whose length of the shortest linear recurrence is l in T_n, is $(q-1)q^{2l-1}$ if $2l \leq n+1$.*

Proof. We know that if $2l \leq n+1$ and the shortest linear recurring length of the sequence $\{a_0, a_1, \cdots, a_n\}$ is l then A_{l-1} is reversible and the coefficient is unique through lemma 1. There are $(q-1)q^{2l-2}$ reversible A_{l-1} and a_{2l-1} can be any value in q kinds of choices through lemma 2. We can choose one of them and then the coefficient can be determined. $\{a_0, a_1, \cdots, a_n\}$ can be confirmed by $\{a_0, a_1, \cdots, a_{2l-1}\}$ and recurring coefficient. Therefore, theorem 1 has been proved. □

Lemma 3. *The length of the shortest linear recurrence of the sequence $\{a_0, a_1, \cdots, a_n\}$ is l and*

$$2l > n+1 (l \leq n) \Longleftrightarrow M_{n-l+1,l-1} X_{l-2} = 0 \tag{5}$$

has no solution, and $M_{n-l,l-1}$ is full rank. That is to say the rank is $n-l+1$.

Proof. (1) Sufficiency: The linear recurrence of the sequence $\{a_0, a_1, \cdots, a_n\}$, whose length is l, can be expanded to a_{2l-1} by induction because $2l > n+1$, a_{l-1} is reversible. Therefore, its row-vector groups are linear independence through lemma 1.

(2) Necessity: The matrix $M_{n-l,l-1}$ is full rank. Therefore, the matrix equations $M_{n-l,l} X_{l-1} = 0$ has a solution. That is to say the sequence $\{a_0, a_1, \cdots, a_n\}$ has the shortest linear recurrence whose length is l. As the following matrix equations $M_{n-l+1,l-1} X_{l-2} = 0$ have no solution, the sequence $\{a_0, a_1, \cdots, a_n\}$ has the shortest linear recurrence whose length is l. □

Theorem 2. *The number of sequence whose shortest linear recurring length is l in T_n is $(q-1)q^{2(n-l+1)}$ if $2l > n+1$.*

Proof. The matrix $M_{n-l,l-1}$ is full rank through lemma 3. As l is the shortest linear recurrence length, the equations $M_{n-l+1,l-1} X_{l-2} = 0$ has no solution. Therefore, the rank of coefficient matrix must be smaller than the rank of augmented matrix. That is to say the last row of the augmented matrix cannot be the linear combination of the former $n-l+1$ rows. We know $n-l+2 \leq l$ because $2l > n+1$. Therefore, we can get the following result.

$$Rank\{M_{n-l+1,l-1}\} = n-l+2, \qquad Rank\{M_{n-l+1,l-2}\} = n-l+1. \tag{6}$$

We assume k is the smallest one which can make the decreased,

$$Rank\{M_{k,l-2}\} = k. \tag{7}$$

If $k = 0$, then $a_0 = \cdots = a_{l-2} = 0$ and $a_{l-1} \neq 0$. Therefore, there are $q-1$ values can be chosen. As $a_l, \ldots, a_n$ can be any value, we know that the number of this kind of sequences is $(q-1)q^{n-l+1}$. The first k row vectors of the matrix

$M_{k,l-2}$ are linear independent and the last row-vector can be represented by the linear combination of the former k row-vectors if $0 < k \leq n-l+1$. Therefore, we can get that $a_0, \ldots, a_{k+l-2}$ has the shortest linear recurrence whose length is k. Because of $2k \leq k+n-l+1 < k+l$, we know that A_{k-1} is reversible and the number of this kind of sequence $\{a_0, \ldots, a_{k+l-2}\}$ is $(q-1)q^{2k-1}$ based on lemma 1 and theorem 1. However, the matrix $M_{k,l-1}$ is full rank. Therefore, if we choose certain $a_0, \ldots, a_{k+l-2}, a_{k+l-1}$ could only be chosen from $q-1$ values. We choose one certain value for a_{k+l-1} and notice that

$$c_0 a_{k-i} + \cdots + c_{k-1} a_{i-1} + a_i = 0, \ i = k, \cdots, k+l-2 \tag{8}$$

The matrix $M_{n-l+1,l-1}$ will be changed to the other matrix by primary operation:

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 1 & 0 & \cdots & \cdots & 0 \\ c_0 & \cdots & c_{k-1} & 1 & 0 & \cdots & 0 \\ 0 & c_0 & \cdots & c_{k-1} & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_0 & \cdots & c_{k-1} & 1 \end{pmatrix} M_{n-l+1,l-1} = \begin{pmatrix} a_0 & \cdots & a_{k-1} & \cdots & a_{l-1} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ a_{k-1} & \cdots & a_{2k-2} & \cdots & a_{l+k-2} \\ 0 & \cdots & \cdots & 0 & c \\ \vdots & \cdots & \cdots & \cdots & * \\ 0 & \cdots & c & * & * \end{pmatrix}$$

where $c \neq 0$, "$*$" can be any value we have computed. That is to say, $a_{k+l}, \ldots, a_n$ can be any value. Therefore, the number of this kind of sequence $a_0, \ldots, a_n$ is $(q-1)q^{2k-1}(q-1)q^{n+1-k-l} = (q-1)^2 q^{n+k-l}$. As a result, the number of the sequence, whose length of the shortest linear recurrence in T_n is l and $2l > n+1$, is

$$(q-1)q^{n-l+1} + (\sum_{k=1}^{n-l+1} q^{k-1})q^{n-l+1}(q-1)^2 = (q-1)q^{2(n-l+1)}. \tag{9}$$

From theorem 1 and theorem 2, we can see that if the length $n+1$ of the sequence is even, the probability of l equaling to $\lfloor (n+1)/2 \rfloor$ is the biggest and which is $(q-1)/q$. If the length $n+1$ of the sequence is odd, the probability of l equaling to $\lfloor (n+1)/2 \rfloor + 1$ is the biggest and which is also $(q-1)/q$. □

4 The Possible Distribution of the SLRL of Two Sequences

Example: We assume P is a prime number and N is the length of sequences and the number in the axis X is the length of l. In the experiment we use C program language to realize the integrated algorithm of two sequences. We can get the result as shown in Table 1.

From Table 1 we can know that there are $(P^2-1)P^{3l-2}$ kinds of two sequences whose shortest linear recurrence length is l if $2l < n+1$ and there are (P^2-1)

Table 1. Results for two sequences

P	N	0	1	2	3	4	5	6	7	8
2	2	1	6	9						
2	4	1	6	36	21					
2	4	1	6	48	156	45				
2	5	1	6	48	336	540	93			
2	6	1	6	48	384	2160	1308	189		
2	7	1	6	48	384	2880	9840	2844	381	
2	8	1	6	48	384	3072	20928	34416	5916	765
3	2	1	24	56						
3	3	1	24	504	200					
3	4	1	24	648	5256	632				
5	3	1	120	12600	2904					
5	6	1	120	15000	1875000	196515000	45360600	374904		

$(P^N - P + 1)$ kinds of two sequences whose shortest linear recurrence length is l if $l = n$ and the number of two sequences whose shortest linear recurrence length is l has a factor $(P^2 - 1)P^{2(N-l)}$ if $2l \geq n + 1$. We need to prove them in the future research work.

References

1. Wilfried Meidl: On the Stability of 2n-Periodic Binary Sequences. IEEE Trans Inform Theory, vol. IT-51(3) (2005)1151-1155
2. James A. Reeds and Neil J. A. Sloane: Shift-register synthesis (modulo m). SIAM Journal on Computing, vol. 14(3) (1985)505-513
3. Ye Dingfeng and Dai Zongduo: Periodic sequence linear order of complexity under two marks replacing. Fourth session of Chinese cryptology academic conference collection (1996) (in Chinese)
4. Zhou Yujie and Zhou Jinjun: The relationship between two kinds of sequences syntheses algorithm over ring Z/(m). Fifth session of Chinese cryptology academic conference collection (1998) (in Chinese)
5. Gao Liying, Zhu yuefei: The shortest linear recurrence of two sequences. Journal of information engineering university, vol. 2(1) (2001) 17-22 (in Chinese)
6. Feng Guiliang, Tzeng K K.A: generalization of the Berlekamp-Massey algorithm for multi-sequence shift-register synthesis with applications to decoding cyclic codes. IEEE Trans Inform Theory, vol. 37(5) (1991) 1274-1287
7. Feng Guiliang, Tzeng K K.A: generalized Euclidean algorithm for multi-sequence shift-register synthesis. IEEE Trans Inform Theory, vol. 35(3) (1989) 584-594
8. Zhou Jinjun, Qi Wenfeng, Zhou Yujie: Gröbner base extension and Multi-sequences comprehensive algorithm over Z/(m). Science in China (1995)113-120 (in Chinese)

Cryptanalysis of a Cellular Automata Cryptosystem

Jingmei Liu, Xiangguo Cheng, and Xinmei Wang

Key Laboratory of Computer Networks and Information Security,
Xidian University, Ministry of Education,
Xi'an 710071, China
jmliu@mail.xidian.edu.cn, chengxiangguo@hotmail.com,
xmwang@xidian.edu.cn

Abstract. In this paper we show that the new Cellular Automata Cryptosystem (CAC) is insecure and can be broken by chosen-plaintexts attack with little computation. We also restore the omitted parts clearly by deriving the rotating number δ of plaintext bytes and the procedure of Major CA. The clock circle Δ of Major CA and the key S_N are also attacked.

1 Introduction

Confidentiality is necessary for a majority of computer and communication applications. The emergence of the ad-hoc network and common networking requires new generations of security solutions. Essential element of any secure communication is cryptography techniques. A general summary of currently known or emerging cryptography techniques can be found in [1]. Cellular automata are one of such promising cryptography techniques [2]. CA system was proposed for public-key cryptosystems by Guan [3] and Kari [4], and for systems with a secrete key were first studied in [5-10]. Block cipher using reversible and irreversible rules was proposed by Gutowitz[11].

Paper [12] is the latest effort in cellular automata cryptosystems (CAC) design. Due to the affine property, the vulnerability of the previous CAC is removed in the latest effort in cellular automata cryptosystems (CAC) design. The new CAC is much better than those previous cryptosystems for its clearly specifying the 128 bits block size and 128 bits key size, and four transformations included in the encryption and decryption. However, there are still many parts omitted, such as how to derive the secret CA from the secret key, how to derive the rotate number δ of plaintext bytes etc. Feng B [13] has broken this cipher with hundreds of chosen plaintexts by restoring the rotating number δ of plaintext bytes, but he fails to find the key bytes.

In this paper we show that the new CAC is still insecure. The designer has omitted many design parts, but this paper presents the omitted parts clearly by deriving the rotating number δ of plaintext bytes and the procedure of Major CA. The clock circle Δ of Major CA and the key S_N are also attacked.

This paper is organized as follows: In Section 2 we present the new CAC. Section 3 presents an equivalent transform of the CAC. Section 4 presents the attack. Section 5 concludes the paper.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 49 – 54, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Description of the New CAC

The new CAC presented in [12] is based on one-dimensional 16-cell GF(2^8) cellular automata and the local function $f_i, i = 1, 2, \cdots 16$ over GF(2^8). The encryption algorithm is based on two different classes of group CA, 16 cell GF(2^8) Major CA and 16 cell GF(2^8) Minor CA. The Major CA is one group CA with equal cycles of length 32. The Minor CA is a maximum-length CA with equal cycles of length 128. The Minor CA is used to transform a secret key K into a secret state S_N, which controls the four transforms of the CAC, namely T_1, T_2, T_3 and T_4. The CAC is a block cipher with the structure described and explained as follows.

Linear Transformation on Key: The key used for CAC scheme is a bit string of length same as the number of bits used for Minor CA. The input key is used as the initial seed of the Minor CA.

Role of Minor CA. The Minor CA is operated for a fixed number of clock cycles for input of each plaintext. Initially, the seed S_0 of the Minor CA is the key K. For each successive input plaintext, Minor CA generates a new state marked as S_N after running d number of steps from its current state. The Minor CA is a linear 16-cell GF(2^8) CA. The secret key K is taken as the initial state of the minor CA. Run the Minor CA a fixed number of clock cycles d. Denote the obtained state by S_N, which controls the four transforms of the CAC:

T_1: Rotate Linear Transform on Plaintext: Derive a number δ from S_N. Rotate each byte of the input (plaintext) T by δ steps (bits). The result T_1 is the input to the major CA.

T_2: Major CA Affine Transform: The Major CA is generated by an efficient synthesis algorithm from S_N. The Major CA is an affine CA, which has circles equal to length 32. That is, for any initial state, the Major CA return back the initial input state after running 32 clock cycles. S_N also determines a number $1 < \Delta < 32$. T_2 takes the T_1 as input and runs Δ clock cycles. The result T_2 is input to T_3. (In the decryption, the Major CA is run 32-Δ clock cycles.)

T_3: Non-Affine Transform: A non-affine transform is achieved by selective use of Control Major Not (CMN) gate. The CMN gate is a non-linear reversible gate with four inputs (1 data input and 3 control inputs) and one output. We will denote the control bits by c1, c2 and c3. The function is defined as $y = x \oplus \{(c_1 \cdot c_2) \oplus (c_2 \cdot c_3) \oplus (c_1 \cdot c_3)\}$, where $\oplus$ denote the XOR and $\cdot$ denote the AND function. The plaintext T_2 is subjected to CMN gate depending on the result of a function called Majority Evaluation Function (MEN). The Majority Evaluation Function takes the 5 bits, referred to as fixed-bits of T_2 and calculates the number of 1's in these bits. The 5 bit fixed positions are selected depending on S_N. If the number of 1's is greater than 2, then each bit of T_2 except these fixed-bits are subjected to CMN gate. Otherwise, T_2 remains as what it is. In any case, we call the resultant after T_3 transformation as T_3. Two sets of control bits taken from S_N applied to the CMN gate alternately. The fixed-bits have to be remained fixed because during decryption the same fixed-bits will require to get the same result from majority evaluation function.

T_4: Key Mixing: To enhance the security and randomness, we generate final encrypted plaintext Tencr by XORing the Minor CA state S_N with the plaintext T_3. The output of T_4 is denoted by $T_{encrypt} = T_3 \oplus S_N$.

3 An Equivalent Transform of the New CAC

In CAC, the Major CA is an affine CA. That means the local function for the *i*-th cell can be expressed as by its neighbors with operation over GF(2^8):

$$f_i(v_{i-1}, v_i, v_{i+1}) = w_{i-1}v_{i-1} + w_i v_i + w_{i+1}v_{i+1} + u_i,\ w_i, w_{i-1}, w_{i+1} \in GF(2^8)$$

Major CA means that running the Major CA Δ steps is equivalent to Δ iterations of the local function, which lead to another affine function. Formally, the i-th cell is determined by all the (i-Δ)-th,…, (i+Δ)-th cells after Δ steps. We denote the i-th byte of T_1 by T_1[i] and the i-th byte of T_2 by T_2[i], $i = 1, 2, \cdots, 16$ .We have

$$T_2[i] = \sum_{j=1}^{16} w_j[i]T_1[j] + U_{[i]}, w_1, w_2, \cdots w_{16}, U_{[i]} \in GF(2^8)$$

If Δ ≥ 16, each cell affects itself. Otherwise some cells are not affected by all the cells. Here we can restore what is the value of Δ in our attack, and just find all the values W_j[i] and $U_{[i]}$ in the above equation. Paper [13] said that it is impossible to find U. The reason is that it takes the different equivalent transformation of CAC, and T_2 will be xored with S_N, no matter whether it goes through CMN or not. Since the + in GF(2^8) is xor, $U_{[i]}$ will be xored with S_N[i] finally. So it is impossible to separate $U_{[i]}$ and S_N[i] from the $U_{[i]} \oplus S_N$[i], and only find $U_{[i]} \oplus S_N$[i] in [13] attack. But if the equivalent transformation of ours is considered in this paper, S_N can be found and $U_{[i]}$ will be determined respectively. Denote the i-th bit of T_2 by b_i[T2]. The equivalent CAC encryption transform is divided into three steps as follows.

Equivalent transform of CAC (E_k() of CAC with key K)
Initiate the key K.

Step 1: obtain T_1 by rotating each byte of T, T[i],δ steps.
Step 2: obtain T_2 by setting

$$T_2[i] = \sum_{j=1}^{16} w_j[i]T_1[j] + U_{[i]}, if\ (\Delta < 16), w_k[i], w_{16-k}[i] \equiv 0, k = 1, 2, \cdots, \Delta$$

Step 3: key mixing:

$$T_{encrypt} = \begin{cases} T_2 + m_1 & (weight(b_1(T_2)b_2(T_2)\cdots b_5(T_2)) \geq 3)\ \&\ \&(c_1c_2 + c_2c_3 + c_1c_3 = 1) \\ T_2 + m_2 & weight(b_1(T_2)b_2(T_2)\cdots b_5(T_2)) \leq 2 \\ & (weight(b_1(T_2)b_2(T_2)\cdots b_5(T_2)) \geq 3)\ \&\ \&(c_1c_2 + c_2c_3 + c_1c_3 = 0) \end{cases} \quad or$$

Where $b_i[T_2]$ denote *i*-th bits of T_2, c_1, c_2, c_3 are three control bits determined by S_N . $m_1 = S_N + T_{cmn}$, $m_2 = S_N$, $T_{cmn} = 2^{16} - 2^{b_1(T_2)} - 2^{b_2(T_2)} - \cdots - 2^{b_1(T_5)}$. The equivalent transformation [13] of CAC also has three steps, and their difference is in step 2 and step 3. The relationship of $U_{[i]}$ and U is as follows:

$$U = (U_{[1]} \| U_{[2]} \| \cdots \| U_{[16]})$$

4 Chosen Plaintexts Attacks on CAC

In this section we will denote how to attack the CAC under the chosen-plaintext.

4.1 Initial Cryptanalysis of CAC

In CAC, the result value $T_{encrypt}$ is determined by the random 5 bits of T_2 and the result is to chose m_1 or m_2 in step 3. The cryptanalysis of m_1 and m_2 is very useful for us to attack CAC, but it is impossible to find the probability of m_1 and m_2 from the equivalent transformation of CAC. So it is necessary to research the initial CAC to find the probability of m_1 and m_2. The only nonlinear transformation is included in step 3. In this step, if the number of 1's of the 5 fixed-bits is less than 3, then the transformation of step 3 is linear and there is no nonlinear transformation in the whole CAC. The probability of this step is:

$$(C_5^0 + C_5^1 + C_5^2)/(C_5^0 + C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5) = 16/32 = 1/2$$

However if the number of 1's of the 5 fixed-bits is greater than 2, the probability of $c_1c_2 \oplus c_2c_3 \oplus c_1c_3 = 0$ is 1/2, and $T_{encrypt} = T_2 \oplus m_2$. It is calculated that the probability of $T_{encrypt} = T_2 \oplus m_2$ is $1/2 + 1/2 * 1/2 = 3/4$. So the CAC will take the value $T_{encrypt} = T_2 + m_2$ with a high probability, and it is easy to attack the CAC by this property.

4.2 Rotating Restore Rotating Number δ of Plaintext

Randomly choose two plaintexts P_1, P_2. Encrypt P_1, P_2 and denote the two cipher-texts by $T_{encrypt1}, T_{encrypt2}$. There are eight choices of rotating number δ of plaintext in step 1, and denote the eight 128 bit plaintext values by $(T_1[0], T_1[1], \cdots, T_1[7])$, $(T_1'[0], T_1'[1], \cdots, T_1'[7])$ respectively. After the key mixing of step 3, the decryption is:

$$T_{encrypt1}[i] = \sum_{j=1}^{16} w_j[i] T_1[j] + U_{[i]} + m_2[i] = T_2[i] + m_2[i] \quad (1)$$

$$T_{encrypt2}[i] = \sum_{j=1}^{16} w_j[i] T_1'[j] + U_{[i]} + m_2[i] = T_2'[i] + m_2[i] \quad (2)$$

From (1) and (2), it is found that $T_{encrypt1} \oplus T_{encrypt2} = T_2 \oplus T_2'$. The differential of output is equal to that of the output of Major CA. Let the outputs of P_1, P_2 after Major CA be $(T_2[0], T_2[1], \cdots, T_2[7])$, $(T_2'[0], T_2'[1], \cdots, T_2'[7])$, and the testing values should be $T_2[i] \oplus T_2'[i], i = 0,1,\cdots,7$, where $T[i]$ denotes the 128bit data. If the equation

$T_{encrypt1} + T_{encrypt2} = T_2 + T_2'$ holds true, then the rotating number δ of plaintext in step 1 is determined, and δ=i with the data complexity of two chosen plaintexts pair.

4.3 Restore Affine Transformation of Major CA

Randomly choosing 17 plaintexts and their corresponding cipher-texts, we can derive the 17 values of T_2 according to the method of attacking the rotating number δ of plaintext, and constitute the following equation:

$$\begin{cases} T_{2,1}[i] = \sum_{j=1}^{16} w_j[i] T_{1,1}[j] + U_{[i]} \\ T_{2,2}[i] = \sum_{j=1}^{16} w_j[i] T_{1,2}[j] + U_{[i]} \\ \quad\quad\vdots \\ T_{2,17}[i] = \sum_{j=1}^{16} w_j[i] T_{1,16}[j] + U_{[i]} \end{cases} \tag{3}$$

When the rotating number δ of plaintext and affine transformation expression after Δ iteration of Major CA are determined, it is easy to determine the iteration number Δ from the affine transformation expression when $\Delta < 16$.We can deem that the Δ can be resolved from the value $w_j[i]$. If the value $w_j[i] = 0$ and j is the least, then $\Delta = j$. When $\Delta \geq 16$, we iterate the affine transformation of Δ iteration of Major CA by number n and to test if the final result is equal to the plaintext.

4.4 Restore Key S_N

It is necessary to find the key when attacking any cipher and if the key is attacked, the cipher is broken in practice. When attacking the CAC, [13] presented that the key S_N could not be found, and only $S_N + U$ could be found. In section 3.1, it is known that if the expression $T_{encrypt1} + T_{encrypt2} = T_2 + T_2'$ holds true in the process of determining rotating number δ of plaintext, then from the equation $T_{encrypt1}[i] = \sum_{j=1}^{16} w_j[i] T_1[j] + U_{[i]} + m_2[i] = T_2[i] + m_2[i]$, the byte key S_N could be restored.

5 Conclusions

In this paper we break the latest effort on cellular automata cryptosystem. The attack is very efficient, requiring only two chosen plaintexts and a small computation amount with time complexity of two encryptions. Although the designer has omitted many design parts, this paper restores the omitted parts clearly by deriving the rotating number δ of plaintext byte and the procedure of Major CA. The clock circle Δ of Major CA and the key S_N are also broken. So it is insecurity to protect the sense information and should be intensified. It enforces the idea that we should not blindly trust the pseudo randomness brought by mathematical systems.

References

1. Schneier, B.: Applied Cryptography. Wiley, New York (1996)
2. Sarkar, P.: A brief history of cellular automata. ACM Computing Surveys, Vol.32, No.1. ACM Press (2000) 80–107
3. Guan, P.: Cellular automata public key cryptosystem. Complex System, Vol.1. ACM Press (1987) 51–57
4. kari, J.: Cryptosystems based on reversible cellular automata. Personal communication. ACM Press (1992)
5. Wolfram, S.: Cryptography with Cellular Automata. Lecture Notes in Computer Science, Vol.218. Springer-Verlag, Berlin Heidelberg New York (1986) 429-432
6. Habutsu, T. , Nishio, Y. , Sasae, I. and Mori, S.: A Secret Key Cryptosystem by Iterating a Chaotic Map. Proc. of Eurocrypt'91. Springer-Verlag, Berlin Heidelberg New York (1991) 127-140
7. Nandi, S., Kar, B. K. and Chaudhuri, P. P.: Theory and Applications of Cellular Automata in Cryptography. IEEE Trans. on Computers, Vol. 43. IEEE Press (1994) 1346-1357
8. Gutowitz, H.: Cryptography with Dynamical Systems. In: Goles, E. and Boccara, N. (eds.): Cellular Automata and Cooperative Phenomena. Kluwer Academic Press(1993)
9. Tomassini, M. and Perrenoud, M.: Stream Ciphers with One and Two-Dimensional Cellular Automata. In Schoenauer, M. at al. (eds.): Parallel Problem Solving from Nature - PPSN VI, LNCS 1917. Springer-Verlag, Berlin Heidelberg New York (2000) 722-731
10. Tomassini, M. and Sipper, M.: On the Generation of High-Quality Random Numbers by Two-Dimensional Cellular Automata. IEEE Trans. on Computers, Vol. 49, No.10. IEEE Press (2000) 1140-1151
11. Gutowitz, H.: Cryptography with Dynamical Systems, manuscript
12. Sen, S., Shaw, C. R., Chowdhuri, N., Ganguly and Chaudhuri, P.: Cellular automata base cryptosystem (CAC). Proceedings of the 4th International Conference on Information and Communications Security (ICICS02), LNCS 2513. Springer-Verlag, Berlin Heidelberg New York (2002) 303–314
13. Feng, B.: cryptanalysis of a new cellular automata cryptosystem. In ACSP2003, LNCS 2727. Springer-Verlag, Berlin Heidelberg New York (2003) 416-427

A New Conceptual Framework Within Information Privacy: Meta Privacy

Geoff Skinner, Song Han, and Elizabeth Chang

School of Information Systems, Curtin University of Technology,
Perth, WA, Australia
Geoff.Skinner@newcastle.edu.au,
{Song.Han, Elizabeth.Chang}@cbs.curtin.edu.au

Abstract. When considering information security and privacy issues most of the attention has previously focussed on data protection and the privacy of personally identifiable information (PII). What is often overlooked is consideration for the operational and transactional data. Specifically, the security and privacy protection of metadata and metastructure information of computing environments has not been factored in to most methods. Metadata, or data about data, can contain many personal details about an entity. It is subject to the same risks and malicious actions personal data is exposed to. This paper presents a new perspective for information security and privacy. It is termed Meta Privacy and is concerned with the protection and privacy of information system metadata and metastructure details. We first present a formal definition for meta privacy, and then analyse the factors that encompass and influence meta privacy. In addition, we recommend some techniques for the protection of meta privacy within the information systems. Further, the paper highlights the importance of ensuring all informational elements of information systems are adequately protected from a privacy perspective.

1 Introduction

It seems that where ever you go on the Internet today every body wants to know your name or at least your identity. This is usually along with a host of other personal details [1]. It's a scenario that has painted a bleak future for information privacy. As more and more services are being moved online and computerized, the system owners insist on collecting vast amounts of personal information. The need for excessive and increasing data collection habits is the cause for concern for all entities involved. This practise needs to be analysed for its intentions and stopped were it represents serious threats to personal privacy. Most of the time the user entities are not given a reasonable spectrum of choices for what information you provide in order to use the services. It is normally a scenario of filling in all of the required form fields, or do not use the service at all. When an entity does not really have any choice but to use the service they are placed in an uncompromising position. It is a situation where personal privacy is the added and often hidden cost for using the service.

There are a number of solutions that have been proposed that attempt to address the issue of system wide privacy protection [2, 3, 4]. Some solutions are based on techno-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 55 – 61, 2005.
© Springer-Verlag Berlin Heidelberg 2005

logical approaches, and are commonly referred to as Privacy Enhancing Technologies (PETs). Other methods rely on privacy policy electronic representations, regulations, and legal enforcement. The remainder use a combination of techniques both technological and regulatory. An issue that is of major importance is that our research has revealed that no solution considers the security and privacy protection of metadata and metastructure information. To the best of our knowledge, current information security and privacy methods do not protect or even consider metadata and metastructure privacy protection. Both metadata and metastructure information may reveal an entity's identity as well as other personal details. Both forms of data about data and structure are increasingly common in information systems used today. As a result, they should be protected by the same levels of information security and privacy protection techniques afforded to personal data.

This concept and area of research has been termed Meta Privacy. It is the focus of this paper and is explained in greater detail in the following sections. The organization for the rest of the paper is presented as follows: Section 2 provides relevant background material and related work. This is followed by a formal definition of Meta Privacy provided in Section 3. In Section 4 the factors that encompass and influence Meta Privacy are discussed and analysed. Techniques for the protection and support of Meta Privacy are detailed in Section 5. A brief summary is presented in Section 6.

2 Background and Related Work

The two main areas of background material and related work are concerned with the fields of Privacy and Metadata. Privacy is a very broad field of study so only a specific dimension of it is relevant to this paper. The dimension of Information Privacy is discussed in section 2.1. Metadata and Metastructure are discussed in section 2.2 below. These sections have been removed for the 6 page publication. Please contact author for full version of paper.

3 Definition and Understanding of Meta Privacy

The biggest issue with many of the current privacy protection approaches is their inability to provide protection across a broad spectrum of information privacy issues. Most of the privacy tools listed in Section 2 only address specific areas of information privacy. They are applied in an ad-hoc fashion resulting in a piecemeal approach to privacy protection. These methods applied in such a way have proved to be ineffective and inadequate for protecting personal information, or Personally Identifiable Information (PII). This includes attempts at self regulation and schemes using P3P for issuing Privacy Policies. It is often found that entities, normally organizations, do not always do what their privacy policy says they do. So an organizational P3P policy and structure might look good to the user entity at face value but it does not provide any guarantee that the policies are actually being enforced [9]. Self-regulation of privacy protection will always conflict with the economic interests of organizations, and without enforceable laws and regulations it will continue to be ineffective. Therefore we need system privacy controls designed and integrated into the system that entities are

unable to circumvent. These controls and PETs should be modeled on enforceable regulations and guidelines [10]. Included in these controls and regulations should be consideration for the protection of Meta Privacy. That is, the protection of metadata and metastructure information.

The term Meta Privacy does not seem to have been proposed before this paper and therefore needs a formal definition. A common definition for the word Meta is as a prefix and when used in an information systems context means "relating to" or "based on". More formally it is a prefix meaning “information about”. When used in conjunction with the term privacy it formulates the new term Meta Privacy.

Meta Privacy means ensuring the security and privacy of data about privacy and personal data. Meta privacy is concerned with the security and privacy of the information used to support other system services and processors that may impact upon an entities privacy. This encompasses the protection of metadata and metastructure information that may reveal an entities identity and other personal information.

In this context an entity may be an individual, group, or organization. Further, an individual represents a singular entity, most often a human being with may be an information system user. A group is defined as a ‘non-committed’ informal relationship between entities. The members of the group may be individuals, other groups and organizations. An organization is defined as a committed formal relationship between entities. The members of an organization may be individuals, groups, and other organizations.

An example of what the Meta Privacy concept is can be explained by using a commonly used desktop application scenario. It has been found that Microsoft Word generates a number of potentially privacy invasive metadata fields. That is, a typical Word document contains twenty-five different types of hidden metadata [11]. Many of these metadata fields may contain personal information related to the entity, or as discussed above the identity, creating or editing the document. These include such things as Authors (Entity or Identity) name, Organization (Entity or Identity) Name, the date the document was created, last edited and saved. In this example Meta Privacy encompasses the protection, use and management of the metadata associated with the document.

Proper Meta Privacy practises would ensure that none of the personal information contained in the metadata and metastructure is used for any purpose other than that specially agreed upon by the personal information owner. Further, that the metadata is not provided to any third party not authorized to access the data without the owners express permission. In the example provided, if the document is to be shared with other third parties, good Meta Privacy practices would be in place to ensure all metadata of a personal and identifiable nature are stripped from the document before the document is accessible. Where possible as a pre-emptive measure, the entity should also be able to generate and edit the document in a pseudo-anonymous or anonymous way.

It is the metadata and metastructure implementation that can be the source of either privacy enhancing benefits or privacy invasive drawbacks. In either case it is the privacy of an entity that should be the focus of Meta Privacy protection just as it is with the privacy and security of personal information. As mentioned previously entities include individuals, groups and organizations. Therefore, any type of data

pertaining to their identity, and hence subject to classification as personal data, should be protected. This includes descriptive information about an organizations data and data activities which may be classified and metastructure information. For example, Meta Privacy would include the protection of information that defines and enforces an organizations privacy policies and protection techniques.

4 Meta Privacy Components

Meta Privacy is about the protection of metadata and metastructure information that affects the privacy of entities and system privacy management. It is only natural then that the way the metadata and metastructure information is used and managed is a major influence on Meta Privacy. Meta-information and processes making use of metadata and metastructure information can be classified as either a Meta Privacy Risk (MPR) or a Meta Privacy Benefit (MPB). It depends on how the data is utilized. Where metadata provides information about the content, quality, condition, and other characteristics of entity data it can be classified as being in a Meta Privacy Risks (MPR) category. This classification also extends to metastructure information with similar content. Metastructure information containing an entities system's structural data and component details are also classified in the MPR category. Like the personal information they describe, they are exposed to the same risks and malicious attacks. That is, meta-information should be protected by the same measures implemented to protect personal and identifying data.

Protecting metadata and metastructure information should also facilitate the privacy protection objectives of Unlinkability and Unobservability. Unlinkability means that entity's (individuals, groups, organizations, system processors and components) transactions, interactions and other forms of entity influenced system processors and uses are totally independent of each other. From an identity perspective it means an entity may make multiple uses of resources or services without other entities being able to link these uses together [6]. That is, between any numbers of transactions, no correlating identification details can be deduced. Each transaction when examined individually or in a collective group of transactions does not reveal any relationships between the transactions and also the identities who may have initiated them. This also encompasses the condition that one should not be able to link transactions from multiple identities to a single entity.

An approach that is of relevance to Meta Privacy is the use of meta-information for privacy protection. Meta privacy tags and metadata can be used for entity privacy policy preferences representation and enforcement. The use of metadata and metastructure information in this way is classified as Meta Privacy Benefits (MPB). The leading example of use of metadata for representing privacy preferences is P3P [7]. Other approaches have been proposed that use metadata and metastructure information to protect personal data and privacy in a number of alternate operational settings. One such technique is associating and storing metadata for representing individual items of personally identifiable information (PII) [13]. The technique utilizes semantic web languages like OWL [14] and RDFS [15] to better represent personal information building upon the basic formatting provided by P3P. Through the advanced metastructure representations, fine grained control over the release of the individual

personal data items can be obtained. The technique goes further to propose an ontology-based framework for controlled release of PII at both policy writing and evaluation time. Regardless of the semantic language used the metadata and metastructure information generated is a useful method for privacy protection and policy representation. As a result, it is classified as being a MPB.

Another technique makes use of "privacy meta-data" [16] and stores the metadata in a relational database as tables. This is stored in the database along with the personal information collected from the entities. Extending this technique further, stores the exact user privacy preferences for each individual personal data element. The added benefit is that the data is protected by the user privacy policy selections at the time of collection. This is extremely useful for situations in which the privacy policy conditions may have been changed in such a way to decrease the level of privacy protection offered to entities on a whole. By default the information owners do not continually have to be concerned with what level of protection is being provided for their personal data. The privacy metadata stays the same until the information owner elects to modify their privacy policy preferences. Due to its inherent nature to protect an entity's privacy, this technique for metadata use is also a Meta Privacy Benefit.

Meta Privacy therefore encompasses both Meta Privacy Risk and Meta Privacy Benefit categories. Where metadata and metastructure information contains details that reflect some level of knowledge pertaining to an individual's identity or other forms of personal information, then they are a potential risk to privacy.

5 Protecting Meta Privacy in Information Systems

From an operational standpoint the metadata and metastructure information has to be protected by the same levels of security used to protect personal information. System owners need to ensure all metadata personal tags are removed when information is shared. Further, the system owners and the entities providing their personal information need to be aware of metadata generation and usage. Likewise, it should be subjected to the same privacy policy guidelines selected by an entity to protect their personal data. As it is possible that one can learn information by looking at the data that defines what personal data is collected, how it is protected and stored, what privacy policies govern its use.

For example, in certain situations an entity may interact in a virtual collaboration with their true identity, a number of pseudo-anonymous identities, and also an anonymous identity. In each and every case the individual transactions conducted by which ever identity of an entity should not be linked back to the entity or any other identity of that entity. This allows an entity to conduct business or interact with a number of different entities using a different pseudo-anonymous identity for each. With no linkability between pseudo-anonymous identities or linkability back to an entity the privacy of the entity and their business transactions are protected. It is also intended to protect the entities identity against the use of profiling of the operations. So while the entity may already be using a pseudo-anonymous or anonymous identity, unlinkability further ensures that relations between different actions can not be established. For example, if some entity with malicious intent was trying to determine the usage patterns of a particular identity. By utilizing a similar set of metadata and metastructure

privacy protection techniques, transactions and entity system interactions can be made unobservable. Unobservability is like a real time equivalent of unlinkability. Formally it is defined as an entities ability to use a resource or service without other entities, especially third parties, being able to observe that the resource or service is being used [6]. The difference lies in the fact that the objective of unobservability is to hide an entity's use of a resource, rather than the entity's identity. This can be achieved through a number of different techniques that are discussed in the full length paper version.

As unobservability is concerned with not disclosing the use of a resource, it is a very important component of Meta Privacy protection. For example, metadata is data about data, which includes system logs and records of identities use of system resources. It may also include details of an identity's access to and modification of personal data. Metastructure information may contain access control details for identities, data on privacy policies used by identities and other types of processing information influencing identity-system interaction. For that reason all forms of identity related meta-information, including metadata and metastructure, need to remain unobservable to ensure entity privacy. Metadata and metastructure information that needs to remain unobservable and unlinkable can also be classified in the Meta Privacy Risks. By their simple existence and generation the meta-information may be a source of potential risks to entities privacy. That is, proper security and privacy measures need to be taken to ensure the meta-information is well protected. There are a number of ways to achieve this that are discussed throughout this paper.

One way to provide extra privacy protection is to use Privacy Metadata. The metadata 'attaches' itself to individual data elements in order to protect them. As the metadata is being stored in database along with the personal information, any time the personal information is accessed the privacy policies governing its use are readily available for verification. Further, the metadata can even be used to control access to the data, regulate the use of the data, and to enforce accountability with respect to its use. If this done then we need to protect the metadata as well from malicious attack. Like normal personal data, metadata transactions and events should not be linkable or observable. This is due to the fact that is may be possible to combine this data with other information to deduce additional personal information about an entity. Therefore proper protection techniques for metadata are required during both processing and while it is at rest.

6 Conclusion

The concept of Meta Privacy has been formally defined and examined in this paper. Meta Privacy addresses the problem of no metadata and metastructure privacy protection considerations in currently proposed information privacy methods. The analysis of metadata and metastructure information found that they can be divided into one of two main Meta Privacy categories. That is, meta-information containing personal or identifiable information is classified as a Meta Privacy Risk. This type of meta-information should be very well protected. When meta-information is used to represent and enforce entity privacy policies and preferences they are classified as a Meta Privacy Benefit's. The meta-information should remain unlinkable and unobservable.

References

1. Schwartz, P.M.: Privacy and Democracy in Cyberspace. 52 VAND. L. REV. 1609 (1999) 1610-11.
2. Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y.: Hippocratic Databases. Proceedings of the 28th VLDB Conference, Hong Kong, China (2002).
3. Hes, R. and Borking, J.: Privacy-Enhancing Technologies: The path to anonymity. Registratiekamer, The Hague, August (2000).
4. Goldberg, I.: Privacy-enhancing technologies for the Internet, II: Five years later. PET2002, San Francisco, CA, USA 14 - 15 April (2002).
5. Clarke, R.: Introduction to Dataveillance and Information Privacy, and Definitions and Terms. http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html (1999).
6. Common Criteria: Common Criteria for Information Technology Evaluation. January, 2004, http:// www.commoncriteria.org (2004).
7. W3C: The platform for privacy preferences 1.0 (P3P1.0) specification, Jan., 2002. W3C Proposed Recommendation, http://www.w3.org/TR/P3P (2002).
8. Webopedia: Definition of Metadata – What is Metadata? http://www.webopedia.com/ TERM/m/metadata.html (1998).
9. Massacci, F., and Zannone, N.: Privacy is Linking Permission to Purpose. Technical Report University of Trento, Italy (2004).
10. Clarke, R.: Internet Privacy Concerns Confirm the Case for Intervention. ACM 42, 2 (February 1999) 60-67.
11. Rice, F.C.: Protecting Personal Data in your Microsoft Word Documents. MSDN Online Article August 2002. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnword2k2/html/odc_ProtectWord.asp (2002).
12. Extensible Markup Language (XML), World Wide Web Consortium (W3C), http://www.w3.org/ XML/
13. Ceravolo, P., Damiani, E., De Capitani di Vimercati, S., Fugazza, C., and Samarati, P.: Advanced Metadata for Privacy-Aware Representation of Credentials. PDM 2005, Tokyo, Japan (April 9 2005).
14. RDF Vocabulary Description Language (RDFS). World Wide Web Consortium. http://www.w3.org/ TR/rdf-schema/.
15. Web Ontology Language (OWL). World Wide Web Consortium. http://w3.org/2004/ OWL/
16. Agrawal, R., Kini, A., LeFevre, K., Wang, A., Xu, Y., and Zhou, D.: Managing Healthcare Data Hippocratically. Proc. of ACM SIGMOD Intl. Conf. on Management of Data (2004).

Error Oracle Attacks on Several Modes of Operation

Fengtong Wen[1,3], Wenling Wu[2], and Qiaoyan Wen[1]

[1] School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
wftwq@163.com, wqy@bupt.edu.cn
[2] State Key Laboratory of Information Security, Institute of Software, Chinese Academy of Sciences, Beijing 100080, China
wwl@is.iscac.ac.cn
[3] School of Science, Jinan University, Jinan 250022, China

Abstract. In [7] Vaudenay demonstrated side-channel attacks on CBC-mode encryption, exploiting a "valid padding" oracle. His work showed that several uses of CBC-mode encryption in well-known products and standards were vulnerable to attack when an adversary was able to distinguish between valid and invalid ciphertexts. In [2] [5] [6], Black, Paterson,Taekeon et al.generalized these attacks to various padding schemes of CBC-mode encryption and multiple modes of operation. In this paper, we study side-channel attacks on the CFB, CBC|CBC, CFB|CFB, CBC|CBC|CBC, CFB|CFB|CFB modes under the error oracle models, which enable an adversary to determine the correct message with knowledge of ciphertext. It is shown that an attacker can exploit an oracle to efficiently extract the corresponding position plaintext bits of any block if the target plaintext contains some fixed bits in a known position of one block.

1 Introduction

In [7], Vaudenay presented side-channel attacks on block cipher CBC-mode encryption under the padding oracle attacks models. This attack requires an oracle which on receipt of a ciphertext, decrypts it and replies to the sender whether the padding is valid or not. The attacker can know the return of the oracle. The result is that the attacker can recover the part or full plaintext corresponding to any block of ciphertext. Further research has been done by Black, who generalized Vaudenay's attack to other padding schemes and modes operations and concluded that most of padding schemes are vulnerable to this attack. In [5] Paterson employed a similar approach to analyze the padding methods of the ISO CBC-mode encryption standard and showed that the padding methods are vulnerable to this attack. In [6] Taekeon Lee etc. studied the padding oracle attacks on multiple mode of operation and showed that 12 out of total 36 double modes and 22 out of total 216 triple modes were vulnerable to the padding oracle attacks. In [3]Mithell presented another side channel attack on CBC-mode encryption under an error oracle model.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 62–67, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we study side-channel attacks on the CFB[4], CBC|CBC, CFB|CFB, CBC|CBC|CBC, CFB|CFB|CFB[1] mode under an error oracle models. We assume that an attacker has access to an error oracle operating under the fixed key and has intercepted a ciphertext encrypted in these modes under that key. The attacker's aim is to extract the plaintext bits for that ciphertext. We further assume that the attacker is able to choose the initialization vector when submitting ciphertexts to the oracle. Under the above assumption, our result is that the attacker can recover all the pointed position plaintext bits for any block under the error oracle models.

The paper is organized as follows. Preliminaries are presented in Section 2. Section 3 discusses error oracle attacks on several modes of operation. Finally, we summarize our conclusion in Section 4.

2 Preliminaries

2.1 Notations

$X_{i,j}$: The j-th byte of X_i.
X‖Y: The concatenation of strings X and Y.
O: The oracle to distinguish whether the deciphered plaintext has the ascertain status or not.
Valid or Invalid: Output of oracle, whether the deciphered plaintext has the ascertain status or not.

2.2 Error Oracle Attack

We suppose that the target plaintext contains a fixed byte in a known position and the nature of this structure is known to the attacker. In this scenario an attacker submits a target ciphertext to an oracle, the oracle decrypts it and replies to the attacker whether the target plaintext has the ascertain status or not. If the status present, the oracle O will return "valid". Otherwise it will return "invalid". There are many protocols that set certain bytes to fixed value, such as IPv4 and IPv6, which destination Address is a fixed value for a recipient.

2.3 Review of Mode Encryption

In this section, we will describe the CFB-mode and CBC|CBC mode encryption. Cipher feedback(CFB) is a mode of operation for an n-bit block cipher for encrypting data of arbitrary length. Let E_K be encryption operation of the block cipher under key K. Let D_K denote the inverse operation to E_K. CFB-mode encryption operates as follows:

1. P is divided into n-bit blocks $P_1, P_2, \cdots, P_q$.
2. An n bit number is chosen, at random, as the initialization vector IV.
3. Compute ciphertext block $C_1 = E_K(IV) \oplus P_1$, and

$$C_i = E_K(C_{i-1}) \oplus P_i, 2 \leq i \leq q.$$

4. The resulting $C = IV \parallel C_1 \parallel C_2 \cdots \| C_q$ is the CFB-encrypted ciphertext.

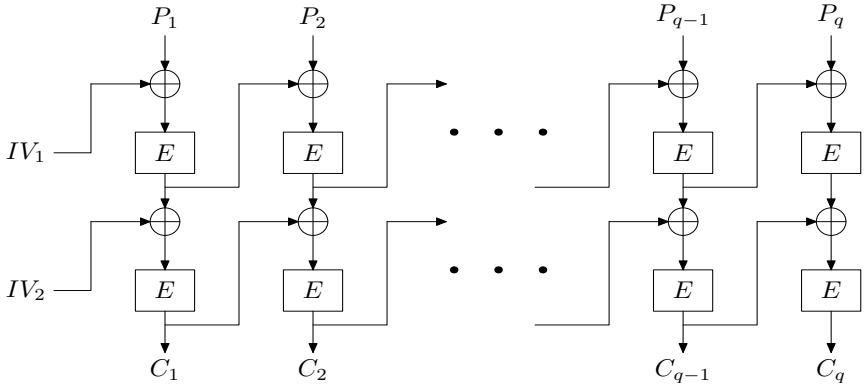


Fig. 1. CBC|CBC mode of operation

CBC|CBC-mode consists of double layers where each layer is the form CBC. Fig.1 provides an illustration of the CBC| CBC-mode. Similarly, CFB|CFB is a double mode of operation where each layer is the form CFB. CBC|CBC|CBC, CFB|CFB|CFB are triple modes of operation where each layer is the form CBC or CFB.

3 Error Oracle Attack on Mode of Operation

In error oracle model, we suppose that the target plaintext $P_1, P_2, \cdots, P_q$ corresponding to the target ciphertext $C_1, C_2, \cdots, C_q$ contains a fixed byte in a known position. Suppose that the fixed byte is the $P_{s,j}$ for some s. This scenario enables the attacker to learn the value of $P_{i,j}, 1 \leq i \leq q, i \neq s$ using a series of error oracle calls.

3.1 Error Oracle Attack on CFB-Mode

Let$C = IV \parallel C_1 \parallel C_2 \cdots \| C_q$ be the target ciphertext. For each $i(1 \leq i \leq q)$, the attacker constructs a series of "ciphertexts" with modification to the blocks C_{s-1} *and* C_s, where the modified ciphertext has the form:

$$C' = C_1, C_2, \cdots, C_{s-2}, C_{i-1}, C_i \oplus Q_t, C_{s+1}, \cdots, C_q \ for\ t = 0, 1, \cdots, 255.$$

The n-bit block Q_t has as its jth byte the 1-byte binary representation of t and zeros elsewhere. The attacker submits these ciphertexts to the error oracle in turn,until the oracle return "valid".i.e. the recovered plaintext $P'_1, P'_2, \cdots, P'_q$ for the manipulated ciphertext has the property that the $P'_{s,j}$ is the fixed value. If this occurs for Q_{t_0}, then the attcker immediately knows that $P_i = P'_s \oplus Q_{t_0}, P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j$. The reason is that $P'_s = C_i \oplus Q_{t_0} \oplus E_K(C_{i-1}) = P_i \oplus Q_{t_0}$. That is , given that the $P'_{s,j}$ is a fixed value, the attacker has discovered the value of the $P_{i,j}, (1 \leq i \leq q)$. The number of calls is about $128q$.

This is performed as follows:

1. For $t = 255$ to 0 do.

$$Q_t = 00, \cdots, 0 \quad \underbrace{(t)_2}_{j-th\ byte} \quad 0, 0, \cdots, 0$$

2. Pick $C' = C_1, C_2, \cdots, C_{s-2}, C_{i-1}, C_i \oplus Q_t, C_{s+1}, \cdots, C_q$
3. If $O(C') = valid$, then $P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j$
   Otherwise, go back to the first step for another t.

In section 3.2, 3.3, the process of the attack is similar to the above except for the construction of the ciphertext C'. So we mainly desctibe how to construct the modified ciphertext C' and how to obtain the $P_{i,j}$ from C' in the following discussion.

3.2 Error Attack on CBC|CBC, CFB|CFB Mode

Let $C = IV_1 \parallel IV_2 \parallel C_1 \parallel C_2 \cdots \| C_q$ be the target ciphertext. In CBC|CBC mode, the modified ciphertext has the form:

$$\begin{aligned}
C' &= C_1, C_2, \cdots, C_{s-3}, C_{i-2} \oplus Q_t, C_{i-1}, C_i, C_{s+1} \cdots, C_q, && i \geq 3 \\
C' &= C_1, C_2, \cdots, C_{s-3}, IV_2 \oplus Q_t, C_1, C_2, C_{s+1} \cdots, C_q, && i = 2 \\
C' &= C_1, C_2, \cdots, C_{s-3}, IV_1 \oplus Q_t, IV_2, C_1, C_{s+1} \cdots, C_q, && i = 1
\end{aligned}$$

When oracle O return "valid"$(t = t_0)$, we can obtain the $P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j$ from the equation:

$$\begin{aligned}
P'_s &= C_{i-2} \oplus Q_{t_0} \oplus D_K(D_K(C_i) \oplus C_{i-1}) \oplus D_K(C_{i-1}) = P_i \oplus Q_{t_0}, && i \geq 3 \\
P'_s &= IV_2 \oplus Q_{t_0} \oplus D_K(D_K(C_2) \oplus C_1) \oplus D_K(C_1) = P_2 \oplus Q_{t_0}, && i = 2 \\
P'_s &= IV_1 \oplus Q_{t_0} \oplus D_K(D_K(C_1) \oplus IV_2) \oplus D_K(IV_2) = P_1 \oplus Q_{t_0}, && i = 1
\end{aligned}$$

In CFB|CFB mode , C' has the form:

$$C' = C_1, C_2, \cdots, C_{s-3}, C_{i-2}, C_{i-1}, C_i \oplus Q_t, C_{s+1}, \cdots, C_q$$

When the oracle O return "valid"$(t = t_0)$, we can obtain the $P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j, 1 \leq i \leq q$ from the equation:

$$P'_s = C_i \oplus Q_{t_0} \oplus E_K(E_K(C_{i-2}) \oplus C_{i-1}) \oplus E_K(C_{i-1}) = P_i \oplus Q_{t_0}$$

We can use the same methods to attack 12 out of total 36 double modes.

3.3 Error Oracle Attack on CBC|CBC|CBC, CFB|CFB|CFB Mode

Let $C = IV_1 \parallel IV_2 \parallel IV_3 \parallel C_1 \parallel C_2 \cdots \| C_q$ be the target ciphertext. In CBC|CBC|CBC mode, the modified ciphertext has the form:

$$\begin{aligned}
C' &= C_1, C_2, \cdots, C_{s-4}, C_{i-3} \oplus Q_t, C_{i-2}, C_{i-1}, C_i, C_{s+1} \cdots, C_q, && i \geq 4 \\
C' &= C_1, C_2, \cdots, C_{s-4}, IV_3 \oplus Q_t, C_1, C_2, C_3, C_{s+1} \cdots, C_q, && i = 3 \\
C' &= C_1, C_2, \cdots, C_{s-4}, IV_2 \oplus Q_t, IV_3, C_1, C_2, C_{s+1} \cdots, C_q, && i = 2 \\
C' &= C_1, C_2, \cdots, C_{s-4}, IV_1 \oplus Q_t, IV_2, IV_3, C_1, C_{s+1} \cdots, C_q, && i = 1
\end{aligned}$$

When the oracle O return "valid"$(t = t_0)$, we can obtain the $P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j, 1 \le i \le q$ from the equation:

$$\begin{aligned}
P'_s &= C_{i-3} \oplus Q_{t_0} \oplus D_K[D_K(D_K(C_i) \oplus C_{i-1}) \oplus D_K(C_{i-1} \oplus C_{i-2})] \\
&\quad \oplus D_K[(D_K(C_{i-1}) \oplus C_{i-2}] \oplus D_K(C_{i-2}) = P_i \oplus Q_{t_0}, \qquad i \ge 4 \\
P'_s &= IV_3 \oplus Q_{t_0} \oplus D_K[D_K(D_K(C_3) \oplus C_2) \oplus D_K(C_2 \oplus C_1)] \oplus \\
&\quad D_K[(D_K(C_2) \oplus C_1] \oplus D_K(C_1) = P_3 \oplus Q_{t_0}, \qquad i = 3 \\
P'_s &= IV_2 \oplus Q_{t_0} \oplus D_K[D_K(D_K(C_2) \oplus C_1) \oplus D_K(C_1 \oplus IV_3)] \oplus \\
&\quad D_K[(D_K(C_1) \oplus IV_3] \oplus D_K(IV_3) = P_2 \oplus Q_{t_0}, \qquad i = 2 \\
P'_s &= IV_1 \oplus Q_{t_0} \oplus D_K[D_K(D_K(C_1) \oplus IV_3) \oplus D_K(IV_3 \oplus IV_2)] \oplus \\
&\quad D_K[(D_K(IV_3) \oplus IV_2] \oplus D_K(IV_2) = P_1 \oplus Q_{t_0}, \qquad i = 1
\end{aligned}$$

In the CFB|CFB|CFB mode, C' has the form:

$$C' = C_1, C_2, \cdots, C_{s-4}, C_{i-3}, C_{i-2}, C_{i-1}, C_i \oplus Q_t, C_{s+1}, \cdots, C_q$$

When the oracle O return "valid"$(t = t_0)$, we can obtain $P_{i,j} = P'_{s,j} \oplus (Q_{t_0})_j$, $1 \le i \le q$ from the equation:

$$\begin{aligned}
P'_s &= C_i \oplus Q_{t_0} \oplus E_K(C_{i-1}) \oplus E(C_{i-1} \oplus E_K(C_{i-2})) \oplus \\
&\quad E_K[E_K(C_{i-2}) \oplus C_{i-1} \oplus E_K(E_K(C_{i-3}) \oplus C_{i-2})] = P_i \oplus Q_{t_0}.
\end{aligned}$$

Similarly, we can attack 38 out of total 216 triple modes under the same oracle model.

4 Conclusions

In this paper,we study side-channel attacks on the CFB, CBC|CBC, CFB|CFB, CBC|CBC|CBC, CFB|CFB|CFB mode under the error oracle model, which enable an adversary to determine the correct message with knowledge of ciphertext. Investigation shows that the attacker can efficiently extract the corresponding position plaintext bits of every block if the target plaintext contains a fixed byte in a known position, these modes are vulnerable to the error oracle attack.Eventually,we conclude that 12 out of total 36 double modes and 38 out of total 216 triple modes are vulnerable to the same error oracle attack.

Acknowledgement

The authors would like to thank the referees for their comments and suggestions.This work was supported partially by the National Natural Science Foundation of China under Grant No.60373059 and 60373047; the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20040013007; the National Basic Research 973 Program of China under Grant No.2004CB318004.

References

1. Biham, E.: Cryptanalysis of Multiple Modes of Operation. Lecture Notes in Computer Science, Vol. 917. Springer-Verlag London.UK(1994) 278–292
2. Black, J. and Urtubia, H.: Side-Channel Attacks on Symmetric Encryption Schemes: The Case for Authenticated Encryption. In proc. of 11th USENIX Security Symposium. USENIX (2002) 327–338
3. Mitchell, Chris J.: Error Oracle Attacks on CBC Mode: Is There a Future for CBC Mode Encryption? . Lecture Notes in Computer Science, Vol. 3650. Springer-Verlag(2005) 244–258
4. National Bureau of Standards, DES Modes of Operation, FIPS-pub.46, National Bureau of Standards U.S Department of Commerce, Washington D.C(1980)
5. Paterson, G. and Arnold, Yau.: Padding Oracle Attacks on the ISO CBC Mode Encryption Standard. Lecture Notes in Computer Science, Vol. 2964. Springer-Verlag(2004) 305–323
6. Taekeon, L., Jongsung, K.:Padding Oracle Attacks on Multiple Modes of Operation. Lecture Notes in Computer Science, Vol. 3506. Springer-Verlag Berlin Heidelberg(2005) 343–351
7. Vaudenenay, S.: Security Flaws Induced by CBC Padding-Applications to SSL, IPSEC, WTLS. . . . Lecture Notes in Computer Science, Vol. 2332. Springer-Verlag (2002) 534–545

Stability of the Linear Complexity of the Generalized Self-shrinking Sequences*

Lihua Dong, Yong Zeng, and Yupu Hu

The Key Lab. of Computer Networks and Information Security, The Ministry of Education, Xidian University, Xi'an, ShaanXi Province 710071, P.R. China
Lih_dong@hotmail.com

Abstract. The stability of the linear complexity of the generalized self-shrinking sequences over GF(2) with period $N=2^{n-1}$ is investigated. The main results follow: The linear complexity of the periodic sequences obtained by either deleting or inserting one symbol within one period are discussed, and the explicit values for the linear complexity are given.

1 Introduction

Traditionally, binary sequences have been employed in numerous applications of communications and cryptography. Depending on the context, these sequences are required to possess certain properties such as long period, high linear complexity, equal number of zeros and ones and two-level autocorrelation[1].

The linear complexity $L(S)$ of an N-periodic sequence $S=(s_0,s_1,s_2,\ldots,s_{N-1})^\infty$ is 0 if S is the zero sequence and otherwise it is the length of the shortest linear feedback shift register that can generate S[2]. The Berlekamp-Massey algorithm needs $2L(S)$ sequence digits in order to determine $L(S)$[3]. Therefore, the linear complexity is a critical index for assessing the strength of a sequence against linear cryptanalytic attacks. However, the linear complexity is known to have such an instability caused by the three kinds of the minimum changes of a periodic sequence, i.e., one-symbol substitution[4], one-symbol insertion[5], or one-symbol deletion[6]. The bounds of the linear complexity have been also reported for the two-symbol-substitution case of an m-sequence[7,8]. In [9], the bounds of the linear complexity is given for a sequence obtained from a periodic sequence over GF(q) by either substituting, inserting, or deleting k symbols within one period. However, the lower bound is not tight.

In this paper, the bounds of the linear complexity are given for the sequence obtained from a generalized self-shrinking sequence by either deleting or inserting one symbol within one period. Here the generalized self-shrinking sequence has been proposed in [10] as follows.

Definition 1. Let $a=\ldots a_{-2}a_{-1}a_0a_1a_2\ldots$ be an m-sequence over $GF(2)$, with the least period 2^n-1. $G=(g_0,g_1,g_2,\ \ldots,g_{n-1})\in GF(2)^n$. Sequence $v=\ldots v_{-2}v_{-1}v_0v_1v_2\ldots$ such that $v_k=g_0a_k+g_1a_{k-1}+\ldots+g_{n-1}a_{k-n+1}$.Output v_k if $a_k=1$, or no output, otherwise $k=0,1,2,\ldots$.

* This work was supported in part by the Nature Science Foundation of China (No. 60273084) and Doctoral Foundation (No. 20020701013).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 68 – 73, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this way, we output a sequence $b_0b_1b_2\ldots$. We denote this sequence as $b(G)$ or $b(v)$. We call sequence $b(G)=b(v)$ a generalized self-shrinking sequence. We call the sequence family $B(a)=\{b(G), G\in GF(2)^n\}$, the family of generalized self-shrinking sequences based on m-sequence a.

The most interesting aspects of the generalized self-shrinking sequence include that the least period of each $b(G)$ is a factor of 2^{n-1}, thus each $b(G)$ has the linear complexity larger than 2^{n-2}; for each k, $0<k<n$, no more than $1/2^{n-k}$ of the sequences have the least period less than 2^k; no more than 1/4 of the sequences have the least period less than 2^{n-1}.

2 Preliminary

Let $A=(a_0,a_1,a_2,\ldots,a_{N-1})^\infty$ be a binary sequence with the period $N=2^{n-1}$. Now, $f(x)$ is called least polynomial of the sequence A if

$$\sum_{i\geq 0} a_i x^{-(i+1)} = \frac{a(x)}{x^N-1} = \frac{a(x)}{(x-1)^{2^{n-1}}} = \frac{g(x)}{f(x)} \tag{1}$$

and $g(x)/f(x)$ be the reduced rational fractions, $\deg(g(x))<\deg(f(x))$, and

$$a(x)=a_0x^{N-1}+a_1x^{N-2}+\ldots+a_{N-1}. \tag{2}$$

It is clear that $L(A^P)\leq N$, $f(x)=(x-1)^{L(A)}$, and $x-1$ does not divide $g(x)$.

Lemma 1. let $\bar{A}$ be the complementary sequence of the binary sequence A, that is $_i=a_i+1$ for all $i\geq 0$, then the sequence $\bar{A}$ and A have the same least polynomial.

Our derivation needs the following lemmas that stem from [11, Char 2, Char 3].

Definition 2. Let $f(x)$ be a nonzero polynomial over $GF(2)$. If $f(0)\neq 0$, then the least positive integer e for which $f(x)$ divides x^e-1 is called the order of f and denoted by $\mathrm{ord}(f)=\mathrm{ord}(f(x))$. If $f(0)=0$, then $f(x)=x^h g(x)$, where h is an integer and $g(x)$ is a nonzero polynomial over $GF(2)$, then $\mathrm{ord}(f)$ is defined to be $\mathrm{ord}(g)$.

Lemma 2. Let c be an integer, $f(x)$ be a polynomial over $F_2[x]$ with $f(0)\neq 0$, then $f(x)$ divides x^c-1 if and only if $ord(f)$ divides c. Let $g_1,\ldots,g_k$ *be* pairwise relatively prime nonzero polynomials over GF(2), and let $f=g_{1\ldots}g_k$. Then $ord(f)$ is equal to the lease common multiple of $ord(g_1),\ldots,ord(g_k)$.

Lemma 3. The number of monic irreducible polynomials in $F_q[x]$ of degree m and order e is equal to $\phi(e)/m$ if $e\geq 2$, and m is the multiplicative order of q *modulo e,* equal to 2 if $m=e=1$ and equal to 0 in all other cases. In particular, the degree of an irreducible polynomials in $F_q[x]$ of order e must be equal to the multiplicative order of q *modulo e.*

Lemma 4. For $1\leq e\leq m$, if $\gcd(2e,m)=\gcd(e,m)$, then $\gcd(2^e+1,2^m-1)=1$; else if $\gcd(2e,m)=2\gcd(e,m)$, then $\gcd(2^e+1,2^m-1)=2^{\gcd(e,m)}+1$.

3 One-Symbol Deletion

In this section, firstly, the bounds of the linear complexity are given for the sequence $\{\tilde{a}_i\}_{i\geq 0}$ obtained from a binary sequence $\{a_i\}_{i\geq 0}$ (for which the least period is $N=2^{n-1}$, the linear complexity L is larger than 2^{n-2}, thus $L>N/2$) by deleting one symbol within one period.

Theorem 1. Let the sequence $\{\tilde{a}_i\}_{i\geq 0}$ be obtained by one-symbol deletion at the end of one period of a generalized self-shrinking sequence $\{a_i\}_{i\geq 0}$, then we have

- $L(\{\tilde{a}_i\})=N-2$ if $ord(g(x))$ is relatively prime to $N-1$,
- $LC(\{\tilde{a}_i\})=(N-1)-1-k(n-1)=N-k(n-1)-2$ if $ord(g(x))$ is not relatively prime to $N-1$ and $N-1$ is a prime, and $LC(\{\tilde{a}_i\})=(N-1)-1-v(n-1)-\sum_{i=1}^{u} p_i$ if $ord(g(x))$ is not relatively prime to $N-1$ and $N-1$ is not a prime.

Where

k is the number of the irreducible factors for the polynomial $h(x)$.

u is the number of the monic irreducible factors $h_j(x)$ for the polynomial $h(x)$ that $\mathrm{ord}(h_j(x))$ divides $N-1$ and is of the type as $2^{pj}-1$, with $j=1,2,\ldots,u$.

v is the number of the monic irreducible factors $h_t(x)$ for the polynomial $h(x)$ that $\mathrm{ord}(h_t(x))$ divides $N-1$ and is not of the type as $2^{pt}-1$, with $t=1,2,\ldots,v$.

Proof :

1. if $a_{N-1}=0$, since $\tilde{a}(x)=a_0x^{N-2}+a_1x^{N-3}+\cdots+a_{N-2}=x^{-1}a(x)$, and from (1) and (2), we know that $x|g(x)$, that is $g(x)=x^ih(x)$, where $(x-1)\nmid h(x)$, $i\geq 1$, thus

$$\frac{\tilde{a}(x)}{x^{N-1}-1}=\frac{a(x)}{x(x^{N-1}-1)}=\frac{g(x)\dfrac{x^N-1}{f(x)}}{x(x^{N-1}-1)}=\frac{x^{i-1}h(x)\dfrac{x^N-1}{f(x)}}{x^{N-1}-1}, \tag{3}$$

Where $\gcd(x^N-1,x^{N-1}-1)=x-1$.

- $L(\{\tilde{a}_i\})=N-2$ if $\mathrm{ord}(h(x))$ is relatively prime to $N-1$.
- $LC(\{\tilde{a}_i\})=(N-1)-1-k(n-1)=N-k(n-1)-2$ and $\mathrm{ord}(h(x))=N-1$ if $N-1$ is a prime (for example $2^{31}-1,2^{127}-1$) and $ord(h(x))$ is not relatively prime to $N-1$.
  - Let $h(x)=h_1(x)h_2(x)\ldots h_k(x)$ and $h_1(x),\ldots,h_k(x)$ be monic irreducible polynomials. Then we have $\mathrm{ord}(h_1(x))=\mathrm{ord}(h_2(x))=\ldots=\mathrm{ord}(h_k(x))=N-1$. For if there are some $i\neq j$, and $\mathrm{ord}(h_i(x))\neq\mathrm{ord}(h_j(x))$, according to *lemma* 2 $\mathrm{ord}(h_i(x))$ and $\mathrm{ord}(h_j(x))$ are proper factors of $N-1$, but $N-1$ is a prime, contradiction. $\deg(h_i(x))=n-1$ with $i=1,2,\ldots,k$ and there are $\phi(2^{n-1}-1)/(n-1)$ monic irreducible polynomials over $F_2[x]$ of degree $n-1$ and order $2^{n-1}-1$ according to *lemma* 3. Thus from (4) the result follows.
- $LC(\{\tilde{a}_i\})=(N-1)-1-v(n-1)-\sum_{i=1}^{u} p_i$ if $N-1=2^{n-1}-1$ is not a prime and $ord(h(x))$ is not relatively prime to $N-1$ (that is, $h(x)$ is not relatively prime to $x^{N-1}-1$).

  Let $h(x)=h_1(x)h_2(x)\ldots h_k(x)$, here $h_i(x)$ is monic irreducible polynomial, $\deg(h_i(x))=m_i$ and $\mathrm{ord}(h_i(x))=e_i$ with $i=1,2,\ldots,k$. Let $n-1=p_1\,p_2\ldots p_w$, here p_j is a prime, then $2^{pj}-1$ divides $N-1$, with $j=1,2,\ldots,w$.

- $m_i=p_i$ and there are $\phi(2^{pi}-1)/p_i$ monic irreducible polynomials over $F_2[x]$ of degree p_i and order $2^{pi}-1$ if e_i is of the type as $2^{pi}-1$ according to *lemma* 3.
- $m_i=n-1$ and there are $\phi(e_i)/(n-1)$ monic irreducible polynomials over $F_2[x]$ of degree $n-1$ and order e_i if e_i is a factor of $N-1=2^{n-1}-1$ and of other types according to *lemma* 3.

According to the analysis above the result follows if the factors of $h(x)$ are satisfied that there are u monic irreducible polynomials $h_{ij}(x)$ such that $\mathrm{ord}(h_{ij}(x))$ divides $N-1$ and is of the type as $2^{pj}-1$, with $j=1,2,\ldots,u$ and v monic irreducible polynomials $h_{it}(x)$ such that $\mathrm{ord}(h_{it}(x))$ divides $N-1$ and is not of the type as $2^{pt}-1$, with $t=1,2,\ldots,v$.

2. If $a_{N-1}=1$, then $_{N-1}=0$, thus we may estimate the linear complexity $L(\{\widetilde{\widetilde{a}}_i\})$ for the complement sequence $\{\bar{a}_i\}_{i\geq 0}$ of the generalized self-shrinking sequence $\{a_i\}_{i\geq 0}$, from lemma 1 we know that $L(\{\widetilde{\widetilde{a}}_i\})$ has the same distribution as $L(\{\tilde{a}_i\})$. #

Lemma 5[7]**.** The shifted equivalent sequences have the same least polynomial and thus the same linear complexity.

Used *lemma* 5, the same results can be obtained if the deleted symbol does not lie in the end of one period of the sequence.

4 One-Symbol Insertion

In this section, the bounds of the linear complexity are given for the sequence $\{\tilde{a}_i\}_{i\geq 0}$ obtained from a generalized self-shrinking sequence $\{a_i\}_{i\geq 0}$ by inserting one symbol within one period.

Theorem 2. Let $\{\tilde{a}_i\}_{i\geq 0}$ be a one-symbol insertion of a generalized self-shrinking sequence $\{a_i\}_{i\geq 0}$ (for which the least period is $N=2^{n-1}$), then we have

- $LC(\{\tilde{a}_i\})=N$ if $ord(g(x))$ is relatively prime to $2^{n-1}+1$,
- $LC(\{\tilde{a}_i\})=(N+1)-1-2k(n-1)=N-2k(n-1)$ if $ord(g(x))$ is not relatively prime to $2^{n-1}+1$ and $2^{n-1}+1$ is a prime, and $LC(\{\tilde{a}_i\})=(N+1)-1-2v(n-1)-2\sum_{i=1}^{u} p_i$ if $ord(g(x))$ is not relatively prime to $2^{n-1}+1$ and $2^{n-1}+1$ is not a prime.

Where

k is the number of the irreducible factors for the polynomial $g(x)$,

u is the number of the monic irreducible factors $g_j(x)$ for the polynomial $g(x)$ that $\mathrm{ord}(g_j(x))$ divides $N+1$ and is of the type as $2^{pj}+1$, with $j=1,2,\ldots,u$,

v is the number of the monic irreducible factors $g_t(x)$ for the polynomial $g(x)$ that $\mathrm{ord}(g_t(x))$ divides $N+1$ and is not of the type as $2^{pt}+1$, with $t=1,2,\ldots,v$.

Proof :

1. If $b=0$, since $\tilde{a}(x)=a_0x^N+a_1x^{N-1}+\cdots+a_{N-1}x+b=xa(x)$, thus

$$\frac{\tilde{a}(x)}{x^{N+1}-1}=\frac{xa(x)}{x^{N+1}-1}=\frac{xg(x)\dfrac{x^N-1}{f(x)}}{x^{N+1}-1}$$

Where $\gcd(x^N-1,x^{N+1}-1)=x-1$ and $x-1$ does not divide $g(x)$.

- $LC(\{\tilde{a}_i\})=N$ and $g(x)$ is relatively prime to x^{N+1}-1 if gcd($ord(g(x))$,N+1)=1,
- $LC(\{\tilde{a}_i\})=(N+1)-1-2k(n-1)= N-2k(n-1)$ if $ord(g(x))$ is not relatively prime to N+1 and N+1 is a prime.
  - Let $g(x)=g_1(x)g_2(x)\ldots g_k(x)$ and $g_i(x)$ be monic irreducible polynomial, with i=1,2,…,k. Then we have ord($g_i(x)$)=N+1 with i=1,2,…,k. For if there are some $i\neq j$, and ord($g_i(x)$)≠ord($g_j(x)$), according to *lemma* 2 ord($g_i(x)$) and ord($g_j(x)$) are proper factors of N+1, but N+1 is a prime, a contradiction. Assuming that deg($g_i(x)$)=m_i, then m_i is the multiplicative order of 2 module 2^{n-1}+1 according to *lemma* 3, and m_i=2(n-1) according to *lemma* 4, with i=1,2,…,k and there are $\phi(N+1)/2(n-1)$ monic irreducible polynomials over $F_2[x]$ of degree 2(n-1) and order N+1. Thus $LC(\{\tilde{a}_i\})=(N+1)-1-2k(n-1)=N-2k(n-1)$.
- $LC(\{\tilde{a}_i\})=(N+1)-1-2v(n-1)-2\sum_{i=1}^{u} p_i$ if $ord(g(x))$ is not relatively prime to N+1, (that is, $g(x)$ is not relatively prime to x^{N+1}-1), and N+1=2^{n-1}+1 is not a prime.
  Let $g(x)=g_1(x)g_2(x)\ldots g_k(x)$, deg($g_i(x)$)=$m_i$, and ord($e_i$) with i=1,2,…,k, and n-1=p_1 $p_2\ldots p_w$, here p_j is a prime, then 2^{pj}+1 divides N+1, with j=1,2,…,w.
  - m_i is the multiplicative order of 2 module 2^{pi}+1 if e_i is of the type as 2^{pi}+1 according to *lemma* 3 and thus $m_i=2p_i$ according to lemma 4 with i=1,2,…,k and there are $(2^{pi}+1)/2p_i$ monic irreducible polynomials over $F_2[x]$ of degree $2p_i$ and order 2^{pi}+1.
  - m_i=2(n-1) if e_i is a factor of 2^{n-1}+1 and not of the type as 2^{pi}+1, and there are $\phi(e_i)/2(n-1)$ monic irreducible polynomials over $F_2[x]$ of degree 2(n-1) and order e_i.

  According to the analysis above the result follows if the factors of $g(x)$ are satisfied that there are u monic irreducible polynomials $g_j(x)$ such that ord($g_j(x)$) divides N+1 and is of the type as 2^{pj}+1, with j=1,2,…,u and v monic irreducible polynomials $g_t(x)$ such that ord($g_t(x)$) divides N+1 and is not of the type as 2^{pt}+1, with t=1,2,…,v.

2. If the bit inserted into the sequence $\{a_i\}_{i\geq 0}$ is equal to 1, then the bit inserted into the sequence $\{\tilde{a}_i\}_{i\geq 0}$ is equal to 0. Following the lemma 1, $L(\{\tilde{\tilde{a}}_i\})$ has the same distribution as that of $L(\{\tilde{a}_i\})$. #

The same results can be obtained if the inserted symbol does not lie in the end of one period of the sequence. Using *lemma* 5 we know that all one needs to do is to transform the sequence into a shifted equivalent sequence such that the deleting symbol lies in the end of one period of the sequence.

5 Conclusions

The linear complexity of the sequence $\{\tilde{a}_i\}_{i\geq 0}$ obtained from a generalized self-shrinking sequence $\{a_i\}_{i\geq 0}$ by one-symbol insertion, one-symbol deletion within one period is given. The generalization of the results to arbitrary sequences with large linear complexity is desirable.

References

1. E. L.Key, "An analysis of the structure and complexity of nonlinear binary sequence generators," *IEEE Trans. on Information Theory*, 22, (1976) 732-736
2. R. A. Rueppel, "Stream ciphers," in Contemporary Cryptology: The Science of *Information Integrity*, G. J. Simmons, Ed. New York: IEEE, 1992, pp. 65–134.
3. J. Massey, Shift-register synthesis and BCH decoding, *IEEE Trans. on Information Theory*, 15, (1969)122-127
4. Z. Dai and Imamura K, "Linear complexity for one-symbol substitution of a periodic sequence over *GF(q)*," *IEEE Trans. on Information Theory*, 44, (1998)1328-1331
5. Uehara S and Imamura K, "Linear complexity of periodic sequences obtained from *GF(q)* sequences with period q^n-1 by one-symbol insertion," *IEICE Trans. Fundamentals*, vol. E79-A, (1996)1739-1740
6. Uehara S and Imamura K, "Linear complexity of periodic sequences obtained from *GF(q)* sequences with period q^n-1 by one-symbol deletion," *IEICE Trans. Fundamentals*, vol. E80-A, (1997)1164-1166
7. Ye. D and Z Dai, "Linear complexity of periodic sequences under two symbol substitutions," in *Advances in Cryptology-China Crypto'96*. Beijing, China: Science Press, pp.7-9, in Chinese.
8. Uehara S and Imamura K, "Linear complexity of periodic sequences obtained from an m-sequence by two-symbol substitution," in *Proc. 1998 Int. Symp. Information Theory and It's Applications (ISITA'98)*, Mexico City, Mexico, Oct. 1998, pp.690-692.
9. Shaoquan Jiang, Zongduo Dai, and Kyoki Imamura, "Linear complexity of periodic sequences obtained from a Periodic Sequence by Either Substituting, Inserting, or Deleting *k* Symbols Within One Period", *IEEE Trans Inform. Theory*, May 2000, Vol 46, No 3, pp: 1174-1177
10. Yupu Hu and Guozhen Xiao. Generalized Self-shrinking Generator. *IEEE Transaction on Information Theory*, 50(4), (2004)714-719
11. Lidl. R and Niederreiter. H, "Finite fields," in *Encyclopedia of Mathematics and Its Applications*, vol. 20. Reading, MA: Addison-Wesley, 1983.

On the Construction of Some Optimal Polynomial Codes

Yajing Li and Weihong Chen

Department of Applied Mathematics, Information Engineering University,
P.O. Box 1001-7410, Zhengzhou 450002, China
yajingli2003@163.com

Abstract. We generalize the idea of constructing codes over a finite field $\mathbf{F}_q$ by evaluating a certain collection of polynomials at elements of an extension field of $\mathbf{F}_q$. Our approach for extensions of arbitrary degrees is different from the method in [3]. We make use of a normal element and circular permutations to construct polynomials over the intermediate extension field between $\mathbf{F}_q$ and $\mathbf{F}_{q^t}$ denoted by $\mathbf{F}_{q^s}$ where s divides t. It turns out that many codes with the best parameters can be obtained by our construction and improve the parameters of Brouwer's table [1]. Some codes we get are optimal by the Griesmer bound.

1 Introduction

Let q be a prime power and $\mathbf{F}_q$ be the finite field with q elements. Constructions of generalized Reed-Solomon codes over $\mathbf{F}_q$ only employ elements of $\mathbf{F}_q$, hence their lengths are at most q. Firstly, we recall an alternative description of generalized Reed-Solomon codes. Let $1 \le n \le q$, $1 \le k \le n$ and let

$$P_k = \{f(x) \in \mathbf{F}_q[x]:\ \deg f(x) < k\}.$$

Choose n distinct elements $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbf{F}_q$, and n non-zero elements $v_1, v_2, \ldots, v_n$ in $\mathbf{F}_q$. The generalized Reed-Solomon code is defined by

$$GRS_q(n, k) = \{\ (v_1 f(\alpha_1), v_2 f(\alpha_2), \ldots, v_n f(\alpha_n)):\ f(x) \in P_k\ \}.$$

It is well known that the code $GRS_q(n, k)$ is a linear code with the parameters $[n, k, n-k+1]$ over $\mathbf{F}_q$. Therefore it is an MDS code. However, its length is limited by q. In order to increase the length of the codes, the idea of the construction described in [2] is to choose a special collection of polynomials over $\mathbf{F}_q$ that return elements in $\mathbf{F}_q$ when they are evaluated at elements of $\mathbf{F}_{q^2}$. A generalization of the idea was considered in [6] and obtained results for extensions of prime degrees. Another generalization of the idea of constructing linear codes that uses symmetric polynomials was given in [3]. The codes constructed with their method have as good a lower bound on the minimum distance as the ones listed in Brouwer's table [1].

We generalize the idea of constructing codes over a finite field $\mathbf{F}_q$ by evaluating a certain collection of polynomials at elements of an extension field of $\mathbf{F}_q$. The method we employ for extensions of arbitrary degrees is different from the approach in [3]. We make use of a normal element and circular permutations to construct polynomials over the intermediate extension field between $\mathbf{F}_q$ and $\mathbf{F}_{q^t}$ denoted by $\mathbf{F}_{q^s}$ where s divides t. We will show that many codes with the best parameters can be obtained through our construction and improve the parameters of Brouwer's table [1]. Some codes we get are optimal by the Griesmer bound.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 74 – 79, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Preliminaries

Let $\mathbf{F}_q$ be the finite field with q elements and let $\mathbf{F}_{q^t}$ be the finite extension of $\mathbf{F}_q$ of degree t. Two elements α, β are said to be conjugates over $\mathbf{F}_q$ if they are roots of the same irreducible polynomial over $\mathbf{F}_q$, equivalently if $\beta = \sigma^i(\alpha)$ for some integer i where σ is the Frobenius map $\sigma(x) = x^q$ which generates the Galois group of the extension. For an irreducible polynomial $f(x) \in \mathbf{F}_q[x]$ of degree n, let α be a root of $f(x)$ in an extension field of $\mathbf{F}_q$. Then the conjugacy class of α is the set $\{\alpha, \alpha^q, \ldots, \alpha^{q^{n-1}}\}$ of size n.

Definition 1. (see [7]) *Let* $\mathrm{K} = \mathbf{F}_q$ *and* $\mathrm{F} = \mathbf{F}_{q^m}$. *Then a basis of* F *over* K *of the form* $\{\alpha, \alpha^q, \ldots, \alpha^{q^{m-1}}\}$, *consisting of a suitable element* $\alpha \in \mathrm{F}$ *and its conjugates with respect to* K, *is called a normal basis of* F *over* K. *The element* α *is called a normal element.*

For any finite field K and any finite extension F of K, there exists a normal basis of F over K. Let s be a divisor of t and $\mathbf{F}_{q^s}$ be the extension over $\mathbf{F}_q$ of degree s. We have $\mathbf{F}_q \subseteq \mathbf{F}_{q^s} \subseteq \mathbf{F}_{q^t}$.

The circular n – permutation $\odot a_1 a_2 \ldots a_n$ can be denoted by an infinite sequence $a_1 a_2 \ldots a_n a_{n+1} a_{n+2} \ldots$, where $a_{n+i} = a_n$ for all $i \in \mathbf{N}$.

Definition 2. *The period of the circular* n *– permutation* $\odot a_1 a_2 \ldots a_n$ *is defined to be the least positive integer* d *such that* $a_{d+i} = a_i$ *for all* $i \in \mathbf{N}$ *in the infinite sequence* $a_1 a_2 \ldots a_n a_{n+1} a_{n+2} \ldots$

It is easy to see that the period d divides n and the circular n – permutation $\odot a_1 a_2 \ldots a_n$ of period d can give rise to d different linear n – permutations.

Lemma 1. *Let S be a multi-set with r distinct objects each with an infinite repetition number. Then the number of circular* n *– permutations* $\odot a_1 a_2 \ldots a_n$ *of S is* $C_r(n) = \sum_{d \mid n} \frac{1}{d} \left(\sum_{d' \mid d} \mu\left(\frac{d}{d'}\right) r^{d'} \right)$ *where* μ *is the Möbius function.*

Moreover, $C(d) = \frac{1}{d} \sum_{d' \mid d} \mu\left(\frac{d}{d'}\right) r^{d'}$ *is the number of the circular n– permutations* $\odot a_1 a_2 \ldots a_n$ *of period d.*

For an integer $0 \le m \le q$, let the distinct objects of S be the m integers 0, 1, …, m–1 so that $S = \{\infty \cdot 0, \infty \cdot 1, \ldots, \infty \cdot (m-1)\}$. Let A denote the set of circular t– permutations of S. We denote the set of the circular t – permutations of S of period d by $A(d)$. It is easy to see that d divides t and $A = \bigcup_{d \mid t} A(d)$.

Let τ denote the left-cyclic-shift operation: $\tau(i_1, i_2, \ldots, i_t) = (i_2, \ldots, i_t, i_1)$ and τ^n denote the n-fold composition of τ, cyclic shift by n positions . We know that the circular t – permutation $\odot a_1 a_2 \ldots a_t$ of period d can give rise to d different linear t –permutations which are the cyclic shift of $a_1 a_2 \ldots a_t$. The set containing the d different linear t – permutations $a_1 a_2 \ldots a_t$, $a_2 a_3 \ldots a_1$, …, $a_d a_1 \ldots a_{d-1}$ is denoted by $M_{a_1 a_2 \cdots a_t}(d)$.

Choose an arbitrary circular t – permutation $\odot i_1 i_2 \ldots i_t \in A(d)$ where d is divisible by s and let $\alpha \in \mathbf{F}_{q^s}$ be a normal element, i.e., $\{\alpha, \alpha^q, \ldots, \alpha^{q^{s-1}}\}$ is a normal basis of $\mathbf{F}_{q^s}$ over $\mathbf{F}_q$. For $k = 0, 1, \ldots, s-1$, we define s polynomials as follows :

$$e^{(k)}_{i_1 i_2 \cdots i_t}(x) = \sum_{j=0}^{d-1} \alpha^{q^{j+k}} x^{(q^{t-1}, q^{t-2}, \cdots, q, 1) \bullet \tau^j (i_1, i_2, \cdots, i_{t-1}, i_t)}$$

where $\bullet$ denotes the usual dot product.

For $d = 1$, i.e., $i_1 = i_2 = \ldots = i_t$, let

$$e_{i_1 i_2 \cdots i_t}(x) = x^{q^{t-1} i_1 + q^{t-2} i_2 + \cdots + q i_{t-1} + i_t} = x^{\frac{q^t - 1}{q-1} i_1}$$

Remark. When $s = 1$ and $t = 2$, it is easy to know $\alpha = 1$ for this case. The polynomials we defined above are exactly the same as the polynomials constructed in [2].

Proposition 1. (1) *For* $k = 0, 1, \ldots, s-1$, $e^{(k)}_{i_1 i_2 \cdots i_t}(x) \in \mathbf{F}_{q^s}[x]$, *and*

$$\deg e^{(k)}_{i_1 i_2 \cdots i_t}(x) = \max_{j_1 j_2 \cdots j_t \in M_{i_1 i_2 \cdots i_t}(d)} \{q^{t-1} j_1 + q^{t-2} j_2 + \cdots + j_t\} \le \frac{q^t - 1}{q-1}(m-1).$$

(2) $e^{(k)}_{i_1 i_2 \cdots i_t}(\beta)$ *is an element of* $\mathbf{F}_q$ *for any* $\beta \in \mathbf{F}_{q^t}$.

Proof. (1) It is easy to know from the definition.

(2) For any $\beta \in \mathbf{F}_{q^t}$,

$$\left(e^{(k)}_{i_1 i_2 \cdots i_t}(\beta)\right)^q = \left(\sum_{j=0}^{d-1} \alpha^{q^{j+k}} \beta^{(q^{t-1}, q^{t-2}, \cdots, q, 1) \bullet \tau^j (i_1, i_2, \cdots, i_{t-1}, i_t)}\right)^q$$

$$= \sum_{j=0}^{d-1} \alpha^{q^{j+k}} \beta^{(q^{t-1}, q^{t-2}, \cdots, q, 1) \bullet \tau^j (i_1, i_2, \cdots, i_{t-1}, i_t)} = e^{(k)}_{i_1 i_2 \cdots i_t}(\beta)$$

So $e^{(k)}_{i_1 i_2 \cdots i_t}(\beta) \in \mathbf{F}_q$.

Proposition 2. (1) *For* $k = 0, 1, \ldots, s-1$, $e^{(k)}_{i_1 i_2 \cdots i_t}(x)$ *are* $\mathbf{F}_q$ *– linearly independent.*

(2) *For* $\odot i_1 i_2 \ldots i_t \neq \odot j_1 j_2 \ldots j_t$, *the set of* $2s$ *polynomials* $\{ e^{(k_1)}_{i_1 i_2 \cdots i_t}(x), e^{(k_2)}_{j_1 j_2 \cdots j_t}(x) \mid 0 \le k_1, k_2 \le s-1 \}$ *is* $\mathbf{F}_q$ *– linearly independent.*

Proof. (1) It is easy to see that $e^{(k)}_{i_1 i_2 \cdots i_t}(x)$, $k = 0, 1, \ldots, s-1$ have the same degree $\max_{j_1 j_2 \cdots j_t \in M_{i_1 i_2 \cdots i_t}(d)} \{q^{t-1} j_1 + q^{t-2} j_2 + \cdots + j_t\}$. If α is the leading coefficients of $e^{(i)}_{i_1 i_2 \cdots i_t}(x)$ for some $0 \le i \le s-1$, then α^q is the leading coefficients of $e^{(i+1)}_{i_1 i_2 \cdots i_t}(x)$. It means that the leading coefficients of $e^{(k)}_{i_1 i_2 \cdots i_t}(x)$, $k = 0, 1, \ldots, s-1$ is the permutation

of $\{\alpha, \alpha^q, \ldots, \alpha^{q^{s-1}}\}$. Since $\{\alpha, \alpha^q, \ldots, \alpha^{q^{s-1}}\}$ is a normal basis of $\mathbf{F}_{q^s}$ over $\mathbf{F}_q$. It follows that $e^{(k)}_{i_1 i_2 \cdots i_t}(x)$, $k = 0, 1, \ldots, s-1$ are $\mathbf{F}_q$ – linearly independent.

(2) It suffices to show that $e^{(k)}_{i_1 i_2 \cdots i_t}(x)$ and $e^{(k)}_{j_1 j_2 \cdots j_t}(x)$ have different degrees. We suppose that $\deg(e^{(k)}_{i_1 i_2 \cdots i_t}(x)) = \deg(e^{(k)}_{j_1 j_2 \cdots j_t}(x))$. In other words, there exists $u_1 u_2 \ldots u_t$ and $v_1 v_2 \ldots v_t$ such that

$$q^{t-1}u_1 + q^{t-2}u_2 + \cdots + u_t = q^{t-1}v_1 + q^{t-2}v_2 + \cdots + v_t$$

It follows that $u_t \equiv v_t \bmod q$. Since $0 \le u_t, v_t \le m-1 \le q-1$, we have that $u_t = v_t$. Then it follows that $u_{t-1} \equiv v_{t-1} \bmod q$ and hence $u_{t-1} = v_{t-1}$. Continuing the same argument, we obtain $u_i = v_i$ for all $1 \le i \le t$. This is a contradiction. Thus, the set of $2s$ polynomials $\{ e^{(k_1)}_{i_1 i_2 \cdots i_t}(x),\ e^{(k_2)}_{j_1 j_2 \cdots j_t}(x) \mid 0 \le k_1, k_2 \le s-1 \}$ is $\mathbf{F}_q$– linearly independent.

3 Construction of Codes

Let $\mathbf{F}_{q^t}$ be the finite extension of $\mathbf{F}_q$ of arbitrary degree t. Let p be a prime divisor of t and $\mathbf{F}_{q^s}$ be the extension over $\mathbf{F}_q$ of degree $s = t/p$. There are no intermediate extensions (strictly) between $\mathbf{F}_{q^s}$ and $\mathbf{F}_{q^t}$. Firstly, we list all elements of $\mathbf{F}_{q^s}$ as $\mathbf{F}_{q^s} = \{\alpha_1, \alpha_2, \ldots, \alpha_{q^s}\}$. For $\beta \in \mathbf{F}_{q^t} - \mathbf{F}_{q^s}$, its conjugacy class is the set $\{\beta, \beta^{q^s}, \cdots, \beta^{q^{s(p-1)}}\}$ of size p. There are exactly $(q^t - q^s)/p$ such conjugacy classes. Therefore, we can list the elements of $\mathbf{F}_{q^t}$ as follows:

$$\{\alpha_1, \alpha_2, \cdots, \alpha_{q^s}, \beta_1, \beta_1^{q^s}, \cdots, \beta_1^{q^{s(p-1)}}, \cdots, \beta_r, \beta_r^{q^s}, \cdots, \beta_r^{q^{s(p-1)}}\}$$

where $r = (q^t - q^s)/p$.

We choose the circular t – permutation $\odot i_1 i_2 \ldots i_t \in A(d)$ where d is divisible by s, i.e., $d = s$ or $d = t$. For $0 \le m \le q$, we define the $\mathbf{F}_q$ – linear space V_m to be the $\mathbf{F}_q$ – linear span of the set

$$E_m = \bigcup_{\substack{\odot i_1 i_2 \cdots i_t \in A(d) \\ s \mid d}} \{e^{(k)}_{i_1 i_2 \cdots i_t}(x), 0 \le k \le s-1\} \bigcup \bigcup_{d=1} e_{i_1 i_2 \cdots i_t}(x)$$

V_m has the property that $f(\beta) \in \mathbf{F}_q$ for any $f(x) \in V_m$ and $\beta \in \mathbf{F}_{q^t}$ since $e^{(k)}_{i_1 i_2 \cdots i_t}(\beta) \in \mathbf{F}_q$. As the polynomials in E_m are $\mathbf{F}_q$ – linearly independent, we have

$$\dim(V_m) = |E_m| = \left(\frac{1}{s} \sum_{d' \mid s} \mu\left(\frac{s}{d'}\right) m^{d'} + \frac{1}{t} \sum_{d'' \mid t} \mu\left(\frac{t}{d''}\right) m^{d''} \right) s + m.$$

For $0 \le l \le q^s$, $1 \le b \le r = (q^t - q^s)/p$ and $1 \le m < 1 + \dfrac{bp(q-1)}{q^t - 1}$, we define a linear code $C_q(l, b, m)$ over $\mathbf{F}_q$ as follows:

$$C_q(l,b,m)=\{(f(\alpha_1),f(\alpha_2),\cdots,f(\alpha_l),f(\beta_1),f(\beta_2),\cdots,f(\beta_b))\mid f(x)\in V_m\} \quad (3.1)$$

This is obviously a linear code of length $n = l + b \le q^s + (q^t - q^s)/p$ over $\mathbf{F}_q$.

Theorem 1. *For* $1\le m<1+\frac{bp(q-1)}{q^t-1}$, *let* $C_q(l, b, m)$ *be the linear code defined in* (3.1). *Then the code* $C_q(l, b, m)$ *has parameters* $[n, k, d]$, *where*

$$n = l + b,$$

$$k=\left(\frac{1}{s}\sum_{d'\mid s}\mu\left(\frac{s}{d'}\right)m^{d'}+\frac{1}{t}\sum_{d''\mid t}\mu\left(\frac{t}{d''}\right)m^{d''}\right)s+m,$$

$$d\ge b-\frac{1}{p}\left(\frac{(m-1)(q^t-1)}{q-1}-l\right).$$

Proof. Let $f(x)$ be an arbitrary nonzero polynomial of V_m. Suppose that $f(x)$ has r_1 roots among $\{\alpha_1, \alpha_2, \ldots, \alpha_l\}$ and r_2 roots among $\{\beta_1, \beta_2, \ldots, \beta_b\}$. Then $r_1 \le l$ and $r_1 + p\, r_2 \le \deg f(x) \le (m-1)(q^t-1)/(q-1)$. This implies that, for $c_f = (f(\alpha_1),f(\alpha_2),\cdots,f(\alpha_l),f(\beta_1),f(\beta_2),\cdots,f(\beta_b)) \in C_q(l, b, m)$, its hamming weight $\mathrm{wt}(c_f)$ satisfies

$$\mathrm{wt}(c_f)\ge n-(r_1+r_2)\ge l+b-\frac{1}{p}\left(\frac{(m-1)(q^t-1)}{q-1}+(p-1)l\right)$$

$$=b-\frac{1}{p}\left(\frac{(m-1)(q^t-1)}{q-1}-l\right).$$

hence $d\ge b-\frac{1}{p}\left(\frac{(m-1)(q^t-1)}{q-1}-l\right)$.

Consider the $\mathbf{F}_q$ – linear map

$$\phi:\ V_m \to C_q(l, b, m)$$

$$f(x) \mapsto (f(\alpha_1),f(\alpha_2),\cdots,f(\alpha_l),f(\beta_1),f(\beta_2),\cdots,f(\beta_b))$$

Suppose $\phi(f)$ is the zero vector. Then the number of zeros of $f(x)$ is at least pb which is greater than $\frac{q^t-1}{q-1}(m-1)\ge\deg(f(x))$. Therefore, $f(x)$ must be the zero polynomial. Hence ϕ is one-to-one and the dimension of $C_q(l, b, m)$ is equal to the dimension of V_m. Thus we have

$$\dim(C_q(l, b, m)) = |E_m| = \left(\frac{1}{s}\sum_{d'\mid s}\mu\left(\frac{s}{d'}\right)m^{d'}+\frac{1}{t}\sum_{d''\mid t}\mu\left(\frac{t}{d''}\right)m^{d''}\right)s+m.$$

4 Examples

We compute for various values of the parameters and compare the codes obtained from the construction above with the codes from Brouwer's table [1]. It turns out that many codes with the best parameters can be obtained from our construction and improve the parameters of Brouwer's table [1]. Some codes we get are optimal by the Griesmer bound. Here are some examples.

Let $q = 3$, $s = 1$ and $t = 3$. Taking $m = 2$, $b = (3^3 - 3) / 3 = 8$ and we get a ternary [8, 4, 4]-linear code which is optimal by the Griesmer bound. By taking $l = 1$ and $b = 8$ we also find a ternary [9, 4, 5]-linear code which is optimal by the Griesmer bound.

Let $q = 5$, $s = 1$ and $t = 3$. Taking $m = 2$, $b = (5^3 - 5) / 3 = 40$, $l = 0$ and we get a 5-ary [40, 4, 30]-linear code which is optimal by the Griesmer bound.

Let $q = 4$, $s = 2$ and $t = 4$. Taking $m = 3$, $b = (4^4 - 4^2) / 2 = 120$, $l = 3$ and we get a quaternary [123, 45, 37] which gives an improvement on the parameters of Brouwer's table [1] for Brouwer's $d_B = 36$.

We list some other examples in the table below:

Table 1. Parameters of codes obtained by the method described in Section 3

q	s	t	m	b	n	k	d	d_B
4	2	4	2	120	120	10	78	76
4	2	4	2	120	123	10	79	78
4	2	4	3	120	120	45	35	34
4	2	4	3	120	125	45	38	37
4	2	4	3	120	127	45	39	38
4	2	4	3	120	131	45	41	40
5	1	3	3	40	40	11	20	20
5	1	3	4	40	41	24	10	9

Remark. We can introduce another parameters δ to change the dimension of the code such that the dimension depends on both m and δ. We can carry out the analysis for the dimension and the minimum distance similarly.

References

1. Brouwer, A.E.: Bounds on the Minimum Distance of Linear Codes (on-line server). **http://www.win.tue.nl/~aeb/voorlincod.html**
2. Xing, C., Ling, S.: A Class of Linear Codes with Good Parameters. IEEE Trans. Inform. Theory, Vol. 46 (2000) 2184–2188
3. Ling, S., Niederreiter, H., Xing, C.: Symmetric Polynomials and Some Good Codes. Finite Fields and Their Applications 7 (2001) 142–148
4. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes. Amsterdam, The Netherlands: North-Holland (1977)
5. van Lint, J.H.: Introduction to Coding Theory 3rd ed. Springer-Verlag, Berlin Heidelberg New York (2003)
6. Li, C., Feng, K., Hu, W.: Construction of A Class of Linear Codes with Good Parameters. Acta Electronica Sinica, Vol. 31 (2003) 51–53
7. Lidl, R., Niederreiter, H.: Finite fields. Number 20 in Encyclopedia of mathematics and its applications. Addison-Wesley, Reading, MA (1983)

Perceptual Hashing of Video Content Based on Differential Block Similarity

Xuebing Zhou[1], Martin Schmucker[1], and Christopher L. Brown[2]

[1] Fraunhofer IGD, Fraunhoferstr. 5, 64283 Darmstadt, Germany
[2] Institute of Telecommunications, Darmstadt University of Technology, Merckstrasse 25, 64283 Darmstadt

Abstract. Each multimedia content can exist in different versions, e.g. different compression rates. Thus, cryptographic hash functions cannot be used for multimedia content identification or verification as they are sensitive to bit flips. In this case, perceptual hash functions that consider perceptual similarity apply. This article describes some of the different existing approaches for video data. One algorithm based on spatio-temporal color difference is investigated. The article shows how this method can be improved by using a simple similarity measure. We analyze the performance of the new method and compare it with the original method. The proposed algorithm shows increased reliability of video identification both in robustness and discriminating capabilities.

1 Introduction

The huge volume of available multimedia content requires technologies for its efficient (automatic) identification. The perceptual hashing techniques, also called fingerprinting technologies, extract the most important features to calculate compact digests that allow efficient content identification. In comparison with watermark techniques which can also be used for identification, they do not modify the content.

Perceptual hash functions can be compared with cryptographic hash functions: Both extract a short digest from large data [1]. However, perceptual hashes must be robust against content manipulation resulting in perceptual similar content. This makes it distinct from cryptographic hashes. Perceptual hashes can be considered as robust hashes. Moreover, they are related to human perception.

Perceptual hashing technologies find application in areas such as (broadcast) monitoring, content verification, content-based searching, media asset management, archiving of media data in databases and video indexing.

In section 2, we discuss existing perceptual hashes algorithms. Section 3 introduces a perceptual hashing algorithm based on the similarity of video frames. In section 4 we show that this algorithm extracts features with increased reliability. The robustness and discernibility of the perceptual hashes are enhanced. An outlook for further improvements and acknowledgment are given in section 5 and Acknowledgment.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 80–85, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 State of the Art

Different approaches exist for the extraction of features from video content. Most perceptual hashing algorithms extract features in the spatial domain. One method reported in [2] is based on image hashing and the detection of key frames. Key frames - defined as groups of successive (visually) similar frames - are detected in each video shot. Perceptual image hashing technology based on radial variance projections is applied on these individual key frames. Other image hashing algorithms as in [3], [4] can also be used. In [5], statistical features are extracted from the single frame. Each frame is tiled in blocks and in each block the minimum of the variance matrices is chosen to represent a single frame. In [6] a small binary image is compressed to represent the original frame in the video set.

The second possibility to extract perceptual hashes is using its spatio-temporal information. In [7], each incoming frame of a video clip is divided into small blocks. The mean luminance is calculated for each block. The hash value is the sign of the spatio-temporal difference of mean luminance. The entropy of the resulted digests is higher than the value extracted from highly correlated space information. It performs well both in discriminative power and robustness for most video manipulations, remarkably even for rapidly varying video.

One generic difficulty in perceptual hashing of videos is the ensuing complexity of matching a candidate hash value with the stored hash values of videos in a (potentially very large) database. Efficient database structures and searching methods are needed to limit searching space.

To summarize, the perceptual hashing algorithms based on the spatial information extract feature from individual frames. They offer good robustness to manipulations based on the common image processing, in particular for still images. However, they need a process to vanquish the high temporal correlation of the video signal. These algorithms can be used in content-based searching, indexing and so on.

By contrast, perceptual hashing using spatio-temporal features performs well for fast varying video clips. Due to its simple implementation and searching process, we choose the spatio-temporal information to extract perceptual hashes. In the following section, we propose to extract robust features based on the similarity of frames to enhance the robustness of the perceptual hashes against noise and slowly varying clips.

3 Video Perceptual Hashes Using Differential Block Similarity

Through the processes of video acquisition, recording, processing and transmission, a video can be corrupted by noise. This significantly impacts the effectiveness of the perceptual hashing algorithm developed at Philips [7]. In this section we keep the structure of the Philip's algorithm and focus on the process of extracting the similarity features in order to enhance the robustness of the perceptual hashes.

The similarity is a statistical characteristic of the video signal. It is dependent on their perceptual similarity, whether two videos can be regarded as identical

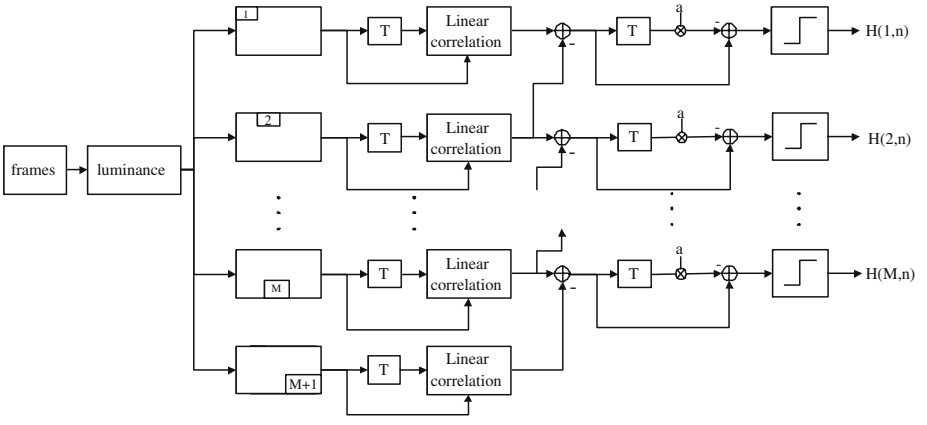


Fig. 1. Block diagram of the algorithm

material or not. A measure of video similarity can be found in many ways [8]. We have chosen the linear correlation, because of its low computation complexity. Additionally, it is robust to global manipulation of video files, such as scaling, shifting and so on. More important is that the linear correlation has strong robustness to noise, especially to white noise.

In the proposed algorithm, each frame is divided into $M + 1$ blocks. The luminance of pixel j in the block i of frame n is denoted as $x(j, i, n)$ with $i \in [1, \cdots, M + 1]$. The similarity of temporal consecutive blocks is $S(i, n)$ with:

$$S(i,n) = \sum_{j} x(j,i,n) \cdot x(j,i,n-1) \tag{1}$$

Then the spatio-temporal difference are calculated. Zigzag order, which enhance the robustness for the geometric manipulations such as scaling and shifting, is utilized to calculate the spatial differencing. In the temporal differencing, the subtrahend is damped with factor a, that is slightly smaller than 1, so that the DC- components of spatial difference is not fully suppressed. As a consequence, the robustness for still scenes is increased as well as the robustness against noise. Finally, the resulting value $B(i, n)$ is compared with the fixed threshold 0 resulting in the perceptual hash value $H(i, n)$ (see block diagram in figure 1):

$$H(i,n) \begin{cases} 1 & \text{if} \quad B(i,n) \geq 0 \\ 0 & \text{if} \quad B(i,n) < 0 \end{cases} \tag{2}$$

with $B(i,n) = (S(i,n) - S(i+1,n)) - a \cdot (S(i,n-1) - S(i+1,n-1))$.

4 Analysis and Evaluation

We implemented the Philips perceptual hashing algorithm [9] in Matlab and compare it with the proposed algorithm in section 3. In the simulation the following parameter values were chosen: 25 blocks per frame, $a = 0.95$. The perceptual

Table 1. Bit error rates for different types of processing

Video operations		**Algorithm using similarity**	**Philips' perceptual hashes algorithm**
histogram equalisation of 64 levels		7.26%	7.25%
median filter 3×3 neighborhoods		2.71%	3.09%
contrast	75%	0.23%	0.40%
	125%	1.68%	1.21%
brightness	85%	3.68%	2.48%
	125%	5.17%	1.89%
noise			
gaussian (zero mean and variance 0.01)		*9.02%*	*14.88%*
salt&pepper (noise density of 0.05)		*11.11%*	*18.01%*
speckle (zero mean and variance 0.04)		*9.87%*	*13.97%*
horizontal scaling	80%	4.03%	4.91%
	90%	3.93%	4.82%
	98%	2.62%	3.15%
	102%	1.95%	2.53%
	110%	1.97%	2.54%
	120%	2.34%	3.02%
horizontal shifting	1 pixel	1.58%	2.03%
	4 pixel	5.16%	6.03%
	8 pixel	9.13%	10.37%
	12 pixel	12.57%	13.86%
	20 pixel	18.09%	18.93%
	32 pixel	24.56%	24.62%
MPEG2	600kbit/s	8.61%	9.79%

hashes were extracted from 17 videos of 8 sec long run-time (7 NTSC videos and 10 PAL videos) and every perceptual hashes comprised 24×24 bits (one hash block corresponds to 1 second video). The video material has been downloaded from Communications Research Centre Canada (ftp://ftp.crc.ca/crc/vqeg/).
In table 1, the matching performance of the original perceptual hashes and the perceptual hashes extracted from manipulated videos is given for the new algorithm and Philips' algorithm. The bit error rate (BER) is the rate of the mismatching bits. The new algorithm has improved the BER for nearly all the video operations. In the cases of noise the BER using similarity are more than 4% lower than Philips'. The average BER for noise is reduced to around 10%. For additive white noise, the BER of individual video clips by the different algorithms is plotted in figure 2. It can be seen, that the algorithm using similarity decrease the BER by around 8% for the slowly varying video clips (video 5, 10 and 13). The boxplots in figure 3 show the spread of BER for different videos in different types of noise contamination. Comparing the two algorithms, the range of BER is also strongly suppressed. For MPEG2-compression at 600 kbits/sec we have also 1.2% gain. This enhancement is not so much as for noise, because the compression noise is colored. This algorithm also has an improvement of robustness for geometric processsing like scaling and shifting. However this improvement falls off with expanding geometric change. But, brightness and histogram equal-

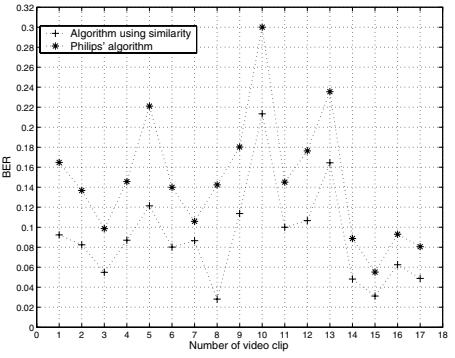


Fig. 2. BER of different video clips by white additive Gaussian noise

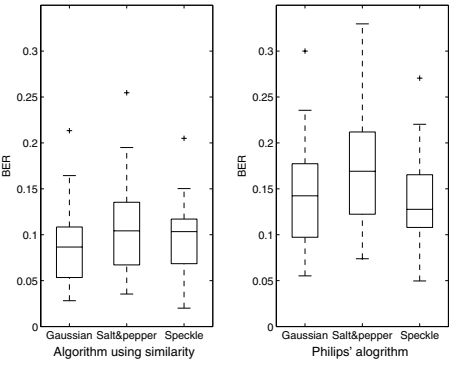


Fig. 3. Boxplots showing the spread of BER for different videos in different types of noise contamination

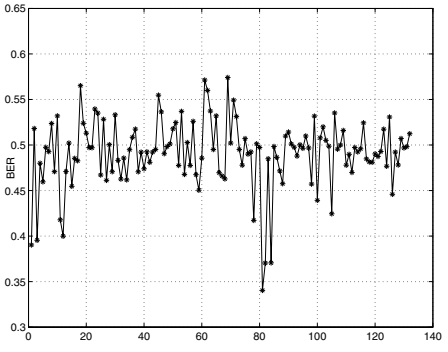


Fig. 4. BER for comparison with completely different clips

ization the BER are slightly higher than Philips' algorithm. This can stem from the high degree of nonlinearity of these processing.

Figure 4 shows the discriminability of the new algorithm. Most of the BER are near the desired rate of 50%. The extreme value of 34.5% is the result of comparison of the video clips that consist of still frames. For such video clips, the spatio-temporal information is too little to do the identification. The average BER of the algorithm using similarity is 48.61% and is 0.57% better than the Philips' 48.04%.

5 Conclusion and Outlook

The proposed perceptual hashing algorithm based on inter frame similarity shows increased robustness, especially to noise. Moreover it provides better discrimi-

nation than the Philips algorithm. Its complexity is a little higher in comparison with the original method.

For videos consisting of still images this algorithm does not discriminate well, because the resulting hashes are too strongly correlated to get enough information about the original video clips. To circumvent this drawback, we can calculate the perceptual hashes using spatial information, for example video annotations method in [6], as a supplement to still images.

Acknowledgment

The work presented in this paper was supported in part by the European Commission under Contract IST-2004-511299 AXMEDIS. The authors would like to thank the EC IST FP6 for the partial funding of the AXMEDIS project (www.axmedis.org), and to express gratitude to all AXMEDIS project partners internal and external to the project including the Expert User Group and all affiliated members, for their interests, support and collaborations.

References

1. Champskud J. Serepth, Andreas Uhl: Robust hash function for visual data: an experimental comparison (2003)
2. C. De Roover, C. De Vleeschouwer, F. Lefebvre, B.Macq: Key-frame radial projection for robust video hashing (2004)
3. R. Venkatesan, S.-M. Koon, M. H. Jakubowski, and P. Moulin: Robust image hashing (2000)
4. V. Fotopoulos and A.N. Skodras: A new fingerprinting method for digital images (2000)
5. A.Mucedero, R.Lancini and F. Mapelli: A novel hashing algorithm for video sequences (2004)
6. Yaron Caspi, David Bargeron: Sharing video annotations (2004)
7. Job Oostveen, Ton Kalker, Jaap Haitsma: Visual hashing of digital video: applications and techniques (2001)
8. Sen-ching S. Cheung, Avideh Zakhor: Efficient Video Similarity Measurement with Video Signature (2003)
9. Job Oostveen, Ton Kalker, Jaap Haitsma: Feature Extraction and a Database Strategy for Video Fingerprint (2002)

Secure Software Smartcard Resilient to Capture

Seung Wook Jung and Christoph Ruland

Institute for Data Communications Systems, University of Siegen,
Hölderlinstraße 3, 57076 Siegen, Germany
{seung-wook.jung, christoph.ruland}@uni-siegen.de

Abstract. We present a simple secure software smartcard that can be immunized against offline dictionary attack when the adversary captures the device. The proposed scheme also provides proactive security for the device's private key, i.e., proactively updates to the remote server and device to eliminate any threat of offline dictionary attacks due to previously compromised devices.

1 Introduction

The public key infrastructure (PKI) is a common mechanism for public key distribution in order to authenticate one device to another device. However, the security of the private key is still an important issue. The most basic threat is stealing a private key that is stored on a disk. Usually, such a key is stored in a software key container such as a PKCS#5– a file wherein the keys are encrypted by a password. This is vulnerable to dictionary attacks when the adversary *capture* the device. The tamper-resistant device is a solution for protecting against such attacks. However, such a device is still expensive (in terms of buying and management). Also, it can be weak against a virus that infects the user's computer to modify messages exchanged with the device[1]. Another solution for such a treat is software smartcards[1][2][3][4] that is a software-only scheme. In the existing software smartcard schemes[1][2][3][4], the mobile adversary who can capture some resources can generates the signature after subsequently capturing the other resource. In order words, the existing schemes do not provide proactive security(e.g., [5][6]). Intuitively, proactive security encompasses techniques for periodically refreshing the cryptographic secrets held by various components of a system, thereby rendering any cryptographic secrets captured before the refresh process is useless to the adversary.

The goal of this paper is to present a new efficient secure software smartcard for open PKI which is resilient to capture resources and offers a means for realizing the proactive security. The proposed scheme supports proactive security so that the resources captured by the adversary are useless after the refresh. Like [4], the proposed scheme is used for an ElGamal-like signature scheme, but our scheme is much more efficient than [4] in term of communication and computation.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related work in the literature. In section 3, we describe some notations

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 86–95, 2005.
© Springer-Verlag Berlin Heidelberg 2005

used throughout this paper and goals of this paper. In section 4, secure software smartcard scheme is presented. In section 5, the security is analyzed. Finally, we conclude with some discussions on other properties of the protocol in section 6.

2 Related Work

[1] proposed the first software smartcard scheme that is secure against off-line dictionary attacks when the user device is captured by the adversary. Unfortunately, in [1] the public key must be confidential, so [1] cannot be used in an open PKI.

The authors of [2][3] proposed software smartcard schemes that can be used in open PKI. However, the schemes are not used for ElGamal-like signature schemes. [4] aimed for ElGamal-like signature scheme. However, [4] is based on a threshold scheme and is still expensive, since it requires four rounds of communication and several public key encryption operations and decryption operations for only one signature even if it is more efficient than other threshold schemes for the protecting of the private key. Our scheme is used for ElGamal-like signature schemes that is much more efficient than that of [4].

In the existing software smartcard schemes[1][2][3][4], the mobile adversary who can capture some resources can generate the signature after subsequently capturing of the other resource. That is, the existing schemes do not provide a means of proactive security. Our scheme provides a means for realizing proactive security.

The existing software smartcard schemes [1][2][3][4] are based on PKI, while our scheme is based on Privilege Management Infrastructure (PMI) and PKI. The purpose of authentication is to access the resource of the system. For access control in an open network, PMI was proposed.

3 Preliminaries

3.1 Notation

In this paper, we use the following notation

- $a \in A$ denotes that the value a is in the set A.
- $a \in_R A$ denotes the choice of a uniform and independent random value a from the set A.
- Z_p denotes the set $\{0, 1, \ldots, p-1\}$, Z_p^* denotes the set $\{1, 2, \ldots, p-2, p-1\}$ with prime p.
- $S_k(m)$ denotes a signature on a message m with key k and outputs σ.
- $V_y(\sigma, m)$ denotes a verification algorithm of a signature σ using a public key y and the outputs true or false.
- $\mathcal{K}(), \mathcal{E}_y()$ and $\mathcal{D}_x()$ denote an asymmetric encryption scheme, where $\mathcal{K}()$ is a key generation algorithm, $\mathcal{E}_y()$ is an encryption algorithm using a public key y and $\mathcal{D}_x()$ denotes a decryption algorithm using a private key x.

3.2 Goals

The proposed software smartcard scheme consists of a user, a device dvc, an Attribute Authority(AA), and an adversary ADV. The user remembers the password π, controls the dvc, and types password into the dvc. The dvc holds a part of the certificate. AA holds the other part of the certificate. The dvc and AA communicate over an open network. The adversary has a dictionary of passwords and is a network wire, so that the adversary can watch all messages, modify messages, and initiate protocols. Also, the adversary knows all public information such as cryptographic function, domain parameters, and so on. Also, the computation power of the adversary is bounded by the Probabilistic Polynomial Time (PPT). Moreover, an adversary can *capture* certain resources[2]. The possible resources that may be captured by the adversary are the dvc, AA.u and the password, where AA.u denotes the user specific data on AA. Assume that the private key of AA, AA.p, cannot be compromise because the private key may be protected by the tamper-resistant device, while all user specific data may not be protected by the tamper-resistant device. Let us utilize the following definition for analyzing the weak point [2].

Definition 1. *ADV(S) is adversary who succeed capturing S where* $S \subseteq (dvc, AA.u, \pi)$. *It must hold* $ADV(S_1) \subseteq ADV(S_2)$ *if* $S_1 \subseteq S_2$.

Definition 2. *The software smartcard scheme is considered a secure software smartcard if the scheme meets following security goals.*

G.1 *Any adversary in* $ADV(AA.u, dvc)$ *can forge signatures if and only if an offline dictionary attack is successful on the user's password.*

G.2 *Any adversary in* $ADV(dvc)$ *can forge signatures with a probability at most* $q/|\mathbb{D}|$ *after q invocation of the server, where q is the threshold of the online dictionary attack and* $|\mathbb{D}|$ *denotes the size of dictionary.*

G.3 *Any adversary in* $ADV(AA.u, \pi)$ *cannot forge signatures for* ID_u, *where* ID_u *is the identity of the given user.*

G.4 *Any adversary in* $ADV(dvc, \pi)$ *can forge signatures for the device only until the device key is disabled and subsequently cannot forge signatures for the device.*

4 Secure Software Smart Card

The proposed scheme consists of setup, registration, issuing of a short-lived certificate (SLC), non-interactive password change, interactive password change, and instant revocation. The entities of the proposed scheme are user, CA, AA, and dvc.

4.1 Setup

The setup procedure of the CA is as follows:

1. The CA chooses large primes p, q such that $q|p-1$, a generator g of a multiplicative subgroup of Z_p^* with order q and a collision resistant hash function $h()$ where $h() : \{0,1\}^* \rightarrow Z_q^*$.

2. The CA chooses $x_{CA} \in_R Z_q^*$ as a private key and computes $y_{CA} = g^{x_{CA}} \bmod p$ as a public key.
3. The CA must keep securely x_{CA}. The public key, parameters $(p, q, g, y_{CA}, ID_{CA})$, and $h()$ are publicly known in the network, where ID_{CA} is the identity of CA.

After the setup of the CA, the AA sets up the system as follows:

1. The AA chooses a private key $x_{AA} \in_R Z_q^*$ and generates corresponding public key $y_{AA} = g^{x_{AA}} \bmod p$.
2. The AA gets a public key certificate $Cert_{AA}$ for y_{AA} from CA such as the X.509 certificate.
3. The AA chooses a public key encryption scheme $\mathcal{AE} = (\mathcal{K}, \mathcal{E}, \mathcal{D})$. The encryption algorithm $\mathcal{E}$ should be a randomized algorithm.
4. The AA generates public key pair $(x_{AA.e}, y_{AA.e})$ for public key encryption using $\mathcal{K}$. The security parameter of a private/public key pair for public key encryption should be as long as the length of the security parameter of the CA's private/public key pair.
5. The AA gets a public key certificate $Cert_{AA.e}$ such as the X.509 certificate for encryption from CA.
6. The AA publishes $Cert_{AA}$ and $Cert_{AA.e}$. Also, the public key encryption scheme must be known to all network entities.

4.2 Registration

The CA/AA and users communicate through an authenticated and confidential channel such as SSL/TLS and the CA properly authenticates users.

1. A user types a password pw, chooses a salt $v \in_R Z_q^*$ and a random number $a \in_R Z_q^*$. The user computes $\pi = h(pw||v)$ and $r' = g^{a+\pi} \bmod p$.
2. The user sends ID_u and r' to the CA where ID_u is the identifier of the user.
3. The CA chooses $k \in_R Z_q^*$ and computes $r = g^k \cdot r' \bmod p$.
4. The CA computes $s' = -x_{CA} \cdot h(CI||r) - k \bmod q$, where certificate information CI is the concatenation of the user identity ID_u, the CA's identity ID_{CA}, the unique certificate identifier $UCID$, and so on. That is, $CI = ID_u||ID_{CA}||UCID||etc.$ The CA returns s', r and CI to the user.
5. The user computes $s = s' - a \bmod q$ and checks $g^{\pi} \stackrel{?}{=} g^s \cdot y_{CA}^{h(CI||r)} \cdot r \bmod p$. Note, $s = -x_{CA} \cdot h(CI||r) - k - a \bmod q$ and $r = g^{k+a+\pi} \bmod p$. If true the user stores $s^b = s' - a - \pi \bmod q$ in backup device for instant revocation.
6. The user registers himself with CI and r to AA through an authenticated and confidential channel. At the end of the procedure, the user stores s and v in the user device dvc confidential and AA stores (r, CI) confidential.

4.3 Issuing a Short-Lived Certificate

1. The user types the password pw and generates $\pi = h(pw||v)$.
2. The user chooses a random number $b \in_R Z_q^*$ and computes $x = \pi + b \bmod q$ and $s'' = s + b \bmod q$.

User		CA
$a, v \in_R Z_q^*$ $r' = g^{a+h(pw\|\|v)} \bmod p$		
	$\rightarrow ID_u, r'$	
		$k \in_R Z_q^*$ $r = g^k \cdot r' \bmod p$ $s' = -x_{CA} \cdot h(CI\|\|r) - k \bmod q$
	$s', r, CI \leftarrow$	
$s = s' - a \bmod q$ checks $g^{h(pw\|\|v)} \stackrel{?}{=} g^s \cdot y_{CA}^{h(CI\|\|r)} \cdot r$		
		AA
$b \in_R Z_q^* \bmod q$ $s'' = s + b \bmod q, x = \pi + b \bmod q$ $e = E_{x_{AA.e}}(T\|\|ID_{AA}\|\|s''\|\|r\|\|CI)$ $\sigma_1 = S_x(e)$		
	$e, \sigma_1 \rightarrow$	
		Decrypts e, checks T and ID_{AA} $y = g^{s''} \cdot y_{CA}^{h(CI\|\|r)} \cdot r \bmod p$ $V_y(\sigma_1, e)$ $\sigma_2 = S_{y_{AA}}(T+1\|\|ID_u\|\|Success)$
	$\leftarrow \sigma_2$	
Verifies σ_2 Stores (s, v)		Store (r, CI)

Fig. 1. Registration

3. The user encrypts $e = \mathcal{E}_{y_{AA.e}}(s'', u)$.
4. The user generates a signature $\sigma = \mathcal{S}_x(m)$ as a proof of possession of x, where m includes at least a time stamp, e, attribute such as role name, and the identity of the AA.
5. The user sends m, σ to the AA.
6. The AA checks that the receiver is himself and that the timestamp is in the acceptance window. If not, AA aborts further processing.
7. The AA decrypts e using $x_{AA.e}$.
8. The AA checks whether requested attributes are belonging to ID_u. If not, the AA aborts further processing.
9. The AA searches CI, r using ID_u. Note CI includes the user's identifier.
10. The AA generates the user public key $y = g^{s''} \cdot y_{CA}^{h(CI||r)} \cdot r \bmod p$.
11. The AA verifies the signature $\mathcal{V}_y(\sigma, m)$. If the result of the verification is false, the AA increases the number of consecutive error and the number of the total error by 1. If the number of consecutive errors reaches the limitation or the number of total errors reaches the limitation, the AA considers the request as an online dictionary attack and blocks the user account. If the result of the verification is false, and the number of consecutive errors or total errors does not reach the limitation, the AA aborts further processing.

If the result of verification is true, the AA resets the number consecutive errors and processing continues.

12. The AA generates certificate $Cert$ that is a signature over (CI, y) using his certificate issuing private key x_{AA}, where CI includes user attribute, valid period, and so on.
13. The AA sends $Cert$ to the user.

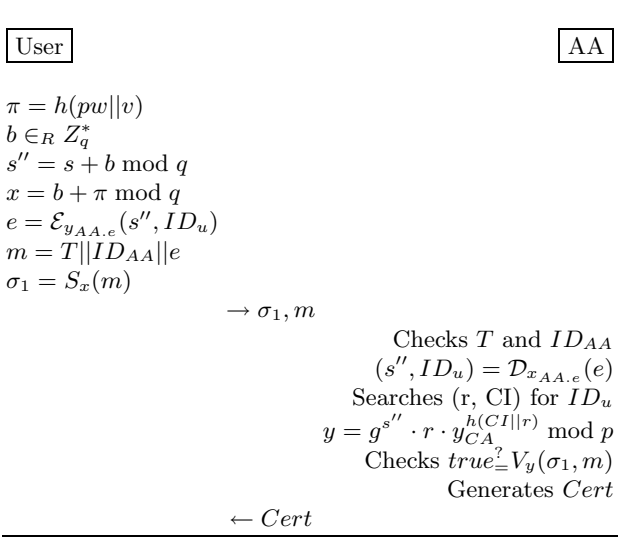


Fig. 2. Short-Lived Certificate

After that, the user will generate the proof of possession of the private key x and authenticated with the proof of possession and short-lived attribute certificate $Cert$ to the service provider that trusts the AA.

4.4 Proactivity

A secret sharing scheme distributes a secret among n devices so that the exposure of secrets from, say, t of these devices will not allow an adversary to "break" the schemes. That is, the adversary cannot generate valid signature as long as the number of captured devices is less than some predetermined security parameters (smaller than the number of devices needed to generate a valid signature). The proactive scheme [6][7] can tolerate a strong mobile adversary. That is the proactive scheme allows the changing of a shared secret so the capture secrets are not valid after changing the shared secret. The proposed scheme supports non-interactive proactive security and interactive proactive security.

The proposed scheme supports non-interactive proactive security so that user does not need to communicate with the AA, but changes the password at the user side as follows:

1. A user sets a new s as $s - h(pw||v) + h(pw'||v') \bmod p$, where pw' is a new password and sets a new salt v as v'.
2. A user keeps s, v.

In practice, the user's password can be disclosed by various means e.g., by being observed. The adversary who captures the user's device and password simultaneous can impersonate the user. If the adversary captures the password and later captures the user's device, then the user can change the password before the capturing of the user's device, so that ADV(dvc) has to conduct an online dictionary attack.

Interactive proactive security can be supported in the proposed scheme as follows.

1. A user picks a new password pw' and a new salt v' and computes π'.
2. The user chooses $b, c \in_R Z_q^*$ and computes $s'' = s+b \bmod q$, $x = \pi+b \bmod q$, and $\alpha = -\pi + \pi' + c \bmod q$.
3. The user computes $e = E_{y_{AA.e}}(ID_u, s'', \alpha)$.
4. The user generates a signature $\sigma = S_x(m)$, where $m = e||T||ID_{AA}||IPC$, T is a timestamp, and IPC denotes the interactive password change request.
5. The user sends σ, m to the AA.
6. The AA checks whether the timestamp is in an acceptance window and the receiver is ID_{AA}. If not, AA aborts further procedures.
7. The AA decrypts e and gets (ID_u, s'', α).
8. The AA generates a public key $y = g^{\pi+b}$ and verifies σ. If the result of verification is false, the AA increases the number of consecutive errors and the number of the total errors by 1. If the number of the consecutive errors reaches the limitation or the number of the total errors reaches the limitation, the AA considers the request as an online dictionary attack and blocks the user account. If the result of the verification is false, and the number of the consecutive errors or the total errors does not reach the limitation, the AA aborts further processing. If the verification is true, the AA resets the consecutive error and processing continues.
9. The AA computes $r' = g^{\alpha} \bmod p$ stores (r, r', CI). If r' exist, then AA replaces the old r' by the new r' that is multiplication of old r' by g^{α} modular p.
10. The user changes s as $s - c \bmod q$ and stores it and the new salt v'.

When the user generates a signature, the user picks $b \in_R Z_q^*$ and computes $s'' = s + b \bmod p$ and $x = \pi' + b \bmod q$, and finally signs over a message with x. For the verification, the AA generates a public key $y = g^{s''} \cdot y_{CA}^{h(CI||r)} \cdot r \cdot r' \bmod p$. Note $s = -x_{CA} \cdot h(CI||r) - k - a - c + b$ and $r \cdot r' = g^{k+a+\pi} g^{-\pi+c+\pi'}$. Therefore, $y = g^{\pi'+b} \bmod p$.

4.5 Instant Revocation

1. A user gets s^b, which is stored at the 5th step of registration, from backup device, where $s^b = -x_{CA} \cdot h(CI||r) - k - a - \pi_i$.

User		AA
Chooses pw'		
$v', b \in_R Z_q^*$		
$\pi' = h(pw'\|\|v')$		
$\pi = h(pw\|\|v)$		
$s'' = s + b \bmod q$		
$x = b + \pi \bmod q$		
$c \in_R Z_q^*$		
$\alpha = -\pi + \pi' + c \bmod q$		
$e = E_{y_{AA.e}}(ID_u, s'', \alpha)$		
$m = e\|\|T\|\|ID_{AA}\|\|IPC$		
$\sigma = S_x(m)$		
	$\rightarrow \sigma, m$	
		Checks T and ID_{AA}
		$(s'', ID_u, \alpha) = \mathcal{D}_{x_{AA.e}}(e)$
		Searches (r, CI) for ID_u
		$y = g^{s''} \cdot r \cdot y_{CA}^{h(CI\|\|r)} \bmod p$
		Checks $true \stackrel{?}{=} V_y(\sigma_1, m_1)$
		$r' = r' \cdot \alpha$
Stores $(s = s - c, v = v')$		Stores (r, r', CI)

Fig. 3. Interactive Password Update

User		AA
Gets s_b from backup device		
$\pi = h(pw\|\|v)$		
$b \in_R Z_q^*$		
$s'' = s + b \bmod q$		
$x = b + \pi \bmod q$		
$e = \mathcal{E}_{y_{AA.e}}(s'', s_b, ID_u)$		
$m = e\|\|T\|\|\|IR\|\|ID_{AA}$		
$\sigma = S_x(m)$		
	$\rightarrow \sigma, m$	
		Checks T and ID_{AA}
		$(s'', s_b, ID_u) = \mathcal{D}_{x_{AA.e}}(e)$
		Searches (r, CI) of ID_u
		$y = g^{s''} \cdot r \cdot y_{CA}^{h(CI\|\|r)} \bmod p$
		Checks $true \stackrel{?}{=} V_y(\sigma_1, m_1)$
		Checks $r \stackrel{?}{=} y_{CA}^{h(CI\|\|r)} \cdot g^{-s^b} \bmod p$
		Disable (r, CI)

Fig. 4. Instant Revocation

2. The user chooses $b \in_R Z_q^*$ and computes $s'' = s + b \bmod q$ and $x = \pi + b \bmod q$.
3. The user computes $e = \mathcal{E}_{y_{AA.e}}(s^b, s'', u)$.

4. The user computes a signature σ over $(e||T||IR||ID_{AA})$ with x, where T is a timestamp, and IR denotes an instant revocation message.
5. The user sends σ and (e, T, IR, ID_{AA}) to the AA.
6. The AA checks whether timestamp is in an acceptance window and the receiver is AA. If not, the AA stops following procedures.
7. The AA decrypts e and gets (s^b, s'', u).
8. The AA generates a public key $y = g^{\pi+b}$ and verifies σ. If the result of the verification is false, the AA increases the number of the consecutive errors and the number of the total errors by 1. If the number of the consecutive errors reaches the limitation or the number of the total errors reaches the limitation, the CA considers the request as an online dictionary attack and blocks the user account. If the result of the verification is false, and the number of the consecutive errors or total errors does not reaches the limitation, then AA aborts further processing.
9. The AA verifies Schnorr signature in a way that $r \stackrel{?}{=} y_{CA}^{h(CI||r)} \cdot g^{-s^b} \bmod p$, where $s^b = -x_{CA} \cdot h(CI||r) - k - a - \pi \bmod q$ and $r = g^{k+a+\pi}$.

Only the user can get the unblinded Schnorr signature (s^b, r), because no one except the user knows the random number a in the certificate issuing procedures. Therefore only the legitimate user can request instant revocation.

5 Security (Sketch)

An ADV(dvc, AA.u), who knows s, v, CI, and r, forges signatures after an offline dictionary attack on $g^{\pi}(= g^s \cdot y_{CA}^{(CI||r)} \cdot r \bmod p)$ has succeeded, because the adversary needs π to forge signatures of the user. Therefore, **G.1** is achieved.

Here, we explain why $s'' = s + b \bmod q$ must be encrypted. If s'' is not encrypted, the PPT $ADV(dvc)$, who knows s, v and public data including $y = g^{\pi+b}$, can conduct offline dictionary attacks as follows: $ADV(dvc)$ computes $b = s'' - s \bmod q$ and $g^{\pi} = g^{\pi+b} \cdot g^{-b} \bmod p$ after capturing the device. Therefore, the PPT $ADV(dvc)$ can conduct offline dictionary attacks on g^{π}. Without knowing s'', $ADV(dvc)$ has to conduct online dictionary attacks to forge a signature. The success probability of online dictionary attack is $q/|\mathbb{D}|$, where q is the maximum number of online dictionary attacks and the AA blocks the user's account after q failures. Therefore, **G.2** is achieved.

$ADV(AA.u, \pi)$, who knows CI, r, and π, has to know s to request a SLC from the AA with $s'' = s + b$. The probability to guess s'' is negligible $(1/q)$. Therefore, **G.3** is achieved.

After instant revocation, the AA will not generate any SLC for the user. Moreover, the adversary in $ADV(dvc, \pi)$ cannot know the x for any session. Therefore, the adversary in $ADV(dvc, \pi)$ cannot generate any signature after instant revocation. Therefore, **G.4** is achieved.

6 Conclusion

Offline dictionary attacks against password-protected private keys are a significant threat if the device holding those keys may be captured. In this paper, we have presented a secure software smartcard to eliminate such attacks. Furthermore, the proposed scheme provides the feature of proactive security, so the resource captured is useless after the refresh process. In addition to the above properties, the feature of instant revocation is provided. This enables the user to disable a device's private key immediately, even after the device has been captured and even if the adversary has guessed the user's password.

References

1. Hoover, D., Kausik, B.: Software smart cards via cryptographic camouflage. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy (SSP '99), Washington - Brussels - Tokyo, IEEE (1999) 208–215
2. MacKenzie, P., Reiter, M.: Networked cryptographic devices resilient to capture. International Journal of Information Security **2** (2003) 1–20
3. Kwon, T.: Robust software tokens: Towards securing a digital identity (2001)
4. MacKenzie, P., Reiter, M.K.: Two-party generation of DSA signatures. Lecture Notes in Computer Science **2139** (2001) 137–154
5. Herzberg, A., Jarecki, S., Krawczyk, H., Yung, M.: Proactive secret sharing or: How to cope with perpetual leakage. In Coppersmith, D., ed.: CRYPTO. Volume 963 of Lecture Notes in Computer Science., Springer (1995) 339–352
6. Ostrovsky, R., Yung, M.: How to withstand mobile virus attacks. In: Proc. 10th ACM Symp. on Principles of Distributed Computation. (1991) 51–59
7. Herzberg, A., Jakobsson, M., Jarecki, S., Krawczyk, H., Yung, M.: Proactive public key and signature systems. In: ACM Conference on Computer and Communications Security. (1997) 100–110

Revised Fischlin's (Blind) Signature Schemes*

Kewei Lv

State Key Laboratory of Information Security, Graduate School of Chinese Academy of Sciences, P.O. Box 4588, Beijing 100049, China
conwaylu@tom.com

Abstract. The representation problem based on factoring gives rise to alternative solutions to a lot of cryptographic protocols in the literature. Fischlin applies the problem to identification and (blind) signatures. Here we show some flaw of Fischlin's schemes and present the revision.

1 Introduction

An RSA n-bit modulus $N = pq$ is a product of two distinct primes p, q. A corresponding RSA exponent e is relatively prime to Euler's totient function $\phi(N)$. The RSA representation problem deals with the problem of finding a decomposition of a value into an RSA-like representation. Specifically, We presume that there is an efficient index generator **RSAIndex** for the representation problem which, on input 1^n, returns an n-bit RSA modulus N, a corresponding RSA exponent e and a random element $g \in_R \mathbf{Z}_N^*$. Let $(N, e, g) \leftarrow \mathbf{RSAIndex}(1^n)$ denote the sampling process. An RSA representation for a value $X \in \mathbf{Z}_N^*$ with respect to a tuple (N, e, g) is a pair (x, r) with $x \in \mathbf{Z}_e$ and $r \in \mathbf{Z}_N^*$ such that $X = g^x r^e \bmod N$. Every $X \in \mathbf{Z}_N^*$ has exactly e representations with respect to (N, e, g), because for each $x \in \mathbf{Z}_e$ there is a unique $r \in \mathbf{Z}_N^*$ such that $Xg^{-x} = r^e \bmod N$. We usually omit mentioning the reference to (N, e, g) if it is clear from the context, and simply say that (x, r) is a representation of X. So we have the following RSA Representation problem.

Definition 1. (RSA Rep. Problem) *Given* $(N, e, g) \leftarrow \mathbf{RSAIndex}(1^n)$ *return some* $X \in \mathbf{Z}_N^*$ *as well as two different representations* $(x_1, r_1), (x_2, r_2) \in \mathbf{Z}_e \times \mathbf{Z}_N^*$.

In contrast, ordinary RSA problem asks to compute e-th root $g^{1/e} \bmod N$ given $(N, e, g) \leftarrow \mathbf{RSAIndex}(1^n)$. The task is widely assumed to be intractable, i.e., no polynomial-time algorithm solves the RSA problem with more than negligible success probability. This implies that factoring N, too, is believed to be intractable. Yet, it is an open problem if RSA is indeed equally hard as factoring (see [3]). Provided one can solve the RSA problem, then the RSA representation problem becomes tractable, e.g., for any $r \in \mathbf{Z}_N^*$ both $(0, r)$ and $(1, rg^{-1/e} \bmod N)$ are representations of $X = r^e \bmod N$. The converse holds as well [9], and the equivalence reveals that both problems can be solved with

* Supported by President's Foundation of Graduate School of CAS (yzjj2003010).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 96–103, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the same success/running time characteristics, neglecting minor extra computations (in the sequel we keep on disregarding the effort for such additional minor computations).

At present, the RSA representation problem has a vast number of applications: for instance, Okamoto [9] constructs an identification protocol secure against (parallel) active attacks, which Pointcheval and Stern [10] subsequently turn into a secure signature scheme and a blind signature scheme. Fischlin and Fischlin [7] as well as Di Crescenzo et al. [4] use the RSA representation problem to derive efficient non-malleable commitment schemes based on RSA.

We now address the factoring representation problem by replacing the RSA exponent E by some power of 2. Interestingly, there is a seemingly less popular analogue to the RSA representation problem relying on the assumed hardness of factoring integers. In this case, a representation of X with respect to N, g and a number t is a pair $(x,\ r) \in \mathbf{Z}_{2^t} \times \mathbf{Z}_N^*$ with $X = g^x r^{2^t} \bmod N$. One reason for the unpopularity of the factoring representation problem may stem from the fact that Okamoto's previously proposed identification scheme based on this problem is flawed. It is sufficient to solve the RSA problem to pass the identification scheme with constant probability, without necessarily being able to factor the modulus. Fortunately, the bug in Okamoto's scheme is fixable, and Fischlin and Fischlin [8] devise an identification scheme using the factoring representation problem.

Apparently, given the factorization of N one can easily come up with two different representations. The converse does not hold in general: for example, $(x,\ r)$ and $(x,\ -r)$ represent the same X. Since we are mainly interested in finding distinct x-components we therefore call representations $(x_1,\ r_1)$ and $(x_2,\ r_2)$ different or distinct if and only if $x_1 \neq x_2$. Observe that this subsumes the RSA case where distinct x-components imply different r's and vice versa.

For any odd modulus $N = \prod_{i=1}^{r} p_i^{e_i}$, where $p_1, p_2, \cdots, p_r$ are distinct odd primes and $e_1, e_2, \cdots, e_r \geq 1$, let η denote the smallest integer such that $2^{\eta+1}$ does not divide any $\phi(p_i^{e_i})$. Then squaring is a permutation on the subgroup

$$\mathbf{HQR}_N := \{x^{2^\eta} \mid x \in \mathbf{Z}_N^*\} = \{x \in \mathbf{Z}_N^* \mid \mathrm{ord}_N(x) \textit{ is odd}\}$$

of the highest quadratic residues, namely the 2^η-th powers. In the following, we demand that $\tau \geq \eta - 1$ and thus τ may reveal some information about the factors of N. Let **FactIndex** denote an efficient index generator that outputs an n-bit RSA modulus N, $\tau \geq \eta - 1$, $t \geq 1$ and an independently chosen element $g \in_R \mathbf{HQR}_N$ for input 1^n, and write $(N,\ \tau,\ t,\ g) \leftarrow \mathbf{FactIndex}(1^n)$ for the sampling process. So we have the following factoring representation problem.

Definition 2. (Factoring Rep. Problem) *Given* $(N, \tau, t, g) \leftarrow \mathbf{FactIndex}(1^n)$ *return some* $X \in \mathbf{Z}_N^*$ *as well as two different representations* $(x_1, r_1), (x_2, r_2) \in \mathbf{Z}_{2^t} \times \mathbf{Z}_N^*$ *of* X*, i.e., with* $x_1 \neq x_2$*.*

Furthermore, Fischlin and Fischlin [7] prove the equivalence of the factoring representation problem to the factoring problem. With the techniques introduced in [10], they give (blind) signature schemes in random oracle model. But their scheme is flawed since it could be forged without the information of secret key.

Contribution: In this paper, we first give some attacks on Fischlin's schemes, then present some revision of them, which are provable secure in the random oracle model with the assumption that factoring representation problem is intractable in polynomial time.

Organization: In section 2, we show Fischlin's Signature Schemes. In section 3, we analyze the security of Fischlin's Schemes and give the attacks on them. In section 4, we present the revisions of Fischlin's Schemes.

2 Fischlin's Signature Schemes

Using the Fait-Shamir heuristic, the signature scheme is given in Figure 1. The challenge is generated by applying a hash function H to the message and the initial commitment of the signer. More specifically, publish $(N,\ \tau,\ t,\ g)$ as public parameter, X as public key and use the presentation $(x,\ r)$ as the secret key of the signer $\mathcal{S}$. In order to sign a message m, pick $y \in_R \mathbf{Z}_{2^{\tau+t}}$ and $s \in_R \mathbf{Z}_N^*$ at random, calculate $Y = g^y s^{2^{\tau+t}} \bmod N$, $c = H(Y,\ m)$ and compute $z = y + cx \bmod 2^{\tau+t}$, $W = sr^c g^{\lfloor (y+cx)/2^{\tau+t} \rfloor} \bmod N$. The signature to m becomes $\sigma(m) = (Y,\ W,\ z)$. Verification is straightforward.

Signer $\mathcal{S}$	$(N,\ \tau,\ t,\ g)$	**User** $\mathcal{U}$
representation $x,\ r$ of X	$X = g^x r^{2^{\tau+t}} \bmod N$	
pick $y \in_R \mathbf{Z}_{2^{\tau+t}},\ s \in_R \mathbf{Z}_N^*$		
$Y := g^y s^{2^{\tau+t}} \bmod N$	$\xrightarrow{Y}$	
	$\xleftarrow{c}$	$c := H(Y,\ m)$
$z := y + cx \bmod 2^{\tau+t}$		
$W := sr^c g^{\lfloor (y+cx)/2^{\tau+t} \rfloor} \bmod N$	$\xrightarrow{W,z}$	$YX^c \stackrel{!}{=} W^{2^{\tau+t}} g^z$

Fig. 1. Signature Scheme using Factoring Representation Problem

An important variant of signature schemes are blind signatures. In this case, the user $\mathcal{U}$ blinds the actual message m and requests a signature from the signer $\mathcal{S}$, which $\mathcal{U}$ later turns into a valid signature for the message m while the signer $\mathcal{S}$ cannot infer something about m.

The blind signature scheme based on the factoring problem is given in Figure 2; it is heavily influenced by the discrete-log and RSA protocols of Pointcheval and Stern [10]. In the first step, the signer $\mathcal{S}$ commits to Y. Then the user $\mathcal{U}$ blinds Y by multiplying with $g^\alpha \beta^{2^{\tau+t}} X^\gamma$ for random values α, β and γ. The actual challenge $c \in \mathbf{Z}_{2^t}$ is hidden by $c' = c + \gamma \in \mathbf{Z}_{2^{\tau+t}}$. $\mathcal{S}$ replies the challenge by sending W', z' subject to $YX^{c'} = (W')^{2^{\tau+t}} g^{z'}$. Now, $\mathcal{U}$ undoes the blinding and finally retrieves the signature $\sigma(m) = (Y',\ W,\ z)$:

$$\begin{aligned} W^{2^{\tau+t}} g^z &= (W' \beta g^{-z"} X^{c"})^{2^{\tau+t}} g^{z'} g^{z-z'} = YX^{c'} (\beta g^{-z"} X^{c"})^{2^{\tau+t}} g^{z-z'} \\ &= YX^{c+\gamma} \beta^{2^{\tau+t}} g^\alpha = Y' X^c. \end{aligned}$$

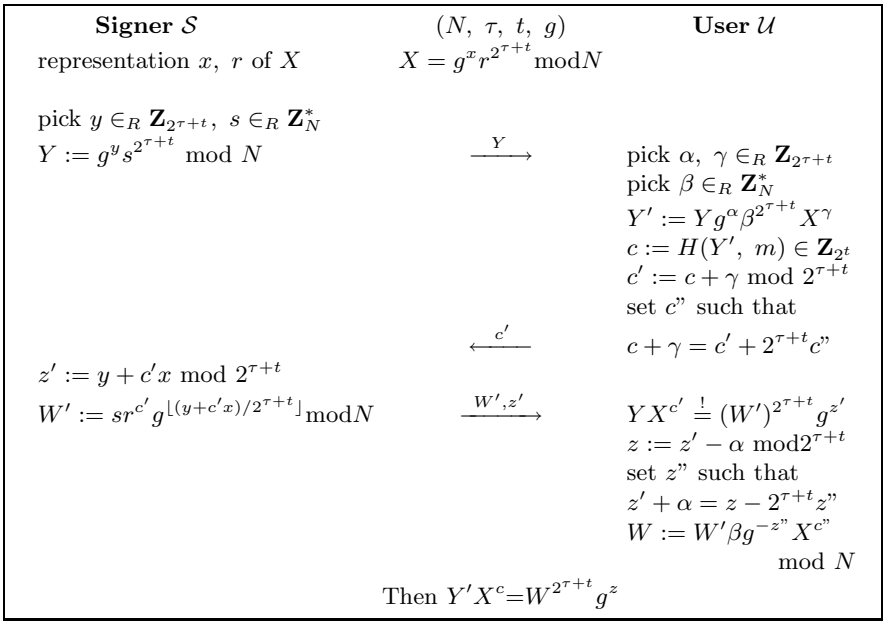


Fig. 2. Blind Signature Scheme using Factoring Representation Problem

Of course, the scheme is perfectly blind as $(Y,\ c',\ W',\ z')$ and $(Y',\ c,\ W,\ z)$ are independently distributed.

Although, as they were claimed in [7], the signature scheme in Figure 1 and the blind signature scheme in Figure 2 are secure respectively against chosen-message attacks and against a "one-more" forgery under a parallel attack (where up to poly-logarithmic signature generations are executed concurrently) relative to the hardness of factoring. But there are some flaw in them as we show in the next section, since the signatures from them are deniable as far as the signer is concerned or forgeable for man-in-the-middle.

3 Analysis of the Signature Schemes

We first show the flaw of the blind signature scheme in Figure 2, then we show that in the signature scheme in Figure 1.

As shown in Figure 3, for the blind signature scheme, when the signer $\mathcal{S}$ commits to Y, an adversary $\mathcal{A}$ between the signer and the user can compute $Y^* = g^l Y$ for some $1 \le l \le 2^{\tau+t} - 1 - c'x$, and sends it to user $\mathcal{U}$. Then the user $\mathcal{U}$ blinds Y^* by multiplying with $g^\alpha \beta^{2^{\tau+t}} X^\gamma$ for random values α, β and γ. The actual challenge $c \in \mathbf{Z}_{2^t}$ is hidden by $c' = c + \gamma \in \mathbf{Z}_{2^{\tau+t}}$. At this time, $\mathcal{A}$ does nothing on receiving c'. When $\mathcal{S}$ replies the challenge by sending W', z' subject to $Y X^{c'} = (W')^{2^{\tau+t}} g^{z'}$, $\mathcal{A}$ change the receiving W' and z' into $W^* = W'$ and $z^* = z' + l$ for the same l as above and sends them to $\mathcal{U}$. Since, for $1 \le l \le 2^{\tau+t} - 1 - c'x$, $\lfloor \frac{y+l+c'x}{2^{\tau+t}} \rfloor = \lfloor \frac{y+c'x}{2^{\tau+t}} \rfloor$. It is easy to see that

$Y^*X^{c'} = (W^*)^{2^{\tau+t}} g^{z^*}$. Now $\mathcal{U}$ undoes the blinding and finally retrieves the signature $\sigma(m) = (\hat{Y},\ \hat{W},\ \hat{z})$:

$$\hat{W}^{2^{\tau+t}} g^{\hat{z}} = (W^*\beta g^{-z"} X^{c"})^{2^{\tau+t}} g^{z^*} g^{z-z^*} = Y^* X^{c'} (\beta g^{-z"} X^{c"})^{2^{\tau+t}} g^{z-z^*}$$
$$= Y^* X^{c+\gamma} \beta^{2^{\tau+t}} g^{\alpha} = \hat{Y} X^c.$$

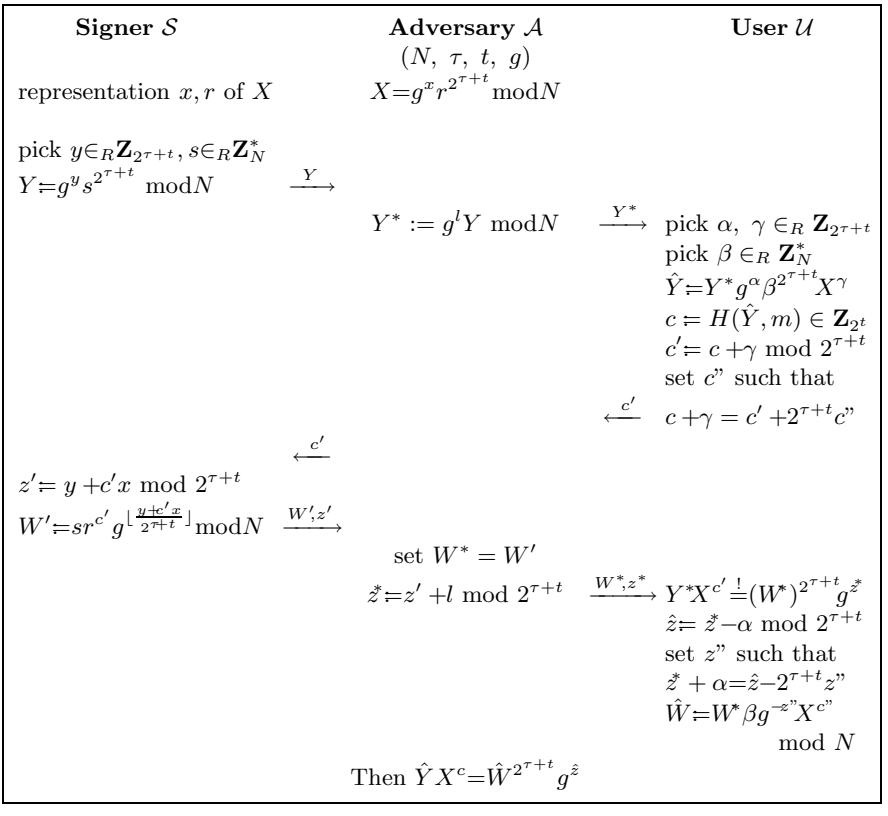


Fig. 3. Attack on Blind Signature Scheme

It is obvious that signer would deny the signature $\sigma(m) = (\hat{Y},\ \hat{W},\ \hat{z})$, and the user could not detect the malicious behavior. Now we analyze the probability of the adversary's success. Indeed, we only need to get the probability that $Y^*X^{c'} = (W^*)^{2^{\tau+t}} g^{z^*}$, i.e., $\lfloor \frac{y+l+c'x}{2^{\tau+t}} \rfloor = \lfloor \frac{y+c'x}{2^{\tau+t}} \rfloor$ for $1 \le l + c'x \le 2^{\tau+t} - 1$ (for instance, $l = 1$). So the probability that the adversary $\mathcal{A}$ succeeds is

$$Pr[\mathcal{A} \text{ is successful}] = Pr[\lfloor \frac{l + z'}{2^{\tau+t}} \rfloor = \lfloor \frac{z'}{2^{\tau+t}} \rfloor] \ge 1 - \frac{z'}{2^{\tau+t}} > \frac{1}{2}\ .$$

So the adversary can succeed in attacking the scheme easily.

For the signature scheme in Figure 1, we have the similar attack. Indeed, the adversary change the Y and z respectively into $\hat{Y} = g^l Y \bmod N$ and

$\hat{z} = z + l \bmod 2^{\tau+t}$, then it would pass the verification of the user and the signature to m is forged into $(\hat{Y},\ \hat{W} = W,\ \hat{z})$ with a probability $\geq \frac{1}{2}$.

4 The Revision of Fischlin's Signature Schemes

In this section, we present some revision of Fischlin's signature schemes as efficient as the original, and its security follows as in [10].

For revision of the blind signature scheme given in Figure 4, in the first step, in addition to commit to Y, the signer $\mathcal{S}$ binds the Y and X as $B = (YX)^a g^{2^{\tau+t}}$ for some $a \in_R \mathbf{Z}_{2^{\tau+t}}$, which could defend the signature scheme against malicious behavior of the adversary $\mathcal{A}$, since it is useless for $\mathcal{A}$ to get a later. Then the user $\mathcal{U}$ blinds Y by multiplying with $g^\alpha \beta^{2^{\tau+t}} X^\gamma$ for random values α, β and γ. The actual challenge $c \in \mathbf{Z}_{2^t}$ is hidden by $b = c + \gamma \in \mathbf{Z}_{2^{\tau+t}}$. $\mathcal{S}$ replies the challenge by sending W', z' and a subject to $B(XY)^b = (XY)^{a+b} g^{2^{\tau+t}}$ and $YX^{c'} = (W')^{2^{\tau+t}} g^{z'}$. Now $\mathcal{U}$ undoes the blinding and finally retrieves the signature $\sigma(m) = (Y',\ W,\ z)$: For $c' \equiv a + b \bmod 2^{\tau+t}$,

$$\begin{aligned} W^{2^{\tau+t}} g^z &= (W' \beta g^{-z"} X^{c"})^{2^{\tau+t}} g^{z'} g^{z-z'} = Y X^{c'+2^{\tau+t}c"} \beta^{2^{\tau+t}} g^{z-z'-2^{\tau+t}z"} \\ &= Y X^{c+\gamma} \beta^{2^{\tau+t}} g^\alpha = Y' X^c \ . \end{aligned}$$

Of course, the scheme is also perfectly blind as $(Y,\ c',\ W',\ z')$ and $(Y',\ c,\ W,\ z)$ are independently distributed. Following as in [10], we can get the security of the revised blind signature scheme:

Theorem 1. *In the random oracle model, the revised blind scheme in Figure 4 based on the factoring representation problem is secure against a "one-more" forgery under a parallel attack (where up to poly-logarithmic signature generations are executed concurrently) relative to the hardness of factoring.*

Similarly, revision of the signature scheme given in Figure 1 follows as in Figure 5: In the first step, in addition to commit to Y, the signer $\mathcal{S}$ binds the Y and X as $B = (YX)^a g^{2^{\tau+t}}$ for some $a \in_R \mathbf{Z}_{2^{\tau+t}}$, which could defend the signature against malicious behavior of the adversary $\mathcal{A}$, since it is useless for $\mathcal{A}$ to get a later. The challenge is generated by applying a hash function H to the message and the initial commitment of the signer. More specifically, publish $(N,\ \tau,\ t,\ g)$ as public parameter, X as public key and use the presentation $(x,\ r)$ as the secret key of the signer $\mathcal{S}$. In order to sign a message m, pick $a,\ y \in_R \mathbf{Z}_{2^{\tau+t}}$ and $s \in_R \mathbf{Z}_N^*$ at random, signer calculates $Y = g^y s^{2^{\tau+t}} \bmod N$ and $B = (YX)^a g^{2^{\tau+t}} \bmod N$. When receives $c = H(Y,\ m)$, signer computes $z = y + cx \bmod 2^{\tau+t}$, $W = s r^c g^{\lfloor (y+cx)/2^{\tau+t} \rfloor} \bmod N$ and sends the user z, W and a subject to $YX^c = W^{2^{\tau+t}} g^z$ and $B(XY)^b = (XY)^{a+b} g^{2^{\tau+t}}$, where $c = a + b \bmod 2^t$. The signature to m is also $\sigma(m) = (Y,\ W,\ z)$.

Easily, provided the hash function H behaves like a random function, then even with the help of an signature oracle any adversary fails to come up with a valid signature for a new message of his own choice [10, Sec.3.2]:

Signer $\mathcal{S}$	$(N,\ \tau,\ t,\ g)$	**User** $\mathcal{U}$
representation $x,\ r$ of X	$X = g^x r^{2^{\tau+t}} \bmod N$	
pick $a,\ y \in_R \mathbf{Z}_{2^{\tau+t}},\ s \in_R \mathbf{Z}_N^*$		
$Y := g^y s^{2^{\tau+t}} \bmod N$		
$B := (YX)^a g^{2^{\tau+t}} \bmod N$	$\xrightarrow{Y,\ B}$	pick $\alpha,\ \gamma \in_R \mathbf{Z}_{2^{\tau+t}}$
		pick $\beta \in_R \mathbf{Z}_N^*$
		$Y' := Y g^{\alpha} \beta^{2^{\tau+t}} X^{\gamma}$
		$c := H(Y',\ m) \in \mathbf{Z}_{2^t}$
		$b := c + \gamma \bmod 2^{\tau+t}$
		set c" such that
	$\xleftarrow{b}$	$c + \gamma = b + 2^{\tau+t} c$"
set $c' := a + b \bmod 2^{\tau+t}$		
$z' := y + c'x \bmod 2^{\tau+t}$		
$W' := s r^{c'} g^{\lfloor \frac{y + c'x}{2^{\tau+t}} \rfloor} \bmod N$	$\xrightarrow{W', z', a}$	$c' := a + b \bmod 2^{\tau+t}$
		$B(XY)^b \stackrel{!}{=} (XY)^{a+b} g^{2^{\tau+t}}$
		$YX^{c'} \stackrel{!}{=} (W')^{2^{\tau+t}} g^{z'}$
		$z := z' - \alpha \bmod 2^{\tau+t}$
		set z" such that
		$z' + \alpha = z - 2^{\tau+t} z$"
		$W \leftarrow W' \beta g^{-z}$" X^{c}" $\bmod N$
	Then $Y' X^{c'} = W^{2^{\tau+t}} g^z$	

Fig. 4. Revision of Blind Signature Scheme

Signer $\mathcal{S}$	$(N,\ \tau,\ t,\ g)$	**User** $\mathcal{U}$
representation $x,\ r$ of X	$X = g^x r^{2^{\tau+t}} \bmod N$	
pick $a,\ y \in_R \mathbf{Z}_{2^{\tau+t}},\ s \in_R \mathbf{Z}_N^*$		
$Y := g^y s^{2^{\tau+t}} \bmod N$		
$B := (YX)^a g^{2^{\tau+t}} \bmod N$	$\xrightarrow{Y}$	
	$\xleftarrow{c}$	$b := H(Y,\ m)$
$z := y + bx \bmod 2^{\tau+t}$		
$c := a + b \bmod 2^t$		
$W := s r^c g^{\lfloor (y + cx)/2^{\tau+t} \rfloor} \bmod N$	$\xrightarrow{W, z, a}$	$c := a + b \bmod 2^t$
		$B(XY)^b \stackrel{!}{=} (XY)^{a+b} g^{2^{\tau+t}}$
		$YX^c \stackrel{!}{=} W^{2^{\tau+t}} g^z$

Fig. 5. Revision of Signature Scheme

Theorem 2. *In the random oracle model, the revision of signature scheme in Figure 5 based on the factoring representation problem is secure against existential forgery under adaptive chosen-message attacks relative to the hardness of factoring.*

5 Conclusion

In this paper, we give some attacks on Fischlin's blind signature schemes and revise it. Moreover, we can see that the identification scheme [8] has the same flaw, which is omitted here for simplicity.

References

1. Bellare, M., Fischlin, M., Glodwasser, S., Micali, S.: Indentification protocols secure against reset attacks. In Advance in Cryptology-Proc. Eurocrypt'01, LNCS. Vol. 2045, (2001) 495–511
2. Bellare, M., Rogaway, P.: Random oracle are practical: a paradigm for designing efficient protocols. In First ACM conference on Computer and Communication Security. ACM Press (1993) 62–73
3. Boneh, D. and Venkatesan, R.. Breaking RSA may not be equivalent to factoring. In Advance in Cryptology-Proc. Eurocrypt'98, LNCS, Vol.1403 (1998) 59–71
4. Di Crescenzo, G., Katz, J., Ostrovsky, R., Smith, A.: Efficient and non-interactive non-malleable commitment. In Advance in Cryptology-Proc. Eurocrypt'01, LNCS, Vol.2045 (2001) 40–59
5. Dolev, D., Dwork, C., Noar, M.: Non-malleable cryptography. In SIAM J. on Computing, Vol.30(2) (2000) 391–437
6. Fiat, A., Shamir, A.: How to prove yourself: practical solutions to identification and signature schemes. In Advance in Cryptology-Proc. Crypto'86, LNCS, Vol.263 (1986) 186–194
7. Fischlin, M., Fischlin, R.: Efficient non-malleable commitment schemes. In Advance in Cryptology-Proc. Crypto'00, LNCS, Vol.1880 (2000) 414–432
8. Fischlin, M., Fischlin, R.: The representation problem based on factoring. In RSA security 2002 cryptographer's track, LNCS, Vol.2271 (2002) 96–113
9. Okamoto, T.: Provable secure and practical identification schemes and corresponding signature schemes. In Advance in Cryptology-Proc. Crypto'92, LNCS, Vol.740 (1992) 31–53
10. Pointcheval, D., Stern, J.: Security arguments for digital signatures and blind signatures. In J. of Cryptology, **13(3)** (2000) 361–396

Certificateless Threshold Signature Schemes

Licheng Wang, Zhenfu Cao, Xiangxue Li, and Haifeng Qian

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
{wanglc, zfcao, xxli, ares}@sjtu.edu.cn

Abstract. We analyze the relationship and subtle difference between the notion of certificateless public key cryptography (CL-PKC) and identity-based schemes without a trusted private key generator (PKG), then propose a certificateless threshold signature scheme based on bilinear pairings. The proposed scheme is robust and existentially unforgeable against adaptive chosen message attacks under CDH assumption in the random oracle model.

1 Introduction

Today, the management of infrastructure supporting certificates, including revocation, storage, distribution and the computation cost of certificate verification, incurs the main complaint against traditional public key signature(PKS) schemes. These situations are particularly acute in processor or bandwidth limited environments. ID-based systems [1], which simplify the key and certificate management procedure of CA-based PKI, are good alternatives for CA-based systems, especially when efficient key management and moderate security are required[2]. However, since an ID-based signature(IBS) scheme still needs a PKG to manage the generation and distribution of the user's private key, the inherent key escrow problem attracts rigorous criticism.

To tackle the certificate management problem in traditional PKS and the inherent key escrow problem in IBS, many wonderful ideas come to frontal arena of modern cryptography[3]. Limited by the space, we will only address three of them, here. They are threshold technique, ID-based scheme without a trusted PKG, and certificateless scheme.

A $(t, n) - threshold$ signature scheme is a protocol that allows any subset of t players out of n to generate a signature, but that disallows the creation of a valid signature if fewer than t players participate in the protocol[4].

At ASIACRYPT'2003, Al-Riyami and Paterson proposed the notion of certificateless public key cryptography(CL-PKC)[5]. They also presented a certificateless signature scheme based on bilinear mapping that are implemented using Weil or Tate pairings on elliptic curves. A certificateless signature is a new digital signature paradigm that simplifies the public key infrastructure, retains the efficiency of Shamir's identity-based scheme while it does not suffer from the inherent private key escrow problem. Motivated by this idea, Yum and Lee describe the generic schemes for construction of certificateless signature[6] and encryption[7] in 2004. Instead of based on bilinear pairings, they build certificate-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 104–109, 2005.
© Springer-Verlag Berlin Heidelberg 2005

less schemes from a generic framework: a public key signature/encryption scheme Π_{PK} plus an identity-based scheme Π_{IB}. In addition, they presented the extended construction of signature scheme that achieves Girault's trust level 3[3, 6].

However, some people do not think CL-PKC is a new concept. They take it just for a variation of ID-based schemes without a trusted PKG. In fact, Al-Riyami and Paterson had also pointed out that their original CL-PKC scheme is derived from Bonech and Franklin's ID-based systems[8] by making a very simple modification. In 2004, Chen et.al[9] proposed ID-based signature schemes without a trusted PKG using a special binding technique which is similar to that of CL-PKC. But in their threshold schemes, the user is in charge of splitting private key among n servers. In their scenario, this is correct since the user can always trust itself. But it can not transfer easily from their scenario to the general threshold scenario. In the general threshold scenario, for example, in a organization or a temporary group, we can not take someone for the fixed trusted dealer. Another similar work has been done by Duan et.al[10]. But the escrow problem in their scheme are still serious. For example, the PKG clerk can decrypt Bob's messages. He can also forge a signcryption on the organization's behalf if he can conspire with the organization's clerk.

Therefore, in our opinion, the original idea of certificateless primitive is still much significative. Although from the viewpoint of algorithm details, these ID-based schemes are much similar to some CL-PKC schemes, the notion of certificateless scheme is not a trivial variation of ID-based system because the primitive itself does not limit any implemental technique. Of course, these new proposed ID-based schemes can help us understand certificateless signature schemes from the implementation perspective. These schemes also suggest us consider certificateless schemes under the general threshold scenario. This is our first motivation for the contribution. Additionally, the generic construction of certificateless schemes in [6] and [7] is to combine an IBS scheme and a PKS scheme together, i.e. to operate a sequential double signing. So, the efficiency of certificateless schemes using this generic method will be inevitably abated. Moreover, if someone combines IBS and PKS carelessly, both the certificate management problem and the inherent key escrow problem would be two obstacles. Can we implement a certificateless threshold signature scheme using only one kind of technique? This is the second motivation. The last motivation is: the general threshold scenario itself is significant, even if the inherent key escrow problem is not in consideration. As far as we know, there is no certificateless threshold scheme has been introduced, yet.

2 Contributions

The main contribution of this paper is to present a certificateless threshold signature scheme as follows[1]:

[1] Since space limitation, all preliminaries and proofs in this paper are abridged. We assume that the readers are familiar with the contents on bilinear maps, pairings and the basic provably secure models.

Let $\mathbb{G}_1$ and $\mathbb{G}_2$ be additive and multiplication groups of the same large prime order q. Assume G_1 be a GDH group, and $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$ be a bilinear pairing map. Define three cryptographic hash functions: $H_1 : \{0,1\}^* \times \mathbb{G}_1 \to \mathbb{G}_1$, $H_2 : \{0,1\}^* \to \mathbb{G}_1$ and $H_3 : \mathbb{G}_2^4 \to \mathbb{Z}_q$.

[Setup]

- The PKG clerk *CLK-p* does the followings:
  - Choose $r_i \in_R \mathbb{Z}_q$ for $0 \leq i \leq u-1$.
  - Set $s = r_0$ as the master key, compute $P_{pub} = sP \in \mathbb{G}_1$ and publish

$$params =< \mathbb{G}_1, \mathbb{G}_2, e, q, P, P_{pub}, H_1, H_2, H_3 >$$

    as the parameter list of the system.
  - Set $R(x) = r_0 + r_1 x + r_2 x^2 + \cdots + r_{u-1} x^{u-1} \in \mathbb{Z}_q[X]$, then compute $s_i = R(i) \in \mathbb{Z}_q$ and $P_{pub}^{(i)} = s_i P \in \mathbb{G}_1$ for $1 \leq i \leq u-1$ and send $(s_i, P_{pub}^{(i)})$ to the corresponding PKG PKG_i via a secure channel.
- When receiving $(s_i, P_{pub}^{(i)})$, each PKG_i verifies the validity of $P_{pub}^{(i)}$ and publishes it public while keeps s_i secret.
- To prevent attacks on the master key, *CLK-p* destroys s after all above have been done. This means that *CLK-p* must destroy s, all s_i and all r_i.

[SetSecretValue]

- After *CLK-p* publishes *params*, the parameter list of the system, the organization clerk *CLK-s* does the followings:
  - Choose $a_i \in_R \mathbb{Z}_q$ for $0 \leq i \leq t-1$.
  - Set $x_A = a_0$ as the secret value of the organization A, compute the concealed secret value $cs_A = x_A P \in \mathbb{G}_1$ and make cs_A publicly known.
  - Set $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{t-1} x^{t-1} \in \mathbb{Z}_q[X]$, then compute the first part of signing key $fsk_i = f(i) \in \mathbb{Z}_q$ and the first part of verify key $fvk_i = f(i)P \in \mathbb{G}_1$ and send (fsk_i, fvk_i) to the corresponding member i via a secure channel, for $1 \leq i \leq t-1$.
- When receiving (fsk_i, fvk_i), each member i in A verifies the validity of fvk_i and publishes it public while keeps fsk_i secret.
- To prevent attacks on the organization's secret value, *CLK-s* destroys x_A after all above have been done. Similarly, this means that *CLK-s* must destroy x_A, all fsk_i and all a_i.

[ExtractPublicKey]

- After *CLK-s* publishes cs_A, the concealed secret value of the organization A, any entity can extract A's public key $pk_A = H_1(id_A, cs_A) \in \mathbb{G}_1$ whenever it is necessary. Note that the organization's identity id_A is always accessible.

[ExtractPartialPrivateKey]

- After *CLK-s* publishes cs_A, $l(\geq u)$ PKGs (without loss of generality, assume they are $PKG_1, \cdots, PKG_l$, respectively) are going to jointly generate the partial private key for A. Each of them independently computes A's partial private key share $psk_A^{(i)} = s_i \cdot pk_A = s_i \cdot H_1(id_A, cs_A) \in \mathbb{G}_1$ and sends it to *CLK-p*. And *CLK-p* accepts $psk_A^{(i)}$, the i^{th} share of A's partial private key, if and only if the following equation

$$e(psk_A^{(i)}, P) = e(P_{pub}^{(i)}, pk_A)$$

  holds.
- After u (maybe less than l) acceptable shares of A's partial private key are collected (without loss of generality, assume they are $psk_A^{(1)}, \cdots, psk_A^{(u)}$, respectively), *CLK-p* constructs A's partial private key

$$psk_A = \sum_{i=1}^{u} \lambda_{0,j} \cdot psk_A^{(i)} = s \cdot pk_A$$

  where $\lambda_{x,j} = \prod_{\substack{i=1 \\ i \neq j}}^{u} \frac{x-i}{j-i} \pmod q$.
- *CLK-p* sets $A_0 = psk_A \in \mathbb{G}_1$, chooses $A_i \in_R \mathbb{G}_1$ for $1 \leq i \leq t-1$ and sets[2]

$$F(x) = A_0 + xA_1 + x^2 A_2 + \cdots + x^{t-1} A_{t-1},$$

  then computes the second part of signing key $ssk_i = F(i) \in \mathbb{G}_1$ and the second part of verify key $svk_i = e(P, ssk_i) \in \mathbb{G}_2$ for each member in A. Finally, *CLK-p* sends (ssk_i, svk_i) to corresponding member i in A.
- When receiving (ssk_i, svk_i), each member i in A verifies the validity of svk_i and publishes it public while keeps ssk_i secret.
- To enhance the robustness of our scheme and to prevent eternal occupation of A's partial private key, *CLK-p* destroys psk_A, including all $psk_A^{(i)}$, all ssk_i and all A_i, after all above have been done. Then, whenever *CLK-p* wants to rebuild A's partial private key, he/she must recollect at least u valid shares of psk_A from the PKG group. Meanwhile, if some members in A suspect the PKG group, they can urge *CLK-s* reset x_A, i.e. the secret value of the organization, by resuming the *SetSecretValue* algorithm(Of course, all corresponding algorithms followed *SetSecretValue* need to be resumed, too.).

[Sign]

- Denote $\Phi = \{i_1, \cdots, i_k\} \subset \{1, \cdots, n\}$ as the set of presented members in this signature task, $|\Phi| \geq t$. Each member j in Φ performs the following steps to jointly create a signature for a message m.

[2] In practice, choosing $A_i \in_R \mathbb{G}_1$ can be implemented by computing $r_i P$ for $r_i \in_R \mathbb{Z}_q$.

- Choose $Q_j \in \mathbb{G}_1$ at random.
- Compute and broadcast

$$T_j = fsk_j \cdot H_2(m), \gamma_j = e(H_2(m), ssk_j),$$

$$\alpha_j = e(P, Q_j), \beta_j = e(H_2(m), Q_j)$$

- Compute

$$\alpha = \prod_{j \in \Phi} \alpha_j^{\lambda_{0,j}^{\Phi}}, \beta = \prod_{j \in \Phi} \beta_j^{\lambda_{0,j}^{\Phi}}, \gamma = \prod_{j \in \Phi} \gamma_j^{\lambda_{0,j}^{\Phi}},$$

$$\delta = e(H_2(m), pk_A), c = H_3(\alpha, \beta, \gamma, \delta), W_j = Q_j + c \cdot ssk_j$$

and broadcast W_j, where $\lambda_{x,j}^{\Phi} = \prod_{\substack{i \in \Phi \\ i \neq j}} \frac{x-i}{j-i} \pmod q$.

– Each member $i \in \Phi$ checks whether

$$e(T_j, P) = e(H_2(m), fvk_j),$$

$$e(P, W_j) = \alpha_j \cdot svk_j^c, e(H_2(m), W_j) = \beta_j \cdot \gamma_j^c$$

for $j \in \Phi$ and $j \neq i$. If the equations fail for some j, then the member i broadcasts **COMPLAINT** against member j.

– If at least t members in Φ are honest, a dealer(selected at random from these honest members) performs calculation as follows:

$$T = \sum_{j \in \Phi} \lambda_{0,j}^{\Phi} T_j, W = \sum_{j \in \Phi} \lambda_{0,j}^{\Phi} W_j.$$

And then, the dealer outputs T and the corresponding proof $(\alpha, \beta, \gamma, W)$ as the signature on the message m.

[Verify]

– When receiving a message m and corresponding signature $sig = (T, \alpha, \beta, \gamma, W)$, the verifier does the following:

- Compute

$$pk_A = H_1(id, cs_A), \delta = e(H_2(m), pk_A),$$

$$\mu = e(P_{pub}, pk_A), c = H_3(\alpha, \beta, \gamma, \delta).$$

- Accept the signature if and only if the following equations hold:

$$e(T, P) = e(H_2(m), cs_A),$$

$$e(P, W) = \alpha\mu^c, e(H_2(m), W) = \beta\gamma^c.$$

It is not difficult to prove that the proposed scheme is correct, robust and existentially unforgeable against adaptive chosen message attacks under CDH assumption in the random oracle model.

3 Conclusions

The security of a certificateless scheme lies on three parameters: the user's secret value x, the system's master key s and the user's partial private key psk. In

order to enhance the system's robustness and security, debase the risk of single point failure and limit the signer's power, we present a certificateless threshold signature scheme by employing threshold technique to deal with s, x and psk in three phases. The proposed implementation is secure in an appropriate model, assuming that the Computational Diffie-Hellman Problem is intractable.

Acknowledgments

This work was supported by the National Natural Science Foundation of China for Distinguished Young Scholars under Grant No. 60225007, the National Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20020248024, and the Science and Technology Research Project of Shanghai under Grant No. 04JC14055 and No. 04DZ07067.

References

1. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakley, G.R., Chaum, D. (eds.): Advances in Cryptology - CRYPTO '84. Lecture Notes in Computer Science, Vol. 196. Springer-Verlag, Berlin Heidelberg New York (1985) 47–53
2. Ratna, D., Rana, B., Palash, S.:. Pairing-Based Cryptographic Protocols : A Survey. Cryptology ePrint Archive: Report 2004/064.
3. Girault, M.: Self-certified public keys. In: Davies, D.W. (ed.): Advances in Cryptology - EUROCRYPT '91. Lecture Notes in Computer Science, Vol. 547. Springer-Verlag, Berlin Heidelberg New York (1991) 490–497
4. Pedersen, T.P.: Non-Interactive and Information-Theorectic secure verifiable secret sharing. In: Feigenbaum, J. (ed.): Advances in Cryptology - CRYPTO '91. Lecture Notes in Computer Science, Vol. 576. Springer-Verlag, Berlin Heidelberg New York (1992) 129–140
5. Sattam, S.A., Kenneth, G.P.: Certificateless Public Key Cryptography. In: Laih, C.S. (ed.): ASIACRYPT 2003. Lecture Notes in Computer Science, Vol. 2894. Springer-Verlag, Berlin Heidelberg New York (2003) 452–473
6. Dae, H.Y., Pil, J.L.: Generic Construction of Certificateless Signature. In: Wang, H. et al. (eds.): ACISP 2004. Lecture Notes in Computer Science, Vol. 3108. Springer-Verlag, Berlin Heidelberg New York (2004) 200–211
7. Dae, H.Y., Pil, J.L.: Generic Construction of Certificateless Encryption. In: Lagana, A. et al. (eds.): ICCSA 2004. Lecture Notes in Computer Science, Vol. 3043. Springer-Verlag, Berlin Heidelberg New York (2004) 802–811
8. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In Kilian, J. (ed.): CRYPTO 2001. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York (2001) 213–229
9. Chen, X., Zhang, F., Divyao, M.K., Kwangjo, K.: New ID-Based Threshold Signature Scheme from Bilinear Pairing. In: Canteaut, A., Viswanathan, K. (eds.): INDOCRYPT 2004. Lecture Notes in Computer Science, Vol. 3348. Springer-Verlag, Berlin Heidelberg New York (2004) 371–383
10. Duan, S., Cao, Z., Lu, R.: Robust ID-based Threshold signcryption Scheme From Pairings. In ACM InfoSecu '04 (2004) 33–37

An Efficient Certificateless Signature Scheme

M. Choudary Gorantla and Ashutosh Saxena

Institute for Development and Research in Banking Technology,
Castle Hills, Masab Tank, Hyderabad 500057, India
gmchoudary@mtech.idrbt.ac.in, asaxena@idrbt.ac.in

Abstract. Traditional certificate based cryptosystem requires high maintenance cost for certificate management. Although, identity based cryptosystem reduces the overhead of certificate management, it suffers from the drawback of key escrow. Certificateless cryptosystem combines the advantages of both certificate based and identity based cryptosystems as it avoids the usage of certificates and does not suffer from key escrow. In this paper, we propose a pairing based certificateless signature scheme that is efficient than the existing scheme.

1 Introduction

In traditional public key cryptosystems (PKC), the public key of a signer is essentially a random bit string. This leads to a problem of how the public key is associated with the signer. In these cryptosystems, the binding between public key and identity of the signer is obtained via a digital certificate, issued by a trusted third party(TTP) called certifying authority (CA). The traditional PKC requires huge efforts in terms of computing time and storage to manage the certificates [8].

To simplify the certificate management process, Shamir [9] introduced the concept of identity (ID) based cryptosystem wherein, a user's public key is derived from his identity and private key is generated by a TTP called Private Key Generator (PKG). Unlike traditional PKCs, ID-based cryptosystems require no public key directories. The verification process requires only user's identity along with some system parameters which are one-time computed and available publicly. These features make ID-based cryptosystems advantageous over the traditional PKCs, as key distribution and revocation are far simplified. But, an inherent problem of ID-based cryptosystems is key escrow, i.e., the PKG knows users' private key. A malicious PKG can frame an innocent user by forging the user's signature. Due to this inherent problem, ID-based cryptosystems are considered to be suitable only for private networks [9]. Thus, eliminating *key ecrow* in ID-based cryptosystems is essential to make them more applicable in the real world.

Recently, Al Riyami and Paterson [1] introduced the concept of certificateless public key cryptosystem(CL-PKC), which is intermediate between traditional PKC and ID-based cryptosystem. CL-PKC makes use of a trusted authority which issues partial private keys to the users. The partial private key for a user

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 110–116, 2005.
© Springer-Verlag Berlin Heidelberg 2005

A is calculated using trusted authority's master secret key and A's identity ID_A. The user A then computes his private key by combining the partial private key with some chosen secret information. Thus, the trusted authority in CL-PKC does not know the user's private key. The user A also computes his public key by combining his secret information with the trusted authority's public parameters. The system is not ID-based as the public key is no longer computable from ID alone. The verification of a signature is done using both the public key and ID of the signer. CL-PKC has less overhead compared to the traditional certificate based PKC as there is no need of certificate management. Hence, it finds applications in the low-bandwidth and low-power environments such as security applications in mobile environment, where the need to transmit and check certificates has been identified as a significant limitation [6].

We note that the signature scheme in [1] is not efficient as it requires a few costly pairing operations in the signing and verification processes. The low-power appliances may not have enough computational power for calculating pairing operations in the signature generation. Recently, Yum and Lee [10] proposed generic construction of certificateless signature schemes. Though their construction yields good security reduction, it results in inefficient schemes.

In this paper, we propose a certificateless signature scheme based on bilinear pairings. Our signature scheme requires no pairing operations in the signing phase and requires one less operation than the scheme in [1] in the verification phase. The size of public key in our scheme is also less when compared to the scheme of [1].

The rest of the paper is organized as follows: Section 2 gives the background concepts on bilinear pairings and some related mathematical problems. Section 3 presents the model of our scheme. Section 4 presents the proposed signature scheme. Section 5 analyzes our scheme with respect to security and efficiency. Finally, we conclude our work in Section 6.

2 Background Concepts

In this section, we first briefly review the basic concepts on bilinear pairings and some related mathematical problems.

2.1 Bilinear Pairings

Let G_1 and G_2 be additive and multiplicative cyclic groups of prime order q respectively and let P be an arbitrary generator of G_1. A cryptographic bilinear map is defined as $e : G_1 \times G_1 \rightarrow G_2$ with the following properties.

Bilinear: $\forall\ R, S, T \in G_1,\ e(R + S, T) = e(R, T)e(S, T)$ and $e(R, S + T) = e(R, S)e(R, T)$.

Non-degenerate: There exists $R, S \in G_1$ such that $e(R, S) \neq I_{G_2}$ where I_{G_2} denotes the identity element of G_2.

Computable: There exists an efficient algorithm to compute $e(R, S)\ \forall R, S \in G_1$.

In general implementation, G_1 will be a group of points on an elliptic curve and G_2 will denote a multiplicative subgroup of a finite field. Typically, the mapping e will be derived from either the Weil or the Tate pairing on an elliptic curve over a finite field. We refer to [3] for more comprehensive description on how these groups, pairings and other parameters are defined.

2.2 Mathematical Problems

Here, we discuss some mathematical problems, which form the basis of security for our scheme.

Discrete Logarithm Problem (DLP): Given q, P and $Q \in G_1^*$, find an integer $x \in Z_q^*$ such that $Q = xP$.
Computational Diffie-Hellman Problem (CDHP): For any $a, b \in Z_q^*$, given $\langle P, aP, bP \rangle$, compute abP.

Decisional Diffie-Hellman Problem(DDHP): For any $a, b, c \in Z_q^*$, given $\langle P, aP, bP, cP \rangle$, decide whether $c \equiv ab \bmod q$.

Gap Diffie-Hellman Problem(GDHP): A class of problems where CDHP is hard while DDHP is easy.

We assume the existence of Gap Diffie-Hellman groups and a detailed description of building these groups is given in [4].

3 The Model

The proposed signature scheme consists of seven algorithms: Setup, Partial-Private-Key-Extract, Set-Secret-Value, Set-Private-Key, Set-Public-Key, Sign and Verify.

Setup: The TA selects a `master-key` and keeps it secret. It then specifies the system parameters `params` and publishes them. The `params` include description of the bilinear map, hash functions, TA's public key, message space $\mathcal{M}$ and signature space $\mathcal{S}$.

Partial-Private-Key-Extract: The TA calculates a partial private key D_A from its `master-key` and identity ID_A of a user `A`. The TA then communicates D_A to `A` in a secure channel.

Set-Secret-Value: The user `A` selects a secret value, which will be used to calculate his private and public keys.

Set-Private-Key: The user calculate his private key S_A from the chosen secret value and partial private key D_A.

Set-Public-Key: The user calculates his public key P_A from the chosen secret value and TA's public key.

Sign: The user `A` signs a message m using his private key S_A and produces a signature $Sig \in \mathcal{S}$. Sig is sent to the verifier along with `A`'s public key.

Verify: The verifier verifies the signature of `A`, using `A`'s identity ID_A and public key P_A after checking P_A's correctness.

3.1 Trust Levels

A public key cryptosystem, involving a TTP, can be classified into three levels based on the trust assumptions [7].

- At level 1, the TTP knows (or can easily compute) the users' private keys and therefore can impersonate any user at any time without being detected.
- At level 2, the TTP does not know (or cannot easily compute) the users' private keys. But, the TTP can still impersonate a user by generating a false public key (or a false certificate) without being detected.
- At level 3, the TTP does not know (or cannot easily compute) the users' private keys. Moreover, it can be proved that the TTP generate false public keys of users if it does so.

3.2 Adversarial Model

CL-PKC uses public directories to store the public keys of the users. As given in [1], it is more appropriate for encryption schemes to have the public keys published in a directory. For signature schemes, the public keys can be sent along with the signed message. We put no further security on the public keys and allow an active adversary to replace any public key with a public key of its own choice. There are two types of adversaries who can replace the public keys: Type I adversary who does not have access to `master-key` and Type II adversary having access to `master-key`. Clearly, the Type II adversary is the TA itself or has the cooperation of TA otherwise. An adversary (Type I or II) will be successful in its attempts if it can calculate the correct private key. It is obvious that with access to `master-key`, Type II adversary can always calculate private keys corresponding to a public key of its choice. For the moment, we assume that Type II adversary does not participate in such type of attacks, thus achieving trust level 2. Later, we show how our scheme can attain trust level 3.

With these assumptions in place, we say that even though an adversary replaces a public key, the innocent user can not be framed of repudiating his signature. Hence, the users have the same level of trust in the TA as they would have in a CA in the traditional PKC. The trust assumptions made in our scheme are greatly reduced compared to ID-based schemes where, the PKG knows the private key of every user.

4 Proposed Scheme

In this section, we present a certificateless signature scheme based on the ID-based signature scheme of [5]. The proposed signature scheme involves three entities: trusted authority (TA), signer and verifier.

4.1 The Signature Scheme

Setup: The TA performs the following steps:

1. Specifies $\langle G_1, G_2, e\rangle$, as described in Section 2.1
2. Chooses an arbitrary generator $P \in G_1$

3. Selects a `master-key` t uniformly at random from Z_q^* and sets TA's public key as $Q_{TA} = tP$
4. Chooses two cryptographic hash functions $H_1 : \{0,1\}^* \times G_1 \rightarrow Z_q^*$ and $H_2 : \{0,1\}^* \rightarrow G_1$

The system parameters are `params` $= \{G_1, G_2, q, e, P, Q_{TA}, H_1, H_2\}$, the message space is $\mathcal{M} = \{0,1\}^*$ and the signature space is $\mathcal{S} = G_1^* \times G_1^*$.

Partial-Private-Key-Extract: The TA verifies the identity ID_A of the user A and calculates $Q_A = H_2(ID_A)$. It then calculates the partial private key $D_A = tQ_A$ and sends it to A in a secure channel.

Set-Secret-Value: The user A selects a secret value $s \in Z_q^*$ based on a given security parameter.

Set-Private-Key: The user A calculate his private key as $S_A = sD_A$.

Set-Public-Key: The user A calculates his public key as $P_A = sQ_{TA}$.

Sign: To sign a message $m \in \mathcal{M}$ using the private key S_A, the signer A performs the following steps:

1. Chooses a random $l \in Z_q^*$
2. Computes $U = lQ_A + Q_{TA}$
3. Computes $h = H_1(m, U)$
4. Computes $V = (l + h)S_A$

After signing, A sends $\langle U, V \rangle$ as the signature along with the message m and the public key P_A to the verifier.

Verify: On receiving a signature $\langle U, V \rangle \in \mathcal{S}$ on a message m from user A with identity ID_A and public key P_A, the verifier performs the following steps:

1. Computes $h = H_1(m, U)$
2. Checks whether the equality $e(P, V)e(P_A, Q_{TA}) = e(P_A, U + hQ_A)$ holds. Accepts the signature if it does and rejects otherwise

Note that in the verification process the validity of P_A is implicitly checked against the TA's public key Q_{TA}. The verification fails if P_A does not have the correct form. The scheme of [1] requires an explicit verification for validating the public key, resulting in an additional pairing operation.

5 Analysis

The security and efficiency of our certificateless signature scheme is analyzed in this section.

5.1 Security

Forgery Attacks: We consider the case of forging the signature by calculating the private key. An attacker trying to forge the signature by calculating the private key of a user is computationally infeasible because given `params` and the publicly transmitted information Q_A, $P_A (= sQ_{TA})$, calculating the private key S_A (i.e. stQ_A) is equivalent to CDHP, which is assumed to be computationally hard.

Given the public information Q_A, Q_{TA} and P_A, forging our signature without the knowledge of private key is also equivalent CDHP. Note that the security of our certificateless signature scheme is linked to that of the ID-based signature scheme of [5], which is existentially unforgeable under adaptively chosen message and ID attacks in the random oracle model [2].

Replacement Attack: An attack of concern in CL-PKC is public key replacement as the public keys are not explicitly authenticated (i.e. no certificates are used). An active adversary is allowed to replace any public key with a choice of its own. However, such a replacement attack is not useful unless the adversary has the corresponding private key, whose production requires the knowledge of the partial private key and hence the `master-key`.

Achieving Trust Level 3. As stated earlier, the scheme described above is in trust level 2. Although the TA has no access to the private key of any user, it still can forge a valid signature by replacing a user's public key with a false public key of its own choice. Since the TA can calculate the component tQ_A(the partial private key), it can create another valid key pair for the identity ID_A as $S'_A = s'tQ_A$ and $P'_A = s'tP$, using its own chosen secret s'. Using this key pair, the TA can forge a valid user with identity ID_A without being detected. This case arises due to the fact that the user also can calculate another key pair from the partial private key $D_A = tQ_A$. Note that this type of an attack can be launched only by an adversary who has access to the `master-key` t i.e. Type II adversary.

The scheme can be elevated to trust level 3 by using the alternate key generation technique given in [1]. Using this technique the component Q_A is calculated as $Q_A = H_2(ID_A||P_A)$, binding the identity and the public key. Note that user cannot create another key pair as he cannot compute the component $tH_1(ID_A||P''_A)$ using a secret value s'' of his choice. But, the TA still can create another pair of keys for the same identity. However, such an action can easily be detected as the TA is the only entity having that capability. Note that this scenario is equivalent to a CA forging a certificate in a traditional PKC as the existence of two valid certificates for a single identity would implicate the CA. Hence, our scheme enjoys the trust level 3.

5.2 Efficiency

Table 1 gives comparison of computational efforts required for our scheme with that of the signature scheme in [1], in the Sign and Verify algorithms. It is evident

Table 1. Comparison of Computational Efforts

	Scheme in [1]	Our Scheme
Sign	1P,3S_m,1H_1,1P_a	2S_m,1H_1, 1S_a
Verify	4P,1H_1,1E_{G_2}	3P,1H_1,1P_a,1S_m

P: Pairing Operation
S_m: Scalar Multiplication in G_1
H_1: H_1 hash operation
P_a: point addition of G_1 elements
S_a: addition of two scalar variables(negligible)
E_{G_2}: exponentiation in the group G_2

from the table that the Sign of [1] is costly as it requires the expensive pairing and point addition operations whereas that of our scheme requires none. Our Verify algorithm is also efficient than that of [1] as it requires one less pairing operation. It may also be noted that the public key of our scheme requires only one elliptic curve point whereas that of [1] requires two elliptic curve points.

6 Conclusions

In this paper, we proposed a certificateless signature scheme based on bilinear pairings. Our scheme is computationally efficient than the existing certificateless signature scheme. The proposed efficient certificateless signature scheme can be used in low-bandwidth, low-power situations such as mobile security applications where the need to transmit and check certificates has been identified as a significant limitation.

References

1. Al-Riyami, S., Paterson, K.: Certificateless Public Key Cryptography. In Asiacrypt'03, LNCS **2894**, Springer-Verlag. (2003) 452-473
2. Bellare, M., Rogaway, P.: Random Oracles are Practical: a Paradigm for Designing Efficient Protocols. In 1st ACM Confernece on Computer and Communications Security. ACM Press. (1993) 62-73
3. Boneh, D., Franklin, M.: Identity-based Encryption from the Weil pairing. In Crypto'01. LNCS **2139**, Springer-Verlag, (2001) 213-229
4. Boneh, D., Lynn, B., Shacham, H.: Short Signatures from the Weil Pairing. In Asiacrypt '01. LNCS **2248**, Springer-Verlag. (2001) 514-532
5. Cha, J., Cheon, J.H.: An Identity-Based Signature from Gap Diffie-Hellman Groups. In PKC'03. LNCS **2567**, Springer-Verlag. (2003) 18-30
6. Dankers, J., Garefalakis, T., Schaffelhofer, R., Wright, T.: Public key infrastructure in mobile systems. IEE Electronics and Commucation Engineering Journal, **14(5)** (2002) 180-190
7. Girault, M.: Self-certified public keys. In Eurocrypt'91, LNCS **547**, Springer-Verlag. (1991) 490-497
8. Guttman, P.: PKI: Its not dead, just resting. IEEE Computer, **35(8)** (2002) 41-49.
9. Shamir, A.: Identity-based Cryptosystems and Signature Schemes. In Crypto'84, LNCS **196**, Springer-Verlag. (1984) 47-53
10. Yum, D.H., Lee, P.J.: Generic construction of certificateless signature. In ACISP'04, LNCS **3108**, Springer. (2004) 200-211

ID-Based Restrictive Partially Blind Signatures*

Xiaofeng Chen[1], Fangguo Zhang[2], and Shengli Liu[3]

[1] Department of Computer Science,
Sun Yat-sen University, Guangzhou 510275, China
isschxf@zsu.edu.cn
[2] Department of Electronics and Communication Engineering,
Sun Yat-sen University, Guangzhou 510275, China
isdzhfg@zsu.edu.cn
[3] Department of Computer Science,
Shanghai Jiao Tong University, Shanghai 200030, China
liu-sl@cs.sjtu.edu.cn

Abstract. Restrictive blind signatures allow a recipient to receive a blind signature on a message not know to the signer but the choice of message is restricted and must conform to certain rules. Partially blind signatures allow a signer to explicitly include necessary information (expiration date, collateral conditions, or whatever) in the resulting signatures under some agreement with receiver. Restrictive partially blind signatures incorporate the advantages of these two blind signatures. The existing restrictive partially blind signature scheme was constructed under certificate-based (CA-based) public key systems. In this paper we follow Brand's construction to propose the first identity-based (ID-based) restrictive blind signature scheme from bilinear pairings. Furthermore, we first propose an ID-based restrictive partially blind signature scheme, which is provably secure in the random oracle model.

1 Introduction

Blind signatures, introduced by Chaum [7], allow a recipient to obtain a signature on message m without revealing anything about the message to the signer. Blind signatures play an important role in plenty of applications such as electronic voting, electronic cash schemes where anonymity is of great concern.

Restrictive blind signatures, firstly introduced by Brands [5], which allow a recipient to receive a blind signature on a message not known to the signer but the choice of the message is restricted and must conform to certain rules. Furthermore, he proposed a highly efficient electronic cash system, where the bank ensures that the user is restricted to embed his identity in the resulting blind signature.

The concept of partially blind signatures was first introduced by Abe and Fujisaki [1] and allows a signer to produce a blind signature on a message for a recipient and the signature explicitly includes common agreed information which

* Supported by National Natural Science Foundation of China (No. 60403007) and Natural Science Foundation of Guangdong, China (No. 04205407).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 117–124, 2005.
© Springer-Verlag Berlin Heidelberg 2005

remains clearly visible despite the blinding process. This notion overcomes some disadvantages of fully blind signatures such as the signer has no control over the attributes except for those bound by the public key. Partial blind signatures paly an important role in design efficient electronic cash systems. For example, the bank does not require different public keys for different coins values. On the other hand, the size of the database that stored the previously spent coins to detect double-spending would not increase infinitely over time.

Maitland and Boyd [11] first incorporated these two blind signatures and proposed a provably secure restrictive partially blind signature scheme, which satisfies the partial blindness and restrictive blindness. Their scheme followed the construction proposed by Abe and Okamoto [2] and used Brand's restrictive blind signature scheme. However, their scheme was constructed under the CA-based public key systems. There seems no such schemes under the ID-based public key systems to the best of our knowledge.

The concept of ID-based public key systems, proposed by Shamir in 1984 [12], allows a user to use his identity as the public key. It can simplify key management procedure compared to CA-based systems, so it can be an alternative for CA-based public key systems in some occasions, especially when efficient key management and moderate security are required. Many ID-based schemes have been proposed after the initial work of Shamir, but most of them are impractical for low efficiency. Recently, the bilinear pairings have been found various applications in constructing ID-based cryptographic schemes [3, 4].

Recently, Chow *et al* first presented an ID-based partially blind signature scheme [9]. In this paper, we utilize their scheme to propose an ID-based restrictive partially blind signature scheme from bilinear pairings. Our contribution is two folds: (1) We first propose an ID-based restrictive blind signature scheme using the ID-based knowledge proof for the equality of two discrete logarithm from bilinear pairings. (2) We first introduce the notion of ID-based restrictive partially blind signatures and propose a concrete signature scheme from bilinear pairings. Furthermore, we give a formal proof of security for the proposed scheme in the random oracle.

The rest of the paper is organized as follows: Some preliminaries are given in Section 2. The definitions associated with ID-based restrictive partially blind signatures are introduced in Section 3. The proposed ID-based restrictive blind signature scheme is given in 4. The proposed ID-based restrictive partially blind signature scheme and its security analysis are given in Section 5. Finally, conclusions will be made in Section 6.

2 Preliminaries

In this section, we will briefly describe the basic definition and properties of bilinear pairings and gap Diffie-Hellman group.

2.1 Bilinear Pairings

Let G_1 be a cyclic additive group generated by P, whose order is a prime q, and G_2 be a cyclic multiplicative group of the same order q. Let a, b be elements

of Z_q^*. We assume that the discrete logarithm problem (DLP) in both G_1 and G_2 are hard. A bilinear pairings is a map $e : G_1 \times G_1 \rightarrow G_2$ with the following properties:

1. Bilinear: $e(aP, bQ) = e(P, Q)^{ab}$;
2. Non-degenerate: There exists P and $Q \in G_1$ such that $e(P, Q) \neq 1$;
3. Computable: There is an efficient algorithm to compute $e(P, Q)$ for all $P, Q \in G_1$.

2.2 Gap Diffie-Hellman Group

Let G be a cyclic multiplicative group generated by g, whose order is a prime q, assume that the inversion and multiplication in G can be computed efficiently. We first introduce the following problems in G.

1. Discrete Logarithm Problem (DLP): Given two elements g and h, to find an integer $n \in Z_q^*$, such that $h = g^n$ whenever such an integer exists.
2. Computation Diffie-Hellman Problem (CDHP): Given g, g^a, g^b for $a, b \in Z_q^*$, to compute g^{ab}.
3. Decision Diffie-Hellman Problem (DDHP): Given g, g^a, g^b, g^c for $a, b, c \in Z_q^*$, to decide whether $c \equiv ab \bmod q$.

We call G a gap Diffie-Hellman group if DDHP can be solved in polynomial time but there is no polynomial time algorithm to solve CDHP with non-negligible probability.

Throughout the rest of this paper we define G_1 be a gap Diffie-Hellman group of prime order q, G_2 be a cyclic multiplicative group of the same order q and a bilinear pairing $e : G_1 \times G_1 \rightarrow G_2$. Define four cryptographic secure hash functions $H : \{0,1\}^* \rightarrow G_1$, $H_1 : {G_2}^4 \rightarrow Z_q$, $H_2 : \{0,1\}^* \times G_1 \rightarrow Z_q$ and $H_3 : {G_1}^2 \times {G_2}^4 \rightarrow Z_q$.

3 Definitions

Abe and Okamoto first present the formal definition of partially blind signatures. Restrictive partially blind signatures can be regarded as partially blind signatures which also satisfies the property of restrictiveness. In the context of partially blind signatures, the signer and user are assumed to agree on a piece of information, denoted by info. In real applications, info may be decided by the negotiation between the signer and user. For the sake of simplicity, we omit the negotiation throughout this paper. In the following, we follow this definitions of [2, 5, 10, 9] to give a formal definition of ID-based restrictive partially blind signatures.

Definition 1. *(ID-based Restrictive Partially Blind Signatures) A restrictive partially blind signature scheme is a four-tuple* $(\mathcal{PG}, \mathcal{KG}, \mathcal{SG}, \mathcal{SV})$.

- **System Parameters Generation $\mathcal{PG}$:** *On input a security parameter k, outputs the common system parameters* Params.
- **Key Generation $\mathcal{KG}$:** *On input* Params *and an identity information ID, outputs the private key $sk = S_{ID}$.*
- **Signature Generation $\mathcal{SG}$:** *Let U and S be two probabilistic interactive Turing machines and each of them has a public input tape, a private random tape, a private work tape, a private output tape, a public output tape, and input and output communication tapes. The random tape and the input tapes are read-only, and the output tapes are write-only. The private work tape is read-write. Suppose* **info** *is agreed common information between U and S. The public input tape of U contains ID and* **info**. *The public input tape of S contains* **info**. *The private input tape of S contains sk, and that for U contains a message m which he knows a representation with respect to some bases in* Params. *The lengths of* **info** *and m are polynomial to k. U and S engage in the signature issuing protocol and stop in polynomial-time. When they stop, the public output of S contains either completed or not-completed. If it is completed, the private output tape of U contains either $\perp$ or $(\textbf{info}, m, \sigma)$.*
- **Signature Verification $\mathcal{SV}$:** *On input $(ID, \textbf{info}, m, \sigma)$ and outputs either accept or reject.*

An ID-based restrictive partial blind signature scheme should satisfy the properties of *completeness, restrictiveness, partially blind, unforgeability.* Due to the space consideration, we omit these definitions here. Please refer to the full version of this paper (Cryptology ePrint Archive, Report 2005/319).

4 ID-Based Restrictive Blind Signature Scheme

Brand's restrictive blind signature scheme is mainly based on Chaum-Pedersen's knowledge proof of common exponent [8]. Maitland and Boyd [11] presented a general construction based on Brand's original scheme. In this paper, we first propose an ID-based restrictive blind signature scheme by using the ID-based knowledge proof for the equality of two discrete logarithm from bilinear pairings [6].

- PKG chooses a random number $s \in Z_q^*$ as the *master-key* and set $P_{pub} = sP$. The system parameters are $params = \{G_1, G_2, e, q, P, P_{pub}, H, H_1\}$.
- The signer submits his/her identity information ID to PKG. PKG computes $Q_{ID} = H(ID)$, and returns $S_{ID} = sQ_{ID}$ to the user as his/her private key. For the sake of simplicity, define $g = e(P, Q_{ID}), y = e(P_{pub}, Q_{ID})$.
- Suppose the signed message be $M \in G_1$.[1] The signer generates a random number $Q \in_R G_1$, and sends $z = e(M, S_{ID}), a = e(P, Q)$, and $b = e(M, Q)$ to the receiver.

[1] In applications, if the signed message m is not an element of G_1, we can use a cryptographic secure hash function to map m into an element M of G_1.

– The receiver generates random numbers $\alpha, \beta, u, v \in_R Z_q$ and computes

$$M' = \alpha M + \beta P, A = e(M', Q_{ID}), z' = z^{\alpha} y^{\beta}, a' = a^u g^v, b' = a^{u\beta} b^{u\alpha} A^v.$$

The receiver then computes $c' = H_1(A, z', a', b')$ and sends $c = c'/u \mod q$ to the signer.
– The signer responds with $S = Q + cS_{ID}$.
– The receiver accepts if and only if $e(P, S) = ay^c, e(M, S) = bz^c$. If the receiver accepts, computes $S' = uS + vQ_{ID}$.

(z', c', S') is a valid signature on M' if the following equation holds:

$$c' = H_1(e(M', Q_{ID}), z', e(P, S')y^{-c'}, e(M', S')z'^{-c'}).$$

This is because

$$\begin{aligned} A &= e(M', Q_{ID}) \\ e(P, S') &= e(P, uS + vQ_{ID}) = e(P, S)^u e(P, Q_{ID})^v \\ &= (ay^c)^u g^v = a^u g^v y^{cu} \\ &= a' y^{c'} \\ e(M', S') &= e(M', uS + vQ_{ID}) = e(M', S)^u e(M', Q_{ID})^v \\ &= e(\alpha M + \beta P, S)^u A^v = (bz^c)^{u\alpha} (ay^c)^{u\beta} A^v \\ &= a^{u\beta} b^{u\alpha} (z^{\alpha} y^{\beta})^{c'} A^v \\ &= b' z'^{c'} \end{aligned}$$

Thus, the receiver obtains a signature on the message M' where $M' = \alpha M + \beta P$ and (α, β) are values chosen by the receiver. In addition, in the particular case where $\beta = 0$, the above signature scheme achieves the restrictiveness [11]. For designing a electronic cash system, the system parameters consist of another two random generators P_1 and P_2. A user chooses a random number u as his identification information and computes $M = uP_1 + P_2$. He then with the bank performs the signature issuing protocol to obtain a coin. When spending the coin at a shop, the user must provide a proof that he knows a representation of M' with respect to P_1 and P_2. This restricts M' must be the form of αM. For more details, refer to [5].

5 ID-Based Restrictive Partially Blind Signatures

In this section, we incorporate our ID-based restrictive blind signature scheme and Chow *et al*'s ID-based partially blind signature scheme [9] to propose an ID-based restrictive partially blind signature scheme.

5.1 ID-Based Restrictive Partially Blind Signature Scheme

– **System Parameters Generation $\mathcal{PG}$:** Given a security parameter k. The system parameters are $Params = \{G_1, G_2, e, q, P_{pub}, k, H, H_3\}$.

- **Key Generation $\mathcal{KG}$:** On input Params and the signer's identity information ID, outputs the private key $S_{ID} = sQ_{ID} = sH(ID)$ of the signer.
- **Signature Generation $\mathcal{SG}$:** Let the shared information $\boldsymbol{info} = \Delta$, and a message M from the receiver. Define $g = e(P, Q_{ID}), y = e(P_{pub}, Q_{ID})$.
  - The signer randomly chooses an element $Q \in_R G_1$, and computes $z = e(M, S_{ID})$, $a = e(P, Q)$, and $b = e(M, Q)$. He also randomly chooses a number $r \in_R Z_q^*$, and computes $U = rP$, and $Y = rQ_{ID}$. He then sends (z, a, b, U, Y) to the receiver.
  - The receiver generates random numbers $\alpha, \beta, u, v, \lambda, \mu, \gamma \in_R Z_q$, and computes $M' = \alpha M + \beta P, A = e(M', Q_{ID})$, $z' = z^{\alpha} y^{\beta}$, $a' = a^u g^v$, $b' = a^{u\beta} b^{u\alpha} A^v$, $Y' = \lambda Y + \lambda\mu Q_{ID} - \gamma H(\Delta)$, $U' = \lambda U + \gamma P_{pub}$, $h = \lambda^{-1} H_3(M', Y', U', A, z', a', b') + \mu$, and $c' = hu$. He then sends h to the signer.
  - The signer responds with $S_1 = Q + hS_{ID}, S_2 = (r + h)S_{ID} + rH(\Delta)$.
  - If the equations $e(P, S_1) = ay^h$ and $e(M, S_1) = bz^h$ hold, the receiver computes $S_1' = uS_1 + vQ_{ID}$, and $S_2' = \lambda S_2$.

  The resulting signature for Δ and message M' is a tuple $(Y', U', z', c', S_1', S_2')$.
- **Signature Verification $\mathcal{SV}$:** Given the message M', the shared information Δ and the tuple $(Y', U', z', c', S_1', S_2')$, the verifier computes $A = e(M', Q_{ID})$, $a' = e(P, S_1')y^{-c'}$, and $b' = e(M', S_1')z'^{-c'}$. He accepts the signature if the following equation holds:

$$e(S_2', P) = e(Y' + H_3(M', Y', U', A, z', a', b')Q_{ID}, P_{pub})e(H(\Delta), U').$$

5.2 Security Analysis

Theorem 1. *The proposed scheme achieves the property of completeness.*

Proof. Note that

$$e(P, S_1') = e(P, S_1)^u e(P, Q_{ID})^v = (ay^h)^u g^v = a'y^{c'}$$

$$e(M', S_1') = e(M', S_1)^u e(M', Q_{ID})^v = e(\alpha M + \beta P, S_1)^u A^v = b'z'^{c'}$$

and

$$\begin{aligned}
e(S_2', P) &= e(\lambda S_2, P) \\
&= e((\lambda r + \lambda h)S_{ID} + \lambda r H(\Delta), P) \\
&= e((\lambda r + H_3(M', Y', U', A, z', a', b') + \lambda\mu)S_{ID}, P)e(H(\Delta), \lambda r P) \\
&= e((\lambda r + H_3(M', Y', U', A, z', a', b') + \lambda\mu)Q_{ID}, P_{pub})e(H(\Delta), U' - \gamma P_{pub}) \\
&= e((\lambda r + H_3(M', Y', U', A, z', a', b') + \lambda\mu)Q_{ID} - \gamma H(\Delta), P_{pub})e(H(\Delta), U') \\
&= e(Y' + H_3(M', Y', U', A, z', a', b')Q_{ID}, P_{pub})e(H(\Delta), U') \\
&= e(Y' + H_3(M', Y', U', A, z', a', b')Q_{ID}, P_{pub})e(H(\Delta), U')
\end{aligned}$$

Thus, the proposed scheme achieves the property of completeness.

Theorem 2. *The proposed scheme achieves the property of restrictiveness.*

Proof. Similar to [5, 11], the restrictiveness nature of the scheme can be captured by the following assumption: The recipient obtains a signature on a message that can only be the form $M' = \alpha M + \beta P$ with α and β randomly chosen by the recipient. In addition, in the particular case where $\beta = 0$, if there exists a representation (μ_1, μ_2) of M with respect to bases P_1 and P_2 such that $M = \mu_1 P_1 + \mu_2 P_2$ and if there exists a representation (μ_1', μ_2') of M' with respect to g_1 and g_2 such that $M' = \mu_1' P_1 + \mu_2' P_2$, then the relation $I_1(\mu_1, \mu_2) = \mu_1/\mu_2 = \mu_1'/\mu_2' = I_2(\mu_1', \mu_2')$ holds.

Theorem 3. *The proposed scheme is partially blind.*

Proof. Suppose S^* is given $\perp$ in step 5 of the game in definition 4, S^* determines b with a probability 1/2.

If in step 5, the shared information $\Delta_0 = \Delta_1$. Let $(Y', U', z', c', S_1', S_2', M')$ be one of the signatures subsequently given to S^*. Let $(Y, U, z, a, b, h, S_1, S_2, M)$ be data appearing in the view of S^* during one of the executions of the signature issuing protocol at step 4. It is sufficient to show that there exists a tuple of random blinding factors $(\alpha, \beta, u, v, \lambda, \mu, \gamma)$ that maps $(Y, U, z, a, b, h, S_1, S_2, M)$ to $(Y', U', z', c', S_1', S_2', M')$.

Let $S_2' = \lambda S_2$, $U' = \lambda U + \gamma P_{pub}$ and $Y' = \lambda Y + \lambda\mu Q_{ID} - \gamma H(\Delta)$. The unique blinding factors (λ, μ, γ) are always exist.[2]

Let $u = c'/h$, we know there exists a unique blinding factor v which satisfies the equation $S_1' = uS_1 + vQ_{ID}$. Determine a representation $M' = \alpha M + \beta P$, which is known to exist. Note that $z' = A^s$ and $z = e(M, Q_{ID})^s$ have been established by the interactive proof and the fact that the signature is valid. Therefore, $z' = e(M', Q_{ID})^s = z^\alpha y^\beta$. Since $e(P, S_1) = ay^h$ and $e(M, S_1) = bz^h$, we have $a' = e(P, S_1')y^{-c'} = a^u g^v$ and $b' = e(M', S_1')(z')^{-c'} = a^{u\beta} b^{u\alpha} A^v$.

Thus, the blinding factors always exist which lead to the same relation defined in the signature issuing protocol. Therefore, even an infinitely powerful S^* succeeds in determining b with probability 1/2.

Theorem 4. *The proposed scheme is secure against on the existential adaptively chosen message and ID attacks under the assumption of CDHP in G_1 is intractable and the random oracle.*

Proof. The proof follows the security argument given by Chow *et al* [9].

6 Conclusions

Restrictive partially blind signatures incorporate the advantages of restrictive blind signatures and partially blind signatures, which play an important role in electronic commerce. In this paper we first propose an ID-based restrictive partially blind signature scheme from bilinear pairings. Furthermore, we give a formal proof of security for the proposed schemes in the random oracle.

[2] Though it is difficult to compute (λ, μ, γ), we only need to exploit the existence of them.

References

1. Abe, M., Fujisaki, E.: How to Date Blind Signatures. In: Kim, K., Matsumoto, T. (eds.): Advances in Cryptology-Asiacrypt. Lecture Notes in Computer Science, Vol. 1163. Springer-Verlag, Berlin Heidelberg New York (1996) 244-251
2. Abe, M., Okamoto, T.: Provably Secure Partially Blind Signature. In: Bellare, M. (ed.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 1880. Springer-Verlag, Berlin Heidelberg New York (2000) 271-286
3. Boneh, D., Franklin, M.: Identity-based Encryption from the Weil Pairings. In: Kilian, J. (ed.), Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York (2001) 213-229
4. Boneh, D., Lynn, B., Shacham, H.: Short Signatures from the Weil Pairings. In: Boyd, C. (ed.): Advances in Cryptology-Asiacrypt. Lecture Notes in Computer Science, Vol. 2248. Springer-Verlag, Berlin Heidelberg New York (2001) 514-532
5. Brands, S.: Untraceable Off-line Cash in Wallet with Observers. In: Stinson, D.R. (ed.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 773. Springer-Verlag, Berlin Heidelberg New York (1993) 302-318
6. Baek, J., Zheng, Y.: Identity-based Threshold Decryption. In: Bao, F., Deng, R., Zhou, J. (eds.): Public Key Cryptography. Lecture Notes in Computer Science, Vol. 2947. Springer-Verlag, Berlin Heidelberg New York (2004) 248-261
7. Chaum, D.: Blind signature for Untraceable Payments. In: Chaum, D., Rivest, R.L., Sherman, A.T. (eds.): Advances in Cryptology-Eurocrypt. Plenum Press, (1982) 199-203
8. Chaum, D., Pedersen, T.P.: Wallet Databases with Observers. In: Brickell, E.F. (ed.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 740. Springer-Verlag, Berlin Heidelberg New York (1992) 89-105
9. Chow, S.M., Hui, C.K., Yiu, S.M., Chow, K.P.: Two Improved Partially Blind Signature Schemes from Bilinear Pairings. In: Boyd, C., Nieto, J.M. (eds.): Australasian Information Security and Privacy Conference. Lecture Notes in Computer Science, Vol. 3574. Springer-Verlag, Berlin Heidelberg New York (2005) 316-328
10. Juels, A., Luby, M., Ostrovsky, R.: Security of Blind Signatures. In: Kaliski, B. (ed.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 1294. Springer-Verlag, Berlin Heidelberg New York (1997) 150-164
11. Maitland, G., Boyd, C.: A Provably Secure Restrictive Partially Blind Signature Scheme. In: Naccache, D., Paillier, P. (eds.): Public Key Cryptography. Lecture Notes in Computer Science, Vol. 2274. Springer-Verlag, Berlin Heidelberg New York (2002) 99-114
12. Shamir, A.: Identity-based Cryptosystems and Signature Schemes. In: Blakley, G.R., Chaum, D. (eds.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 196. Springer-Verlag, Berlin Heidelberg New York (1984) 47-53

Batch Verification with DSA-Type Digital Signatures for Ubiquitous Computing*

Seungwon Lee[1], Seongje Cho[2], Jongmoo Choi[2], and Yookun Cho[1]

[1] School of Computer Science and Engineering, Seoul National University, Seoul, Korea 151-742
[2] Division of Information and Computer Science, Dankook University, Seoul, Korea 140-714

Abstract. We propose a new method for verifying bad signature in a batch instance of DSA-type digital signatures when there is one bad signature in the batch. The proposed method can verify and identify a bad signature using two modular exponentiations and $n+3\sqrt{n}/2$ modular multiplications where n is the number of signatures in the batch instance. Simulation results show our method reduces considerably the number of modular multiplications compared with the existing methods.

1 Introduction

There are many security issues in the ubiquitous environment, including confidentiality, authentication, integrity, authorization, and non-repudiation. A proper security strategy has to be devised and implemented depending on the type of data and the cost of possible loss, modification, and stolen data. In this paper, we investigate a method that verifies efficiently digital signatures in batches for supporting secure transactions in ubiquitous computing system including m-commerce.

Batch verification offers an efficient verification of a collection of related signatures. Batch verification improves the system performance by reducing number of modular exponentiation operations in signature verification. However, batch verification brings about another issue of identifying bad signatures when there are invalid signatures. Some good work has been also conducted to verify multiple signatures in batches or to identify which one is the bad signature [1, 2].

One existing method, divide-and-conquer verifier, takes a contaminated batch and split it repeatedly until all bad signatures are identified. Another existing method, hamming verifier, divides a batch instance into sub-instances based on Hamming Code for identifying a single bad signature [1].

In this paper, we propose a new method of identifying the bad signature efficiently when there is a bad signature in batches. Our method first assigns a unique exponent to each signature according to its position in the batch and compute an equation. By looking at the resulting value, we can tell whether

* The present research was conducted by the research fund of Dankook University in 2003.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 125–130, 2005.
© Springer-Verlag Berlin Heidelberg 2005

there exist one or more bad signatures. If there exist one bad signature, we can find out its location.

2 Related Work

In this paper, we mainly focus on DSA-type signatures. In a DSA-type signature scheme with a message m and its signature s, the following equation must hold.

$$g^m \equiv s \bmod p \tag{1}$$

In the equation, g is a generator for a cyclic group of order q, p and q are large prime numbers where q divides $p-1$. The signature s can be verified by checking whether it satisfies the condition of Equation (1).

GIVEN : A batch instance $x = ((m_1, s_1), \cdots, (m_n, s_n))$ of length $n = 2^r - 1$ for some r.

Generic Test ($GT(x, n)$) : $GT(x, n)$ tries to

i) Return "true" whenever all signatures are valid. The test never makes mistakes for this case.

ii) Return "false" whenever there is at least one bad signature. In this case the test makes mistakes with probability 2^{-l}.

Divided-and-Conquer Verifier ($DCV_\alpha(x, n)$) : $DCV_\alpha(x, n)$ tries to

i) If n=1, then run $GT(x, 1)$. If $GT(x, 1)$ is "true", return "true" and exit. Otherwise return x as the bad signature.

ii) If $n \neq 1$, then run $GT(x, n)$. If $GT(x, n)$ is "true", return "true" and exit. Otherwise go to the step 3.

iii) Divide instance x into α batch instances$(x_1, x_2, \cdots, x_\alpha)$ containing $\frac{n}{\alpha}$ signatures each. Apply DCV_α to each α sub instances, i.e. $DCV_\alpha(x_1, \frac{n}{\alpha}), \cdots, DCV_\alpha(x_\alpha, \frac{n}{\alpha})$.

Hamming Code Verifier ($HCV(x, n)$) : $HCV(x, n)$ tries to

i) Apply GT on the input instance. If $GT(x, n) =$ "true", exit. Otherwise, go to the next step.

ii) Create r sub instances. i.e.

$$x_i = \{(m_j, s_j) | h_{i,j} = 1\}$$

for $i = 1, \cdots, r$ where x_i is a sub instance composed from elements of x chosen whenever $h_{i,j}$ is equal to 1(elements of x for which $h_{i,j} = 0$ are ignored).

iii) Apply $GT(x_i, 2^{r-1}) = \sigma_i$ for $i = 1, \cdots, r$, if the test returns "true" then set $\sigma_i = 0$, otherwise set $\sigma_i = 1$. The $syndrome(\sigma_1, \cdots, \sigma_r)$[3] identifies the position of the bad signature.

iv) Apply GT on the input instance without the bad signature. If the batch instance is accepted, return the index of the bad signature. Otherwise, return "false" and exit.

Fig. 1. Batch Verification Methods for DSA

A batch instance $x = ((m_1, s_1), (m_2, s_2), \cdots, (m_n, s_n))$ with n signatures can be verified by checking if it satisfies the following equation [4, 1].

$$g^{\sum_{i=1}^{n} m_i} \equiv \prod_{i=1}^{n} s_i \bmod p \tag{2}$$

In the equation, m_i is a message and s_i is m_i's signature. Typically, the calculation of $\sum_{i=1}^{n} m_i$ is done modulo q. When all of the signatures $s_1, s_2, \cdots, s_n$ in x are valid, Equation (2) is satisfied. However, even though it is rare, there is a possibility that two or more bad signatures in a batch instance affect mutually and make a side effect that Equation (2) is satisfied.

To decrease the possibility, Bellare et al. presented batch verification tests [5]. These batch verification tests, denominated as GT (Generic Test) in previous study [1], make use of a probabilistic algorithm and is defined in Fig.1.

By applying GT, we can decide whether the batch instance contains bad signatures or not. When the test fails, (i.e. $GT(x, n) =$ "false"), the verifier is faced with the problem of identifying bad signatures. Identification of bad signatures can be performed by so called divide-and-conquer verifier (DCV_α), which is defined in Fig.1.

When there is one bad signature in a batch instance, the division in the step *iii*) in Fig.1 is executed by $\log_\alpha n$ times and GT is invoked α times at each division. Then GT is invoked once additionally at the outset. So, DCV_α calls GT $\alpha log_\alpha n + 1$ times [1].

Another method, Hamming verifier, divides a batch instance into sub instances based on Hamming Code for identifying a single bad signature [1]. Hamming Code is described with a block length n and r parity check equations where $n = 2^r - 1$. It allows a quick identification of the error position as the error *syndrome* is the binary index of the position in which the error occurs [3]. Hamming Code verifier(HCV) is defined as follows: HCV calls GT $\log_2 n + 2$ times [1].

3 Proposed Method

We present a method, denominated as EBV (Exponent Based Verifier), that identifies a bad signature for the case when there exists one bad signature in a batch instance. EBV is defined in Fig.2.

If there is one bad signature in a batch instance of size n, EBV calls GT only twice to identify it, while DCV_α calls GT $\alpha log_\alpha n + 1$ times and HCV calls GT $\log_2 n + 2$ times. EBV can identify a bad signature according to step *v*), that is Equation (3). We will elaborate what Equation (3) implies in Theorem 1.

Theorem 1. *Given a batch instance of length n which has one bad signature. If an integer k* ($1 \leq k \leq n$) *satisfies Equation (3), then* (m_k, s_k) *is the bad signature in the batch instance.*

Proof: Assuming that the j-th signature s_j is invalid in a batch instance x. Since all signatures except (m_j, s_j) satisfies Equation (1), we can reduce the fractions($\frac{\prod_{i=1}^{n} s_i}{g^{\sum_{i=1}^{n} m_i}}$ and $\frac{\prod_{i=1}^{n} {s_i}^i}{g^{\sum_{i=1}^{n} i m_i}}$) to their lowest terms as follows:

GIVEN : A batch instance $x = ((m_1, s_1), \cdots, (m_n, s_n))$ of length n.

Exponent Based Verifier ($EBV(x, n)$) : $EBV(x, n)$ tries to

i) Apply GT on the input batch instance and store the intermediate values of $\sum_{i=1}^{n} m_i$ and $\prod_{i=1}^{n} s_i$.

ii) If $GT(x, n) =$ "true", return "true" and exit.Otherwise, go to the next step.

iii) Compute $\sum_{i=1}^{n} im_i$ by addition of all intermediate values of $\sum_{i=1}^{n} m_i$ like as $m_n + (m_n + m_{n-1}) + \cdots + (m_n + m_{n-1} + \cdots + m_2 + m_1)$ and $\prod_{i=1}^{n} {s_i}^i$ by multiplication of all intermediate values of $\prod_{i=1}^{n} s_i$ like as $s_n \times s_n s_{n-1} \times \cdots \times s_n s_{n-1} \cdots s_2 s_1$.

iv) Compute $g^{\sum_{i=1}^{n} im_i}$.

v) Find a integer k which satisfied the following equation where $1 \leq k \leq n$.

$$\left(\frac{\prod_{i=1}^{n} s_i}{g^{\sum_{i=1}^{n} m_i}} \right)^k \equiv \frac{\prod_{i=1}^{n} {s_i}^i}{g^{\sum_{i=1}^{n} im_i}} \bmod p \tag{3}$$

If the integer k exist, go to the next step. Otherwise, return "false" and exit.

vi) Apply GT on the input instance without the k-th signature $x' = ((m_1, s_1), \cdots, (m_{k-1}, s_{k-1}), (m_{k+1}, s_{k+1}), \cdots, (m_n, s_n))$. If $GT(x', n-1)$ is "true", return k as the index of the bad signature. Otherwise, return "false" and exit.

Fig. 2. Batch Identification Method for DSA

$$\frac{\prod_{i=1}^{n} s_i}{g^{\sum_{i=1}^{n} m_i}} \equiv \frac{s_1}{g^{m_1}} \cdot \frac{s_2}{g^{m_2}} \cdots \frac{s_j}{g^{m_j}} \cdots \frac{s_{n-1}}{g^{m_{n-1}}} \cdot \frac{s_n}{g^{m_n}} \equiv \frac{s_j}{g^{m_j}} \bmod p$$

$$\frac{\prod_{i=1}^{n} {s_i}^i}{g^{\sum_{i=1}^{n} im_i}} \equiv \frac{s_1^1}{g^{m_1}} \cdot \frac{s_2^2}{g^{2m_2}} \cdots \frac{s_j^j}{g^{jm_j}} \cdots \frac{s_{n-1}^{n-1}}{g^{(n-1)m_{n-1}}} \cdot \frac{s_n^n}{g^{nm_n}} \equiv \left(\frac{s_j}{g^{m_j}} \right)^j \bmod p.$$

By substituting $\frac{s_j}{g^{m_j}}$ and $(\frac{s_j}{g^{m_j}})^j$ for $\frac{\prod_{i=1}^{n} s_i}{g^{\sum_{i=1}^{n} m_i}}$ and $\frac{\prod_{i=1}^{n} {s_i}^i}{g^{\sum_{i=1}^{n} im_i}}$ in Equation (3) respectively, we can derive the equation $\left(\frac{s_j}{g^{m_j}} \right)^k \equiv \left(\frac{s_j}{g^{m_j}} \right)^j$. This implies that k is equal to j, which is the index of the bad signature. ■

4 Performance Analysis

The EBV reduces significantly the number of GTs invoked compared with the DCV_α and HCV. However, another factor has to be considered to evaluate the overall performance. That is, the EBV causes some additional overhead to solve Equation (3) which is not incurred in the DCV_α and HCV. Consequently, to compare the EBV with the DCV_α and HCV fairly, we measure the performance of each method in terms of the number of modular multiplications. First, we present the computational cost of EBV,DCV_α and HCV in terms of the computational cost of GT. The cost of GT depends on the length of a given batch instance and a security parameter. So, let $CostGT_l(n)$ denote the compu-

Table 1. The Computational cost of EBV, HCV and DCV_α according to the number of GTs invoked

	Number of GTs invoked	**Computational Cost of GTs invoked**
EBV	2	$2CostGT_l(n)$
HCV	$\log_2 n + 2$	$2CostGT_l(n) + \log_2 n \times CostGT_l(\frac{n}{2})$
DCV_α	$\alpha log_\alpha n + 1$	$CostGT_l(n) + \alpha \sum_{i=1}^{\log_\alpha n} CostGT_l(\frac{n}{\alpha^i})$

tational cost of GT when the length of a given batch instance is n and security parameter is l. Table 1 shows the computational cost of EBV,DCV_α and HCV.

Next, we convert the computational cost of GT into the number of modular multiplications and compute the additional overhead of EBV. Since GT is the random subset test [1], GT requires l modular exponentiations and $2 \times l \times n$ modular multiplications where l is a security parameter and n is the size of a batch instance. The additional overhead of EBV is incurred in the step iii), iv) and v) of Fig.2.

In step iii), we can compute $\sum_{i=1}^{n} im_i$ and $\prod_{i=1}^{n} {s_i}^i$ at cost of n modular multiplications. Specifically, while computing $\sum_{i=1}^{n} m_i$ and $\prod_{i=1}^{n} s_i$ in step i),we can obtain the intermediate values. Then by adding all the intermediate values of $\sum_{i=1}^{n} m_i$ like as $m_n + (m_n + m_{n-1}) + \cdots + (m_n + m_{n-1} + \cdots + m_2 + m_1)$ and multiplying all the intermediate values of $\prod_{i=1}^{n} s_i$ like as $s_n \times s_n s_{n-1} \times \cdots \times s_n s_{n-1} \cdots s_2 s_1$, we can obtain $\sum_{i=1}^{n} im_i$ and $\prod_{i=1}^{n} {s_i}^i$. In step iv), we need one modular exponentiation.

We compute the overhead incurred in step v). Let A denote $\prod_{i=1}^{n} s_i / g^{\sum_{i=1}^{n} m_i}$ and B denote $\prod_{i=1}^{n} {s_i}^i / g^{\sum_{i=1}^{n} im_i}$. Then Equation (3) can be abbreviated as $A^k \equiv B \bmod p$. This problem is a kind of the *discrete logarithm problem* (DLP) [6]. The EBV is in a restricted domain of DLP, that is, k lies in the certain interval of integer, say [1,n]. So, we employ Shanks' baby-step giant-step algorithm [7] which can compute k in at most $2\sqrt{n}$ modular multiplications ($3\sqrt{n}/2$ on the average case) [6]. As a result, EBV requires one modular exponentiation and $n + 3\sqrt{n}/2$ modular multiplications additionally.

Table 2 shows the number of modular operations to perform these methods. In this table,we assign 2 to α in DCV_α (2 is optimal value when there is one bad signature in a batch instance [1]). The bold characters in Table 2 represents the additional modular operations of EBV.

One modular exponentiation in DSA can be computed by 208 modular multiplications where exponent length is 160 bits. Then l is commonly set to 60 [2, 8].

Table 2. The number of modular operations to perform EBV, HCV and DCV_2 where $CostGT_l(n)$ consists of l modular exponentiations and $2ln$ modular multiplications

	Computational Cost	
	modular exponentiation	**modular multiplication**
EBV	$2l + \mathbf{1}$	$4ln + \mathbf{n} + \mathbf{3\sqrt{n}/2}$
HCV	$l \log_2 n + 2l$	$l \times n \log_2 n + 4ln$
DCV_2	$2l \log_2 n + l$	$6ln - 4l$

Table 3. The number of modular multiplications of EBV, HCV and DCV_2 where l is 60

	Number of modular multiplications			
	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
EBV	148,594	272,000	518,804	1,012,400
HCV	536,640	1,009,920	2,005,440	4,106,880
DCV_2	421,200	630,480	1,024,080	1,786,320

Table 3 shows the comparison of average computational cost for performing the methods. The table shows that EBV outperforms conventional methods in terms of the number of modular multiplication operations. For example, EBV reduces the number of modular multiplications by 49.4% compared with DCV_2 and by 74.1% compared with HCV when n is 2048.

5 Conclusion

In this paper, we have proposed a new batch verification method EBV. EBV finds a bad signature by calling GT twice when there is one bad signature in a batch instance. The measure of performance shows that EBV is more efficient than DCV_α and HCV in terms of the number of modular multiplications.

For the case when there exist more than one bad signature, we can extend the proposed method using a divide-and-conquer approach to find out the bad signatures. Details of the extension will be left as future research.

References

1. J.Pastuszak, D.Michalek, J.Pieprzyk, J.Seberry: Identification of bad signatures in batches. In: PKC'2000. (2000)
2. J-S.Coron, D.Naccache: On the security of rsa screening. In: PKC99. (1999) 197–203
3. Berlekamp, E.: Algebraic Coding Theory. McGraw-Hill (1968)
4. Harn, L.: Batch verifying multiple dsa-type digital signatures. In: Electronics Letters. (1998) 870–871
5. Bellare, M., Garay, J., Rabin, T.: Fast batch verification for modular exponentiation and digital signatures. Advances in Cryptology-EUROCRYPT'98 (1998) 236–250
6. E.Teske: Computing discrete logarithms with the parallelized kangaroo method. Discrete Applied Mathematics **130** (2003) 61–82
7. Buchmann, J., Jacobson, M., Teske, E.: On some computational problems in finite abelian groups. Mathematics of Computation **66** (1997) 1663–1687
8. Bellare, M., Garay, J.A., Rabin, T.: Batch verification with applications to cryptography and checking. Lecture Notes in Computer Science **1380** (1998)

On Anonymity of Group Signatures*

Sujing Zhou and Dongdai Lin

SKLOIS Lab, Institute of Software, Chinese Academy of Sciences,
4# South Fourth Street, Zhong Guan Cun, Beijing 100080, China
{zhousujing, ddlin}@is.iscas.ac.cn

Abstract. A secure group signature is required to be anonymous, that is, given two group signatures generated by two different members on the same message or two group signatures generated by the same member on two different messages, they are indistinguishable except for the group manager. In this paper we prove the equivalence of a group signature's anonymity and its indistinguishability against chosen ciphertext attacks if we view a group signature as an encryption of member identity. Particularly, we prove ACJT's group signature is IND-CCA2 secure, so ACJT's scheme is anonymous in the strong sense. The result is an answer to an open question in literature.

1 Introduction

Group Signatures. A group signature, which includes at least algorithms of Setup, Join, Sign, Verify, Open and Judge (defined in Section 2), is motivated by enabling members of a group to sign on behalf of the group without leaking their own identities; but the signer's identity could be opened by the group manager, i.e., GM, on disputes.

Models of Group Signatures. In formally, a secure group signature scheme satisfies traceability, unforgeabilty, coalition resistance, exculpability, anonymity and unlinkability [1]. Formal models [2, 3, 4, 5] of secure group signatures compressed the above requirements into redefined *anonymity*, *traceability*, *non-frameability*.

Anonymity. In [1], anonymity means similarly to IND-CPA (indistinguishable against chosen plaintext attacks, [6]), but in [2, 3, 4, 5], it means similar to IND-CCA2 (indistinguishable against chosen ciphertext attacks, Section 2). We mean anonymity in the later strong sense hereafter.

An anonymous generic group signature is constructed based on any IND-CCA2 public encryption scheme [3]. The question is whether an IND-CCA2 public encryption is the minimum requirement to construct an anonymous group signature.

Some group signatures adopting ElGamal encryption are considered not anonymous and it is pointed out that after replacing the ElGamal encryption

* Supported by 973 Project of China (No. 2004CB318004), 863 Project of China (No. 2003AA144030) and NSFC90204016.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 131–136, 2005.
© Springer-Verlag Berlin Heidelberg 2005

with a double ElGamal encryption scheme, an IND-CCA2 public encryption scheme, the group signatures will become anonymous (e.g. [4, 7]). In [8], it is further presented as an open question that whether ACJT's scheme [1] utilizing a single ElGamal encryption scheme provides anonymity. We explore this problem in this paper and answer the open question positively.

We point out that the problem lies in the behavior,specifically Open, of GM or OA (decryption oracle in the case of public encryption scheme).

Take an ordinary ElGamal encryption [9] as an example, let $(T_1 = m_i y^r, T_2 = g^r)$ be a challenge, an adversary can easily change it into a new ciphertext $(my^s T_1, g^s T_2)$ and feed it to the decryption oracle, who definitely would reply with $my^s m_i$ since the query is valid and different from challenge, then the adversary can resolve the challenge problem. In other word, ElGamal encryption is IND-CPA[6].

It is well known that an IND-CCA2 encryption scheme is available by double encrypting the same message under an IND-CPA encryption scheme [10]. The resulting IND-CCA2 ElGamal ciphertext consists of two independent ElGamal encryptions and a proof that the same plaintext is encrypted. The strong security of double-encryption transformed IND-CCA2 schemes comes from the difficulty of composing a valid ciphertext relating to the challenge by an computation bounded adversary, while a uncorrupted decryption oracle only decrypts queried valid ciphertexts.

Nevertheless a half corrupted decryption oracle might just ignore the invalidity of a ciphertext, decrypt any one of the two ciphertext pieces and reply to adversaries. It is possible in reality, for instance, a not well designed decryption software might misuse its decryption key by decrypting whatever it has got before checking the validity of the ciphertext, throw away decryption outputs inadvertently when they are found meaningless.

When ElGamal encryption is embedded in a group signature, e.g., ACJT scheme [1], the intuition is that it is difficult for an adversary to forge a new valid group signature from a challenge group signature, and the open oracle would firstly check the validity of a query before replying with the decrypted content.

In anonymous group signature schemes adopting double ElGamal encryption [4, 7, 8], if GM(OA) is half corrupted, i.e., it would directly open any queried group signature no matter whether the proof included in the ciphertext is correct or not, or the whole group signature is valid or not, anonymity of the group signature scheme is hard to guarantee.

So in case of half corrupted GM(OA), not all IND-CCA2 encryption will ensure anonymity of the group signatures; but for uncorrupted GM(OA) an IND-CPA secure encryption might be enough to ensure anonymity. The point is that GM(OA), i.e., the open oracle should check the validity before applying its private key instead of misusing it.

Our Contribution: We prove the equivalence between anonymity of a group signatures and IND-CCA2 of it, if we view the group signature as a public key encryption of group member identities. Particularly, we prove the ACJT's group signature is IND-CCA2 secure under the DDH assumption, so ACJT's scheme is

anonymous in the strong sense of [3]. The result is an answer to an open question proposed in [8].

2 Formal Definitions

Group Signature [3]. Group manager GM is separated into issuer authority IA and opener authority OA. A group signature $\mathcal{GS}$ is composed of the following algorithms:

Setup. It includes a group key generation algorithm Gkg, and a user key generation algorithm Ukg.
- Gkg: a probabilistic polynomial-time algorithm for generating the group public key gpk and IA's secret key ik, as well as OA's secret key ok, given security parameter 1^{k_g};
- Ukg: a probabilistic polynomial-time run by a group member candidate to obtain a personal public and private key pair (upk_i, usk_i), given security parameter 1^k.

Join. A probabilistic polynomial-time interactive protocol between IA and a member candidate with user public key upk_i that results in the user becoming a new group member in possession of secret signing key gsk_i, i.e., a certificate signed by group issuer. They follow a specified relation R: $R(ID_i,upk_i,usk_i,gsk_i) = 1$. Set $Grp = Grp \cup \{ID_i\}$, where Grp denotes the set of valid group members, with initial value $NULL$.

Sign. A probabilistic polynomial-time algorithm which, on input a message M, gsk_i,upk_i,usk_i,$ID_i \in Grp$, and gpk, returns a group signature σ on M. (m, σ) can also be written as (m, ρ, π), where ρ is a blinding of member identity, π is a proof of correctness of ρ.

Verify. A deterministic polynomial-time algorithm which, on input a message-signature pair (M, σ), and gpk, returns 1 (*accept*) or 0 ($reject$); a group signature (M, σ) is valid if and only if Verify$(M, \sigma) = 1$.

Open. A deterministic polynomial-time algorithm that on input a message-signature pair (M, σ), OA's secret key ok, returns an ID, and a proof π showing its correctness in decryption.

Judge. A deterministic polynomial-time algorithm that takes (M, σ, ID, π) as input, returns 1 (*accept*) or 0 ($reject$) indicating a judgement on output from Open.

Anonymity [3]. A group signature scheme is anonymous if for any polynomial-time adversary $\mathcal{A}$, large enough security parameter k, Adv_A^{anon} is negligible: $Adv_A^{anon} = P[Exp_{\mathcal{A}}^{anon-1}(k) = 1] - P[Exp_{\mathcal{A}}^{anon-0}(k) = 1]$, where experiments Exp^{anon-b}, $b = \{0, 1\}$ are defined as in Table 1. Oracles Ch, $Open$, $SndToU$, $WReg$, USK, $CrptU$ are defined as:

Ch: It randomly chooses $b \in \{0, 1\}$ and generates a valid group signature σ on a given m under keys $(ID_u, upk_b, usk_b, gsk_b)$, where $b \in_R \{0, 1\}$.

$Open$: If input (σ, m) is not valid, it returns $reject$; else it open σ, outputs (ID, π). We emphasize that $Open$ oracle is fully reliable, i.e, decrypts a group signature if and only if it is valid, in analyzing anonymity through this paper.

Table 1. Definitions of Experiments

$Exp_{\mathcal{A}}^{anon-b}(k)$: $(gpk, ik, ok) \leftarrow GKg(1^k)$, $d \leftarrow \mathcal{A}(gpk,ik,Ch,Open,SndToU,$ $WReg, USK, CrptU)$, return d.	$Exp_{\mathcal{A}}^{IND-CCA2-b}(k)$: $(pk, sk) \leftarrow Gk(1^k)$, $d \leftarrow \mathcal{A}(pk, Ch, Dec, Enc)$, return d.

$SndToU$ plays as IA in Join, i.e., generating valid certificates (secret signing keys) gsk_u on queries. $WReg$ resets any entry in registration table (storing Join transcripts) to specified value. USK returns usk_i, gsk_i of specified member i. $CrptU$ sets a corrupted member's upk_i to specified value.

Public Key Encryption [6]. Specify key space $\mathcal{K}$, message space $\mathcal{M}$ and ciphertext space $\mathcal{C}$, a public key encryption scheme based on them consists of the following algorithms:

–Gen: a probabilistic polynomial-time algorithm that on input 1^k outputs a public/secret key pair $(pk, sk) \in \mathcal{K}$;

–Enc: a probabilistic polynomial-time algorithm that on input 1^k, a message $m \in \mathcal{M}$, pk, returns a ciphertext $c \in \mathcal{C}$;

–Dec: a deterministic polynomial-time algorithm that on input 1^k, a ciphertext $c \in \mathcal{C}$, sk, returns a message $m' \in \mathcal{M}$ or a special symbol $reject$.

IND-CCA2 [6]. A public key encryption is indistinguishable against chosen ciphertext attacks if for any polynomial time adversary $\mathcal{A}$, large enough security parameter k, $Adv_{\mathcal{A}}^{IND-CCA2}$ is negligible: $Adv_{\mathcal{A}}^{IND-CCA2} = P[Exp_{\mathcal{A}}^{IND-CCA2-1}(k) = 1] - P[Exp_{\mathcal{A}}^{IND-CCA2-0}(k) = 1]$, where experiments $Exp_{\mathcal{A}}^{IND-CCA2-b}$, $b = \{0, 1\}$ are defined as in Table 1. Oracles $Ch, Open, Enc$ are defined as:

Ch: It randomly chooses $b \in \{0, 1\}$ and generates a valid encryption c of m_b on input (m_0, m_1).

Dec: On a query ciphertext c, it firstly checks its validity and returns decrypted plaintext if valid, else returns $reject$.

Enc: It generates a ciphertext c of queried m.

3 Equivalence of Anonymity and IND-CCA2

Abdalla et al. constructed a public key encryption scheme from any group signature [11], and proved that if the adopted group signature is secure, i.e., fully anonymous (same as anonymous in [3]) and fully traceable [2] , their construction is an IND-CPA secure public key encryption, furthermore it is IND-CCA2 if the message space is restricted to $\{0, 1\}$, but they did not investigate the inverse direction.

It is evident that an IND-CCA2 secure public key encryption alone is impossible to produce a secure group signature because of lack of non-traceability and non-frameability. Nevertheless we show the existence of an equivalence between anonymity of group signatures and IND-CCA2 of corresponding public key encryptions.

An Encryption Scheme of Member Identity. Suppose there exists a group signature $\mathcal{GS}$ as defined in Section 2, let $\mathcal{K} = \{gpk, ik, ok : (gpk, ik, ok) \leftarrow \mathrm{Gpk}(1^{k_g})\}$, $\mathcal{M} = \{ID : R(ID, upk_u, usk_u, gsk_u) = 1 : \exists upk_u \leftarrow Ukg(1^k)$, $gsk_u \leftarrow \mathrm{Join}\ (upk_u\ , ik, gpk)\}$ and $\mathcal{C}$, the following algorithms compose a public key encryption scheme $\mathcal{EI}$:

–Gen: i.e., Gkg, outputs $pk = (gpk, ik)$, $sk = ok$;

–Enc: to encrypt an ID, firstly generate upk_u, usk_u, gsk_u such that $R(ID, upk_u, usk_u, gsk_u) = 1$, select a random $r \in_R \{0,1\}^*$, then run Sign on r, return (σ, r);

–Dec: given a ciphertext (σ, r), run Open, and return an ID and a proof π.

Theorem 1. *If $\mathcal{GS}$ is anonymous, then $\mathcal{EI}$ is IND-CCA2 secure.*

Proof. Suppose $\mathcal{A}$ is an IND-CCA2 adversary of $\mathcal{EI}$, we construct $\mathcal{B}$ to break anonymity of $\mathcal{GS}$.

$\mathcal{B}$ has inputs gpk, ik and accesses of oracles $Ch, Open, SndToU, WReg, USK, CrptU$. It publishes $\mathcal{M}$ and corresponding (upk_u, usk_u, gsk_u), for $ID_u \in \mathcal{M}$. It simulates oracles of $\mathcal{EI}$ as follows:

Decryption Oracle $\mathcal{EI}.Dec$: after getting query ciphertext (m, ρ, π), transfers to $Open$ oracle. If it is valid, $Open$ would return corresponding plaintext, i.e., member's identity ID. $\mathcal{B}$ transfers the reply to $\mathcal{A}$.

Challenge Oracle $\mathcal{EI}.Ch$: after getting query $ID_0, ID_1 \in \mathcal{M}$, selects $m \in_R \{0,1\}^*$ and sends (ID_0, ID_1, m) to its oracle Ch. Ch would choose $b \in_R \{0,1\}$ and generate a group signature of m by (upk_b, usk_b, gsk_b): (m, ρ_b, π_b).

$\mathcal{B}$ may continue to answer queries to $\mathcal{EI}.Open$ except (m, ρ_b, π_b).

$\mathcal{B}$ transfers (m, ρ_b, π_b) to $\mathcal{A}$ who is able to figure out b with probability more than $1/2$. $\mathcal{B}$ outputs whatever $\mathcal{A}$ outputs. □

Theorem 2. *If $\mathcal{EI}$ is IND-CCA2 secure, then the underlying $\mathcal{GS}$ is anonymous.*

Proof. Suppose $\mathcal{A}$ is a adversary against anonymity of $\mathcal{GS}$, we construct $\mathcal{B}$ to break IND-CCA2 security of $\mathcal{EI}$.

$\mathcal{B}$ has access to oracles Ch, Dec. It simulates $\mathcal{GS}$'s oracles $\mathcal{GS}.Ch$, $\mathcal{GS}.Open$, $\mathcal{GS}.\{SndToU, WReg, USK, CrptU\}$ as follows:

Open Oracle $\mathcal{GS}.Open$: after getting query (m, ρ, π), transfers to its decryption oracle Dec. If it is a valid ciphertext, Dec would return the corresponding plaintext, i.e., member's identity ID and π. $\mathcal{B}$ transfers the reply to $\mathcal{A}$.

Oracles of $\mathcal{GS}.\{SndToU, WReg, USK, CrptU\}$: since $\mathcal{B}$ has the private keys of issue authority, it can simulate these oracles easily.

Challenge Oracle $\mathcal{GS}.Ch$: after getting challenge query $(ID_0, upk_0, usk_0, gsk_0)$, $(ID_1, upk_1, usk_1, gsk_1)$ and m, $\mathcal{B}$ transfers them to its challenge oracle Ch, who chooses $b \in_R \{0,1\}$ and generates a valid encryption of ID_b using random m: (m, ρ_b, π_b), i.e., a valid signature of m under $(ID_b, upk_b, usk_b, gsk_b)$. Subsequent proof is the same as in Theorem 1. □

4 Anonymity of ACJT's Group Signature

ACJT's scheme [1] dose not conform to the model of [3] (Section 2) completely, but such aspects are beyond our consideration of anonymity here. The following is a rough description of ACJT's scheme:

–**Setup**. IA randomly chooses a safe RSA modulus n and a, a_0, g, h, specifies two integer intervals Δ, Γ. OA chooses x, sets $y = g^x$. $gpk = \{n, a, a_0, y, g, h\}$, ik is factors of n, $ok = x$.

–**Join**. User selects $usk_i = x_i$,$upk_i = a^{x_i}$, where $x_i \in_R \Delta$, gets $gsk_i = (A_i, e_i), e_i \in_R \Gamma$ from IA. A relation is defined $R : A_i = (a^{x_i} a_0)^{1/e_i} \bmod n$.

–**Sign**. A group signature $(T_1, T_2, T_3, s_1, s_2, s_3, s_4, c)$ is a zero-knowledge proof of knowledge of A_i, x_i, e_i, and T_1, T_2 is a correct encryption of A_i.

–**Open**. OA decrypts $A := T_1/T_2^x$, and a proof of knowledge of decryption key x.

–**Verify**&**JUDGE**. Verification of corresponding zero-knowledge proof.

Theorem 3. *ACJT's scheme is IND-CCA2 secure encryption of* $\mathcal{M} = \{A \in QR_n | \exists e \in \Gamma, \overline{x} \in \Delta, A^e = a^{\overline{x}} a_0\}$*, under DDH assumption in random oracle model.*

The proof is standard [6], so it is omitted for space limit. It follows that:

Theorem 4. *ACJT's group signature is anonymous under DDH assumption in random oracle model.*

References

1. G. Ateniese, J. Camenisch, M. Joye, and G. Tsudik: A practical and provably secure coalition-resistant group signature scheme, Crypto'00, Springer-Verlag, LNCS 1880 (2000) 255–270,
2. M. Bellare, D. Micciancio, and B. Warinschi: Foundations of group signatures: Formal definitions, simplified requirements, and a construction based on general assumptions, Eurocrypt'03, Springer-Verlag, LNCS 2656 (2003) 614–629,
3. M. Bellare, H. Shi, and C. Zhang: Foundations of group signatures: The case of dynamic groups, CT-RSA'05, Springer-Verlag, LNCS 3376 (2005)136–153
4. A. Kiayias and M. Yung: Group signatures: Provable security, efficient constructions and anonymity from trapdoor-holders, http://eprint.iacr.org/2004/076/
5. A. Kiayias and M. Yung: Group signatures with efficient concurrent join, Eurocrypt'05, Springer-Verlag, LNCS 3494 (2005)198–214
6. R. Cramer and V. Shoup: Design and analysis of practical public-key encryption schemes secure against adaptive chosen ciphertext attack, SIAM J. Comput. 33(2004) 167–226
7. A. Kiayias, Y. Tsiounis, and M. Yung: Traceable signatures, Eurocrypt'04, Springer, LNCS 3027 (2004) 571–589
8. L. Nguyen and R. Safavi-Naini: Efficient and provably secure trapdoor-free group signature schemes from bilinear pairings, Asiacrypt'04, Springer-Verlag, LNCS 3329 (2004) 372–386
9. T. E. Gamal: A public key cryptosystem and a signature scheme based on discrete logarithms, Crypto'84, Springer, LNCS 196 (1985)10–18
10. M. Naor and M. Yung: Public-key cryptosystems provably secure against chosen ciphertext attacks, 22nd Annual ACM Symposium on the Theory of Computing, ACM Press (1990) 427–437
11. M. Abdalla and B. Warinschi: On the minimal assumptions of group signature schemes, Information and Communications Security (ICICS 2004), Springer-Verlag, LNCS 3269(2004)1–13

The Running-Mode Analysis of Two-Party Optimistic Fair Exchange Protocols

Yuqing Zhang[1], Zhiling Wang[2], and Bo Yang[3]

[1] State Key Laboratory of Information Security, GUCAS, Beijing 100049, China
zhangyq@gucas.ac.cn
[2] National Key Laboratory of ISN, Xidian University, Xi'an 710071, China
ywswang@163.com
[3] College of Information, South China Agricultural University, Guangzhou 510642, China
Byang0531@163.com

Abstract. In this paper, we present a method of running-mode to analyze the fairness of two-party optimistic fair exchange protocols. After discussing the premises and assumptions of analysis introduced in this technique, we deduce all the possible running modes that may cause attack on the protocols. Then we illustrate our technique on the Micali's Electronic Contract Signing Protocol (ECS1), and the checking results show that there are three new attacks on the protocol.

1 Introduction

Fair exchange protocol is a crucial theoretical basis to make secure electronic commerce and electronic business transactions possible. An exchange is fair if at the end of the exchange, either each player receives the item it expects or neither player does. Early efforts are mainly focused on contracting signing schemes with computational fairness: Both parties exchange their commitments/secrets "bit-by-bit". However, this approach is unsatisfactory in real life. Most modern protocols rely on the intervention of a TTP (trusted third party), which to some extent solves the fairness issue. But since the visible TTP is involved during each session of the protocol, it often generates congestion/bottlenecks and costs more. A more realistic approach for contract signing is to make optimistic use of a trusted third party, where the trusted third party is invoked only if one of the participants misbehaves, or in case of a network error. It hence helps to reduce cost and eliminate bottlenecks, and improves the fairness, timeliness and communication overhead.

The method of running-mode analysis has been proved to be an effective way to analyze security protocols. Now there are many researches on the running-mode analysis of two-party and three-party classical security protocols [1,2], but little attention was paid to the analysis of the fair exchange protocols. By principles of optimistic exchange protocol, we make significant changes to the method available in [1], providing the method of running-mode of two-party optimistic fair exchange protocol. By applying this technique, we experimented with Micali's ECS1 protocol [3] and proved its effectiveness.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 137 – 142, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Differences Between Optimistic Fair Exchange Protocols and Classical Security Protocols

In this section, we will briefly highlight several major differences between optimistic fair exchange protocols and classical security protocols.

(1) The Intruder Model and Properties
Classically the properties of the protocol are such that the entities participating in the protocol want to achieve them together. In such protocols, all intruders have identical abilities, even when they work together. Concerning fair exchange protocols, the malicious party may be one of the participating entities, and each participant knows that the other entities may be potentially dishonest. Thus, protecting players from each other is as important as protecting them from outside intruders. When malicious party colludes with intruder, they could have more power.

(2) The Communication Model
In fair exchange protocols, we need to make stronger assumptions about the underlying network. In 1999, Pagnina and Gartner gave a proof of the impossibility of fair exchange without a trusted third party [4]. The result also implies that it is impossible to build a fair exchange protocol with only unreliable channels. Otherwise, all messages to the TTP could be lost, and the protocol would be equivalent to a protocol without TTP. Due to the impossible result, achieving fairness requires that at least the communication channels with the TTP need to ensure eventual delivery.

(3) The Structure of Protocols
Comparing with classical security protocols, optimistic fair exchange protocols are divided into several sub-protocols, making branching possible, although they should be executed in a given order by an honest entity. Hence, it is not always easy to foresee all possible interleaving of these protocols, or even be sure at which moment in the protocol these sub-protocols can be launched.

Due to the differences stated above, the currently available running-mode techniques cannot be directly applied to analyze the fair exchange protocol without modifications according to the characteristics of fair exchange protocol. Here we make corresponding modification to the current running-mode technique so that it is able to analyze the fairness of the two-party optimistic fair exchange protocols.

3 Premises and Assumptions

(1) Types of attack on the protocol
   ① Single attack launched by intruder or malicious party;
   ② Collusive attack launched by intruder in company with malicious party.
(2) The ability of intruder
   ① For the messages between entities running protocol, the intruder can :
      – Intercept and tamper the messages, he also can decrypt the messages that are encrypted with his key ;
      – Replay any message he has seen (possibly changing plain-text parts), even if he does not understand the contents of the encrypted part;
      – Make use of all he has known and can generate new temporary value.

② The intruder is incapable of intercepting, modifying or replaying the messages between the entities and TTP, but may send his available information to TTP;
③ Besides the abilities stated above, the intruders in collusive attack are also capable of sharing each other's private information, such as secret key.

(3) The communication channels

It is assumed that the communication channel between the entities is unreliable. The messages through it may be lost and delayed. But the communication channels with the TTP should be reliable and the messages through them need to ensure eventual delivery.

All the assumptions made above are different from that of the running-mode analysis method for classical security protocols. And the same assumptions are omitted here.

4 The Running-Modes of the Protocols

Although the optimistic fair exchange protocol relies on a TTP, TTP is only needed in case where one player crashes or attempts to cheat. Besides, according to assumption (3) in section 3, the communication channel with TTP is reliable and secured so that intruders cannot impersonate TTP to receive and send any message. Hence the running modes involving TTP are very simple and have very limited styles. Therefore, we only list the running modes that the attacks may occur. While looking into each running mode, we consider the following cases separately: with and without TTP participation. When considering the running mode with TTP participation, we should insert corresponding sub-protocol modes involving TTP whenever a sub-protocol is possibly involved.

4.1 The Small System Model

Although what we discuss in this paper is the method of analyzing two-party optimistic fair exchange protocol, it indeed involves TTP also. So the small system model is similar to that in [2]. Since it is required to analyze the collusive attack in the fair exchange protocol, the small system model developed herein includes a collusive entity, which is different from that in [2]. As we are interested in the running-mode of the protocol, only the set of identities of all involving parties in the small system is contained.

In our small system, the set of identities of all parties is ID={InitSet,RespSet,ServSet}. The set of the initiators is InitSet={A,$\mathcal{A}$,I,I (A) }.The set of the responders is RespSet={B,$\mathcal{B}$,I,I(B) },and the set of the server is ServSet={S}.The meanings and possible actions of A,B,I, I(A),I(B) and S are the same as that in [2]. The reason why our small system differs from that in [2] lies in the fact that collusive attack needs to be considered in analysis of the fair exchange protocol. When a malicious party makes attacks on the protocol by himself, he is equal to an outside intruder. But in a collusive attack, it is another meaning. So we add $\mathcal{A}$ which represents malicious initiator colluding with outside intruder in InitSet, whereas $\mathcal{B}$ represents malicious responder colluding with outside intruder in RespSet. In a two-party fair exchange protocol, $\mathcal{A}$ and $\mathcal{B}$ can only collude with I respectively. In collusive attack, malicious party can send messages not only in himself identity but also in other collusive party's identity. According to

assumption (3) in section 3, the communication channel with the TTP is security. It means that the intruder cannot impersonate the server S. Therefore, we delete I(S) in ServSet.

4.2 The Running-Modes of the Protocol

From previous discussion, we can list all the possible running-modes between the entities. Our small system has a single honest initiator and a single honest responder, each of who can run the protocol precisely once. Moreover, every run of the protocol has at least an honest entity. Under these hypotheses, the number of simultaneous runs of the two-party optimistic fair exchange protocol is no more than two [2]. In this paper, we only consider the modes of single run and two simultaneous run.

According to the combination of initiators and responders and above restriction of small system, we may conclude that the two-party optimistic fair exchange protocols have only 7 possible running modes when the protocol runs once:

(1) A⇔B ;(2) A⇔ I; (3) A⇔I(B); (4) A⇔$\mathcal{B}$; (5) I ⇔B; (6) I(A)⇔B ;(7) $\mathcal{A}$⇔ B .

Two simultaneous run of the protocol is the combination of once running modes. There are seven once running modes in all, so the number of all the possible modes of two simultaneous runs of the protocol is 49.

Since each honest party can only run the protocol once, we can then discard 37 modes. Each party cannot be honest and malicious at the same time. For collusive attack, we then do not need to check the running modes between malicious party and intruder. Therefore, the other four modes can be discard also. The last eight modes are really four different modes. So there are only four modes remain:

(1)A⇔I I⇔B; (2)A⇔I I(A)⇔B;(3)A⇔I(B) I⇔B;(4)A⇔I(B) I(A)⇔B.

With these running modes, the verification of the two-party optimistic fair exchange protocols can be done by hand.

5 Case Study: The Micali's ECS1 Protocol

In this section we utilize the results obtained in section 4 to verify Micali's fair contract signing protocol (ECS1). Besides the known attacks, we find several new attacks.

5.1 Brief Description of ECS1 Protocol

In order to simplify the formalized description of the protocol, we define the following symbols to describe the protocol ECS1: A represents the initiator Alice, B represents the responder Bob, C represents the contract to be signed, M represents the nonce selected by initiator, $SIG_X(M)$ represents X's signature on a message M and Ex(M) represents encryption of a message M with party X's public key. We assume, for convenience, that M is always retrievable from $SIG_X(M)$. We also explicitly assume (as it is generally true) that ,given the right secret key, M is computable from $E_X(M)$.

We now review Micali's protocol ECS1 as follows (for details see [3]).

Main protocol: ($Z=E_{TTP}(A,B,M)$)
Message 1 A—>B: $SIG_A(C,Z)$
Message 2 B—>A: $SIG_B(C,Z)$, $SIG_B(Z)$
Message 3 A—>B: M
Dispute resolution protocol:
Message 1 B—>TTP: $SIG_A(C,Z)$, $SIG_B(C,Z)$, $SIG_B(Z)$
Message 2 TTP—>A : $SIG_B(C,Z)$,$SIG_B(Z)$
Message 3 TTP—>B: M

5.2 Premises of Analysis of Protocol ECS1

The author of [3] did not say clearly that Z must satisfy $Z=E_{TTP}(A,B,M)$ in their commitment. In this paper we redefine that in each party's commitment Z must satisfy $Z=E_{TTP}(A,B,M)$.Otherwise, the commitment is invalid. Further discussions are based on this redefinition.

In [3], Micali did not describe what TTP is supposed to do if it receives a value Z that does not match the cipher of desired (A, B, M), even though $SIG_A(C,Z)$, $SIG_B(C,Z)$ and $SIG_B(Z)$ are valid signatures. In other words, TTP is not given a dispute resolution policy when $D_{TTP}(Z)\neq(A,B,M)$ for any M. Essentially, it has only the following two choices of dispute resolution policy:

(1) TTP rejects Bob's request, and sends nothing to Bob (and related parties).
(2) TTP forwards M to Bob, and ($SIG_B(C,Z)$, $SIG_B(Z)$) to Alice (and related parties).

In the following parts, we will analyze the ECS1 protocol according the two policies respectively and show that no policy is fair.

5.3 Analysis of Micali's Protocol ECS1

(1) When signature is right but $D_{TTP}(Z)\neq(A,B,M)$,TTP select the first policy.
After analyzing the seven once running modes in 4.2, we find out a new attack at modes I⇔B:

Attack 1 （mode I⇔B） : $Z= E_{TTP}(B,I,M)$
Message 1.1 I—>B: $SIG_I(C, Z)$
Message 1.2 B—>I: $SIG_B(C,Z)$, $SIG_B(Z)$
Message 1.3 I—>B： None

(2) When signature is right but $D_{TTP}(Z)\neq(A,B,M)$,TTP select the second policy.
After analyzing the seven once running modes in 4.2, we do not find any new attack.
In case of policy (2), after analyzing the four modes of two simultaneous run of the protocol in 4.2, we find two new attacks on the protocol at modes (A⇔I,I⇔B) and (A⇔I,I(A) ⇔B)：

Attack 2 （mode A⇔I, I⇔B） ： $Z= E_{TTP}(A,I,M)$
Message 1.1 A—>I: $SIG_A(C,Z)$
Message 2.1 I—>B : $SIG_I(C',Z)$
Message 2.1 B—>I : $SIG_B(C',Z)$, $SIG_B(Z)$

Message 2.3 I—>TTP : $SIG_B(C',Z)$, $SIG_I(C',Z)$, $SIG_I(Z)$
Message 2.4 TTP—>B: $SIG_I(C',Z)$, $SIG_I(Z)$
Message 2.5 TTP—>I : M

Attack 3 （mode A⇔I, I(A)⇔B） : $Z= E_{TTP}(A,I,M)$
Message 1.1 A—>I: $SIG_A(C,Z)$
Message 2.1 I(A)—>B : $SIG_A(C,Z)$
Message 2.1 B—>I(A) : $SIG_B(C,Z)$, $SIG_B(Z)$
Message 2.3 I —>TTP : $SIG_B(C,Z)$, $SIG_I(C,Z)$, $SIG_I(Z)$
Message 2.4 TTP—>B : $SIG_I(C,Z)$, $SIG_I(Z)$
Message 2.5 TTP—>I : M

The first attack on the protocol in [5] can be avoided by the hypotheses in this paper. With this method, we can also find the other two attacks on the protocol that have been found in [5].And these attacks are omitted here (for details see [5]).

6 Conclusions

Based on the characters of fair exchange protocols, we extends the method of running-mode analysis of the two-party security protocol so that it can be used to analyze two-party optimistic fair exchange protocols.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Grant Nos. 60102004 and 60373040).

References

1. Zhang, Y.Q., Li,J.H., Xiao,G.Z.: An approach to the formal verification of the two-party cryptographic protocols.ACM Operating Systems Review, (1999) 33(4): 48–51
2. Zhang,Y.Q., Liu,X.Y.: An approach to the formal verification of the three-principal cryptographic protocols. ACM Operating Systems Review,(2004) 38(1):35–42
3. Micali,S.: Simple and fast optimistic protocols for fair electronic exchange. In:Proc. of 22th Annual ACM Symp. on Principles of Distributed Computing, ACM Press (2003):12–19
4. Pagnia,H., Gartner,C.: On The Impossibility of Fair Exchange without a Trusted Third Party. Technical Report TUD-BS-1999-02, Darmstadt University of Technology (1999)
5. Bao,F., Wang,G.L., Zhou,J.Y.,Zhu,H.F.: Analysis and Improvement of Micali's Fair Contract Signing Protocol. LNCS 3108. Springer-Verlag (2004)176–187

Password-Based Group Key Exchange Secure Against Insider Guessing Attacks*

Jin Wook Byun, Dong Hoon Lee, and Jongin Lim

Center for Information Security Technologies (CIST),
Korea University, Anam Dong, Sungbuk Gu, Seoul, Korea
{byunstar, donghlee, jilim}@korea.ac.kr

Abstract. Very recently, Byun and Lee suggested two provably secure group Diffie-Hellman key exchange protocols using n participant's distinct passwords. Unfortunately, the schemes were found to be flawed by Tang and Chen. They presented two password guessing attacks such as off-line and undetectable on-line dictionary attacks by malicious insider attacker. In this paper, we present concrete countermeasures for two malicious insider attacks, and modify the two group Diffie-Hellman key exchange protocols to be secure against malicious insider password guessing attacks. Our countermeasures do not require additional round costs, hence they are efficient.

1 Introduction

To communicate securely over an insecure public network it is essential that secret keys are exchanged securely. Password-based authenticated key exchange protocol allows two or more parties holding a same memorable password to agree on a common secret value (a session key) over an insecure open network. Most password-based authenticated key exchange schemes in the literature have focused on a *same password authentication* (SPWA) model which provides password-authenticated key exchange using a shared common password between a client and a server [2, 3, 5, 13]. Normally two parties, client and server, use a shared password to generate a session key and perform key confirmation with regard to the session key. Bellovin and Merrit first proposed Encrypted Key Exchange (EKE) scheme secure against dictionary attacks [3]. EKE scheme has been the basis for many of the subsequent works in the SPWA model.

1.1 Related Works and Our Contribution

Recently, many protocols have been proposed to provide password-based authenticated key exchange between clients with their different passwords and some of them have easily broken and re-designed in 3-party and N-party settings

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 143–148, 2005.
© Springer-Verlag Berlin Heidelberg 2005

[12, 9, 7, 10]. In this *different password authentication* (DPWA) model clients generate a common session key with their distinct passwords by the help of a server.

In N-party setting, Byun and Lee suggested two provably secure N-party encrypted Diffie-Hellman key exchange protocols using different passwords [6]. One is an N-party EKE-U in the unicast network and the other is an N-party EKE-M in the multicast network. However, the schemes were found to be insecure by Tang and Chen. In [11], They showed that N-party EKE-U and N-party EKE-M protocols suffered from off-line dictionary attack and undetectable on-line guessing attack by malicious insider attackers, respectively. In this paper, we suggest concrete countermeasures for the malicious insider attacks by Tang and Chen, and present the modified N-party EKE-U and N-party EKE-M protocols to be secure against malicious insider attacks. Our countermeasures do not require additional round costs, hence they are efficient.

2 Attack on N-Party EKE-U and Its Countermeasure

2.1 Overview of N-Party EKE-U

Let G=$\langle g \rangle$ be a cyclic group of prime order q. In N-party EKE-U protocol, three types of functions are used. All clients or server contribute to generation of a common session key by using function $\phi_{c,i}$, $\pi_{c,i}$, and $\xi_{s,i}$ for positive integer i. The description of functions are as follows:

$$\phi_{c,i}(\{\alpha_1, .., \alpha_{i-1}, \alpha_i\}, x) = \{\alpha_1^x, .., \alpha_{i-1}^x, \alpha_i, \alpha_i^x\},$$
$$\pi_{c,i}(\{\alpha_1, .., \alpha_i\}) = \{\alpha_1, .., \alpha_{i-1}\},$$
$$\xi_{s,i}(\{\alpha_1, \alpha_2, .., \alpha_i\}, x) = \{\alpha_1^x, \alpha_2^x, .., \alpha_i^x\}.$$

In the up-flow, C_1 first chooses two numbers in Z_q^* randomly, calculates $X_1 = \phi_{c,1}(X_0, x_1) = \{g^{v_1},\ g^{v_1 x_1}\}$, and sends m_1 to C_2, which is an encryption[1] of X_1 with the password pw_1. Upon receiving m_1, C_2 executes a **TF** protocol with server S. In the **TF** protocol, C_2 sends m_1 to S. Then S selects a random number v_2 and calculates $X_1' = \xi_{s,2}(X_1, v_2)$. Since S knows all clients' passwords, it can construct $m_1' = \mathcal{E}_{pw_2}(X_1')$ and sends it back to C_2. This is the end of **TF** protocol.

On receiving $m_1' = \mathcal{E}_{pw_2}(X_1')$, C_2 decrypts it to get X_1'. Next C_2 chooses its own random number x_2 and computes $X_2 = \phi_{c,2}(X_1, x_2)$. Finally C_2 sends a ciphertext $m_2 = \mathcal{E}_{pw_2}(X_2)$ to the next client C_3. The above process is repeated up to C_{n-2}. The last client C_{n-1} chooses a random number x_{n-1}, and calculates $X_{n-1} = \pi_{c,n-1}(\phi_{c,n-1}(X_{n-2}^{'}, x_{n-1}))$. The function $\pi_{c,n-1}$ only eliminates the last element of $\phi_{c,n-1}(X_{n-2}^{'}, x_{n-1})$. Finally the client C_{n-1} encrypts X_{n-1} with pw_{n-1}, and sends the ciphertext, m_{n-1} to the server S.

In the down-flow, S first decrypts m_{n-1} to get X_{n-1}, chooses a random number v_n, and computes $m_n = \xi_{s,n}(X_{n-1}, v_n)$. For $1 \leq i \leq n-1$, let $m_{n,i} = (g^{x_1 \ldots x_{i-1} x_{i+1} \ldots x_{n-1}})^{v_1 \ldots v_n}$ which is the i-th component of m_n. S encrypts each

[1] We assume that an encryption algorithm of N-party EKE protocols is an ideal cipher $\mathcal{E}$ which is a random one-to-one function such that $\mathcal{E}_K : M \rightarrow C$, where $|M| = |C|$.

$m_{n,i}$ with password pw_i and sends the resulting ciphertexts to the clients. Each client C_i decrypts $\mathcal{E}_{pw_i}(m_{n,i})$ to obtain $m_{n,i}$. Next, C_i computes session key $sk = \mathcal{H}(Clients||K)$ where $K = (m_{n,i})^{x_i} = (g^{x_1 \cdots x_{n-1}})^{v_1 \cdots v_n}$ and $Clients = \{C_1, ..., C_{n-1}\}$.

2.2 Off-Line Dictionary Attack on N-Party EKE-U

In [11], Tang and Chen first presented an off-line dictionary attack on N-party EKE-U protocol by malicious insider attacker as follows.

- **Step 1:** A malicious user $\mathcal{U}_j$ first selects two random values α, β, and sends m_j to its neighbor $\mathcal{U}_{j+1}$ where

$$m_j = \mathcal{E}_{pw_i}(X_j),\ X_j = \{g^{\alpha}, g^{\alpha\beta}, g^{\gamma_3}, ..., g^{\gamma_j}, g^{V_j \xi_j}\}$$
$$\gamma_k = V_j(\xi_j / x_k) \text{ where } x_k \in Z_q^* \text{ and } 3 \le k \le j$$
$$V_j = v_1 \cdot v_2 \cdots v_j,\ \xi_j = x_1 \cdot x_2 \cdots x_j.$$

- **Step 2:** $\mathcal{U}_{j+1}$ just forwards m_j to server $\mathcal{S}$. $\mathcal{S}$ decrypts m_j with password pw_j and computes m_{j+1} with a password pw_{j+1} and a randomly selected value v_{j+1} where

$$m_{j+1} = \mathcal{E}_{pw_{j+1}}(X_{j+1}),\ X_{j+1} = \{g^{\alpha v_{j+1}}, g^{\alpha\beta v_{j+1}}, g^{\gamma_3}, ..., g^{\gamma_{j+1}}, g^{V_{j+1}\xi_j}\}$$
$$\gamma_k = g^{V_{j+1}(\xi_{j+1}/x_k)} \text{ where } x_k \in Z_q^* \text{ and } 3 \le k \le j+1$$
$$V_{j+1} = v_1 \cdot v_2 \cdots v_{j+1},\ \xi_{j+1} = x_1 \cdot x_2 \cdots x_{j+1}.$$

  $\mathcal{S}$ sends m_{j+1} to $\mathcal{U}_{j+1}$
- **Step 3:** $\mathcal{U}_{j+1}$ mounts an off-line dictionary attack on pw_{j+1} with the message m_{j+1}. $\mathcal{U}_{j+1}$ chooses an appropriate password pw'_{j+1} and decrypts m_{j+1} as

$$\mathcal{D}_{pw'_{j+1}}(m_{j+1}) = \{g_1, g_2, ..., g_{j+1}\} \text{ where } g_l \in \mathrm{G} \text{ and } 1 \le l \le j+1$$

  $\mathcal{U}_{j+1}$ checks $g_1^{\beta} = g_2$. This relation leads to an off-line dictionary attack.

2.3 Countermeasure

The main idea to prevent the malicious insider off-line dictionary attacks is that we apply an ephemeral session key instead of password to encrypt keying material between server and clients. In the protocol, we use two encryption functions; one is an ideal cipher $\mathcal{E}$ which is a random one-to-one function such that $\mathcal{E}_K : M \to C$, where $|M| = |C|$ and the other function is a symmetric encryption E which has adaptively chosen ciphertext security. $\mathcal{H}$ is an ideal hash function such that $\mathcal{H} : \{0,1\}^* \to \{0,1\}^l$. The detail descriptions are as follows.

[Description of the modified N-party EKE-U]. In the up-flow, C_1 first chooses two numbers in Z_q^* randomly, calculates $X_1 = \phi_{c,1}(X_0, x_1) = \{g^{v_1},\ g^{v_1 x_1}\}$, and sends m_1 to C_2, which is an encryption of X_1 with the password pw_1.[2] Upon

[2] For $2 \le i \le n-1$, m_i is encrypted with sk_i ephemerally generated between clients and server.

receiving m_1, C_2 executes a **TF** protocol with server S. In the **TF** protocol, C_2 sends m_1 and $\zeta_{c_2}(= \mathcal{E}_{pw_2}(g^{a_2}))$ to S for a randomly selected value $a_2 \in Z_q^*$. Then S selects a random number v_2, b_2 and calculates $X_1' = \xi_{s,1}(X_1, v_2)$. S also computes $\zeta_{s_2}(= E_{pw_2}(g^{b2}))$, $sk_2(= \mathcal{H}(C_2||S||g^{a_2}||g^{b_2}||g^{a_2 b_2}))$, $\eta_2(=E_{sk_2}(X_1^{'}))$, and $Mac_2 = \mathcal{H}(sk_2||2)$, and then sends $\zeta_{s_2}, \eta_2, Mac_2$ back to C_2. Mac_2 is used for key confirmation of sk_2 on client sides. For a key confirmation on server sides, we can use an additional key confirmation of $Mac_2' = h(sk_2||S)$. This is the end of **TF** protocol.

On receiving $\eta_2 = E_{sk_2}(X_1')$, C_2 first calculates sk_2 by decrypting ζ_{s_2} with password pw_2, and decrypts η_2 to get X_1'. Next C_2 chooses its own random number x_2 and computes $X_2 = \phi_{c,2}(X_1, x_2)$. Finally C_2 sends a ciphertext $m_2 = E_{sk_2}(X_2)$ to the next client C_3. The above process is repeated up to C_{n-2}. The last client C_{n-1} chooses a random number x_{n-1}, and calculates $X_{n-1} = \pi_{c,n-1}(\phi_{c,n-1}(X_{n-2}^{'}, x_{n-1}))$. Finally the client C_{n-1} encrypts X_{n-1} with sk_{n-1}, and sends the ciphertext, m_{n-1} to the server S.

Theorem 1. *The modified N-party EKE-U protocol is secure against off-line dictionary attacks by malicious insider users.*

Proof. Due to the limited space, the proof will be presented in the full paper.

3 Attack on N-Party EKE-M and Its Countermeasure

3.1 Overview of N-Party EKE-M Protocol

N-party EKE-M protocol consists of two rounds. One round for generating an ephemeral session key between client and server. The other round is for distributing a common secret key by using the generated ephemeral key. $\mathcal{H}_i$ is an ideal hash function such that $\mathcal{H}_i : \{0,1\}^* \rightarrow \{0,1\}^l$ for $1 \leq i \leq 4$.

In the first round, the single server S sends $\mathcal{E}_{pw_i}(g^{s_i})$ to $n-1$ clients concurrently. Simultaneously each client C_i, $1 \leq i \leq n-1$, also sends $\mathcal{E}_{pw_i}(g^{x_i})$ to the single-server concurrently in the first round. After the first round finished S and C_i, $1 \leq i \leq n-1$, share an ephemeral *Diffie-Hellman* key, $sk_i = \mathcal{H}_1(sid'||g^{x_i s_i})$ where session identifier $sid' = \mathcal{E}_{pw_1}(g^{x_1})||\mathcal{E}_{pw_2}(g^{x_2})||...||\mathcal{E}_{pw_{n-1}}(g^{x_{n-1}})$.

In the second round, S selects a random value N from Z_q^* and hides it by exclusive-or operation with the ephemeral key sk_i. S sends $N \oplus sk_i$ to C_i, $1 \leq i \leq n-1$, concurrently. After the second round finished all clients can get a random secret N using its sk_i, and generate a common session key, $sk = \mathcal{H}_2(SIDS||N)$ where $SIDS = sid'||sk_1 \oplus N||sk_2 \oplus N||...||sk_{n-1} \oplus N$.

3.2 Undetectable On-Line Dictionary Attack on N-Party EKE-M

Second, Tang and Chen presented undetectable on-line guessing attack on N-party EKE-M protocol. The undetectable on-line guessing attack is first mentioned by Ding and Horster [8]. An malicious insider user first guesses a password pw of one of the users and uses his guess in an on-line transaction. The

malicious user verifies correctness of his guess using responses of server S. Note that a failed guess never be noticed by S and other users. Thus, the malicious user can get sufficient information on pw by participating the protocol legally and undetectably many times. The attack on N-party EKE-M is summarized as follows.

- **Step 1:** In the first round, a malicious insider attacker $\mathcal{U}_j$ impersonates $\mathcal{U}_i$ $(1 \leq i \neq j \leq n-1)$, and broadcasts $\mathcal{E}_{pw'_i}(g^{x_i})$ to a server $\mathcal{S}$ by using an appropriate password pw'_i and randomly selected x_i.
- **Step 2:** After finishing the second round, $\mathcal{A}$ can get $\mathcal{E}_{pw_i}(g^{s_i})$ and $m_i = sk_i \oplus N$ sent by S. $\mathcal{U}_j$ computes ephemeral session key $sk'_i = h(sid'||(\mathcal{D}_{pw'_i}(\mathcal{E}_{pw_i}(g^{s_i})))^{x_i})$.
- **Step 3:** $\mathcal{U}_j$ checks $N = m_i \oplus sk'_i$ where $\oplus$ denotes exclusive-or operator. This relation leads to an undetectable on-line guessing attack.

3.3 Countermeasure

The main idea to prevent an undetectable on-line guessing attack is that we use an authenticator $\mathcal{H}_2(sk_1||C_1)$ for an ephemeral session key between clients and server. The malicious user can not generate the authenticator since he does not get s_i, hence the server can detect on-line guessing attack. The detailed explanation is as follows.

[Description of Modified N-party EKE-M]. $sk_i(= \mathcal{H}_1(sid'||g^{x_i s_i}))$ is an ephemeral key generated between S and client C_i in the first round and $sk = \mathcal{H}_3(SIDS||N)$ is a common group session key between clients.

- In the first round, the single server S sends $\mathcal{E}_{pw_i}(g^{s_i})$ to $n-1$ clients concurrently. Simultaneously each client C_i, $1 \leq i \leq n-1$, also sends $\mathcal{E}_{pw_i}(g^{x_i})$ to the single-server concurrently in the first round. After the first round finished S and C_i, $1 \leq i \leq n-1$, share an ephemeral *Diffie-Hellman* key, $sk_i = \mathcal{H}_1(sid'||g^{x_i s_i})$.
- In the second round, S selects a random value N from Z_q^* and hides it by exclusive-or operation with the ephemeral key sk_i. S broadcasts $N \oplus sk_i$ and authenticator $\mathcal{H}_2(sk_i||S)$ to C_i for $1 \leq i \leq n-1$. Concurrently, clients broadcast authenticators $\mathcal{H}_2(sk_i||C_i)$ for sk_i, respectively. S and C_i checks that its authenticator is valid by using sk_i. After the second round finished all clients can get a random secret N using its sk_i, and generate a common session key, $sk = \mathcal{H}_3(SIDS||N)$.

To add the mutual authentication (key confirmation) to N-party EKE-M protocol, we can use the additional *authenticator* $\mathcal{H}_4(sk||i)$ described in [4].

Theorem 2. *The modified N-party EKE-M protocol is secure against undetectable on-line dictionary attacks by malicious insider users.*

Proof. Due to the limited space, the proof will be presented in the full paper.

4 Conclusion and Future Works

We presented countermeasures for off-line and undetectable on-line dictionary attacks against N-party EKE-U and N-party EKE-M protocols, respectively. It would be a good future work to design a generic construction of password-authenticated key exchange protocols in the N-party setting based on any secure and efficient 2-party protocols.

Acknowledgement

We very thank Ik Rae Jeong and Qiang Tang for valuable discussions.

References

1. M. Abdalla, D. Pointcheval: Interactive Diffie-Hellman Assumptions With Applications to Password-Based Authentication, In Proceedings of FC 2005, Springer-Verlag, LNCS Vol. 3570 (2005) 341-356
2. M. Bellare, D. Pointcheval, P. Rogaway: Authenticated key exchange secure against dictionary attacks, In Proceedings of Eurocrypt'00, Springer-Verlag, LNCS Vol.1807(2000) 139-155
3. S. Bellovin, M. Merrit: Encrypted key exchange: password based protocols secure against dictionary attacks, In Proceedings of the Symposium on Security and Privacy (1992) 72-84
4. E. Bresson, O. Chevassut, D. Pointcheval, J. J. Quisquater: Provably authenticated group diffie-hellman key exchange, In proceedings of 8th ACM Conference on Computer and Communications Security (2001) 255-264
5. V. Boyko, P. MacKenzie, S. Patel, Provably secure password-authenticated key exchange using diffie-hellman, In Proceedings of Eurocrypt'00, Springer-Verlag, LNCS Vol. 1807(2000) 156-171
6. J. W. Byun, D. H. Lee: N-party Encrypted Diffie-Hellman Key Exchange Using Different Passwords, In Proc. of ACNS05', Springer-Verlag, LNCS Vol. 3531 (2005) 75-90
7. J. W. Byun, I. R. Jeong, D. H. Lee, C. Park: Password-authenticated key exchange between clients with different passwords, In Proceedings of ICICS'02, Springer-Verlag, LNCS Vol. 2513(2002) 134-146
8. Y. Ding, P. Horster: Undetectable On-line Password Guessing Attacks, ACM Operating System Review 29(1995) 77-86
9. R. C.-W. Phan, B. Goi, "Cryptanalysis of an Improved Client-to-Client Password-Authenticated Key Exchange (C2C-PAKE) Scheme, In Proceedings of ACNS 2005, Springer-Verlag, LNCS Vol. 3531(2005) 33-39
10. M. Steiner, G. Tsudik, M. Waider: Refinement and extension of encrypted key exchange, In ACM Operation Sys. Review 29(1995) 22-30
11. Q. Tang, L. Chen: Weaknesses in two group Diffie-Hellman Key Exchange Protocols, Cryptology ePrint Archive (2005)2005/197
12. S. Wang, J. Wang, M. Xu: Weakness of a password-authenticated key exchange protocol between clients with different passwords, In Proceedings of ACNS 2004, Springer-Verlag, LNCS Vol. 3089(2004) 414-425
13. T. Wu: Secure remote password protocol, In Proceedings of the Internet Society Network and Distributed System Security Symposium (1998)97-111

On the Security of Some Password-Based Key Agreement Schemes

Qiang Tang and Chris J. Mitchell

Information Security Group, Royal Holloway, University of London,
Egham, Surrey TW20 0EX, UK
{qiang.tang, c.mitchell}@rhul.ac.uk

Abstract. In this paper we show that three potential security vulnerabilities exist in the strong password-only authenticated key exchange scheme due to Jablon. Two standardised schemes based on Jablon's scheme, namely the first password-based key agreement mechanism in ISO/IEC FCD 11770-4 and the scheme BPKAS-SPEKE in IEEE P1363.2 also suffer from some of these security vulnerabilities. We further show that other password-based key agreement mechanisms, including those in ISO/IEC FCD 11770-4 and IEEE P1363.2, also suffer from these security vulnerabilities. Finally, we propose means to remove these security vulnerabilities.

1 Introduction

Password-based authenticated key agreement has recently received growing attention. In general, such schemes only require that a human memorable secret password is shared between the participants. In practice, password-based schemes are suitable for implementation in a wide range of environments, especially those where no device is capable of securely storing high-entropy long-term secret keys. Password-based key agreement schemes originate from the pioneering work of Lomas et al. [8]. Subsequently many password-based key establishment schemes have been proposed (for example, those in [1, 2, 3, 4, 5]). Of course, this is by no means a complete list of existing protocols.

The password used in such protocols is often generated in one of the following two ways. Firstly, the password might be randomly selected from a known password set by a third party. In this case the need for users to be able to memorise the password will limit the size of the password set. As a result, the password will possess low entropy. Secondly, a user might be required to select his password from a known password set. In this case, the user is very likely to choose the password based on his personal preferences (such as name, birth date) again in order to memorise the password easily. As a result, even if the password set is large, the password will still possess low entropy. Moreover, for convenience, many users select the same passwords with different partners. For example, in a client-server setting, the client might choose to use the same password with several different servers.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 149–154, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Because of this low password entropy, despite their implementation convenience, password-based key agreement schemes are potentially prone to password guessing attacks, including online dictionary attacks and offline dictionary attacks. In the password-based key agreement protocols described in the literature, much effort has been devoted to prevent such attacks. To restrict online dictionary attacks, the commonly used measure is to set a certain interval between two consecutive protocol executions, and at the same time to limit the number of consecutive unsuccessful executions of the protocol. It is clear that an adversary can easily mount a denial of service (DoS) attack against an honest user. However, means of preventing such attacks are beyond the scope of this paper.

In this paper, we first show that three potential security vulnerabilities exist in Jablon's strong password-only authenticated key agreement scheme [2]. The first password-based key agreement mechanism specified in a draft ISO standard [6] and the scheme BPKAS-SPEKE given in an IEEE standard draft [7], which are both based on Jablon's scheme, also suffer from some of these security vulnerabilities. Other password-based key agreement schemes also suffer from these vulnerabilities. Finally, we show how to remove these vulnerabilities.

2 Description of Jablon's Scheme

In this section, we describe the Jablon scheme. At relevant points we also point out the differences between the Jablon scheme and the first password-based key agreement mechanism (in the discrete logarithm setting) in [6], and the scheme BPKAS-SPEKE (in the discrete logarithm setting) in [7].

In the Jablon protocol, the following parameters are made public. p and q are two large prime numbers, where $p = 2q + 1$. h is a strong one-way hash function. Suppose a user (U) with identity ID_U and a server (S) with identity ID_S share a secret password pw, where pw is assumed to be an integer. When U and S want to negotiate a session key, they first compute $g = pw^2 \bmod p$.

Note that in the first mechanism of ISO/IEC FCD 11770-4 [6] g is instead computed as $h(pw||str)^2$, where str is an optional string. Also, in BPKAS-SPEKE in draft D20 of P1363.2 [7], g is instead computed as $h(salt||pw||str)^2$, where $salt$ is is a general term for data that supplements a password when input to a one-way function that generates password verification data. The purpose of the $salt$ is to make different instances of the function applied to the same input password produce different outputs. Finally, str is an optional string which it is recommended should include ID_S.

U and S perform the following steps.

1. U generates a random number $t_1 \in Z_q^*$, and sends $m_1 = g^{t_1} \bmod p$ to S.
2. After receiving m_1, S generates a random number $t_2 \in Z_q^*$, and sends $m_2 = g^{t_2} \bmod p$ to U. S computes $z = g^{t_2 t_1} \bmod p$, and checks whether $z \geq 2$. If the check succeeds, S uses z as the shared key material, and computes $K = h(z)$ as the shared key.

3. After receiving m_2, U computes $z = g^{t_2 t_1} \bmod p$, and checks $z \geq 2$. If the check succeeds, U uses z as the shared key material, and computes $K = h(z)$ as the shared key. Then U constructs and sends the confirmation message $C_1 = h(h(h(z)))$ to S.

   Note that in both the ISO/IEC FCD 11770-4 and IEEE P1363.2 versions of the mechanism, C_1 is instead computed as:

$$C_1 = h(3||m_1||m_2||g^{t_1 t_2}||g).$$

4. After receiving C_1, S checks that the received message equals $h(h(h(z)))$. If the check fails, S terminates the protocol execution. Otherwise, S computes and sends the confirmation message $C_2 = h(h(z))$ to U.

   Note that in both the ISO/IEC FCD 11770-4 and IEEE P1363.2 versions of the mechanism, C_2 is instead computed as:

$$C_2 = h(4||m_1||m_2||g^{t_1 t_2}||g),$$

5. After receiving C_2, U checks that it equals $h(h(z))$. If the check fails, U terminates the protocol execution. Otherwise, U confirms that the protocol execution has successfully ended.

Finally, note that in the elliptic curve setting the first password-based key agreement mechanism in [6] and the scheme BPKAS-SPEKE in [7] are essentially the same as above, except that g is a generator of the group of points on an elliptic curve.

3 Security Vulnerabilities

In this section we describe three security vulnerabilities in the Jablon protocol, the third of which is of very general applicability. In addition, we show that the standardised password-based key agreement mechanisms in [6, 7] also suffer from certain of these vulnerabilities.

3.1 The First Security Vulnerability

We show that the Jablon protocol suffers from a partial offline dictionary attack, which means that an adversary can try several possible passwords by intervening in only one execution of the protocol.

To mount an attack, the adversary first guesses a possible password pw' and replaces the server's message with $m_2' = (pw')^{2t_2'}$ in the second step of an ongoing protocol instance. The adversary then intercepts the authentication message C_1 in the third step of the same instance and mounts the attack as follows.

1. The adversary sets $i = 1$.
2. The adversary computes $pw'' = (pw')^i$, and checks whether pw'' falls into the password set. If the check succeeds, go to the third step. Otherwise, stop.

3. The adversary checks whether $C_1 = h(h(h((m_1)^{it'_2})))$. If the check succeeds, the adversary confirms that $pw = pw''$. Otherwise, set $i = i + 1$ and go to the second step.

It is straightforward to verify that this attack is valid. We now give a concrete example of how the attack works. Suppose that the password set contains all binary strings of length at most n, where the password pw is made into an integer by treating the string as the binary representation of an integer. Suppose that the adversary guesses a password $pw' = 2$; then he can try $n-1$ passwords $(pw')^i$ $(1 \leq i \leq n-1)$ by intervening in only one execution of the protocol.

However, note that the attack only works when the initial guessed password pw' satisfies $pw' < 2^{n/2}$.

3.2 The Second Security Vulnerability

This security vulnerability exists when one entity shares the same password with at least two other entities. This is likely to occur when a human user chooses the passwords it shares with a multiplicity of servers. Specifically we suppose that a client, say U with identity ID_U, shares a password pw with two different servers, say S_1 with identity ID_{S_1} and S_2 with identity ID_{S_2}. A malicious third party can mount the attack as follows.

Suppose U initiates the protocol with an attacker which is impersonating server S_1. Meanwhile the attacker also initiates the protocol with server S_2, impersonating U. The attacker now forwards all messages sent by U (intended for S_1) to S_2. Also, all messages sent from S_2 to U are forwarded to U as if they come from S_1. At the end of the protocol, U will believe that he/she has authenticated S_1 and has established a secret key with S_1. However S_1 will not have exchanged any messages with U. In fact, the secret key will have been established with S_2.

The above attack demonstrates that, even if the server (S_1) is absent, the attacker can make the client believe that the server is present and that they have computed the same session key as each other. Of course, if U shares the same password with servers S_1 and S_2, then S_1 can always impersonate U to S_2 and also S_2 to U, regardless of the protocol design. However, the problem we have described in the Jablon scheme applies even when U, S_1 and S_2 all behave honestly, and this is not a property that is inevitable (we show below possible ways in which the problem might be avoided).

Based on the descriptions in Section 2, it is straightforward to mount this attack on the first password-based key agreement mechanism in [6]. In fact, this attack also applies to the other key agreement mechanisms in [6]. However, if the identifier of the server is used in computing g, e.g. if it is included in the string str, then this attack will fail. The scheme BPKAS-SPEKE in [7] is thus immune to this attack as long as the recommendation given in [7] to include this identifier in str is followed.

3.3 A Generic Vulnerability

We next discuss a general problem with key establishment protocols, which not only applies to the Jablon protocol, but also all those discussed in this paper. In general this problem may apply if two different applications used by a client-server pair employ the same protocol and also the same keying material. Suppose the client and server start two concurrent sessions (A and B say), both of which need to execute a key establishment protocol.

Suppose also that the protocol instances are running simultaneously, and that an attacker can manipulate the messages exchanged between the client and server. The attacker then simply takes all the key establishment messages sent by the client in session A and inserts them in the corresponding places in the session B messages sent to the server; at the same time all the session B key establishment messages are used to replace the corresponding messages sent to the server in session A. Precisely the same switches are performed on the messages sent from the server to the client in both sessions. At the end of the execution of the two instances of the key establishment protocol, the keys that the client holds for sessions A and B will be the same as the keys held by the server for sessions B and A respectively. That is, an attacker can make the key establishment process give false results without it being detected by the participants.

This problem will arise in any protocol which does not include measures to securely bind a session identifier to the key established in the protocol. In particular, the first password-based key agreement mechanisms specified in FCD 11770-4 [6] and the scheme BPKAS-SPEKE in P1363.2 [7] suffer from this vulnerability, as will many other two-part key establishment schemes, including other schemes in these two draft standards.

4 Countermeasures

The following methods can be used to prevent the two security vulnerabilities discussed above.

1. To prevent the first attack, which only applies to the Jablon protocol, one possible method is to require g to be computed as $g = h(pw||ID_U||ID_S||i) \bmod p$, where i ($i \geq 0$) is the smallest integer that makes g a generator of a multiplicative subgroup of order q in $GF(p)^*$.

   It is straightforward to verify that the proposed method can successfully prevent the first attack.
2. One possible method to prevent the second attack is to include the identities of the participants in the authentication messages C_1 and C_2. In the Jablon scheme, C_1 and C_2 would then be computed as follows:

$$C_1 = h(h(h(z||ID_U||ID_S))),\ C_2 = h(h(z||ID_S||ID_U))$$

Correspondingly, in the first password-based key agreement mechanism in [6], C_1 and C_2 would then be computed as follows:

$$C_1 = h(3||m_1||m_2||g^{t_1 t_2}||g^{t_1}||ID_U||ID_S),$$

and

$$C_2 = h(4||m_1||m_2||g^{t_1 t_2}||g^{t_1}||ID_S||ID_U),$$

3. One possible means of addressing the generic attack described in section 3.3 is to include a unique session identifier in the computation of g in every protocol instance. For example, in the two standardised mechanisms [6, 7] the session identifier could be included in str.

5 Conclusions

In this paper we have shown that three potential security vulnerabilities exist in the strong password-only authenticated key exchange scheme due to Jablon [2] where one of these vulnerabilities is of very general applicability and by no means specific to the Jablon scheme. We have further shown that the first password-based key agreement mechanism in ISO/IEC FCD 11770-4 and the scheme BPKAS-SPEKE in IEEE P1363.2 also suffer from certain of these security vulnerabilities.

References

1. Bellovin, S., Merritt, M.: Encrypted Key Exchange: Password-Based Protocols Secure Against Dictionary Attacks. In: SP '92: Proceedings of the 1992 IEEE Symposium on Security and Privacy, IEEE Computer Society, (1992) 72–84
2. Jablon, D.: Strong Password-Only Authenticated Key Exchange. Computer Communication Review. **26** (1996) 5–26
3. Jablon, D.: Extended Password Key Exchange Protocols Immune to Dictionary Attack. Proceedings of the WETICE '97 Workshop on Enterprise Security. (1997) 248–255
4. Bellare, M., Pointcheval, D., Rogaway, P.: Authenticated Key Exchange Secure against Dictionary Attacks. In Preneel, B., ed.: Advances in Cryptology – EUROCRYPT '00. Volume 1807 of Lecture Notes in Computer Science, Springer-Verlag (2000) 139–155
5. Abdalla, M., Chevassut, O., Pointcheval, D.: One-Time Verifier-Based Encrypted Key Exchange. In Serge, V., ed: Proceedings of the 8th International Workshop on Theory and Practice in Public Key (PKC '05). Volume 3386 of Lecture Notes in Computer Science, Springer-Verlag (2005) 47–64
6. International Organization for Standardization. ISO/IEC FCD 11770–4, Information technology — Security techniques — Key management — Part 4: Mechanisms based on weak secrets. (2004)
7. Institute of Electrical and Electronics Engineers, Inc. IEEE P1363.2 draft D20, Standard Specifications for Password-Based Public-Key Cryptographic Techniques. (2005)
8. Lomas, T., Gong, L., Saltzer, J., Needham, R.: Reducing risks from poorly chosen keys. ACM SIGOPS Operating Systems Review. **23** (1989) 14–18

A New Group Rekeying Method in Secure Multicast

Yong Xu and Yuxiang Sun

College of Mathematics and Computer Science, AnHui Normal University, China
yxu11@mail.ahnu.edu.cn

Abstract. LKH(Logical Key Hierarchy) is a basic method in secure multicast group rekeying. It does not distinguish the behavior of group members even they have different probabilities (join or leave). When members have diverse changing probability or different changing mode, the gap between LKH and the optimal rekeying algorithm will become bigger. If the probabilities of members have been known, LKH can be improved someway, but it can not be known exactly the changing probabilities of members. Based on the basic knowledge of group members' behavior (ex. active or inactive), the active members and inactive members are partitioned in the new method, and they are set on the different location in logical key tree firstly. Then the concept "Dirty Path" is introduced in order to reduce the repeat rekeying overhead in the same path. All these can decrease the number of encryptions in the Group Manager and the network communication overhead. The simulation results indicate that the new method has a better improvement over traditional LKH method even if the multicast group members' behavior could be distinguished "approximately".

1 Introduction

Secure multicast catches more attentions in recent years. With all members in a multicast group to share a common key, secure multicast can keep the advantage of the property of multicast[1]-[7]. The development of multicast services should consider the two constrains, that is forward security and backward security (new joined members can not access the contents of communication before and leave members can not access the contents of communication after). When there is a member changes(join or leave), the group key should be changed, this is called the group rekeying. The efficiency of rekey is an important problem in secure multicast since the dynamic of the multicast group. Factors that determine the efficiency are the storage of group manager and members, the amount of encryption/decryption, and communication overhead. These factors are constrained each other, so almost all the rekey methods till now are tradeoff the above factors in order to satisfy the needs they proposed.

The method of LKH uses key tree (graph) to manage the group keys. In the key tree, the root node is the group key and the leaf nodes are correspond with group members one by one, the key they stored is shared only by group manager and the member. All the other keys are called auxiliary keys and is used to reduce the overhead of group rekey. The overhead of LKH in one rekey procedure is $O(d \log_d N)$, where N is the number of the multicast group, i.e., the group size. Another method called OFT[4] also takes the key tree (binary key tree) to implement the group key

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 155 – 160, 2005.
© Springer-Verlag Berlin Heidelberg 2005

management. Based on the one-way function tree, it can reduce the rekey overhead from $O(d \log_d N)$ to $O(\log N)$. LKH+ [3]and LKH+2[6] optimize the member joining procedure , but they not do any change in member leaving.

In LKH and OFT, one group member holds all the keys in the path from leaf node to root. When the node changes (member joining or leaving), all the keys in the path should be changed. The group manager needs to reproduce some new keys firstly, then encrypts them with corresponding keys from bottom to top, assembles rekey packets, and multicasts it to the whole group. So the size of rekey packets (or number of encryption) and the number of group manager's calculation will be the key factors of which methods are better or not. Based on the analysis of LKH, we present a refined LKH method based on the behavior of group members called R-LKH. When members leave the group, only the group key should be changed and the other keys are keep unchanged. Thus reduces not only the amount of keys need to be produced, but also the overhead of group manager's calculation.

The rest of the paper is organized as follows. In section 2, we describe this new algorithm based on some material analyzing. In section 3, we detail the simulation results and analysis. In section 4, we present some ways to realize this new method. Finally, section 5 is the conclusions and some future works.

2 The New Method Algorithm: R-LKH

The central idea of R-LKH algorithm is to let the auxiliary keys to be changed later without affects the security of multicast. In order to implement it, when there is any member leaves, we only have a group key changed. When there is any member joins, we will rekey the whole path from the leaf node to root. Intuitively, the operation reduces the overhead with combining two paths rekeying into one procedure.

When there are several members leave continually, we can construct a tree contain all keys the leave members hold, and the tree has the same height with the original key tree. Cutting this tree from key tree, we can get several sub-trees of the key tree. If the numbers of sub-trees is i now, when there has any member leaves, the number of sub-tree will add h(h is the height of the sub-tree where the member leaves), so the number of encryption will increase h times in rekey packet. The worst case is the value of h is only 1 less than the height of the key tree.

If we consider that there is not only members leave, but also members join. When members leave and join in turn and we insert the new joining member in the location of the leave member just leaving, this can reduce one whole path rekeying in a leaving-joining procedure.

Those above two cases are special. Actually, although members join and leave alternately, at a period of time, there maybe several consecutive members leave or join since the stochastic behavior of members. So the above algorithm is not adequate to all circumstances.

In order to avoid the limits of the algorithm, we will ameliorate it in some aspects. We introduce a definition at first.

Definition. If the auxiliary key on a path in key tree is not changed when the member leaves, we call the path a "dirty path" or DP for simplicity. If two DP joint at root node, we call the two DP unrelated, otherwise, we call they are related.

From the definition, we know there has no DP in LKH. But in R-LKH, the number of DP is at least 1. In a binary tree, there are 2 unrelated DPs at most. Generally, in the d-ary tree, there are at most d unrelated DPs. To a binary tree with height h, we have conclusion as follows:

Conclusion 1. A binary key tree with 2 unrelated DPs. When the group key need to be rekeyed, the numbers of encryption should be done by key server(or group manager) is $2(h-1)$.

Conclusion 2. If new member joins a binary key tree with one DP, the group manager needs to play *2h* encryptions in rekeying the whole DP. If there exists two irrelated DPs, the encryption will be *3h-1*.

From conclusion 2, we can see that with the increasing of DP, the overhead of rekey increasing too. So in our R-LKH algorithm, we chose the number of DP not exceed 2.

Conclusion 3. In a binary key tree with 2 DPs, when a member leaves, the worst case encryption number of group manager is $4h-9$. For 1 DP, this number becomes $3h-7$.

Conclusion 3 indicates when there is only one member leaves, in the worst case, R-LKH algorithm has no advantage than LKH. Generally, in a binary key tree with two DPs, if the height of sub-tree with the root that of the junction of these two DPs is h_1, then the length of the path they shared is $h-h_1$, so the encryption number of the group manager is: $2h_1-1+h_1-1+h-h_1-1-1+h-1=2h+2h_1-5$. If $h_1<2$, the encryption of R-LKH is lower then LKH. Similarly, for a binary key tree with 1 DPs, when member leaves, the encryption R-LKH needed is $2h_1-1+h_1-1+h-h_1-1-1+1=h+2h_1-3$. If $h_1 \le (h+2)/2$, the encryption number R-LKH needed is less than LKH. Thus, when there has only member leaves, one DP has clearly advantage than two DPs.

From conclusion 2, when there is one DP, the overhead is same as LKH for member joins. But for R-LKH, with the member joins, the DP disappears simultaneously, this reducing the overhead of the member leaving afterward.

Synthetically, from those two aspects, we can see that in a key tree constructing, if we can congregate some type members in a same sub-tree based on some classifying of group members like activity, we will gain a lower overhead on R-LKH algorithm. The late simulations have shown the same results.

3 Simulation Result and Analysis

We select a stochastic 200 member joining and leaving as our example. Consider the effect of leaving members, we design 10 type of leave and join sequences in which the leaving members is from 190 to 100, each step decreases 10. The group size is from 1024 to 32K. During the actual computing, to exploit the randomicity of the changes of members better, we do ten times computing totally for each type and use arithmetic average of them as final results.

3.1 Primary Data Structure of R-LKH(Binary Tree)

```
typedef struct {
bool IsDirty;      // "dirty" flag
bool IsEmpty;      // member is deleted or empty node
int Father;        // no. of father node
int LChild;        // no. of left child node
int RChild;        // no. of right child node
}
```

Key trees of R-LKH and LKH have same construction. We use the full binary tree each for simplicity. The number of node from root to leaf are arranged from 1 to 2*maximum-number-of-group. All nodes in the tree use the same data structure. From the property of binary tree, the node *i*'s father node is [*i*/2], the left child is 2*i* and the right is 2*i*+1. The root's father node and the leaf's child node are all set to 0.

3.2 Simulation Results

The above simulation results indicate that when the changing member is aggregated in the key tree, for different group size, whenever the number of changing member in group, R-LKH has an definite advantage than LKH(see the above figures right-hand). When the member's changing range is double extended (see the above figures left-hand), at the location of "join/leave members" coordinates 4, i.e., the joining member

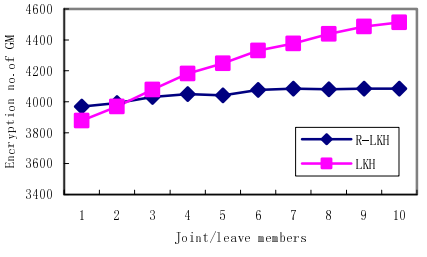


Fig. 1. 4096 members, range in 1/8 key tree

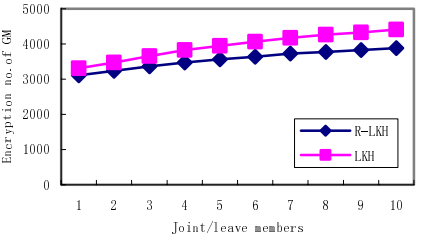


Fig. 2. 4096 members, range in 1/16 key tree

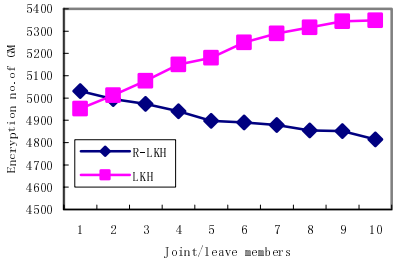


Fig. 3. 16K members, range in 1/16 key tree

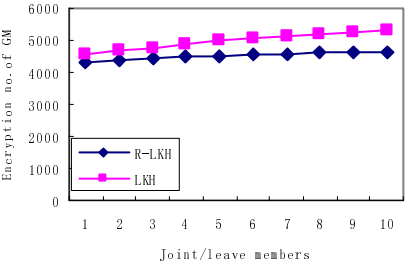


Fig. 4. 16K members, range in 1/32 key tree

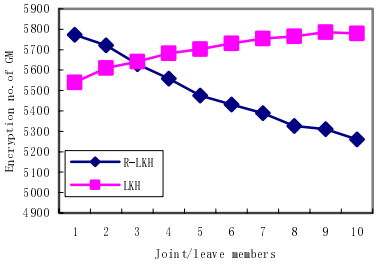


Fig. 5. 32K members, range in 1/16 key tree

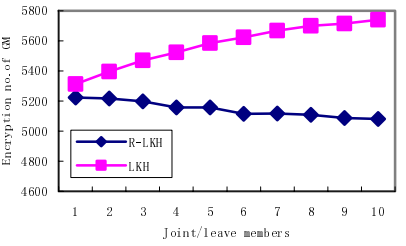


Fig. 6. 32K members, range in 1/32 key tree

is 40 and the leaving member is 160(the proportion of joining and leaving is equal to 25%), R-LKH have general advantage than LKH initiatorily. With the increasing proportion of joining and leaving, the advantage of R-LKH increases gradually, especially after the coordinates 8, i.e., the proportion of member joining and leaving is 66.7%, R-LKH has more advantages. For a variety group size, the advantage is above 8% generally.

When the multicast group becomes steady, the leaving members are almost as many as joining members. Therefore, in the construction of key tree, how to assemble the members that maybe changing in one multicast session in a sub-tree will be a key factor in implementing the R-LKH algorithm efficiently.

4 Implement of R-LKH

It is impossible to know the rigorous leaving probability as the multicast session begins. But at the beginning of multicast group initialization, it is feasible to ask would-be members to tell their behavior in the future. These behaviors mainly include the frequency or duration of members in the session and weather members is enrolled or not. Through the statistical comparison and analysis of member behaviors, we can divide the group members into two categories: active members and inactive members[8]. The active members stay shorter in the group or maybe only the temporary members. Otherwise, the inactive members stay longer or are the enrolled members. We arrange the active members in a smallest tree as we can, the inactive members can be put in the other locations. From the above simulation results, we can find that it is not very strictly for the location of the active members in the key tree. For example, in a group size of 8K, we can get better result when we put the active members in a sub-tree size from 512 to 1024. In a group size of 32K, the size of sub-tree can be 1024 upwards. In the rage of the size, R-LKH is more efficient than LKH in the rekeying. As a result, in R-LKH, it does not need any special computing that we can determine the locations of all members. So R-LKH has more flexibility and availability. Furthermore, in most circumstance of rekeying, only the group key needs to be changed, so the number of new keys the group manager needs to produce is reduced. In the same time, encryption that the manager needs to be done is only encrypting the same entity using different keys, so it has more efficiency.

5 Conclusion and Future Works

In this work we proposed a new algorithm R-LKH for secure multicast group key management. Based on the behaviors of members, R-LKH has less overhead than LKH in group rekeying. Furthermore, R-LKH does not need to known the exactly probability of members. Approximately members' behavior is enough, so R-LKH is easy to implement in practice. Certainly, the more we know the group members' behavior, the better we can gain from R-LKH. It is convincing that with the development of multicast deployment, the model of multicast member behaviors will be known deeply. Therefore, the future work we will do is to consummate the current algorithm at first, then we will design a model of secure multicast with a specific multicast application and have our algorithm validated.

References

1. Wong, C.K., Gouda, M., Lam, S.S.: Secure group communications using key graphs. IEEE/ACM Transactions on Networking, Vol.8(2000)16-30.
2. Wallner, D., Harder, E., Agee, R.. Key management for multicast: Issues and architectures. RFC 2627(1999).
3. Sherman, A.,T., Mcgrew, D.,A.:Key establishment in large dynamic groups using one-way function trees, IEEE Transaction on Software Engineering,Vol. 29(2003) 444-458.
4. Waldvogel, M., Caronni, G., Sun, D., Weiler, N., Plattner,B.:The versaKey framework: versatile group key management. IEEE Journal on Selected Areas in Communications, Vol,17(1999)1614-1631.
5. Rafaeli, S., Mathy, L., Hutchison, D.: LKH+2: An improvement on the LKH+ algorithm for removal operations, Internet draft(work in progress), Internet Eng. Task Force(2002).
6. Pegueroles, J., Bin, W., Rico-Novella, F., Jian-Hua, L.: Efficient multicast group rekeying, In: Proc.of the IEEE Globecom 2003. San Francisco:IEEE Press(2003).
7. Pegueroles, J., Rico-Novella, F., Hernández-Serrano, J., Soriano, M.: Improved LKH for batch rekeying in multicast groups. In:IEEE International Conference on Information Technology Research and Education (ITRE2003). New Jersey:IEEE Press(2003)269-273.
8. Setia, S., Zhu, S., Jajodia, S.: A scalable and reliable key distribution protocol for multicast group rekeying, Technical Report, http://www.ise.gmu.edu/techrep/2002/02_02.pdf, Fairfax:George Mason University(2002).

Pairing-Based Provable Blind Signature Scheme Without Random Oracles

Jian Liao, Yinghao Qi, Peiwei Huang, and Mentian Rong

Department of Electronic Engineering, ShangHai JiaoTong University,
Shanghai 200030, China
{liaojian, qyhao, pwhuang, rongmt}@sjtu.edu.cn

Abstract. Blind signature allows the user to obtain a signature of a message in a way that the signer learns neither the message nor the resulting signature. Recently a lot of signature or encryption schemes are provably secure with random oracle, which could not lead to a cryptographic scheme secure in the standard model. Therefore designing efficient schemes provably secure in the standard model is a central line of modern cryptography. Followed this line, we proposed an efficiently blind signature without using hash function. Based on the complexity of q-SDH problem, we present strict proof of security against one more forgery under adaptive chosen message attack in the standard model. A full blind testimony demonstrates that our scheme bear blind property. Compared with other blind signature schemes, we think proposed scheme is more efficient. To the best of our knowledge, our scheme is the first blind signature scheme from pairings proved secure in the standard model.

1 Introduction

The concept of blind signatures was introduced by Chaum [1]. In the last couple of years, the bilinear pairings have been found various applications in cryptography. Recently, many signature [3-9] and encryption schemes based on bilinear pairing have been presented. In these cryptographic schemes, some were proved secure in standard model [2-7], whereas to the best of our knowledge, no formal notion of security has ever been studied, nor proved in the context of blind signatures based on pairings without random oracle. Our paper follows this line to present a blind signature schemes based on pairing and prove secure without random oracles.

Juels [2] proposed the first complexity-based proof for blind signature and constructed a secure digital cash scheme not from pairings, but in fact the authors themselves presented the proof of security against existential forgery, not against one more forgery. Boldyreva [3] presented a blind signature based on pairings and proved its security in random oracle model assuming the *Chosen-Target CDH* problem is hard. The security proof they adapted was similar to Chaum's RSA-based blind signature scheme [1]. Zhang presented two ID-based blind signature schemes [4][5]. These schemes were more efficient than other previous blind signature schemes, but they still remained some open problems to be solved in the future. In [4], Zhang pointed out that their scheme were difficult to prove that the security against the one more forgery depend on the difficulty of ROS-problem. In [5], there was no proof on secu-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 161 – 166, 2005.
© Springer-Verlag Berlin Heidelberg 2005

rity against one-more forgery. Based on analyzing the linkability of the ZSS's [5] scheme, Sherman [6] proposed two improved partially blind signature schemes from bilinear pairings and verified the unlinkability of *M*, which user want to be signed. Sherman also proofed that their two improved schemes is secure against existential forgery in random oracle model, therefore more formal proof must be found to verify the security of their schemes. Summarizing the work of Boldyreva, Zhang [7] presented a partially blind signature from pairings based on *Chosen Target Inverse CDH* assumption. The scheme was secure against one more forgery within random oracle, while their scheme was less efficient. Camenisch [8] presented a blind signature scheme without random oracles, but there was no proof of security against one more forgery and strongly unforgeability.

In this paper, we proposed an efficiently blind signature using bilinear pairings and analyzed their security and efficiency. We present the strict proof of security against one more forgery in standard model using a complexity assumption, Strong Diffie-Hellman assumption (SDH) [9]. A full blind testimony demonstrates that our scheme meet with the blind property. The conclusion shows that our blind signature scheme prevents against one-more forgery under adaptive chosen message attack. Compared with other blind signature schemes, we think proposed scheme is more efficient.

The rest of paper is organized as follows: In next section, our blind signature is presented in Section 2. We demonstrate that our scheme is blind, unlinkable and prove secure in Section 3. Finally, Section 4 concludes our paper.

2 Our Blind Signature from Pairings

Let G_1 be a cyclic group generated by *P*, whose order is a prime *p*, and G_2 be a cyclic multiplicative group of the same order *p*. The discrete logarithm problems in both G_1 and G_2 are hard. Let $e: G_1 \times G_1 \to G_2$ be a pairing which satisfies three properties: bilinear, non-degenerate and computability [6]. Our paper adopt Strong *q*-SDH problem. The basic blind signature scheme consists of the following four algorithms: ParamGen, KeyGen, Blind Signature Issuing Protocol, and Verification.

ParamGen: the system parameters are $\{G_1, G_2, e, q, P\}$, P is the generator of G_1, $e: G_1 \times G_1 \to G_2$.

KeyGen: Signer chooses randomly $s \in_R Z_q^*$ and computes $P_{Spub} = sP$. The public/secret key pair of Signer is $\{s, P_{Spub}\}$.

Blind Signature Issuing Protocol: Assumed that $m \in Z_q^*$ is the message to be signed. The signer chooses randomly $b \in_R Z_q^* / s$, computes $w = bP$ and sends commitment w to the user.

Blinding: the user chooses randomly $a \in_R Z_q^*$, computes $U = aw$, $V = aP_{Spub}$, $u = mP_{Spub} + U + V$ and sends u to the signer.

Signing: the signer computes $v = (b+s)^{-1}(u + bP_{Spub})$ and sends v to user.

Unblinding: the user computes $\delta = v - aP$ and then outputs $\{m, U, V, \delta\}$ as a signature.

Verification: Given a signature, verify $e(\delta, U+V) = e(mP, V)e(U, P_{Spub})$.

3 Analysis of Our Blind Signature

3.1 Completeness

The completeness can be justified by the follow equations:

$$
\begin{aligned}
& e(\delta, U+V) = e(v - aP, aw + aP_{Spub}) \\
& = e((b+s)^{-1}(u + bP_{Spub}) - aP, a(b+s)P) \\
& = e((b+s)^{-1}(mP_{Spub} + U + V + bP_{Spub}) - aP, a(b+s)P) \\
& = e((b+s)^{-1}(mP_{Spub} + aw + aP_{Spub} + bP_{Spub}) - aP, a(b+s)P) \\
& = e((b+s)^{-1}(mP_{Spub} + bP_{Spub}) + aP - aP, a(b+s)P) \\
& = e(mP_{Spub} + bP_{Spub}, aP) = e(mP_{Spub}, aP)e(bP_{Spub}, aP) \\
& = e(mP, V)e(U, P_{Spub})
\end{aligned}
$$

3.2 Blindness

Considering the *Blind Signature Issuing Protocol* of our scheme, we can prove that the signer can learn no information on the message to be signed similar to the proof of blindness property in [5, 6]. Due to the limitation of length, the detail proof is omitted here.

3.3 Unforgeability

Our proof of unforgeability without random oracles is based on complexity of strong q-SDH assumption. Similar to [9], we require that the challenger must generate all commitments $(b_1, \cdots, b_q)$ before seeing the public key and sent them to the adversary.

We assumed that the adversary control the user to interact with the signer and adaptively choose message to query the signer. For simplicity, we omit the procedure of blinding and unblinding in *Blind Signature Issuing Protocol*, which have on any assist for the user or adversary (adversary/forger and challenger defined as below) to forge a signature. We assumed that the adversary output $u = m'P$, where the adversary construct m' as his/her wish, but the adversary himself must know the relationship between the messages m and m', it is convenient that the adversary himself can recover the message m from m'. Owing to the transformation between m and m' is not in the interaction between the adversary and signer, the adversary start adaptive chosen message attack means that the adversary can adaptively choose message m' to be signed, instead of m. For convenience, we only discuss m' as the message to be signed.

Lemma: Suppose the forger F is a one more forgery under an adaptive chosen message attack, s/he can (q_S, t, ε) break our scheme. Then there exists a (q, t', ε') challenger C solve the q-SDH problem, where $\varepsilon = \varepsilon'$, $q_S < q$ and $t \le t' - \theta(q^2 T)$.

T is the maximum time for an exponentiation in G_1.

Proof: We assume (q, t, ε) forger F break our blind signature scheme. The challenger C is an efficient (q, t', ε') algorithm to solve the q-SDH problem in G_1. We must prove that the forger F breaks our scheme with a non-negligible advantage well then the challenger C also solve the q-SDH problem with the approximative non-negligible advantage. The challenger C is given an $q+1$ -SDH tuple $(P, sP, s^2P, \cdots, s^qP)$, $A_i = s^i P \in G$ for $i = 0,1,\cdots,q$ and for some unknown $s \in_R Z_p^*$. C's goal is to challenge the strong q-SDH problem to produce a pair $(b, (b+s)^{-1}P)$ for some $b \in_R Z_p^*$. The forger F play a game with the challenger C as follows:

KeyGen: We assumed that the forger F make q signature query to C. The challenger C must firstly random choose $(b_1, \cdots, b_q) \in Z_p^*$, and send them to F. The challenger must promise that when forge a signature they must use these commitments in serial order, or the forger and the challenger may employ different commitment to interact. Let $f(x)$ be the polynomial $f(x) = \prod_{i=1}^{q}(x + b_i) = \sum_{i=0}^{q} \alpha_i x^i$, where $\alpha_0, \cdots, \alpha_q \in Z_p^*$. It is easy to compute α_i, $i = 0,1,\cdots,q$ because the commitments b_i, $i = 0,1,\cdots,q$ give out front. Then C computes:

$$Q \leftarrow f(s)P = \sum_{i=0}^{q} \alpha_i s^i P = \sum_{i=0}^{q} \alpha_i A_i$$

$$Q_{Spub} \leftarrow sQ = sf(s)P = \sum_{i=0}^{q} \alpha_i s^{i+1} P = \sum_{i=1}^{q+1} \alpha_{i-1} A_i$$

where C is given the tuple $(A_0, \cdots, A_{q+1})$ of the q-SDH problem. The public key given to F is (Q, Q_{Spub}).

Query: The forger F adaptively chooses messages m_i', computes $u_i = m_i' Q$ and then sends (u_i, m_i') to C, $i = 1, \cdots, q$. As analyzed above, it is for the simplicity that we present the secure proof and for convenience that the challenger forges a signature.

Response: For every query from the forger F, the challenger C must produce a valid signature v_i on message m_i'. C computes as follow:

Let $f_i(x)$ be the polynomial $f_i(x) = f(x)/(x + b_i) = \prod_{j=1, i \ne j}^{q}(x + b_j) = \sum_{i=0}^{q-1} \beta_i x^i$, where $\beta_0, \cdots, \beta_{q-1} \in Z_q^*$ which is easy to be computed in the same way to α_i. Then C produces:

$$v_i \leftarrow (b_i + s)^{-1} u_i = (b_i + s)^{-1}(m_i' + b_i)Q = f(s)(b_i + s)^{-1}(m_i' + b_i)P$$
$$= (m_i' + b_i) f_i(s) P = (m_i' + b_i) \sum_{i=0}^{q-1} \beta_i s^i P = (m_i' + b_i) \sum_{i=0}^{q-1} \beta_i A_i$$

Observe that v_i is a valid signature on m_i' under the public key (Q, Q_{Spub}). For every signature query from the forger, the challenger must response without any hesitation. After received v_i, the forger recovers the message m_i and signature δ_i.

Output: After q query, the forger must output a forge signature (m_*, b_*, δ_*), namely (m_*', b_*, v_*) where $b_* \notin \{b_1, \cdots, b_q\}$. The signature must be valid, so $e(\delta_*, b_* P + Q_{Spub}) = e((m_* + b_*)Q, P)$ will be hold. And below equation also hold.

$$v_* = (m' + b_*)_* (s + b_*)^{-1} Q$$

Let $f(x)$ be the polynomial $f(x) = \gamma(x)(x + b_*) + \gamma_{-1}$, using long division we get that $\gamma(x) = \sum_{i=0}^{q-1} \gamma_i x^i$ and $\gamma_{-1} \in Z_q^*$, where $\gamma_{-1}, \gamma_0, \cdots, \gamma_{q-1} \in Z_q^*$ which is easy to be computed in the same way to α_i. And then we get $f(x)/(x + b_*) = \gamma(x) + \gamma_{-1}/(x + b_*)$ $= \sum_{i=0}^{q-1} \gamma_i x^i + \gamma_{-1}/(x + b_*)$, hence the signature will be:

$$\begin{aligned} v_* &= (m_*' + b_*)(s + b_*)^{-1} Q = (m_*' + b_*) f(s)(s + b_*)^{-1} P \\ &= (m_*' + b_*)[\sum_{i=0}^{q-1} \gamma_i s^i P + \gamma_{-1} P/(s + b_*)] \end{aligned}$$

Observe that since $b_* \notin \{b_1, \cdots, b_q\}$, we get $\gamma_{-1} \neq 0$. And then we compute:

$$\begin{aligned} \eta &= ((m_*' + b_*)^{-1} v_* - \sum_{i=0}^{q-1} \gamma_i A_i)/\gamma_{-1} = \{[\sum_{i=0}^{q-1} \gamma_i s^i P + \gamma_{-1} P/(s + b_*)] + \sum_{i=0}^{q-1} \gamma_i A_i\}/\gamma_{-1} \\ &= \{[\sum_{i=0}^{q-1} \gamma_i s^i P + \gamma_{-1} P/(s + b_*)] + \sum_{i=0}^{q-1} \gamma_i A_i\}/\gamma_{-1} = \{\gamma_{-1} P/(s + b_*)\}/\gamma_{-1} = (s + b_*)^{-1} P \end{aligned}$$

The challenger C outputs $\eta = (s + b_*)^{-1} P$ as the solution of the strong q-SDH problem. If the forger F breaks our scheme with a non-negligible advantage ε well then the challenger C also solve the q-SDH problem with the approximative non-negligible advantage ε'. At the same time, q-SDH problem is impossible to be solved in actual situation, which lead to a paradox.

3.4 Efficiency

We consider the costly operations that include point addition on G_1 (denoted by Ga), point scalar multiplication on G_1 (Gm), addition in Z_q (Za), multiplication in Z_q (Zm), inversion in Z_q (Zi), hash function (H) and pairing operation (e). We only present the computation in the blind signature issuing protocol. In Sherman's scheme (the first scheme in [6]), the computation of the signer is $Ga + 4Gm + H$, the computation of the user is $3Ga + 6Gm + Zm + Zi + H$, and the computation of verification is $Ga + Gm + H + 3e$. Correspondently the computation of our scheme is $Ga + Gm + Zi + Za$, $4Gm + 3Ga$ and $Ga + Gm + 3e$. So we draw a conclusion that our scheme is more efficient.

4 Conclusions

We proposed an efficiently blind signature without using hash function. The randomness of signature is brought by commitment b. Based on the complexity of strong q-SDH assumption, we present the strict proof of security against one more forgery in standard model. Different from formal one more forgery, we require that the challenger must generate all commitments before seeing the public key and sent them to the adversary. The conclusion shows that our blind signature scheme prevents against one-more forgery under adaptive chosen message attack. A full blind testimony demonstrates that our scheme meet with blind property. Compared with other blind signature schemes, we think proposed scheme is more efficient. To the best of our knowledge, our scheme is the first blind signature scheme from pairings proved secure in the standard model.

References

1. Chaum, D.: Blind Signatures for Untraceable Payments. In: Crypto '82, Plenum (1983) 199-203
2. Juels, A., Luby, M., Ostrobsky, R.: Security of Blind Digital Signatures. In: B.S. Kaliski Jr. (eds.): Cachin, C., Camenisch, J. (eds.): Advanced in Cryptology-CRYPTO'97. Lecture Notes in Computer Science, Vol. 1294. Springer-Verlag, Berlin Heidelberg New York (1997) 150–164
3. Boldyreva, A.: Efficient threshold signature, multisignature and blind signature schemes based on the Gap-Diffie-Hellman group signature scheme. In: Y.G. Desmedt (eds.): Public Key Cryptography-PKC 2003. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York (2003) 31–46
4. Fangguo, Zhang., Kwangjo, K.: ID-Based Blind Signature and Ring Signature from Pairings. In: Y. Zheng (eds.): Advance in Cryptology - ASIACRYPT 2002. Lecture Notes in Computer Science, Vol. 2501. Springer-Verlag, Berlin Heidelberg New York (2002) 533–547
5. Fangguo, Zhang., Kwangjo, K.: Effcient ID-Based Blind Signature and Proxy Signature from Bilinear Pairings. In: Proceedings ACISP 2003. Lecture Notes in Computer Science, Vol. 2727. Springer-Verlag, Berlin Heidelberg New York (2003) 12–323
6. Sherman S.M., Lucas, C.K., Yiu, S.M., Chow, K.P.: Two Improved Partially Blind Signature Schemes from Bilinear Pairings. Available at: http://eprint.iacr.org/2004/108.pdf
7. Fangguo, Zhang., Reihaneh, S.N., Willy, S.: Efficient Verifiably Encrypted Signature and Partially Blind Signature from Bilinear Pairings. In: Thomas Johansson, T., Maitra, S. (eds.): Progress in Cryptology-INDOCRYPT 2003. Lecture Notes in Computer Science, Vol. 2904. Springer-Verlag, Berlin Heidelberg New York (2003) 191–204
8. Camenisch, J., Koprowski, M., Warinschi, B.: Efficient Blind Signatures without Random Oracles. In: Blundo, C., Cimato, S. (eds.): Security in Communication Networks-SCN 2004. Lecture Notes in Computer Science, Vol. 3352. Springer-Verlag, Berlin Heidelberg New York (2005) 134–148
9. Boneh, D., Boyen, X.: Short Signature without Random Oracles. In: Cachin, C., Camenisch, J. (eds.): Advances in Cryptology: EUROCRYTP'04. Lecture Notes in Computer Science, Vol. 3027. Springer-Verlag, Berlin Heidelberg New York (2004) 56–73

Efficient ID-Based Proxy Signature and Proxy Signcryption Form Bilinear Pairings

Qin Wang and Zhenfu Cao

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200030, China
chhwangqin@sjtu.edu.cn, cao-zf@cs.sjtu.edu.cn

Abstract. In this paper, based on bilinear pairings, we would like to construct an identity based proxy signature scheme and an identity based proxy signcryption scheme without secure channel. We also analyze the two proposed schemes from efficiency point of view and show that they are more efficient than the existed ones. What's more, our proposed schemes satisfy all of the security requirements to proxy signature and proxy signcryption schemes assuming the CDH problem and BDH problem are hard to solve.

1 Introduction

An *identity based cryptosystem* is a novel type of public cryptographic scheme in which the public keys of the users are their identities or strings derived from their identities. For this to work there is a Key Generation Center (KGC) that generates private keys using some master key related to the global parameters for the system.

The *proxy signature* primitive and the first efficient solution were introduced by Mambo, Usuda and Okamoto (MUO) [10]. Proxy signature has found numerous practical applications, particularly in distributed computing where delegation of rights is quite common, such as e-cash systems, global distribution networks, grid computing, mobile agent applications, and mobile communications. Many proxy signature schemes have been proposed [1, 2, 10, 11, 12].

In 1997, Zheng proposed a primitive that he called *signcryption* [8]. The idea of a signcryption scheme is to combine the functionality of an encryption scheme with that of a signature scheme. It must provide privacy; must be unforgeable; and there must be a method to settle repudiation disputes. This must be done in a more efficient manner than a composition of an encryption scheme with a signature scheme. After that, some research works on signcryption have been done [5, 6, 7, 8].

The *proxy signcryption* primitive and the first scheme were proposed by Gamage, Leiwo, and Zheng (GLZ) in 1999 [9]. The scheme combines the functionality of a proxy signature scheme and a signcryption scheme. It allows an entity to delegate its authority of signcryption to a trusted agent. The proxy signcryption scheme is useful for applications that are based on unreliable datagram style network communication model where messages are individually signed and not serially linked via a session key to provide authenticity and integrity.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 167–172, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, from bilinear pairings, we will give an ID-based proxy signature scheme and an ID-based proxy signcryption schemes. The two schemes are more efficient than existed ones. They need no secure channel and satisfy all of the security requirements to proxy signature and proxy signcryption schemes assuming the CDH problem and BDH problem are hard to solve.

The rest of this paper is organized as follows: Section 2 gives a detailed description of our ID-based proxy signature scheme and its analysis. Section 3 gives our new ID-based proxy signcryption scheme and its analysis. Section 4 concludes this paper.

2 ID-Based Proxy Signature Scheme from Pairings

Recently, many identity based signature schemes have been proposed using the bilinear parings [3, 4]. In these schemes, Cha-Cheon's scheme [4] is not only efficient but exhibits the provable security relative to CDH problem. In this section, we propose a new ID-based proxy signature scheme based on Cha-Cheon's scheme.

Zhang and Kim proposed an ID-based proxy signature scheme [12] based on Hess's signature [3]. We will compare our scheme with the scheme in [12] from efficiency point of view.

2.1 New ID-Based Proxy Signature Scheme

[**Setup:**]
Let $(\mathbb{G}_1, +)$ denote a cyclic additive group generated by P, whose order is a large prime q, and $(\mathbb{G}_2, \cdot)$ denote a cyclic multiplicative group of the same order q. The bilinear pairing is given by $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$. Define two cryptographic hash functions $H_1 : \{0,1\}^* \to \mathbb{Z}_q$ and $H_2 : \{0,1\}^* \to \mathbb{G}_1$.

The KGC chooses a random number $s \in \mathbb{Z}_q^*$ and sets $P_{pub} = sP$. Then publishes system parameters $\{\mathbb{G}_1, \mathbb{G}_2, e, q, P, P_{pub}, H_1, H_2\}$, and keeps s as the *master-key*, which is known only by itself.

[**Extract:**]
A user submits his/her identity information ID to KGC. KGC computes the user's public key as $Q_{ID} = H_2(ID)$, and returns $S_{ID} = sQ_{ID}$ to the user as his/her private key.

Let Alice be the original signer with identity public key Q_A and private key S_A, and Bob be the proxy signer with identity public key Q_B and private key S_B.

[**Generation of the Proxy Key:**]
To delegate the signing capacity to a proxy signer, the original signer Alice uses Cha-Cheon's ID-based signature scheme [4] to make the signed warrant m_ω. There is an explicit description of the delegation relation in the warrant m_ω. If the following process is finished successfully, Bob gets the proxy key S_P.

1. For delegating his signing capability to Bob, Alice first makes a warrant m_ω which includes the restrictions on the class of messages delegated, the

original signer and proxy signer's identities and public keys, the period of validity, etc.

2. After computing $U_A = r_A Q_A$, where $r_A \in_R \mathbb{Z}_q^*$, $h_A = H_1(m_\omega, U_A)$ and $V_A = (r_A + h_A)S_A$, Alice sends (m_ω, U_A, V_A) to a proxy signer Bob.
3. Bob verifies the validity of the signature on m_ω: check whether $e(P, V_A) = e(P_{pub}, U_A + h_A Q_A)$, where $h_A = H_1(m_\omega, U_A)$. If the signature is valid, Bob computes $S_P = V_A + h_A S_B$, and keeps S_P as the proxy private key. Compute $Q_P = U_A + h_A(Q_A + Q_B)$, and publish Q_P as the proxy public key.

[**Proxy Signature Generation:**]
To sign a message $m \in \{0,1\}^n$, Bob computes $U_P = r_P Q_P$, where $r_P \in \mathbb{Z}_q^*$, $h_P = H_1(m_\omega, m, U_P)$ and $V_P = (r_P + h_P)S_P$. The valid proxy signature will be the tuple

$$< m, U_P, V_P, m_\omega, U_A > .$$

[**Proxy Signature Verification:**]
A verifier can accept the proxy signature if and only $e(P, V_P) = e(P_{pub}, U_P + h_P Q_P)$, where $h_P = H_1(m_\omega, m, U_P)$.

2.2 Efficiency and Security

We compare our ID-based proxy signature scheme with the scheme in [12] from computation overhead. We denote by A a point addition in $\mathbb{G}_1$, by M a scalar multiplication in $\mathbb{G}_1$, by E a modular exponentiation in $\mathbb{G}_2$, and by P a computation of the pairing. We do not take the computation time for hash function H_1.

Table 1. Comparison of our scheme and the scheme in [12]

Schemes	Proposed scheme	The scheme in [12]
Generation of the proxy key	$4A + 5M + 2P$	$2A + 3M + 2E + 3P$
Proxy signature generation	$2M$	$1A + 2M + 1E + 1P$
Proxy signature verification	$1A + 1M + 2P$	$1A + 2E + 2P$

From the above table, it is easy to see that our scheme is more efficient than the scheme in [12]. We note that the computation of the pairing is most time-consuming. Modular exponentiation is also a time-consuming computation.

The proposed ID-based proxy signature scheme satisfies all of the following requirements: verifiability, strong unforgeability, strong identifiability, strong undeniability, prevention of misuse, assuming the hardness of CDH problem in the random oracle model. The details of the discuss is omitted.

3 ID-Based Proxy Signcryption Scheme from Pairings

In this section, we propose a new ID-based proxy signcryption scheme from pairings. Our scheme is based on the proposed proxy signature scheme and the signcryption scheme in [5].

In [13], Li and Chen also proposed an ID-based proxy signcryption scheme from pairings. We will compare our scheme with their scheme from efficiency point of view.

3.1 New ID-Based Proxy Signcryption Scheme

The **Setup**, **Extract**, and **Generation of the proxy key** algorithms are the same as the proposed proxy signature scheme. We detail **Proxy signcryption generation**, **Proxy unsigncryption and verification**, and **Public proxy verification** in the following.

Let Alice be the original signer with identity public key Q_A and private key S_A, Bob be the proxy signer with identity public key Q_B and private key S_B, and Carol be the receiver with identity public key Q_C and private key S_C.

[**Proxy Signcryption Generation:**]
To signcrypt a message $m \in \{0,1\}^n$ to a receiver Carol on behalf of Alice, the proxy signer Bob does the following:

Sign: Compute $U_P = r_P Q_P$, where $r_P \in \mathbb{Z}_q^*$, $h_P = H_1(m_\omega, m, U_P)$ and $V_P = (r_P + h_P)S_P$.

Encrypt: Compute $\omega = e(r_P S_P, Q_C)$ and $C = H_1(\omega) \oplus (V_P||ID_A|ID_B||m)$

The valid proxy signcryption will be the tuple $< C, U_P, m_\omega, U_A >$.

[**Proxy Unsigncryption and Verification:**]
Upon receiving $< C, U_P, m_\omega, U_A >$ from the proxy signer, the receiver Carol does the following:

Decrypt: Compute $\omega = e(U_P, S_C)$, $V_P||ID_A|ID_B||m = C \oplus H_1(\omega)$, and get the proxy signature $< m, U_P, V_P, m_\omega, U_A >$.

Verify: Check whether $e(P, V_P) = e(P_{pub}, U_P + h_P Q_P)$, where $h_P = H_1(m_\omega, m, U_P)$. If so, accepts it; otherwise rejects it.

[**Public Proxy Verification:**]
When public verification, the receiver Carol passes the proxy signature $< m, U_P, V_P, m_\omega, U_A >$ to a third party who can be convinced of the message's origin as follows:

Check whether $e(P, V_P) = e(P_{pub}, U_P + h_P Q_P)$, where $h_P = H_1(m_\omega, m, U_P)$. If so, accepts it; otherwise rejects it.

3.2 Efficiency and Security

Now we compare the efficiency of our new proxy signcryption scheme with the scheme in [13].

Let us first consider the size of the ciphertext of the two schemes. Our ciphertext is $< C, U_P, m_\omega, U_A >$, and their ciphertext is something like $< C, U, m_\omega, S, r >$. The length of the first four elements in their ciphertext are the same as ours, but the last element r which is a random number in Z_q^* isn't needed by our scheme. So our scheme is a little bit more compact than the scheme in [13].

Let us now consider the computation required by the two scheme. The notations are the same as section 2.2.

Table 2. Comparison of our scheme and the scheme in [13]

Algorithm	Proposed scheme	The scheme in [13].
Generation of the proxy key	$4A + 5M + 2P$	$1A + 3M + 1E + 3P$
Proxy signcrypting generation	$3M + 1P$	$2A + 2M + 2E + 2P$
Proxy unsigncrypting and verification	$1A + 1M + 3P$	$4E + 8P$
Public proxy verification	$1A + 1M + 2P$	$2E + 4P$

From the above table, it is easy to see our proposed scheme is more efficient than the scheme in [13]. Our scheme only needs 8 paring computation, while their scheme needs 17 paring computation. Moreover, our scheme needs no modular exponentiation, while their scheme needs 9 modular exponentiation in G_2. In fact, our scheme needs less than half of the computation of the scheme in [13].

In addition, the scheme in [13] needs a secure channel to send the proxy certificate from the origianl signer to the proxy signer, while ours can be done through public channel.

The proposed ID-based proxy signcryption scheme satisfies all of the following requirements: verifiability, strong unforgeability, strong identifiability, prevention of misuse, confidentiality, non-repudiation, assuming the hardness of BDH problem in the random oracle model. The details of the discuss is omitted.

4 Conclusion

We proposed an ID-based proxy signature scheme and an ID-based proxy signcryption scheme from bilinear pairings. The proposed proxy signature scheme was more efficient than the scheme in [12] in computation overhead. The proposed proxy signcryption schemes was more efficient than the scheme proposed in [13] in terms of ciphertext length and computation overhead. Both of the two new schemes need no secure channel. If the CDH problem is hard to solve, the proposed proxy signature scheme would satisfy all of the security requirements to proxy signature schemes. And if the BDH problem is hard to solve, the proposed proxy signcryption scheme would satisfy all of the security requirements to proxy signcryption ones.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China for Distinguished Young Scholars under Grant No. 60225007, the National Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20020248024, and the Science and Technology Research Project of Shanghai under Grant Nos. 04JC14055 and 04DZ07067.

References

1. Kim, S., Park, S., Won, D.: Proxy Signatures, Revisited. In: Han, Y., Okamoto, T., Qing, S. (eds.): International Conference on Information and Communications Security. Lecture Notes in Computer Science, Vol. 1334. Springer-Verlag, Berlin Heidelberg New York (1997) 223-232
2. Lee, J., Cheon, J., Kim, S.: An Analysis of Proxy Signatures:Is a Secure Channel Necessary? In: Joye, M. (ed.): Topics in Cryptology-CT-RSA. Lecture Notes in Computer Science, Vol. 2612. Springer-Verlag, Berlin Heidelberg New York (2003) 68–79
3. Hess, F.: Efficient Identity Based Signature Schemes Based on Pairings. In: Nyberg, K., Heys, H. (eds.): Selected Areas in Cryptography: 9th Annual International Workshop. Lecture Notes in Computer Science, Vol. 2595. Springer-Verlag, Berlin Heidelberg New York (2003) 310–324
4. Cha, J.C., Cheon, J.H.: An Identity-Based Signature from Gap Diffie-Hellman Groups. In: Y.G. Desmedt (ed.): Public Key Cryptography. Lecture Notes in Computer Science, Vol. 2567. Springer-Verlag, Berlin Heidelberg New York (2003) 18–30
5. Chen,L., Lee, J.M.: Improved Identity-Based Signcryption. In: Vaudenay, S. (ed.): Public Key Cryptography. Lecture Notes in Computer Science, Vol. 3386. Springer-Verlag, Berlin Heidelberg New York (2005) 362–379
6. Lee, J.M., Mao, W.: Two Birds One Stone: Signcryption using RSA. In: M. Joye (Ed.): Topics in Cryptology-CT-RSA. Lecture Notes in Computer Science, Vol. 2612. Springer-Verlag, Berlin Heidelberg New York (2003) 211–225
7. Bao, F., Deng, R.H.: A Signcryption Scheme with Signature Directly Verifiable by Public Key. In: Imai, H., Zheng, Y. (eds.): Public Key Cryptography. Lecture Notes in Computer Science, Vol. 1498. Springer-Verlag, Berlin Heidelberg New York (1998) 55–59
8. Zheng,Y.: Digital Signcryption or How to Achieve Cost (Signature and Encryption) ¡¡ Cost(Signature)+Cost(Encryption). In: Kaliski, B.S., Jr. (ed.): Advances in Cryptology-Crypto. Lecture Notes in Computer Science, Vol. 1294. Springer-Verlag, Berlin Heidelberg New York (1997) 165–179
9. Gamage,C., Leiwo, J., Zheng, Y.: An Efficient Scheme for Secure Message Transmission Using Proxy-Signcryption. 22nd Australasian Computer Science Conference. Springer-Verlag, Berlin Heidelberg New York (1999) 420–431
10. Mambo, M., Usuda, K., Okamoto, E.: Proxy Signatures: Delegation of the Power to Sign Messages. IEICE Trans. on Fundamentals, E79-A(9) (1996) 1338-1354
11. Zhang, F.G., Safavi-Naini, R., Lin, C.Y., New Proxy Signature, Proxy Blind Signature and Proxy Ring Signature Schemes from Bilinear Pairing. 2003. http://eprint.iacr.org/2003/104
12. Zhang, F., and Kim, K.: Efficient ID-Based Blind Signature and Proxy Signature from Bilinear Pairings. In: R. Safavi-Naini and J. Seberry (eds.): Australasian Conference on Information Security and Privacy. Lecture Notes in Computer Science, Vol. 2727. Springer-Verlag, Berlin Heidelberg New York (2003) 312–323
13. Li, X., Chen, K.: Identity Based Proxy-Signcryption Sheme from Pairings. Proceedings of the 2004 IEEE International Conference on Services Computing. IEEE Comuter Society (2004) 494-497

An Identity-Based Threshold Signcryption Scheme with Semantic Security

Changgen Peng and Xiang Li

Institute of Computer Science, Guizhou University,
Guiyang 550025, China
{sci.cgpeng, lixiang}@gzu.edu.cn

Abstract. This paper designs a secure identity-based threshold signcryption scheme from the bilinear pairings. The construction is based on the recently proposed signcryption scheme of Libert and Quisquater [6]. Our scheme not only has the properties of identity-based and threshold, but also can achieve semantic security under the Decisional Bilinear Diffie-Hellman assumption. It can be proved secure against forgery under chosen message attack in the random oracle model. In the private key distribution protocol, we adopt such method that the private key associated with an identity rather than the master key is shared. In the threshold signcryption phase, we provide a new method to check the malicious members. This is the first identity-based threshold signcryption scheme that can simultaneously achieve both semantic security and others security, such as unforgeability, robustness, and non-repudiation.

1 Introduction

In 1984, Shamir introduced identity-based (ID) cryptosystems [15]. The idea was to allow using users' identity information serve as public key. These systems have a trusted private key generator (PKG) whose task is to initialize the system and generate master/public key pair. Given the master secret key, the PKG can compute users' private key from their identity information. In such schemes, the key management procedures of certificate-based public key infrastructure (PKI) are simplified. Since then, several practical identity-based signature schemes have been proposed, but a satisfying identity-based encryption scheme from bilinear maps was first introduced in 2001 [1]. Recently, several identity-based signature schemes using pairings were proposed [3, 7, 8, 11].

Since the threshold cryptography approach is presented, there are a large number of works related to this topic. The early threshold signature schemes were proposed based on RSA or discrete-log difficulty. In 2003, Boldyreva [2] proposed a threshold signature scheme based on the Gap Diffie-Hellman (GDH) group. Until recently, the signature schemes that combining threshold technique with identity-based cryptography were proposed. Back and Zheng [3] proposed an identity-based threshold signature (IDTHS) scheme, which important feature is that the private key associated with an identity rather than a master key is shared. They also presented a new computational problem, called the modified

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 173–179, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Generalized Bilinear Inversion (mGBI) problem and formalized the security notion of IDTHS and gave a concrete scheme based on the bilinear parings. After that, another IDTHS scheme was proposed by Chen et al. [8].

Signcryption is a new paradigm in public key cryptography, which was first proposed by Zheng in 1997 [14]. It can simultaneously fulfil both the functions of digital signature and public key encryption in a logically single step, and with a cost significantly lower than that required by the traditional *signature-then-encryption* approach. Since then, several efficient signcryption schemes were proposed. The first identity-based signcryption scheme was proposed in 2002 [12], but this scheme does not have the non-repudiation property. After that, Malone-Lee [5] proposed an identity-based signcryption scheme that can offer the non-repudiation. Recently, Libert and Quisquater [6] proposed an identity-based signcryption scheme that satisfies semantic security. Combining the threshold technique with identity-based signcryption approach, Duan et al. [13] devised an identity-based threshold signcryption scheme. However, the Duan et al.'s scheme still adopts the method that the master key is shared among n parties, and non-repudiation and semantic security cannot be provided.

In this paper, we use the method that the master key of PKG is shared among n parties to design a new identity-based threshold signcryption scheme based on the bilinear maps. This construction is based on the signcryption scheme of Libert and Quisquater [6]. Our scheme can achieve the semantic security under the Decisional Bilinear Diffie-Hellman (DBDH) assumption. The security of private key distribution protocol is based on the mGBI assumption. Running the public verifiability protocol in our scheme can provide the non-repudiation of sender. In addition, our scheme can also offer robustness.

2 Preliminaries

Let us consider two cyclic groups $(G_1, +)$ and $(G_2, \cdot)$, whose have the same prime order $q \geq 2^k$, where k is a secure parameter. Let P be a generator of G_1.

A bilinear pairing be a map $e : G_1 \times G_1 \rightarrow G_2$ satisfying the following properties:

Bilinear: $\forall P, Q \in G_1$ and $\forall a, b \in F_q^*$, we have $e(aP, bQ) = e(P, Q)^{ab}$.
Non-degenerate: for any point $P \in G_1, e(P, Q) = 1$ for all $Q \in G_1$ iff $P = \mathcal{O}$.
Computable: $\forall P, Q \in G_1$, there is an efficient algorithm to compute $e(P, Q)$.

Such bilinear pairing may be realized using the modified Weil pairing or the Tate pairing over supersingular elliptic curves. In the following, we recall several intractable problems about bilinear pairings.

- The ***Computational Diffie-Hellman (CDH) problem.*** Given (P, aP, bP) for unknown $a, b \in F_q$, to compute $abP \in G_1$.
- The ***Decisional Bilinear Diffie-Hellman (DBDH) problem.*** Given (P, aP, bP, cP) for unknown $a, b, c \in F_q$ and $h \in G_2$, to decide whether $h = e(P, P)^{abc}$ or not.
- The ***modified Generalized Bilinear Inversion (mGBI) problem.*** Given $h \in G_2$ and $P \in G_1$, computes $S \in G_1$ such that $e(S, P) = h$.

3 Our Identity-Based Threshold Signcryption Scheme

In this section, we shall show our (t, n) identity-based threshold signcryption scheme based on the Libert and Quisquater's scheme [6]. Let $GA = \{\Gamma_i\}_{i=1,\ldots,n}$ is the signers' group of n parties and the receiver is Bob. Our scheme consists of the following operation phases.

- **Setup:** Given secure parameters k and l, the PKG chooses groups $(G_1, +)$ and $(G_2, \cdot)$ of prime order q, a generator P of G_1, a bilinear map $e : G_1 \times G_1 \rightarrow G_2$, and hash functions $H_1 : \{0,1\}^* \rightarrow G_1, H_2 : G_2 \rightarrow \{0,1\}^l$, and $H_3 : \{0,1\}^* \times G_2 \rightarrow F_q$. (E, D) is a pair secure symmetric encryption/decryption algorithm, where l denotes the sizes of symmetric key. Finally, PKG chooses $s \in F_q^*$ as a master secret key and computes $P_{pub} = sP$. So the system's public parameters are $Params = \{G_1, G_2, l, e, P, P_{pub}, H_1, H_2, H_3\}$.
- **Keygen:** Given the identity ID of group GA, the PKG computes $Q_{ID} = H_1(ID)$ and the private key $d_{ID} = sQ_{ID}$. Similarly, given the identity ID_B of receiver Bob, the PKG computes $Q_{ID_B} = H_1(ID_B)$ and the private key $d_{ID_B} = sQ_{ID_B}$. Finally, the PKG sends private key d_{ID} and d_{ID_B} to GA and Bob via secure channel, respectively.
- **Keydistrib:** In this phase, the PKG plays the role of the trusted dealer. Given the identity ID of group GA, the dealer PKG generates n shares $\{d_{ID}^i\}_{i=1,\ldots,n}$ of d_{ID} and sends the share d_{ID}^i to party Γ_i via secure channel for $i = 1, \ldots, n$. The dealer performs the following distribution protocol.
  1. Generates a random polynomial $f(x) = \sum_{i=0}^{t-1} F_i x^i \in G_1$ which satisfying $F_0 = f(0) = d_{ID}$, where $F_1, F_2, \ldots, F_{t-1} \in G_1^*$. Let $d_{ID}^i = f(i)$ be the share of party Γ_i, $i = 1, \ldots, n$.
  2. Sends d_{ID}^i to party Γ_i for $i = 1, \ldots, n$ secretly and broadcasts $e(F_j, P)$ for $j = 0, 1, \ldots, t-1$.

  The values $e(F_j, P)_{j=0,1,\ldots,t-1}$ as verification keys can be used to check the validity of each share $d_{ID}^i, i = 1, \ldots, n$. Each party Γ_i can verify whether its share d_{ID}^i is valid or not by computing

$$e(d_{ID}^i, P) = \prod_{j=0}^{t-1} e(F_j, P)^{i^j}. \tag{1}$$

  If equation (1) holds, the share d_{ID}^i is valid.
- **Signcrypt:** From the definition of threshold scheme, any t or more out of n parties $(1 \leq t \leq n)$ can represent the group GA to perform the signcryption. Let $\{\Gamma_i\}_{i \in \Phi}$ are such the parties, where $\Phi \subset \{1, 2, \ldots, n\}$ and $|\Phi| = t$. In addition, a party will be randomly selected from $\{\Gamma_i\}_{i \in \Phi}$ as a designated clerk, who is responsible for collecting and verifying the partial signcryption, and then computing the group signcryption. The steps are as follow.
  1. Each of $\{\Gamma_i\}_{i \in \Phi}$ computes $Q_{ID_B} = H_1(ID_B) \in G_1$. Then chooses $x_i \in F_q^*$ at random, computes $K_{i1} = e(P, P_{pub})^{x_i}$ and $K_{i2} = e(P_{pub}, Q_{ID_B})^{x_i}$. Finally, sends K_{i1} and K_{i2} to clerk.
  2. Clerk computes $K_1 = \prod_{i \in \Phi} K_{i1}, K_2 = H_2(\prod_{i \in \Phi} K_{i2}), c = E_{K_2}(m)$, and $R = H_3(c, K_1)$, broadcasts c, R to parties $\{\Gamma_i\}_{i \in \Phi}$.

3. Each of $\{\Gamma_i\}_{i\in\Phi}$ computes partial signcryption $S_i = x_i P_{pub} + R\pi_i d_{ID}^i$ and sends it to clerk, where $\pi_i = \prod_{j\in\Phi, j\neq i} \frac{j}{j-i}$ is a Lagrange coefficient.
4. After clerk receives S_i, he can verifies its validity by

$$e(S_i, P) = K_{i1}\left(\prod_{j=0}^{t-1} e(F_j, P)^{i^j}\right)^{R\pi_i}. \quad (2)$$

If equation (2) holds, S_i is valid. If all the partial signcryptions are valid, clerk computes $S = \sum_{i\in\Phi} S_i$ and sends the signcryption $\sigma = (c, R, S)$ of group GA to receiver Bob.

– **Unsigncrypt:** After receiver Bob receives the signcryption σ, he can verify its validity using public key of group GA and recover the message m using his private key d_{ID_B} by following steps.
1. Computes $Q_{ID} = H_1(ID) \in G_1$.
2. Computes $K_1 = e(P, S)e(P_{pub}, -Q_{ID})^R$.
3. Computes $T = e(S, Q_{ID_B})e(-Q_{ID}, d_{ID_B})^R$ and $K_2 = H_2(T)$.
4. Recovers message $m = D_{K_2}(c)$ and accepts σ if and only if $R = H_3(c, K_1)$.

4 Analysis of Our Scheme

4.1 Correctness

Theorem 1. Equation (1) can check the validity of each share d_{ID}^i and equation (2) can check the validity of partial signcryption S_i.

Proof. If share d_{ID}^i is correct, then $e(d_{ID}^i, P) = e(f(i), P) = e(\sum_{j=0}^{t-1} F_j i^j, P) = \prod_{j=0}^{t-1} e(F_j, P)^{i^j}$. Hence, equation (1) holds.

If partial signcryption S_i is valid, we have $S_i = x_i P_{pub} + R\pi_i d_{ID}^i$. So $e(S_i, P) = e(x_i P_{pub} + R\pi_i d_{ID}^i, P) = e(P, P_{pub})^{x_i} e(d_{ID}^i, P)^{R\pi_i} = K_{i1}\left(\prod_{j=0}^{t-1} e(F_j, P)^{i^j}\right)^{R\pi_i}$, equation (2) holds.

Theorem 2. If all parties of $\{\Gamma_i\}_{i\in\Phi}$ can strictly perform the threshold signcryption steps, the signcryption σ can pass the checking of validity and the designated receiver Bob can also recover the message m.

Proof. Set $x = \sum_{i\in\Phi} x_i$, from the Lagrange formula, we have $\sum_{i\in\Phi} \pi_i d_{ID}^i = \sum_{i\in\Phi} \pi_i f(i) = d_{ID}$. If all parties of $\{\Gamma_i\}_{i\in\Phi}$ can strictly perform the threshold signcryption steps, then $S = \sum_{i\in\Phi} S_i = \sum_{i\in\Phi} x_i P_{pub} + \sum_{i\in\Phi} R\pi_i d_{ID}^i = xP_{pub} + Rd_{ID}$. Thus, the following equations will be held.

$$\begin{aligned} e(P, S)e(P_{pub}, -Q_{ID})^R &= e(P, xP_{pub} + Rd_{ID})e(P_{pub}, -Q_{ID})^R \\ &= e(P, P_{pub})^x e(P, d_{ID})^R e(P, -d_{ID})^R \\ &= e(P, P_{pub})^x = \prod_{i\in\Phi} e(P, P_{pub})^{x_i} = \prod_{i\in\Phi} K_{i1}. \end{aligned}$$

$$\begin{aligned} T &= e(S, Q_{ID_B})e(-Q_{ID}, d_{ID_B})^R = e(xP_{pub} + Rd_{ID}, Q_{ID_B})e(-Q_{ID}, d_{ID_B})^R \\ &= e(P_{pub}, Q_{ID_B})^x e(Rd_{ID}, Q_{ID_B})e(-Rd_{ID}, Q_{ID_B}) \\ &= e(P_{pub}, Q_{ID_B})^x = \prod_{i\in\Phi} e(P_{pub}, Q_{ID_B})^{x_i} = \prod_{i\in\Phi} K_{i2}. \end{aligned}$$

4.2 Security

– **Semantic security**

In the literature [5], Malone-Lee defined a new security notion, which was called indistinguishability of identity-based signcryption under adaptive chosen ciphertext attack (IND-ISC-CCA). This security notion is just semantic security. Unfortunately, the Malone-Lee scheme cannot achieve the semantic security. However, the Libert and Quisquater's scheme [6] had improved this weakness by replacing $R = H_2(U, m)$ by $R = H_3(c, K_1)$. Furthermore, Libert and Quisquater gave a formal proof of semantic security on their signcryption scheme under DBDH assumption.

If a verification equation includes the plaintext of message m, any adversary can simply verify message m^* whether it is the actual message m from such verification equation, such as $R = H_2(U, m)$ in the Malone-Lee scheme. That is to say, such schemes cannot achieve semantic security of message. In fact, from IND-ISC-CCA security notion, adversary can guess the signature on plaintexts m_0 and m_1 produced during the game and determine which one matches the challenge ciphertext. Many current versions have this weakness. Based on the Libert and Quisquater's scheme, our threshold signcryption scheme has overcome this weakness by replacing plaintext m by ciphertext c in verification equation. Therefore, we can obtain the following result (Theorem 3).

Theorem 3. The proposed identity-based threshold signcryption scheme satisfies semantic security notion under DBDH assumption.

– **Robustness**

It is obvious that the security of private key distribution protocol in **Keydistrib** phase depends on mGBI assumption, where the mGBI problem is computationally intractable. The following Lemma 1 has been formally proved by Baek and Zheng [3] and Theorem 4 shows the robustness of our scheme.

Lemma 1. In **Keydistrib**, the attacker that learns fewer than t shares of d_{ID} cannot obtain any information about d_{ID} under the mGBI assumption.

Theorem 4. The **Keydistrib** and **Signcrypt** protocol can be normally executed even in the presence of a malicious attacker that makes the corrupted parties, that is, the proposed identity-based (t, n) threshold signcryption scheme is robust.

Proof. In **Keydistrib**, each party can verify validity of his private key share using the published verification keys $e(F_j, P)_{j=0,1,\ldots,t-1}$ and equation (1). In **Signcrypt**, any $t-1$ or fewer honest parties cannot rebuild the private key and forge valid signcryption, only t or more parties can rebuild the private key and generate a valid signcryption. If there exist malicious adversary, the partial signcryption will be checked using equation (2) and other honest party can be selected to combine signcryption again.

– **Unforgeability**

The acceptable security notion for signature scheme is unforgeability under chosen message attack. Hess [7] presented an unforgeability notion for identity-based signature against chosen message attack (UF-IDS-CMA). After that, Back and Zheng [3] defined unforgeability notion for IDTHS against chosen message attack (UF-IDTHS-CMA) and gave the relationship between UF-IDS-CMA and UF-IDTHS-CMA through *Simulatability*, and they had proved the following Lemma 2. From these results, we can prove the following Theorem 5.

Lemma 2. If the identity-based threshold signature (IDTHS) scheme is simulatable and the corresponding identity-based signature scheme is UF-IDS-CMA secure, then the IDTHS is UF-IDTHS-CMA secure.

Theorem 5. Our identity-based threshold signcryption scheme satisfies unforgeability under CDH assumption in the random oracle model.

Proof. Our scheme can be regarded as a combination of an IDTHS and an identity-based threshold encryption scheme. In fact, the signature component of IDTHS in our scheme is a variant of Hess's signature scheme [7], in that replacing $R = H_3(c, K_1)$ by $R = H_3(m, K_1)$. An adversary who is able to forge a signcryption must be able to forge a signature for the signature component. The unforgeability of signature component can derive from Hess' signature scheme. Therefore, we only need to prove the IDTHS of our scheme is simulatable. In **Keydistrib**, Back and Zheng [3] had proved the private key distribution is simulatable. Now we prove the signing in **Signcrypt** is simulatable. Given the system common *Params*, the identity ID, a signature (R, S) on message m, $t-1$ shares $\{d_{ID}^i\}_{i=1,\ldots,t-1}$ of d_{ID} and the outputs of **Keydistrib**. The adversary first computes $\gamma^* = e(P, S)e(P_{pub}, -Q_{ID})^R$ and $R = H_3(m, \gamma^*)$. Then computes $S_i = x_i P_{pub} + R\pi_i d_{ID}^i$ for $i = 1, \ldots, t-1$, where π_i be a Lagrange coefficient. Let $F(x)$ be a polynomial of degree $t-1$ such that $F(0) = S$ and $F(i) = S_i, i = 1, \ldots, t-1$. Finally, computes and broadcasts $F(i) = S_i$ for $i = t, \ldots, n$.

– **Non-repudiation**

Once Bob receives the signcryption σ, he can prove to a third party its origin. By computing $K_1 = e(P, S)e(P_{pub}, -Q_{ID})^R$ and checking if the condition $R = H_3(c, K_1)$ holds, any third party can be convinced of the message's origin. However, in order to prove that the group *GA* is the sender of a particular plaintext m, Bob has to forward the ephemeral decryption key K_2 to the third party.

5 Conclusions

In this paper, we have designed an identity-based threshold signcryption scheme with semantic security based on the work of Libert and Quisquater. This semantic security notion is reasonable for cryptosystem due to the fact that the

DBDH problem is hard so far. Besides, our scheme also has the others security, such as unforgeability, robustness, and non-repudiation. Our scheme is the first identity-based threshold signcryption scheme with semantic security.

Acknowledgement

The research was supported by the Natural Science of Guizhou Province of China under Grant No.[2005]2107.

References

1. D. Boneh and M. Franklin. Identity based encryption from the Weil pairing. In *Advances in Cryptology - Crypto'01*, LNCS 2139, pages 213-229. Springer-Verlag, 2001.
2. A. Boldyreva. Threshold signatures, multisignatures and blind signatures based on the Gap-Diffie-Hellman-group signature scheme. In *PKC'2003*, LNCS 2567, pages 31-46. Springer-Verlag, 2003.
3. J. Baek and Y. Zheng. Identity-based threshold signature scheme from the bilinear pairings. In *IAS'04 track of ITCC'04*, pages 124-128. IEEE Computer Society, 2004.
4. J. Baek and Y. Zheng. Identity-based threshold decryption. In *PKC'2004*, LNCS 2947, pages 248-261. Springer-Verlag, 2004.
5. J. Malone-Lee. Identity-based signcryption. Cryptology ePrint Archive, 2002. http://eprint.iacr.org/2002/098/.
6. B. Libert and J.-J. Quisquater. New identity based signcryption schemes from pairings. Cryptology ePrint Archive, 2003. http://eprint.iacr.org/2003/023/.
7. F. Hess. Efficient identity based signature schemes based on pairings. In *Selected Areas in Cryptography - SAC'2002*, LNCS 2595, pages 310-324. Springer-Verlag, 2002.
8. X. Chen, F. Zhang, D. M. Konidala, and K.Kim. New ID-based threshold signature scheme from bilinear pairings. In *Progress in Cryptology - INDOCRYPT 2004*, LNCS 3348, pages 371-383. Springer-Verlag, 2004.
9. L. Chen and J. Malone-Lee. Improved identity-based signcryption. Cryptology ePrint Archive, 2004. http://eprint.iacr.org/2004/114/.
10. B. Libert and J.-J. Quisquater. Efficient signcryption with key privacy from gap Diffie-Hellman groups. In *PKC'2004*, LNCS 2947, pages 187-200. Springer-Verlag, 2004.
11. J. Cha and J. Cheon. An identity-based signature from gap Diffie-Hellman groups. In *PKC'2003*, LNCS 2567, pages 18-30. Springer-Verlag, 2003.
12. B. Lynn. Authenticated identity-based encryption. Cryptology ePrint Archive, 2002. http://eprint.iacr.org/2002/072/.
13. S. Duan, Z. Cao, and R. Lu. Robust ID-based threshold signcryption scheme form pairings. In *Proceedings of the 3rd international conference on Information security (Infosecu'04)*, pages 33-37. ACM Press, 2004.
14. Y. Zheng. Digital signcryption or how to achieve cost (signature & encryption)$\ll$ cost (signature) + cost (encryption). In *Advances in Cryptology - Crypto'97 Proceedings*, LNCS 1294, pages 165-179. Springer-Verlag, 1997.
15. A. Shamir. Identity based cryptosystems and signature schemes. In *Advances in Cryptology - Crypto'84*, LNCS 196, pages 47-53. Springer-Verlag, 1984.

A Token-Based Single Sign-On Protocol*

Li Hui and Shen Ting

Key Laboratory of Ministry of Education for Computer and Information Security,
Xidian University, Xi'an 710071, China
lihui@mail.xidian.edu.cn

Abstract. A token based single sign-on protocol for distribution systems is proposed in this paper. When a user C logs on a system, a centralized authentication server A will authenticate C and issue C a token which is signed by A and includes a session key generated by A as well as a time stamp. C can use the token to access any application server S.S will send the C's request to the A. Then A will verify the validity of the token. There are two advantages of this protocol: 1) Time synchronization between severs S and the user C is not necessary. 2) All authentication state information such as session key is stored in the token rather than in the memory of A, thus the performance of A can be promoted effectively.We have used SVO logic to do formal analysis of this protocol.

1 Introduction

Single Sign On(SSO)[1]is an effective authentication mechanism in distribution network system. User need only one authentication procedure to access resources of the network. Up to now, the most typical SSO solution is Kerberos[2]. Although a user need only one identity authentication in Kerberos, he still needs to request a new service ticket from ticket grant sever (TGS) when he wants to access an application server. Our authentication system consists of three principals: user C, authentication server A and application servers S. A is used to identify C. Authentication protocol is designed to base on digital certificates. When the authentication is passed, A issues a token rather than a ticket to C. The token is verified by A itself. With this method, C can access all servers in a secure domain with only one service token. A shared session key between C and A will be stored in the token. Thus, A will not maintain user's authentication state information and the performance can be promoted effectively.

Usually, there are many application servers in a secure domain. When a server S receives an access request of C including a token, it will send the token to A. At this time, A is the trusted third party between C and S. Because the token is generated by A, only A can verify its validity. After the verification, A will distribute a session key which is used to encrypt communications between the C and S.

* This paper is supported by NSFC grant 60173056.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 180–185, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The token based SSO protocol should satisfy the following secure requirement: (1) Mutual authentication between C and A within a secure domain. The shared session key between C and A should be securely stored in the token in order to reduce the A's cost of maintain the authentication state information of the user. (2) It can prevent token replaying attack. (3) Key distribution. As a trusted third party, A will generate the session key for S and C. The following content is arranged as follows: section 2 describes the protocol. Section 3 is formal analysis of the protocol using SVO logic. Section 4 is the conclusion.

2 Token Based Single Sign-On Protocol

2.1 Notations

$Cert_c, Cert_a, Cert_s$	Public key certificate of user, A and application server, which are issued by a certificate authority (CA).
K_c, K_c^{-1}	Public key and secret key of user.
K_s, K_s^{-1}	Public key and secret key of application server.
K_a, K_a^{-1}	Public key and secret key of A.
K	Symmetry key of A, which is used to encrypt the token.
$K_{a,c}$	Shared session key between A and user
$K_{c,s}$	Shared session key between user and application server
$Token_c$	Token of user issued by A
$N_c, N_c^{'}, N_c^{''}$	Nonce generated by user
$N_s, N_s^{'}$	Nonce generated by application server
N_a	Nonce generated by A
T_a	Time stamp generated by A
$lifetime$	lifetime of $Token_c$
$[M]_K$	Signature or HMAC of message M generated by key K. It is the concatenation of a message M and the signature or HMAC.
$\{M\}_K$	M is encrypted with key K

2.2 Single Sign-On Protocol

The SSO model is illustrated in Fig.1. The protocol can be divided into three sub-protocols.

User Login Protocol:

$$
\begin{array}{lll}
M1: & C \rightarrow A & Cert_c, N_c \\
M2: & A \rightarrow C & Cert_a, [N_c + 1, N_a]_{K_a^{-1}} \\
M3: & C \rightarrow A & [N_a + 1, N_c^{'}]_{K_c^{-1}} \\
M4: & A \rightarrow C & \{[Token_c, N_c^{'} + 1, K_{a,c}]_{K_a^{-1}}\}_{K_c} \\
 & & Token_c = [\{ID_C, K_{a,c}, T_a, lifetime\}_K]_{K_a^{-1}}
\end{array}
$$

Message $M1$ and $M2$ is a challenge-response authentication. N_c in $M1$ is the C's challenge to A. $M2$ is the response of A and N_a in M2 is the A's challenge

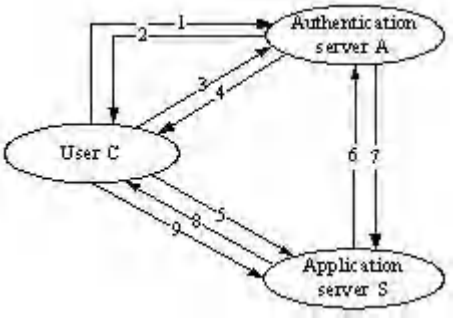


Fig. 1. Authentication model

to C. In $M3$, $N_c^{'}$ is a new nonce to protect the freshness of $M4$. $M3$ is signed by C's secret key. N_a guarantees the freshness of $M3$. $M4$ includes the token which is signed and encrypted by A. Because time stamp T_a is generated and verified by A itself, the time synchronous of C, A and S is not necessary. $K_{a,c}$ is the shared session key between A and C which has the same lifetime as the token.

At the end of this sub-protocol, A and user authenticate each other and share a session key.

Token Verification Protocol:

$$\begin{array}{lll} M5: & C \rightarrow S & Token_c, ID_C, N_c^{''} \\ M6: & S \rightarrow A & Cert_s, [Token_c, N_c^{''}, ID_C, N_s]_{K_s^{-1}} \\ M7: & A \rightarrow S & \{Cert_a, [\{N_c^{''}-1, ID_S, K_{c,s}\}_{K_{a,c}}, ID_C, N_s+1, K_{c,s}]_{K_a^{-1}}\}_{K_s} \end{array}$$

C sends the access request $M5$ to S. Now, A is the trust third party of C and S. A verifies the validity of signature of $Token_c$, then decrypts the token with K, checks T_a and lifetime. If all verifications are passed, A will generate a session key $K_{c,s}$ for C and S , then encrypt $K_{c,s}$ with $K_{a,c}$ and K_s respectively. $N_c^{''}$ is used to protect the freshness of $K_{c,s}$ to C. N_s is used by S to authenticate the A.

At the end of this protocol, C and S share a session key $K_{c,s}$ and S trusts C because A has verified the validity of token.

Application Authentication Protocol:

$$\begin{array}{lll} M8: & S \rightarrow C & \{N_c^{''}-1, ID_S, K_{c,s}\}_{K_{a,c}}, \{N_c^{''}+1, N_s^{'}\}_{K_{c,s}} \\ M9: & C \rightarrow S & \{N_s^{'}+1\}_{K_{c,s}} \end{array}$$

C gets the $K_{c,s}$ through the decryption of message $M8$ with $K_{a,c}$. By $N_c^{''}-1$, C believes that $K_{c,s}$ comes from A and is fresh. And then, use the $K_{c,s}$ to decrypt the $\{N_c^{''}+1, ID_S, N_s^{'}\}_{K_{c,s}}$. By $N_c^{''}+1$, C believes S share the $K_{c,s}$ with him. $M9$ make S believes C get the right $K_{c,s}$ by nonce $N_s^{'}$.

3 SVO Logic Analysis of the Protocol

SVO logic was devised by Syverson and van Oorschot [4, 5].It has two inference rules and twenty axioms.The SVO logic aims to unify several previous logics(BAN,GNY,AT and VO).We shall use it as the basis for our protocol analysis.Readers are referred to [4] for detail of SVO logic and we shall use it directly in this paper.

3.1 Goals of the Protocol

User login protocol is designed to achieve the following goals

G1. C *believes* A *said* $fresh(token_c)$

G2. C *believes* $C \overset{K_{a,c}}{\longleftrightarrow} A \wedge C$ *believes* $fresh(K_{a,c})$

G1 means C believes $token_c$ is fresh and is issued by A.**G2** means C believes $K_{a,c}$ is a fresh shared session key between C and A. At the end of the protocol,the following goals should be achieved

G3. S *believes* $S \overset{K_{c,s}}{\longleftrightarrow} C \wedge S$ *believes* $fresh(K_{c,s})$

G4. C *believes* $C \overset{K_{c,s}}{\longleftrightarrow} S \wedge C$ *believes* $fresh(K_{c,s})$

G5. C *believes* S *says* $C \overset{K_{c,s}}{\longleftrightarrow} S$

G6. S *believes* C *says* $C \overset{K_{c,s}}{\longleftrightarrow} S$

G3-G6 means both C and S believe $K_{c,s}$ is a fresh shared key between them.C or S also believes that the other one believes $K_{c,s}$ is a fresh shared key.

3.2 Verification of the Protocol

Initial States Assumptions

P1. C *believes* $PK_\sigma(A, K_a)$
P2. A *believes* $PK_\sigma(C, K_c)$
P3. C *believes* $PK_\psi(A, K_a)$
P4. A *believes* $PK_\sigma(S, K_s)$
P5. S *believes* $PK_\sigma(A, K_a)$
P6. A *believes* $PK_\psi(S, K_s)$
P7. C *believes* A *controls* $fresh(K_{a,c})$
P8. A *believes* $PK_\psi(C, K_c)$
P9. C *believes* $fresh(N_c)$
P10. C *believes* $fresh(N_c^{'})$
P11. C *believes* $fresh(N_c^{''})$
P12. A *believes* $fresh(N_a)$
P13. S *believes* $fresh(N_s)$
P14. S *believes* $fresh(N_s^{'})$
P15. A *believes* A *has* K
P16. A *believes* $fresh(T_a)$
P17. C *believes* A *controls* $fresh(K_{c,s})$
P18. S *believes* A *controls* $fresh(K_{c,s})$

Received Message Assumptions

P19. A *received* $(Cert_c, N_c)$
P20. C *received* $(Cert_a, [N_c + 1, N_a]_{K_a^{-1}})$

P21. A *received* $[N_a + 1, N_c^{'}]_{K_c^{-1}}$
P22. C *received* $\{[Token_c, N_c^{'} + 1, K_{a,c}]_{K_a^{-1}}\}_{K_c}$
P23. S *received* $(Token_c, ID_C, N_c^{''})$
P24. A *received* $(Cert_s, [Token_c, N_c^{''}, ID_C, N_s]_{K_s^{-1}})$
P25. S *received* $\{Cert_a, [\{N_c^{''} - 1, ID_S, K_{c,s}\}_{K_{a,c}}, ID_C, N_s{+}1, K_{c,s}]_{K_a^{-1}}\}_{K_s}$
P26. C *received* $\{N_c^{''} - 1, ID_S, K_{c,s}\}_{K_{a,c}}, \{N_c^{''} + 1, N_s^{'}\}_{K_{c,s}}$
P27. S *received* $\{N_s^{'} + 1\}_{K_{c,s}}$

Derivations. By P1,P9,P20,A6,A18,we have

$$C \text{ believes } A \text{ said} fresh(N_a) \tag{1}$$

where An are the n-th axiom of SVO logic.
By P2,P8,P21,A6,A10,A18

$$A \text{ believes } C \text{ said } fresh(N_c^{'}) \tag{2}$$

By P3,P8,P10,P22,A9,A10

$$C \text{ received } K_{a,c} \ \wedge C \text{ believes } A \text{ said } fresh(K_{a,c}) \tag{3}$$

$$C \text{ believes} A \text{ said } fresh(token_c) \ (\mathbf{G1}) \tag{4}$$

By (3),P7

$$C \text{ believes } C \overset{K_{a,c}}{\longleftrightarrow} A \wedge C \text{ believes } fresh(K_{a,c}) \ (\mathbf{G2}) \tag{5}$$

By P4, P24,A6

$$A \text{ believes } S \text{ said } (token_c, ID_C, N_s) \tag{6}$$

By P15,P16,P24,A6

$$A \text{ believes } A \text{ said } token_c \wedge A \text{ believes } A \overset{K_{a,c}}{\longleftrightarrow} C \tag{7}$$

$$A \text{ believes } fresh(K_{a,c}) \tag{8}$$

By P4,P5,P13,P18,P25,A6,A9,A10

$$S \text{ believes } A \text{ said } \{N_c^{''} - 1, ID_S, K_{c,s}\}_{K_{a,c}}, ID_C, N_s + 1, K_{c,s} \tag{9}$$

$$S \text{ believes } S \overset{K_{c,s}}{\longleftrightarrow} C \wedge S \text{ believes } fresh(K_{c,s}) \ (\mathbf{G3}) \tag{10}$$

By P11,(5),P17,P26,A5

$$C \text{ believes } C \overset{K_{c,s}}{\longleftrightarrow} S \wedge C \text{ believes } fresh(K_{c,s}) \ (\mathbf{G4}) \tag{11}$$

By P11,(11),P26,A6,A21

$$C \textit{ believes } S \textit{ says } C \overset{K_{c,s}}{\longleftrightarrow} S \ (\mathbf{G5}) \tag{12}$$

By P14,(10),P27,A6,A21

$$S \textit{ believes } C \textit{ says } C \overset{K_{c,s}}{\longleftrightarrow} S \ (\mathbf{G6}) \tag{13}$$

Now we have obtained all goals of our protocol.Both C and S believe $K_{c,s}$ is a good key for communications between them.C or S also believes that the other one believes $K_{c,s}$ is a good key.

4 Conclusion

In this SSO protocol, we assume all principals have a certificate issued by a trusted CA and they can verify the validity of the certificate. In fact, message $M5$ to $M9$ is a modification of Needham-Schroeder Shared-Key(NSSK) Protocol[3].User login protocol establishes a shared key $K_{a,c}$ between A and C, which is used in the following modified NSSK protocol to distribute the shared session key $K_{c,s}$ to C.The most significant feature of this protocol is that $K_{a,c}$ is stored in the token. This reduces the cost of A for maintaining a large user's authentication state database. This protocol can be used in secure web service and grid computing.

References

1. N.Chamberlin. *A Brief Overview of Single Single-on Technology* [EB/OL]. Http://www.gitec.org/ assets/ pdfs, 2000.
2. J.Kohl and C.Neuman. *The Kerberos Network Authentication Service (V5)* [S]. RFC 1510, September 1993.
3. R.M.Needham and M.D.Schroeder. *Using Encryption for Authentication in Large Networks of Computers.* Communications of ACM,21(12):993-999,1978
4. P.Syverson and P.C.van Oorschot.*On unifying some cryptographic protocol logics.*Proceeding of 1994 IEEE Symposium on Research in Security and Privacy, pp14-28,Oakland,California,May, 1994
5. P.Syverson. *Limitations on Design Principles for Public Key Protocols.*Proceedings of 1996 IEEE Symposium on Research in Security and Privacy, 62-72,Oakland,California,May, 1996

Simple Threshold RSA Signature Scheme Based on Simple Secret Sharing

Shaohua Tang

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China
csshtang@scut.edu.cn

Abstract. A new threshold RSA signature scheme is presented, which is based on a newly proposed simple secret sharing algorithm. The private key of RSA algorithm is divided into N pieces, and each piece is delivered to different participant. In order to digitally sign a message, each participant should calculate the partial signature for the message by using its own piece of shadow. Any K or greater than K participants out of N can combine the partial signatures to form a complete signature for the message. At the phase of signature combination, each participant's partial secret (shadow) is not necessary to expose to others and the RSA private key is not required to reconstruct, thus the secret of the private key will not be exposed. Besides, fast computation and simple operation are also the features of this scheme.

1 Introduction

The threshold cryptosystem was first introduced by Desmedt [1] in 1987. The threshold signature is very similar to the threshold cryptosystem. In a (t, n) threshold signature scheme, the signature can only be generated when the number of participating members is not less than the threshold value t. Anyone can use the public key to verify the signature. The most attractive feature of threshold signature is that the private key is never reconstructed but the signature can be calculated. There are some schemes can realize this feature. But almost all related works adopt complex algorithm or narrow the range that the parameters can choose. For example, Frankel's scheme [2] brings the complexity of algorithm design and security proof. Shoup's scheme [3] brings the hardness of computation to the combiner. People may think of designing a threshold RSA signature based upon classical Shamir's secret sharing scheme [6], however, as Desmedt and Frankel briefly addressed in [4], there are some technical obstructions to doing this.

In this paper, we propose a new threshold RSA signature scheme based on a newly proposed simple secret sharing algorithm [5]. The mathematical theory adopted by this simple secret sharing scheme is the union operation in set theory and the addition operation in arithmetic. Since the principle and the operations invoked by our scheme are extremely simple, thus we call it "simple" scheme, which possesses the following features: (1) It is easy to implement. (2) Fast computation of the threshold signature is ensured, because only simple addition and union operations are adopted by the underlying secret sharing algorithm. (3) It requires no strict preconditions and can apply to almost all circumstances requiring threshold signature. (4) It is secure. Though the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 186 – 191, 2005.
© Springer-Verlag Berlin Heidelberg 2005

scheme is simple, it still possesses the security of threshold cryptography. The rest of this paper is organized as follows: A brief review of Tang's simple secret sharing scheme [5] is presented in Section 2. Our proposed threshold RSA signature scheme is described in Section 3. The security, performance, and the comparisons with related works are analyzed in Section 4. Finally, we summarize the results of this paper in Section 5.

2 Review of Simple Secret Sharing Scheme

Recently, Tang proposed a simple secret sharing scheme [5], which invokes only simple addition and union operations. The secret is divided into N pieces, and each piece is delivered to different participants. Any K or greater than K participants out of N can reconstruct the secret, but any (M-1) or less than (M-1) participants would fail to recover the secret, where $M = \lceil N/(N-K+1) \rceil$, and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Tang's scheme is characterized by fast computation as well as easy implementation and secure operation.

Tang's scheme consists of three stages: the shadow generation stage, the shadow distribution stage, and the stage of reconstructing the secret. Suppose F is an arbitrary field, d is the secret data, and $d \in F$. The original N participants are $P_0, P_1, \ldots, P_{N-1}$.

Shadow Generation Stage: (N-2) random numbers $d_0, d_1, \ldots, d_{N-2}$ are selected from F, then d_{N-1} is computed by the equation $d_{N-1} = d - \sum_{i=0}^{N-2} d_i$.

Shadow Distribution Stage: For j=0, 1, …, N-1, the shadow A_j is defined as $A_j = \{(i \bmod N, d_{i \bmod N}) \mid j \le i \le N-k+j\}$. Then each A_j is delivered to the j-th participant P_j.

Stage of Secret Reconstruction: Randomly select K participants whose shadow is not damaged, and then each one should present their own A_j. Any one among the K selected participants can act as a combiner to find $(0, d_0), (1, d_1), \ldots, (n-1, d_{N-1})$ from the presented A_j. Let $d = \sum_{i=0}^{N-1} d_i$, which is the solution.

3 Threshold Signature Scheme

Our proposed threshold RSA signature scheme consists of five stages: the initial stage, the shadow generation stage, the shadow distribution stage, the stage to generate partial signatures, and the stage to combine the partial signatures.

3.1 Initial Stage

Parameters for RSA cryptosystem are generated at the initial stage. Randomly select two large primes p and q, and let $n=p \times q$, n is the public parameter. Compute

$\varphi(n)=(p-1)\times(q-1)$. Randomly choose an integer e with $0<e<\varphi(n)$ and $\gcd(e, \varphi(n))=1$, and compute $d\equiv e^{-1}(\bmod \varphi(n))$. d is the private key and is e the public key. Suppose m is the message to be signed, and $m<n$. $h(\cdot)$ is the one-way hash function.

3.2 Shadow Generation Stage

Suppose K and N are positive integers, and $K\le N$. There are N participants, which are denoted by $P_0, P_1, \ldots, P_{N-1}$. Randomly select $d_0, d_1, \ldots, d_{N-2}\in F$, then compute $d_{N-1}=d-\sum_{i=0}^{N-2} d_i$. Obviously, $d=d_{N-1}+\sum_{i=0}^{N-2} d_i=\sum_{i=0}^{N-1} d_i$. Any single d_i will not expose the secret of d.

We define $A=\{(0,d_0),(1,d_1),\ldots,(N-1,d_{N-1})\}$.

For $j=0, 1, \ldots, N-1$, we define

$$A_j=\{(i \bmod N, d_{i \bmod N}) \mid j\le i\le N-K+j\}$$
$$=\{(j,d_j),\ldots,((N-K+j) \bmod N, d_{(N-K+j) \bmod N})\}.$$

Obviously, $A\supseteq A_j$, and $A=\bigcup_{j=0}^{N-1} A_j$. It is also easy to deduce that $|A_j|=N-K+1$, where $|X|$ denotes the cardinality of the set X. Now, $A_0, A_1, \ldots, A_{N-1}$ are the shadows.

3.3 Shadow Distribution Stage

After the calculation of all A_j ($j=0, 1, \ldots, N-1$), each A_j is delivered to the j-th participant, i.e., A_0 is delivered to the participant P_0, A_1 to P_1, …, A_{N-1} to P_{N-1}.

3.4 Stage to Generate Partial Signatures

According to Tang's secret sharing scheme, the union of A_j of any K (or greater than K) participants out of N can cover A. That is, randomly choose $P_{i_1}, P_{i_2}, \ldots, P_{i_K}$ from the original N participants, and the union of their corresponding partial secret set $A_{i_1}\cup A_{i_2}\cup\cdots\cup A_{i_K}\supseteq A$. Based on the above observation, the following steps can be designed to generate the partial signatures:

Step 1: Randomly select K participants whose shadow is not damaged. Without losing generality, suppose $P_1, P_2, \ldots, P_K$ are the ones being chosen. Any one among these K selected participants can act as a combiner, e.g., P_1 can act as the combiner.

Step 2: The combiner needs to maintain a partial signature set (PSS) and each element of PSS is of the form (i, c_i), where $0\le i\le N-1$ and c_i is the partial signature to be calculated by the selected participants. The initial status of PSS is an empty set, i.e., initially, PSS=ϕ.

Step 3: For each selected P_j (j=1, 2, …, K), P_j sequentially fetch $(i, d_i) \in A_j$, then P_j consults if $(i, \cdot)$ belongs to PSS. If $(i, \cdot) \notin$ PSS, then P_j computes

$$c_i \equiv (h(m))^{d_i} \pmod{n} \tag{1}$$

and delivers (i, c_i) to the combiner. The combiner lets $PSS = PSS \bigcup \{(i, c_i)\}$.

Step 3 continues until the PSS reaches its final status, i.e., until $PSS = \{(i, c_i) | i = 0,1,..., N-1\}$.

3.5 Stage to Combine Partial Signatures

After the PSS reaches its final status $\{(i, c_i) | i = 0,1,..., N-1\}$, the combiner calculates the complete signature for the message m: $c = \prod_{i=0}^{N-1} c_i \pmod{n}$.

***Theorem* 1.** $c = \prod_{i=0}^{N-1} c_i \pmod{n}$ is the digital signature for message m.

Proof: The digital signature for message m is: $s \equiv (h(m))^d \pmod{n}$.

Since $d = \sum_{i=0}^{N-1} d_i$, thus

$$s \equiv (h(m))^d \pmod{n} \equiv (h(m))^{\sum_{i=0}^{N-1} di} \pmod{n} \equiv \prod_{i=0}^{N-1} (h(m))^{d_i} \pmod{n} \ . \tag{2}$$

According to Eqn(1), Eqn(2) can be re-written as $s = \prod_{i=1}^{N-1} c_i \pmod{n}$. Thus c=s.

4 Analysis and Comparison

4.1 Security Analysis

We can derive, according to Proposition 2 of Tang's scheme [5], that the union of A_j from any (M-1) participants cannot cover A, where $M = \lceil N/(N-K+1) \rceil$. Therefore, we can concluded that if the number of participants being attacked is less than M, then the secret d will not be exposed. Similarly, if the number of conspirators among the original participants is less than M, then they cannot derive the secret d.

We can further derive that any (M-1) or less than (M-1) participants would fail to calculate the complete signature. Thus, if the number of participants being attacked is less than M, or the number of conspirators among the original participants is less than

M, then the signature cannot be faked. Therefore, as soon as the number of participants not being damaged is greater than *N*-*M*, the whole system is secure.

At the phase of signature combination, each participant's partial secret (shadow) is not necessary to expose to others and the RSA private key is not required to reconstruct, thus the secret of the private key will not be exposed.

4.2 Performance Analysis

At the stage of shadow generation, (*N*-2) addition operations and one subtraction operation are needed. At the phase of partial signature generation, totally *N* modular exponents are required among *K* selected participants, and averagely *N*/*K* modular exponents are needed for each selected participant. At the stage of partial signature combination, *N* modular multiplication operations are required. Thus, the computation overhead is O(*N*) modular exponents plus O(*N*) modular multiplication operations.

4.3 Comparisons with Related Works

The most attractive feature of threshold signature is that the private key is never reconstructed but the signature can be calculated. There are some schemes that realize this feature. But almost all related works adopt complex algorithm or narrow the range that the parameters can choose. For example, Frankel's scheme [2] requires the polynomial coefficients a_i and x_i should be chosen from a dedicated set or range. Since the selection of parameters is limited dramatically, Frankel's scheme brings the complexity of algorithm design and security proof. As another typical threshold RSA signature algorithm, Shoup's scheme [3] requires strong primes, i.e., the secret primes $p = 2p' + 1, q = 2q' + 1,$ where p' and q' are large primes. All the interpolation equations are computed within the ring modular $m = p'q'$. Shoup's scheme brings the hardness of computation to the combiner. Compared with these related works, our algorithm is easy to implement, requires no strict preconditions, and is fast and secure.

5 Conclusions

A simple threshold RSA signature scheme is proposed in this paper. Compared with other similar schemes, our algorithm owns the following features: 1) Simple, which invokes no complex operations and is easy to implement; 2) Fast computation, it is much fast in computation compared with related works; 3) No strict preconditions, which can apply to almost all circumstances requiring threshold signature; 4) Secure. Though the scheme is simple, it still possesses the security of threshold cryptography.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under grant number 60273064), and by the Projects of Guangdong Government Science and Technology Program, and by the Projects of Guangzhou Government Science and Technology Program. The author would like to acknowledge the valuable comments by the anonymous reviews.

References

1. Y. Desmedt: Society and group oriented cryptography: a new concept, in “Advanced in Cryptology”, Proceeding of the Crypto’87, (1987) 120-127
2. Y. Frankel and P. Gemmel et al, Optimal-resilience proactive public-key cryptosystems, IEEE Symposium on Foundations of Computer Science (1997) 384-393
3. V. Shoup, Practical threshold signatures, Proceedings of the Eurocypt 2000, (2000) 207-220
4. Y. Desmedt and Y. Frankel, Threshold cryptosystems, Advances in Cryptology-Crypto'89, (1989) 307-315
5. Shaohua Tang, Simple Secret Sharing and Threshold RSA Signature Scheme, Journal of Information and Computational Science, vol. 1, no. 2, (2004) 259-262
6. A. Shamir, How to share a secret, Communication of ACM, vol. 22, no.11, (1979) 612-613

Efficient Compilers for Authenticated Group Key Exchange

Qiang Tang and Chris J. Mitchell

Information Security Group, Royal Holloway, University of London,
Egham, Surrey TW20 0EX, UK
{qiang.tang, c.mitchell}@rhul.ac.uk

Abstract. In this paper we propose two compilers which are designed to transform a group key exchange protocol secure against any passive adversary into an authenticated group key exchange protocol with key confirmation which is secure against any passive adversary, active adversary, or malicious insider. We show that the first proposed compiler gives protocols that are more efficient than those produced by the compiler of Katz and Yung.

1 Introduction

The case of 2-party authenticated key exchange has been well investigated within the cryptographic community; however, less attention has been given to the case of authenticated group key exchange protocols, which have more than two participants. A number of authors have considered extending the 2-party Diffie-Hellman protocol [1] to the group setting (e.g. [2, 3]). Unfortunately, most of these schemes only assume a passive adversary, so that they are vulnerable to active adversaries who control the communication network. Recently, several provably secure (against both passive and active adversaries) authenticated group key agreement protocols (e.g. [6, 7]) have been proposed.

In this paper we are particularly interested in protocol compilers which transform one kind of protocols into another. In [8], Mayer and Yung give a compiler which transforms any 2-party protocol into a centralised group protocol which, however, is not scalable. Recently, Katz and Yung [4] proposed a compiler (referred to as the Katz-Yung compiler) which transforms a group key exchange protocol secure against any passive adversary into an authenticated group key exchange protocol which is secure against both passive and active adversaries. Although the Katz-Yung compiler produces more efficient protocols than the Mayer-Yung compiler, it nevertheless still produces rather inefficient protocols. Each participant is required to perform numbers of signatures and verifications proportional to the number of rounds in the original protocol and the number of participants involved, respectively. Additionally, the Katz-Yung compiler also adds an additional round to the original protocol, but does not achieve key confirmation. We propose two new more efficient compilers and prove their security. Both our new compilers result in protocols achieving key confirmation with

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 192–197, 2005.
© Springer-Verlag Berlin Heidelberg 2005

lower computational complexity and round complexity than those produced by the compiler of Katz and Yung. Note that proofs of the two main theorems have been omitted for space reasons — these proofs will be given in the full version of the paper.

The rest of this paper is organised as follows. In section 2, we review the Katz-Yung compiler. In section 3, we propose a new compiler which transforms a group key exchange protocol secure against any passive adversary into an authenticated group key exchange protocol with key confirmation which is secure against any passive adversary, active adversary, or malicious insider. In section 4, we propose a second new compiler which outputs protocols that are both more efficient and provide the same functionality as the first compiler, at the cost of introducing a TTP. In section 5, we conclude the paper.

2 The Katz-Yung Compiler

In this section we review the Katz-Yung compiler. With respect to the protocol P input to the compiler, which must be a group key exchange protocol secure against any passive adversary, we make the following assumptions:

1. There is no key confirmation, and all participants compute the session key after the last round of the protocol.
2. Every protocol message is transported together with the identifier of the source and the round number.

2.1 Description of the Katz-Yung Compiler

Let $\Sigma = (Gen, Sign, Vrfy)$ be a signature scheme which is strongly unforgeable under an adaptive chosen message attack (see, for example, [9] for a definition). If k is a security parameter, $Gen(1^k)$ generates a pair of public/private keys for the signing and verification algorithms $(Sign, Vrfy)$.

Suppose a set $S = \{U_1, \cdots, U_n\}$ of users wish to establish a session key. Let ID_{U_i} represent U_i's identity for every i $(1 \leq i \leq n)$. Given P is any group key exchange protocol secure against any passive adversary, the compiler constructs a new protocol P', in which each party $U_i \in S$ performs as follows.

1. In the initialisation phase, and in addition to all the operations in protocol P, each party $U_i \in S$ generates a verification/signing key pair (PK_{U_i}, SK_{U_i}) by running $Gen(1^k)$, where k is a security parameter.
2. Each user U_i chooses a random $r_i \in \{0,1\}^k$ and broadcasts $ID_{U_i}||0||r_i$, where here, as throughout, $||$ represents concatenation. After receiving the initial broadcast message from all other parties, each U_i sets $nonce_i = ((ID_{U_1}, r_1), \cdots, (ID_{U_n}, r_n))$ and stores this as part of its state information. It is obvious that all the users will share the same nonce, i.e., $nonce_1 = nonce_2 = \cdots = nonce_n$, as long as no attacker changes the broadcast data (or an accidental error occurs).

3. Each user U_i in S executes P according to the following rules.
   - Whenever U_i is supposed to broadcast $ID_{U_i}||j||m$ as part of protocol P, it computes $\sigma_{ij} = Sign_{SK_{U_i}}(j||m||nonce_i)$ and then broadcasts $ID_{U_i}||j||m||\sigma_{ij}$.
   - Whenever U_i receives a message $ID_U||j||m||\sigma$, it checks that: (1) $U \in S$, (2), j is the next expected sequence number for a message from U, and (3) $Vrfy_{PK_U}(j||m||nonce_i, \sigma) = 1$ where 1 signifies acceptance. If any of these checks fail, U_i aborts the protocol. Otherwise, U_i continues as it would in P upon receiving message $ID_U||j||m$.
4. Each non-aborted protocol instance computes the session key as in P.

2.2 Security and Efficiency

Katz and Yung [4] claim that their proposed compiler provides a scalable way to transform a key exchange protocol secure against a passive adversary into an authenticated protocol which is secure against an active adversary. They also illustrate efficiency advantages over certain other provably-secure authenticated group key exchange protocols. With respect to efficiency, we make the following observations on the protocols produced by the Katz-Yung compiler.

1. Each user U_i must store the nonce $nonce_i = ((U_1, r_1), \cdots, (U_n, r_n))$ regardless of whether or not the protocol successfully ends. Since the length of this information is proportional to the group size, the storage of such state information will become a non-trivial overhead when the group size is large.
2. From the second round onwards, the compiler requires each user to sign all the messages it sends in the protocol run, and to verify all the messages it receives. Since the total number of signature verifications in one round is proportional to the group size, the signature verifications will potentially use a significant amount of computational resource when the group size is large.
3. The compiler adds an additional round to the original protocol P; however, it does not provide key confirmation. As Katz and Yung state [4], in order to achieve key confirmation a further additional round is required.

3 A New Compiler Without TTP

In this section we propose a new compiler that transforms a group key exchange protocol P secure against a passive adversary into an authenticated group key exchange protocol P' with key confirmation which is secure against passive and active adversaries, as well as malicious insiders.

We assume that $\Sigma = (Gen, Sign, Vrfy)$ is a signature scheme as specified in section 2.1. We also assume that a unique session identifier S_{ID} is securely distributed to the participants before every instance is initiated. In [5] Katz and Shin propose the use of a session identifier to defeat insider attacks.

Suppose a set $S = \{U_i, \cdots, U_n\}$ of users wish to establish a session key, and h is a one-way hash function. Let ID_{U_i} represent U_i's identity for every i $(1 \leq i \leq n)$.

Given a protocol P secure against any passive adversary, the new compiler constructs a new protocol P', in which each party $U_i \in S$ performs as follows.

1. In addition to all the operations in the initialisation phase of P, each party $U_i \in S$ also generates a verification/signing key pair (PK_{U_i}, SK_{U_i}) by running $Gen(1^k)$.
2. In each round of the protocol P, U_i performs as follows.
   - In the first round of P, each user U_i sets its key exchange history $H_{i,1}$ to be the session identifier S_{ID}, and sets the round number k to 1. During each round, U_i should synchronise the round number k.
   - When U_i is required to send a message m_i to other users, it broadcasts $M_i = ID_{U_i}||k||m_i$.
   - Once U_i has received all the messages M_j $(1 \leq j \leq n, j \neq i)$, it computes the new key exchange history as:

     $$H_{i,k} = h(H_{i,k-1}||S_{ID}||k||M_1||\cdots||M_n)$$

     Then U_i continues as it would in P upon receiving the messages m_j $(1 \leq j \leq n, j \neq i)$. Note that U_i does not need to retain copies of all received messages.
3. In an additional round, U_i computes and broadcasts the key confirmation message $ID_{U_i}||k||\sigma_i$, where $\sigma_i = Sign_{SK_{U_i}}(ID_{U_i}||H_{i,k}||S_{ID}||k)$.
4. U_i verifies the key confirmation messages from U_j $(1 \leq j \leq n, j \neq i)$. If all the verifications succeed, then U_i computes the session key K_i as specified in protocol P. Otherwise, if any verification fails, then U_i aborts the protocol execution.

In addition to the initialisation phase, the above protocol adds one round to the original protocol and achieves key confirmation. Each participant needs to sign one message and verify n signatures, in addition to the computations involved in performing P. In addition, each participant only needs to store the (hashed) key exchange history. Hence this compiler yields protocols that are more efficient than those produced by the Katz-Yung compiler.

Theorem 1. *If h can be considered as a random oracle, the compiler transforms a group key exchange protocol P secure against any passive adversary into an authenticated group key exchange protocol P' with key confirmation which is secure against any passive adversary, active adversary, or malicious insider.*

4 A New Compiler with TTP

We assume that $\Sigma = (Gen, Sign, Vrfy)$ is a signature scheme which is strongly unforgeable under an adaptive chosen message attack. We also assume that a unique session identifier S_{ID} is securely distributed to the participants and the TTP before every protocol instance is initiated. In addition, we assume that the TTP acts honestly and is trusted by all the participants.

Suppose a set $S = \{U_i, \cdots, U_n\}$ of users wish to establish a session key, and h is a one-way hash function. Let ID_{U_i} represent U_i's identity for every i $(1 \leq i \leq n)$.

Given a protocol P secure against any passive adversary, the compiler constructs a new protocol P', in which each party $U_i \in S$ performs as follows.

1. In addition to all the operations in the initialisation phase of P, the TTP generates a verification/signing key pair $(PK_{\mathrm{TA}}, SK_{\mathrm{TA}})$ by running $Gen(1^k)$, and make PK_{TA} known to all the potential participants. Each party $U_i \in S$ also generates a key pair (PK_{U_i}, SK_{U_i}) by running $Gen(1^k)$. The TTP knows all PK_{U_i} $(1 \leq i \leq n)$.
2. In each round of the protocol P, U_i performs according to the following rules.
   - In the first round of P, U_i sets his key exchange history $H_{i,1}$ to be S_{ID}, and sets the round number k to 1. During each round, U_i should synchronise the round number k.
   - When U_i is supposed to send message m_i to other users, it broadcasts $M_i = ID_{U_i}||k||m_i$.
   - When U_i receives the message M_j from user U_j $(1 \leq j \leq n)$, it computes the new key exchange history as:

     $$H_{i,k} = h(H_{i,k-1}||S_{ID}||k||M_1||\cdots||M_n)$$

     Then U_i continues as it would in P upon receiving the messages M_j. As before, U_i does not need to store copies of received messages.
   - In an additional round, U_i computes and sends the key confirmation message $ID_{U_i}||k||H_{i,t}||\sigma_i$ to the TTP, where

     $$\sigma_i = Sign_{SK_{U_i}}(ID_{U_i}||H_{i,t}||S_{ID}||k)$$

3. The TTP checks whether all the key exchange histories from U_j $(1 \leq j \leq n)$ are the same, and verifies each signature. If all these verifications succeed, the TTP computes and broadcasts the signature $\sigma_{\mathrm{TA}} = Sign_{SK_{\mathrm{TA}}}(H_{i,k}||S_{ID}||k)$. Otherwise, the TTP broadcasts a failure message $\sigma_{\mathrm{TA}} = Sign_{SK_{\mathrm{TA}}}(S_{ID}||str)$, where str is a pre-determined string indicating protocol failure.
4. U_i verifies the signature from TA. If the verification succeeds, then U_i computes the session key K_i as required in protocol P. If this check fails, or if U_i receives a failure message from the TTP, then U_i aborts the protocol.

In addition to the initialisation phase, the above protocol adds two rounds to the original protocol and achieves key confirmation. Each participant needs to sign one message and verify one signature, in addition to the computations involved in performing P. In addition, it only needs to store the (hashed) key exchange history. Of course, the TTP needs to verify n signatures and generate one signature.

Theorem 2. *If h can be considered as a random oracle, the compiler transforms a group key exchange protocol P secure against any passive adversary into an authenticated group key exchange protocol P' with key confirmation which is secure against any passive adversary, active adversary, or malicious insider.*

5 Conclusion

In this paper, we have investigated existing methods for building authenticated group key agreement protocols, and proposed two compilers which can generate more efficient protocols than the Katz-Yung compiler.

References

1. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Transactions on Information Theory **IT-22** (1976) 644–654
2. Burmester, M., Desmedt, Y.: A secure and efficient conference key distribution system. In Santis, A.D., ed.: Advances in Cryptology—EUROCRYPT '94. Volume 950 of Lecture Notes in Computer Science, Springer-Verlag (1994) 275–286
3. Kim, Y., Perrig, A., Tsudik, G.: Communication-efficient group key agreement. In: Proc. IFIP TC11 16th Annual Working Conference on Information Security. (2001) 229–244
4. Katz, J., Yung, M.: Scalable protocols for authenticated group key exchange. In Boneh, D., ed.: Advances in Cryptology — Crypto 2003. Volume 2729 of Lecture Notes in Computer Science, Springer-Verlag (2003) 110–125
5. Katz, J., Shin, J.: Modeling Insider Attacks on Group Key-Exchange Protocols. Cryptology ePrint Archive: Report 2005/163. (2005)
6. Bresson, E., Chevassut, O., Pointcheval, D., Quisquater, J.J.: Provably authenticated group Diffie-Hellman key exchange. In: Proceedings of the 8th ACM Conference on Computer and Communications Security. ACM Press (2001) 255–264
7. Bresson, E., Catalano, D.: Constant Round Authenticated Group Key Agreement via Distributed Computation. In: Proc. of PKC 2004. (2004) 115–129
8. Mayer, A., Yung, M.: Secure protocol transformation via "expansion": from two-party to groups. In: Proceedings of the 6th ACM conference on Computer and communications security. ACM Press (1999) 83–92
9. Bellare, M., Neven, G.: Transitive Signatures Based on Factoring and RSA. In Zheng, Y., ed.: Advances in Cryptology — Asiacrypt 2002. Volume 2501 of Lecture Notes in Computer Science, Springer-Verlag (2002) 397–414

Insider Impersonation-MIM Attack to Tripartite Key Agreement Scheme and an Efficient Protocol for Multiple Keys

Lihua Wang[1], Takeshi Okamoto[1], Tsuyoshi Takagi[2], and Eiji Okamoto[1]

[1] Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
wlh@cipher.risk.tsukuba.ac.jp, {ken, okamoto}@risk.tsukuba.ac.jp
[2] School of Systems Information Science, Future University-Hakodate, 116-2 Kamedanakano-cho, Hakodate, Hokkaido 041-8655, Japan
takagi@fun.ac.jp

Abstract. In this paper, we introduce the definition of *insider impersonation-MIM attack* for tripartite key agreement schemes and show that almost all of the proposed schemes are not secure under this attack. We present a new protocol which is much more efficient than the existential secure protocol [13] in terms of computational efficiency and transmitted data size. Moreover, our protocol is the first scheme for *multiple keys* which means that not only a large number of keys but also various kinds of keys can be generated by applying our scheme.

1 Introduction

In 2000, Joux [5] proposed a tripartite Diffie-Hellman key agreement protocol based on the Weil pairing. This protocol is more efficient than the previous ones, because it requires each entity to transmit only a single broadcast message which is also called partial message. However, it suffers from the man-in-the-middle (MIM) attack because it is unauthenticated. Thereafter, in order to resist the MIM attack, several modified protocols [1, 4, 7, 8, 10, 11, 13] were proposed.

From a security viewpoint, a robust key agreement protocol with three (or more than three) participates should prevent insider attacks apart from outsider attacks. Because that anyone who belongs to the three participates might become an attacker to cheat the other two participates. By the way, it is well known that in order to keep security, as the case may be, a session key generated by a key agreement scheme has to be thrown away after being used just for once. So that it is valuable if a large number of session keys can be generated by running a scheme once.

In this paper, we introduce the definition of *insider impersonation-MIM attack* (InIm-MIM attack) for three-party key agreement scheme (see section 3). Under this attack, almost all of the proposed schemes [1, 4, 7, 8, 10, 11] are not secure. Furthermore, we propose a tripartite key agreement protocol for multiple keys. Here, *Multiple keys* mean that not only a large number of keys but also various kinds of keys, which meet with the different practical demands, can be generated by applying our scheme.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 198–203, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The rest of this paper is organized as follows. In section 2, we briefly review the properties of pairing and the security assumption. In section 3, we introduce the definition of the InIm-MIM attack. Then we present our new protocol, analyze its security and give some discussions in section 4, 5 and 6, respectively. Finally, we give a conclusion in section 7.

2 Pairings and the Security Assumptions

Let G_1 be an additive group and G_2 be a multiplicative group. Their order q is some large prime. A map $\hat{e} : G_1 \times G_1 \longrightarrow G_2$ is a bilinear map if it satisfies the following three properties:

(1) *Bilinear.* $\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}$ for all $P, Q \in G_1$ and all $a, b \in Z$.
(2) *Nondegenerate.* If $\hat{e}(P, Q) = 1$ for all $Q \in G_1$ then $P = \infty$, and also if $\hat{e}(P, Q) = 1$ for all $P \in G_1$ then $Q = \infty$.
(3) *Computable.* $\hat{e}(P, Q)$ for any $P, Q \in G_1$ is computed efficiently.

Definition 1. *Let P and g are generators of G_1 and G_2 respectively, and $a, b, c \in Z_q \setminus \{0\}$. The corresponding problems are defined as follows:*

- *The* **Computational Diffie-Hellman (CDH) problem**
  - *in $G_i, i = 1, 2$: $\langle P, aP, bP\rangle \Rightarrow abP$; or $\langle g, g^a, g^b\rangle \Rightarrow g^{ab}$.*
- *The* **Computational Inverse Diffie-Hellman (iCDH) problem** [9]
  - *in G_1: $\langle P, aP\rangle \Rightarrow a^{-1}P$; or $\langle P, aP, abP\rangle \Rightarrow b^{-1}P$.*
  - *in G_2: $\langle g, g^a\rangle \Rightarrow g^{a^{-1}}$; or $\langle g, g^a, g^{ab}\rangle \Rightarrow g^{b^{-1}}$.*
- *The* **BDH problem** *in $\langle G_1, G_2, \hat{e}\rangle$: $\langle P, aP, bP, cP\rangle \Rightarrow \hat{e}(P, P)^{abc}$.*

An algorithm $\mathcal{A}$ is said to solve the BDH problem in $\langle G_1, G_2, \hat{e}\rangle$ with an advantage ϵ if $Pr[\mathcal{A}(P, aP, bP, cP) = \hat{e}(P, P)^{abc}] \geq \epsilon$.

BDH Assumption: We assume that the BDH problem in $\langle G_1, G_2, \hat{e}\rangle$ is hard. In detail, when q is a random k-bit prime, no polynomially bounded algorithm $\mathcal{A}$ has a non-negligible advantage in solving the BDH problem, where "non-negligible" means larger than $1/f(k)$ for some polynomial f.

Lemma 1 ([9]). *The iCDH problem can be solved in polynomial time if and only if CDH problem can be solved in polynomial time.*

Lemma 2 ([3, 6]). *The BDH problem $\langle G_1, G_2, \hat{e}\rangle$ is no harder than the CDH problem in G_i, and the CDH problem in G_i is no harder than the discrete logarithm problem in G_i, respectively, where $i = 1, 2$.*

3 Insider Impersonation-MIM Attack

Definition 2. *Let Σ denote the set of entities in a key agreement scheme. E is an attacker for illegally obtaining session keys. We say E is an insider, if $E \in \Sigma$ (i.e., E has the secret information including the passed session keys, long-term private key and secret partial information); and an outsider, otherwise.*

The attack from an insider was first considered by Shim in [12]. Based on the adversaries' roles, there are passive adversaries and active adversaries in the MIM attack. It is clear that the active attacks from insiders are stronger than the ones from outsiders because that an insider knows more information than an outsider. We focus our attention on an active attack, impersonation attack, from an insider.

Insider Impersonation-MIM Attack Model. Let $\Sigma = \{A, B, C\}$. Insider B acts as an attacker who assumes A's name to ask C to practice tripartite key agreement scheme. He submits two messages X_B and X'_A (which maybe based on some old messages generated by A before) to C. On receipt of the two messages, C is fooled into thinking that he is exchanging messages with A and B, and then transmits his message X_C back when, in fact, B is actually intercepting the message X_C. After checking the validity of messages X'_A and X_B, C computes the session key K_C using the received messages X'_A, X_B, the public information of A, B and his own private information. We call this attack the insider impersonation-MIM attack (InIm-MIM attack).

Informally we say *attacker B succeeds* if (1) C accepts the message he submitted; and (2) he can obtain the same key C generated in a polynomial time.

Definition 3 (Tripartite Key Agreement Scheme). *A tripartite key agreement scheme TriKey is a tuple TriKey$= (\mathcal{K}, \mathcal{V}, \mathcal{SK}, \mathcal{PK}, \mathcal{M})$, where:*

- *$\mathcal{SK}$ and $\mathcal{PK}$ are the secret and public key spaces respectively;*
- *$\mathcal{M}$ is the exchange message space;*
- *$\mathcal{V}$ is the verification algorithm which takes as input a key $pk \in \mathcal{PK}$ and a message $X \in \mathcal{M}$ and output 1, if the pair (pk, X) passes the verification and is accepted; 0, otherwise. Precisely,*

$$\mathcal{V}_{I(J)}(X_J) = \mathcal{V}\left(X_J, pk_J \middle| \begin{array}{c} I \in \Sigma = \{A, B, C\}, \\ J \in \Sigma \setminus \{I\} \end{array}\right) = \begin{cases} 1, & I \text{ accepts } X_J, \\ 0, & \text{otherwise}; \end{cases}$$

- *$\mathcal{K}$ is the session key generation algorithm which takes as input one entity's private key sk and other entities' public keys $pk \in \mathcal{PK}$ and partial messages $X_A, X_B, X_C \in \mathcal{M}$ generated by the entities and output a session key.*

$$K_I^{\Sigma} \longleftarrow \mathcal{K}\left(sk_I, r_I; pk_J, X_A, X_B, X_C \middle| \begin{array}{c} I \in \Sigma = \{A, B, C\}, \\ J \in \Sigma \setminus \{I\} \end{array}\right)$$

denotes the session key computed by I. For the legal partial messages, $K_A^{\Sigma} = K_B^{\Sigma} = K_C^{\Sigma} = K^{\Sigma}$.

Definition 4. *We say a scheme is secure against the InIm-MIM attack if there is no algorithm $\mathcal{B}$ in polynomial time, s.t. advantage $Adv_{\mathcal{B}}^{TriKey} = Pr[\{\mathcal{V}_{C(A)}(X'_A) = 1\} \bigwedge \{K_B^{\{A',B,C\}} = K_C^{\{A',B,C\}}\}]$ is non-negligible.*

4 Our Basic Protocol

In this section, we propose an efficient tripartite key agreement protocol which includes three steps as follows.

(1) **System Key Generation:** The Key Generation Center (KGC) generates the system public/secret key pair (R_{pub}, s), where $R_{pub} = sP$ and P is a generator of G_1; then the KGC publishes the system parameters $\mathcal{P} = (G_1, G_2, q, \hat{e}, P, R_{pub}, H)$, where $H : \{0,1\}^* \rightarrow Z_q^*$ is a hash mapping.

(2) **Individual Long-term Key Generation:** The KGC generates the private keys for user I by selecting number $r_I \in_R Z_q^*$ to compute $R_I = r_I P \in G_1, s_I = r_I + g_I s \in Z_q^*$, where $g_I = H(R_I)$; publishes R_I as I's public key; and then distributes s_I to user I as I's private key in a secure way.

(3) **Three-Party Session Keys Generation:**

- *Partial Messages Computation*: User $I(\in \Sigma)$ secretly select $x_I \in_R Z_q^*$ respectively and compute $i = h_I s_I + x_I, X_I = x_I P, Y_I = i^{-1} P$, where $h_I = H(X_I, A, B, C)$ and $i = a, b, c$ corresponding to $I = A, B, C$. And then I sends partial message (X_I, Y_I) to $\Sigma \setminus \{I\}$.
- *Verification*: $I(\in \Sigma)$'s partial message is verified by the other two entities $\hat{e}(Y_I, X_I + h_I(R_I + g_I R_{pub})) \stackrel{?}{=} \hat{e}(P, P)$.
- *One-Round Session Key Computation*: If the above equation holds, let $Z_I = X_I + h_I(R_I + g_I R_{pub})$, where $I \in \{A, B, C\}$, then user A computes the agreement key as follows:

$$
\begin{aligned}
&K_A^{(1)} = \hat{e}(X_B, X_C)^{x_A}, K_A^{(2)} = \hat{e}(X_B, Y_C)^{x_A}, K_A^{(3)} = \hat{e}(Y_B, Y_C)^{x_A},\\
&K_A^{(4)} = \hat{e}(Y_B, X_C)^{x_A}; K_A^{(13)} = \hat{e}(Z_B, X_C)^{x_A}, K_A^{(14)} = \hat{e}(Z_B, Y_C)^{x_A},\\
&K_A^{(15)} = \hat{e}(Y_B, Z_C)^{x_A}, K_A^{(16)} = \hat{e}(X_B, Z_C)^{x_A}, K_A^{(17)} = \hat{e}(Z_B, Z_C)^{x_A};\\
&K_A^{(j+4)} = (K_A^{(j)})^{x_A^{-1}a}; K_A^{(j+8)} = K_A^{(j)})^{x_A^{-1}a^{-1}}; j = 1, 2, 3, 4.\\
&K_A^{(j+5)} = (K_A^{(j)})^{x_A^{-1}a}; K_A^{(j+10)} = (K_A^{(j)})^{x_A^{-1}a^{-1}}; j = 13, 14, 15, 16, 17.
\end{aligned}
$$

Symmetrically, B and C compute the keys $K_B^{(i)}$ and $K_C^{(i)}, i = 1, 2, ..., 27$, respectively. Let $\mathcal{K}$ be the set of keys and write $\beta = \{x_A, a, a^{-1}\} \times \{x_B, b, b^{-1}\} \times \{x_C, c, c^{-1}\}$, then $\mathcal{K} = \hat{e}(P, P)^{\beta}$.

5 Security Against Insider Impersonation-MIM Attack

Due to limited space, here we only give a discussion about the security of our scheme under the InIm-MIM attack as follows.

Theorem 1. *Our scheme is secure against the InIm-MIM attack under the BDH assumption.*

Proof: (sketch) In order to generate a session key, the partial messages have to pass the verification beforehand. The advantage of the InIm-MIM adversary B

$$
\begin{aligned}
Adv_B^{TriKey} &= Pr[\{\mathcal{V}_{C(A)}(X'_A) = 1\} \bigwedge \{K_B^{\{A',B,C\}} = K_C^{\{A',B,C\}}\}]\\
&\leq Pr[\mathcal{V}_{C(A)}(X'_A) = 1].
\end{aligned}
$$

Therefore, our scheme is secure against the InIm-MIM attack, if the probability $Pr[\mathcal{V}_{C(A)}(X'_A) = 1]$ is negligible. In order to obtain $\mathcal{V}_{C(A)}(X'_A) = 1$, the attacker

B first has to make a forge message $(X'_A, Y'_A) \in G_1 \times G_1$, s.t., $\hat{e}(Y'_A, X'_A + h'_A(R_A + g_A R_{pub})) = \hat{e}(P, P)$, where $h'_A = H(X'_A, A, B, C), R_A + g_A R_{pub} = s_A P$.

For any X'_A, $\exists \alpha \in Z_q \setminus \{0\}$ s.t. $X'_A + h'_A(R_A + g_A R_{pub}) = \alpha P$ though the value of α cannot be computed in polynomial time, because it is the discrete logarithm problem in G_1. And then, in order to pass the verification, B must find out point Y'_A, s.t. $\hat{e}(Y'_A, \alpha P) = \hat{e}(P, P)$, in other word, $Y'_A = \alpha^{-1} P$. It is the iCDH problem in G_1.

According to Lemma 1, Lemma 2, our basic protocol is secure against InIm-MIM attack under BDH assumption.

6 Discussion

1. Variety. Besides 27 *One-round session keys* in the basic protocol, there are other three types of session keys generated by our scheme as follows:

(1) A *non-interactive session key*: $K^{(0-path)} = \hat{e}(P, P)^{s_A s_B s_C}$ is generated with no partial message's exchanging among three entities. It has the merit of needing no on-line computation.
(2) $9 = 3 \times 3$ *one-path session keys*: they are generated by exchanging partial message with one path. If A acts as the proposer to generate the *one-path session keys* with B and C, the key set $\mathcal{K}^{A \to (1-path)} = \hat{e}(P, P)^{\beta_A^{(1-path)}}$, where $\beta_A^{(1-path)} = \{a, a^{-1}, x_A\} \times \{s_B\} \times \{s_C\}$, which have the merits of being practical and efficient. If user A wants to send an email to B and C, she would rather select a one-path session key than a one-round session key, because that she can send the partial message together with encrypted email in one-path.
(3) Similarly, 9 *two-path session keys* can be generated.

2. Efficiency. The most popular pairing choices are the Weil pairing and the Tate pairing, which are computable with Miller's algorithm, but the Tate pairing is usually more efficiently implementable than the Weil pairing. According to the recent results (e.g., [2]), it is estimated that the computation of a pairing is approximately 10 times of the exponentiation in G_2.

Only consider the one-round session keys. Let "M", "Add", "Hash", "E" "P", "Num" and "Bandwidth" denote scalar multiplication to a point, addition between two points in G_1, hashing, exponentiation in G_2, pairing, the number of session keys and the number of messages transmitted by one entity, respectively. If the computation of a scalar multiplication to a point in G_1 is approximately replaced to an exponentiation in G_2, we compare the on-line computations between Zhang et al.'s scheme and ours by ignoring the computation of addition between two points in G_1 and hash mapping in Table 1.

Table 1. Comparison of the transmitted data sizes and on-line computation

Protocols	M	Add	Hash	P	E	Bandwidth	Num	On-line time
Ours: one-round	4	2	3	11	27	2 elements in G_1	27	4M+11P+27E≈141E
Zhang et al. [13]	6	3	3	8	8	3 elements in G_1	8	6M+ 8P+8E≈94E

Consequently, our protocol is much more efficient in terms of the number of keys, computational efficiency and transmitted data size than the existing secure protocol proposed by Zhang et al., because in our scheme, approximately 1.5 times computations than Zhang et al.'s scheme, more than 3 times numbers of session keys can be generated.

7 Conclusion

In this paper, we introduced the definition of InIm-MIM attack for three-party key agreement scheme and an efficient protocol, which is the first multiple keys tripartite key agreement scheme. Multiple keys mean not only a large number of keys but also various kinds (such as non-interactive session key, one-path session keys etc., which have the merits of being more practical and efficient in the real world.) of keys can be generated by applying our scheme only once.

References

1. S. S. Al-Riyami and K. G. Paterson, *Tripartite authenticated key agreement protocols from pairings*, Cryptology ePrint Archive, Report, No.035, 2002.
2. P. S. L. M. Barreto, H. Y. Kim, B. Lynn and M. Scott, *Efficient algorithms for pairing-based cryptosystmes*, in *Advances in Cryptology - CRYPTO 2002*, LNCS 2442, Springer-Verlag, Berlin, pp. 354-368, 2002.
3. D. Boneh and M. Franklin, *Identity-based encryption from the Weil pairing*, in *Advances in Cryptology - CRYPTO 2001*, LNCS 2139, Springer-Verlag, Berlin, pp. 213-229, 2001.
4. Z. Cheng, L. Vasiu and R. Comley *Pairing-based one-round tripartite key agreement protocols*, Cryptology ePrint Archive, Report, No.079, 2004.
5. A. Joux, *A One round protocol for tripartite Diffie-Hellman*, in *Algorithmic Number Theory, 4th International Symposium*, LNCS 1838, Springer-Verlag, Berlin, pp. 385-394, 2000.
6. A. Joux, *The Weil and Tate pairings as building blocks for public key cryptosystems*, in *Algorithmic Number Theory, 5th International Symposium*, LNCS 2369, Springer-Verlag, Berlin, pp. 20-32, 2002.
7. D. Nalla, *ID-based tripartite key agreement with signatures*, Cryptology ePrint Archive, Report, No.144, 2003.
8. D. Nalla and K. C. Reddy, *ID-based tripartite authenticated key agreement protocols from pairings*, Cryptology ePrint Archive, Report, No.004, 2003.
9. A. Sadeghi and M. Steiner, *Assumptions related to discrete logarithms: why subtleties make a real difference*, in *Advances in Cryptology - EUROCRYPT 2001*, LNCS 2045, Springer-Verlag, Berlin, pp. 244-261, 2001.
10. K. Shim, *A man-in-the-middle attack on Nalla-Reddy's ID-based tripartite authenticated key agreement protocol*, Cryptology ePrint Archive, Report, No.115, 2003.
11. K. Shim, *Efficient one-round tripartite authenticated key agreement protocol from the Weil pairing*, Electronics Letters 39, pp. 208–209, 2003.
12. K. Shim, *Cryptanalysis of Al-Riyami-Paterson's authenticated three party key agreement protocols*, Cryptology ePrint Archive, Report, No.122, 2003.
13. F. Zhang, S. Liu and K. Kim, *ID-based one round authenticated tripartite key agreement protocol with pairings*, Cryptology ePrint Archive, Report, No.122, 2002.

An Immune System Inspired Approach of Collaborative Intrusion Detection System Using Mobile Agents in Wireless Ad Hoc Networks

Ki-Won Yeom and Ji-Hyung Park

CAD/CAM Research Center, Korea Institute of Science and Technology,
39-1, Hawolkog-dong, Seongbuk-gu, Seoul, Korea
{pragman, jhpark}@kist.re.kr

Abstract. Many single points of failure exist in an intrusion detection system (IDS) based on a hierarchical architecture that does not have redundant communication lines and the capability to dynamically reconfigure relationships in the case of failure of key components. To solve this problem, we propose an IDS inspired by the human immune system based upon several mobile agents. The mobile agents act similarly to white blood cells of the immune system and travel from host to host in the network to detect any intrusions. As in the immune system, intrusions are detected by distinguishing between "self" and "non-self", or normal and abnormal process behavior respectively. In this paper we present our model, and show how mobile agent and artificial immune paradigms can be used to design efficient intrusion detection systems. We also discuss the validation of our model followed by a set of experiments we have carried out to evaluate the performance of our model using realistic case studies.

1 Introduction

A computer system's security mechanisms should prevent unauthorized access to its resources and data. However, it is impossible to build a completely secure system for many different reasons: programs and operating systems have vulnerabilities, firewalls can be circumvented, passwords can be cracked, and a system can be abused by insiders.

Ad hoc wireless network is a collection of mobile nodes that establish a communication protocol dynamically. The nodes may join the network at any time and communicate with entire network via the neighboring nodes. There are no base stations, and each member of such a network is responsible for accurate routing of information, and takes part in routing decisions. Due to arbitrary physical configuration of an ad hoc network, there is no central decision making mechanism of any kind - rather, the network employs distributed mechanisms of coordination and management.

In this paper, we propose bio-inspired IDS based on mobile agents and show how artificial and human immune system coupled with the mobile agent technology can realistically improve the security of large and complex networks. The proposed immune system consists of a complex network of cells, which are responsible for the defense of the body against attacks of external invaders, such as bacteria, virus or parasites. To accomplish this task, the immune system must be able to distinguish its

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 204 – 211, 2005.
© Springer-Verlag Berlin Heidelberg 2005

own molecules and cells body (e.g., self) from foreign molecules (e.g., non-self). On the other hand, mobile agent technology allows us to combine desirable characteristics when possible. It also makes it possible to reduce system integration effort. In this work, we define a computational model based on both agents and an artificial immune system paradigm. Our approach generates three groups of data as a result of the network intrusion detection: attacks, security violations and security events. Later those events are analyzed, distributed and stored using the computational and agent based architecture model.

2 Background

The overview presented in this section is largely incomplete as only enough information is provided to understand how its concepts can be applied to an IDS.

2.1 An Overview of the Human Immune System

The human immune system is based on the concept of distinguishing molecules and cells of the body, called "self", from foreign ones, called "non-self", and elimination of the latter. The foreign cells are called antigens and include any invaders, such as bacteria and viruses.

The function of the immune system is implemented through the interactions between a large number of different types of cells rather than by one particular organ. The system has multiple levels of defenses, from the skin (which forms the outermost barrier of protection) to the adaptive immune system. The organs of the adaptive immune system, called lymphoid organs, are positioned throughout the body and store lymphocytes, also known as white blood cells. White blood cells function as small, disposable and independent intrusion detectors that circulate through the body in the blood and lymph systems. Each white blood cell is specialized to ignore self-cells and bind to a small number of structurally related non-self cells. Recognition and binding to non-self cells is accomplished through special receptors on white blood cells. The receptors are structured such that they will bind to a particular peptide (a sequence of amino acids, which make up proteins). As different types of cells contain different proteins, the receptors allow a white blood cell to recognize specific non-self cells and bind to them. After binding, many events still take place, usually resulting in scavenger cells (macrophages) eliminating the antigen.

2.2 Immune System Properties

Based on a study of the human immune system, the properties relevant to the proposed IDS are discussed below. These properties not only enable the natural immune system to effectively detect and eliminate intrusions, but they also make the system fault-tolerant. The system can remain functional even when many of its components have failed.

Distributability: The immune system is highly distributed. No central co-ordination takes place; infections can be locally recognized by lymphocytes. Distribute-ability greatly enhances the system's robustness.

Detection: Detection (or recognition) occurs in an immune system occurs when pathogen fragments and receptor on lymphocytes surface bond chemically.

Diversity: The immune system of each individual in a population is unique. This ensures that not all individuals will be vulnerable to the same pathogen to the same degree, thereby enhancing the survival of the population as a whole. Detection in immune system is related with non-self elements of the organism, thus, the immune system must have diverse receptors to ensure that at least some of lymphocytes will react to the pathogenic element.

Learning: An immune system must be capable of detecting and eliminating pathogens as quickly as possible. It includes principle which allows lymphocytes to learn and adapt themselves to specific foreign protein structures, and to "remember" these structures when needed. These principles are implemented by B-cells.

2.3 Immune System and Wireless Ad Hoc Network Security

The immune system protects the body from pathogens elements in the same way as computer security systems protects the computer from malicious users such as:

Disposability: This aspect allows the body to continue working even under pathogen attack.

Correction: It is a manner to prevent the immune systems from attacking its own body cells.

Integrity: This is a way to guarantee that the genetic codes present in the cells will not be corrupted by any pathogen.

Accountability: represents the means adopted by immune systems to identify, find and destroy the pathological agents.

2.4 Mobile Agents

There are many benefits derived from applying mobile agents for intrusion detection. The benefits relevant to the proposed IDS are discussed in this section by using the work of Jansen [6].

Autonomous and Asynchronous Execution: IDS based on mobile agents can continue to operate in the event of failure of a central controller or a communication link. Unlike message passing routines or remote procedure calls, once a mobile agent is launched from a home platform, it can continue to operate autonomously even if the host platform from where it was launched is no longer available or connected to the network.

Dynamic Adaptation: The ability for mobile agent systems to sense their environment and react to changes is useful in intrusion detection. Agents may move elsewhere to gain better position or avoid danger. They can also clone themselves for redundancy and parallelism, or call other agents for assistance. Agents can also adjust to favorable situations as well as unfavorable ones.

Robust and Fault-Tolerant Behavior: The ability of mobile agents to react dynamically to unfavorable situations and events makes it easier to build robust distributed

systems. Their support for disconnected operation and distributed design paradigms eliminate single points of failure problems and allow mobile agents to offer fault tolerant characteristics.

3 IDS Based on the Immune System and Mobile Agents

We propose an approach that is to implement the adaptive immune system layer by kernel assisted lymphocyte processes that can migrate between computers, making them mobile agents. With help from the kernel, the lymphocyte processes are able to query other processes to determine if they are functioning normally.

3.1 The Intrusion Detection Model

In the human immune system, intrusions (in the form of pathogens) are detected by searching for abnormal or "non-self" peptide patterns. The proposed IDS employs anomaly detection, just as the immune system does, because intrusions are detected by searching for deviations from the normal behavior of programs. The first step of anomaly detection is to learn what the normal behavior of each specific program is. The definition of normal behavior of processes is based on the work of [7]. In our approach, the "peptide" used for recognition of non-self is defined in terms of sequences of system calls executed by privileged processes in a networked operating system.

A separate database of normal behavior for each process of interest is created by examining the sequences of system calls made during a program's normal execution. The database is specific to a particular architecture, software version and configuration, local administrative policies, and usage patterns. Several hosts in the network, which will be called lymphoid hosts, store all the different databases defining self. Once the normal behavior of a program is known, we can use it to monitor its behavior whenever it runs. Monitoring is based on examining the behavior of an executing program and comparing it with the normal behavior which has been learnt earlier.

3.2 The Use of Mobile Agents in the IDS

All lymphoid hosts will have the capability to create mobile agents for intrusion detection. However, at any time only one lymphoid host will be assigned the responsibility to create mobile agents, while the other lymphoid hosts will serve as back-ups. When a mobile agent is created, it will obtain the learned normal behavior of a particular program (as decided by the lymphoid host) from the normal profile database found on the lymphoid host. It will then travel from computer to computer in the network. At each computer, it will determine if the particular program is running. If it is, then the mobile agent will examine the system calls generated by that running program and compare them with the system calls stored by the agent. If this examination will show that there is a high deviation from the normal behavior, then the agent will launch an intrusion alert. When a possible intrusion is detected, the mobile agent sends an intrusion alert to the lymphoid host, which also provides a user interface through which the system administrator can learn of intrusions and issue commands to the IDS. At this stage, the IDS may decide to increase the rate of intrusion detec-

tion for this particular program. This can be done in two ways, depending on how the IDS has been configured. One way is to allow the lymphoid host to create and send new mobile agents, whose type is the same as that of the alerting agent. A second way is to allow the alerting agent to multiply.

4 Prototype Implementation

In this prototype implementation, we only concentrate on addressing two different questions: how to model artificial independent individuals; and how to set up an artificial environment, able to provide feedback and support interaction.

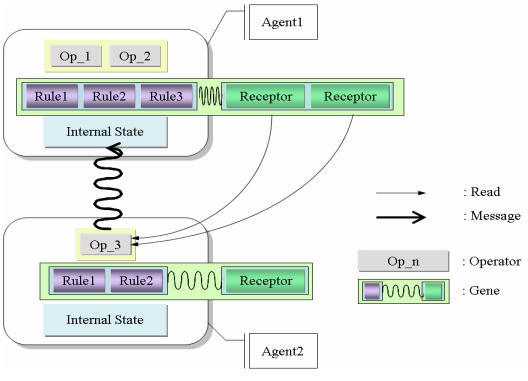


Fig. 1. Internal structure of agents and inter-agent relationships. Particular genes, included in the agent, may be visible to the surrounding environment, acting as receptors. Operators can interfere with other agents using message-passing mechanisms, while receptors may serve as input devices.

Individuals are modeled by way of self-ruling devices, called agents, which can interact with the system as well as with other agents. Comprised within each, lies every mechanism required to express its internal behavior and to respond to system feedback. Since each agent is an independent and closed device, heterogeneous agent populations are directly considered. Communication can be either the result of direct agent interaction, or indirect, using the underlying environment as platform. Each entity has a particular set of rules which determines its external interaction mechanisms, or behavior. These rules are encoded within each gene of the agent's chromosome. Hence, a gene is a partial state controller, where if-then or other type of rules may specify the behavior. Each gene can keep a different format rule, along with other data, such as the conditions under which the rule is to be applied. Therefore, the agent's genotype is the fusion of all the current genes within the chromosome.

Behavior relies on a set of operators which decode specific genetic information and act accordingly. This operators use no explicit fitness function, instead they use environment and genetic information to guide the agent. Thus, global behavior is obtained through each local operator's evaluation of the genetic rules comprised within each agent. Moreover, each agent can present the environment and other agents with a set of external features, which may be also encoded in the genetic code.

The agent population is distributed over a set of sites comprised in the lattice, which, along with the corresponding spatial relationships, can be expressed through an object-oriented cellular automaton. This allows a detachment between each physical location and the lattice that comprises them. Moreover, agents and environment will also become separated. Since each lattice site can be different from all others, it enables the construction of heterogeneous site worlds. Fig.1 shows the internal agents structure and their relationships.

5 Experiments and Results

As an example, suppose the simulation of a system where two different agent classes were comprised. The artificial world is composed of two types of locations: L1, where substances spread to same class sites where its amount is scarcer; in L2 locations substances do not spread but dissolve locally at a certain rate. To simulate this artificial world one needs to define two different features: first, define the behavior of each one of the agents. This is accomplished through the definition of the genetic rules and operators which hold its behavior. An agent is then able to autonomously perform the task subdued to its set of rules; second, define the underlying environment. This is done by creating the local rules for each independent site.

The artificial immune system agents here considered may be separated into two main classes. The first one contains the immune system cells (agents), namely four types of leukocytes: B lymphocytes, CD4 T lymphocytes, CD8 T lymphocytes and macrophages. The other class is a set of pathogenic agents (virus), specifically the HV-I and two other theoretical viruses RB, whose function is to present the system with a series of situations, allowing the study of different immune system responses. The RB virus simply releases specific antigen, leaving the immune system cells unscathed. A subset of immune system soluble mediators is included in this model, comprehending Interleukin-1(IL-1), Interleukin-2(IL-2), Interleukin-6, γ-Interferon and cytotoxic agents produced by T killer cells. Antibody and antigen is produced or secreted accordingly to the pathogen's type.

The first test aimed Interleukin-1 and Interleukin-5, and the obtained results are in Fig.2. At point A, T helper cells were introduced in an otherwise empty system. The system was challenged with antigen upon the addition of RB virus, at point B. No immune response was observed, for T helper cells are not able to recognize antigen in free form, only antigen pairs, thus requiring the help of antigen presenting cells. B cells, were introduced at point C. T helper response was still unobserved, since IL-1 is required to attain a successful activation, and is produced by macrophages, which were not considered in this system. On the other hand, B cells only react in presence of IL-5, which in turn is secreted by activated T CD4 cells. To overcome this deadlock, IL-1 was injected in instant D, which led to the chain of the immune response. As an aftermath, the pathogenic agents were successfully eliminated.

Another simulation presented the immune system the same antigen twice in order to compare both immune responses. A stronger immune response was observer during the second antigen challenge (Fig. 2). At the outset, immune system cells were the only artificial immune system agents. Following the release of RB virus, an immune response was developed, increasing the number of T helper cells and B cells with

specific RB-antigen receptors. T CD8 cell number remained constant, since this virus does not infect cells. When the system was again challenged with the same number of pathogenic RB virus, an immune response developed much stronger T cells more rapidly than the first. This proved that the model possesses one of the key features of a real immune system such as specific antigen memory.

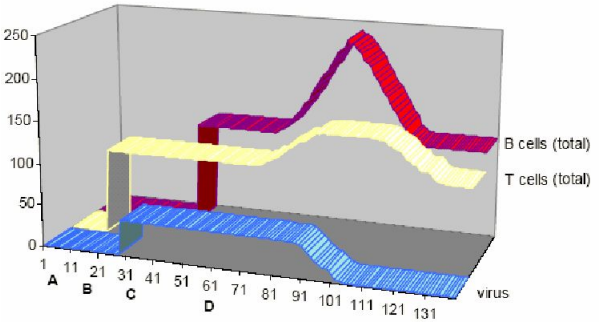


Fig. 2. IL-1and IL-5 as regulators of the immune system. IL-1 was presented in instant D, resulting in IL-5 production. A typical immune response followed. The importance of cell cooperation and immune mediators was asserted.

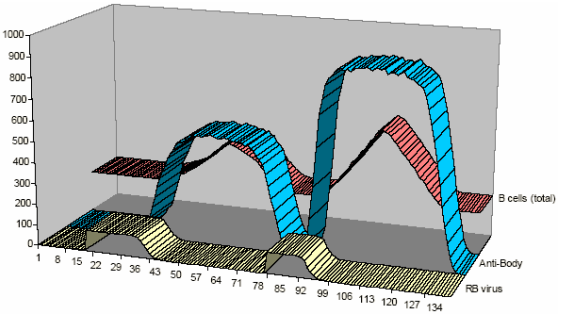


Fig. 3. Secondary immune response. When the immune system is challenged a second time with the same pathogenic agent (around cycle 80), a typical secondary response follows: a stronger cell and antibody development. As can be observed, the second infection was easier to overcome.

6 Conclusion and Future Work

We presented description of how mobile agents and concepts of the human immune system could be applied to an IDS. This paper also discussed the main theoretical aspects of how such an IDS can function as well as the similarities between the human immune system and an IDS that uses mobile agents. The whole aspects of the human immune system have not been applied, either because they would complicate the solution or because they were not considered appropriate for an IDS.

We showed that there was a strong correlation between immune system and a computer network security system. The immune system protects the body from pathogens elements in the same way as computer security system protects the computer from malicious users.

Although this study is still at a theoretical and early implementation stage, we may in future build a prototype to demonstrate aspects of the IDS to determine its effectiveness and practicality. In this sense we showed the simulation results, of course it was not completed work. However, it also has some weaknesses. One weakness results from the use of mobile agents. Mobile code has several security concerns that hinder the widespread use of this technology. There are four broad categories of security threats: (1) agent-to-agent, in which an agent exploits the vulnerabilities of other agents residing on the same agent platform; (2) agent-to-platform, in which an agent exploits the vulnerabilities of its platform; and (3) other-to-platform, in which external entities threaten the security of the agent platform. The ways in which the different threats could be resolved or reduced will need to be investigated.

References

1. Wagner, D., Dean, D.: Intrusion detection via static analysis. IEEE symposium on security and privacy (2001)
2. Crispin, C., Steve, B., John, J, Perry, W.: Pointguard - Protecting pointers from buer over vulnerabilities. Proceedings of the 12th USENIX Security Symposium, Washington, D.C. (2003)
3. Barrantes, E.G., Ackley, D.H. , Forrest, S., Palmer, T.S., et al: Randomized instruction set emulation to disrupt binary code injection attacks. Proceeding of the 10th ACM Conference on Computer and Communications Security (2003)
4. Jon, G., Somesh, J., Bart M.: Efficient context-sensitive intrusion detection. Network and Distributed System Security Symposium (2004)
5. Percus, Jerome K., Percus, Ora E., Alan S., Perelson: Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self non-self discrimination. Vol.90. Proceedings of the National Academy of Sciences of the United States of America (1993) 1691-1695
6. Jansen, W.: Intrusion detection with mobile agents. Vol.25(15). Computer Communications (2002) 1392-1401
7. Forrest, S., Hofmeyr, S., Somayaji A.: Computer Immunology. Vol.40(10). Communications of the ACM (1997)8. Forrest, S., Hofmeyr, S., Somayaji A.: Computer Immunology. Vol.40(10). Communications of the ACM (1997)
8. Forrest, S., Hofmeyr, S., Somayaji A.: Computer Immunology. Vol.40(10). Communications of the ACM (1997)

A New User-Habit Based Approach for Early Warning of Worms

Ping Wang, Binxing Fang, and Xiaochun Yun

Dept. of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, Heilongjiang, China
wangping@hit.edu.cn

Abstract. In the long term usage of the network, users will form certain types of habit according to their specific characteristics, individual hobbies and given restrictions. On the burst-out of worms, the overwhelming flow caused by random scanning will temporarily alter the behavior representation of users. Therefore, it is reasonable to conclude that the statistics and classification of the user habit can contribute to the detection of worms. On the basis of analysis about both users and worms, we construct the model of user-habit and propose a new approach for the early warning of worms. This paper possesses strong direction significance due to its broad applicability since extended models can be derived from the model proposed in this paper.

1 Introduction

Accompanied with the development of computer technology and the ever-growing dependence on the network in the everyday routine life, the worm comes into its prosperous age and poses a serious threat towards the Internet [1]. The spread rate of worms is very fast, as exemplified by Slammer (sometimes called Sapphire) [2].UP to date, Slammer has been the fastest spreading computer worm. As it began to spread throughout the Internet, the worm infected more than 90 percent of vulnerable hosts within 10 minutes, causing significant disruption to finance, transportation, and government institutions. The frequent bursts of worm epidemic attract the attention of the people to address this problem. To deal with the threat from worms has always been a pressing task.

The worms can damage the network so seriously due to their rapid spread and unexpected burst. In a split second after the worm burst, it is impossible for any effective reaction measure to be done. Therefore, the early-warning of worm has become necessary work, since only after finding the worm in time can the fast and automatic defense mechanism be activated to minimize the loss. Nowadays the early warning technologies of worms include: detect worm's scanning using the ICMP T3 messages [3], detect the worm propagation trend using Kalman filter [4], detect the attack in the network using HoneyPot [5], and monitor the network using network telescope, etc. But these means are short of practicability or sluggish, with the early warning of worms remained to be an open problem.

In this paper, we propose a new approach for early warning of worms based on the user-habit. When users use the network and system for a period of time, all kinds of

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 212 – 219, 2005.
© Springer-Verlag Berlin Heidelberg 2005

restraints would make the users develop certain habits. According to this characteristic, we derive user personal habit from the user access history. On the demand of survival and propagation, worms would scan completely blindly [6]. Then on the basis of user habit, we evaluate the network access to find the suspicious new worm, thus the early warning of worm is realized. Detecting the worm epidemic in time saves the valuable time to eliminate worm, thereby keeping the burst of epidemic within limits.

2 User-Habit Based Early Warning of Worms

Nowadays misuse detection is widely adopted in the detection of worms. However, since more and more kinds of new worms have been coded out due to a much deeper understanding of the system and network, traditional signature-based misuse detection can not deal with such complex situations efficiently. In order to detect new worms effectively, we have to use the means of anomaly detection. In this paper, we construct user-habit model, and any mismatch of a user's access information from the normal model spells the worm appearance.

2.1 Behavior of Users

After some one uses certain kinds of tools for a period, the individuality and habit of this person will be exposed. Similarly, in terms of the network, if a user uses the network for a long time range, some regularities of this user will be revealed, or we can say that the user-habit is generated. In general, the reasons why the user-habit can be produced are listed as follows:

- Conventional access. A user would form special interest or even dependence on his favorite sites.
- The interest. For example: book, news, etc.
- Personality. The age and sex also affect the user access habit.
- The constraints. For convenience reasons, the user access is limited by time, space and other factors.

Generally, we consider a different source address as a different user. However, if a user isn't bound with a specific IP address or different users may use the same IP address, in such situations, we can use some additional information to identify the user, for example, the MAC address of user. Consequently, in order to mine the user-habit, a long history of user access information should be collected and analyzed.

Definition 1. User Access History: The set, composed of the destination IP addresses visited by a user in time range T, is referred to as the user access history, denoted by Φ.

In realization, the early warning system proposed in this paper is disposed at the entry of large scale network. The packet header is first extracted from the data captured by sensors, then the source and destination addresses are picked and stored by hash algorithm. In this way, the user access history is obtained.

2.2 Personal Habit of User

The user access history can provide enough information about user's behavior. However, since the access history of each user is infinite, a question of storage and

efficiency is put forward. The storage space is limited and there is little use to record the access information too long ago. Therefore, we use *user-habit* to more efficiently describe the network behavior of a user.

Definition 2. *Access Information*: A quadruple recording specific information of users' access, including destination address, destination port, access frequency and average access, is referred to as the access information, denoted by *AINFO* (dstIP, dstPort, freq, avgTime).

Definition 3. *User Habit*: user habit refers to the set Ψ, which consists of access information frequently produced by users for a long time. Ψ_i denotes one piece of access information where $1 \le i \le n$ and n is the number of access information in Ψ.

Definition 4. *User Access Destination Habit*: we denote the set Θ to be the user access destination habit, with Θ composed of the frequently visited destination IP addresses:

$$\Theta = \{d \mid d = \mathrm{dip}(\Psi_i), (1 \le i \le n, \mathrm{n} = \mathrm{SIZE}(\Psi))$$

where $\mathrm{dip}(\Psi_i)$ extracts the dstIP field from the access information Ψ_i.

The user habit Ψ is a subset of user access history Φ, i.e. $\Psi \subseteq \Phi$. The user access history is an infinite set, whereas the user habit is a limited set. The user access destination habit Θ is the destination information extracted from the user habit Ψ and it is more compact and aggregate. In this paper we focus on the destIP field of *Access Information* exclusively, and to facilitate our discussion below, we call user access destination habit *destination habit* for short. The destination habit has two simple properties:

a) Aggregate. The destination IP addresses in user access destination habit are relatively a minority in the user access history in quantity, but their frequency is much higher than the common destinations.
b) Long-term Effectiveness. After we get the user access destination habit, this habit will not be changed easily and will be effective for a long time.

Therefore, by picking out those highly visited destinations from latest user access history, we can obtain the user access destination habit list to approximately describe the user access habit recently and this list can be effective in quite a long period.

2.3 Algorithm and Realization

Nevertheless, there exists a problem: perhaps one scan from the host can flush out the whole user habit list and thus destroy the user habit extracted from a wealth of access information a long time before. To avoid this update-storm, we create cache queue and add destinations into user habit list according to certain evaluation mechanism.

It is much more complex to generate user habit list, since rather than recording all the accessed destinations in full detail as adopted in recording user access history, the generation of user habit list requires an evaluation implemented by evaluation mechanism. To gain the user habit list, we first analyze the captured packets and then store the source IPs into a hash table. Each item of the hash table points to a list and a cache queue. The list stores the habit destination and the queue holds the information necessary to generate the habit list, e.g., the potential user-habit DIPs, the access frequency, etc. The generation of user destination habit is shown in Algorithm 1.

Algorithm 1: generate a user habit list for every user

```
Input: the access history of n days before
Output: user habit list
for each active sip in the network
  read all destinations accessed by sip
  aggregate the destination addresses DIPs
  count the number of DIPs
  use Algorithm 2 to calculate sip habit list length m
  if m = = max_habit_len then continue
  else
    use Algorithm 3 to obtain m items
    export the user habit list
  end if
end
```

Algorithm 2: calculate the appropriate length of user habit list, the upper limit is *max_habit_len*

```
Input: Array A[n] of access history n days before
Output: an appropriate length of user habit m
m = 0
for each access history A_i in A
  if len(A_i)<max_habit_len and len(A_i)>m then m=len(A_i)
end for
if m = = 0 then
  return max_habit_len
else return m
end if
```

Algorithm 3: select *m* items of the heaviest weight from list *L* of length *n*

```
Input: an unsorted list L of length n
Output: m items of the heaviest weight
create an empty heap hp
for each item L_i in L
  if size(hp)< m then
    insert L_i into hp
    continue
  else if min(hp) < L_i
    delete min(hp) from hp
    insert L_i into hp
  else drop this item
       continue
  end if
end for
export hp as user habit list
```

After generating the user habit list, we use this list as a tool for early warning of worms. The procedure is as follows: If the destination IP visited by a user is not in this user's habit list, then increase the user's habit destroy degree. When the accumulative degree reaches the predefined threshold, the system gives out warning. This warning mechanism is implemented in Algorithm 4. Notice that *WCOFF* in step5 is the warning coefficient, with a value ranging in the interval [1, 2] commonly.

Algorithm 4: give warning according to the habit destroy degree

```
Input: network data: packet
Output: a Boolean judgment of warning or not
warn = false
extract the sip and dip from packet
search user's habit list hb_list identified by sip
if dip in hb_list then return
else
    the sip's damage degree dd++
    if dd > WCOEF * len (hb_list) then warn = true
end if
```

According to the damage degree of user habit, we can give early warning of worms, whereas if the damage degree doesn't increase markedly in a long period, we can reset the degree to zero. The strategy of degree reset can be derived from the situation in the network. We can reset the users with small damage degrees when the user activities are not active, it's late at night or wee hours commonly.

The critical link in this early warning system is the generation of user habit destination list. The longer period the user access information is collected, the more precisely can we describe the real user access habit and thus give the warning accurately and timely. From the experiment results, we can see that the habit list is closely related to user characteristics. For an enthusiastic Internet lover, the habit could be a surprisingly long list, whereas for a not so active one, the habit list is short and the time to construct the list is correspondingly short. Generally, for a common user, the habit list tends to be stable after 5 days. After generating user habit lists, we come down to the core function of our early warning system - warning of the worm. If a user is infected by worm, the damage degree of user's habit will increase rapidly in a very short time as a result of the blind probes made by worm across a wide range of random IP addresses beyond the habit list. When the degree finally reaches the threshold, the System will give the warning and the early warning of worm is implemented in this way.

3 Experiments

3.1 Experiment A (Quick Warning of Worm)

The topology of network in experiment is shown in Fig.1. Three PCs in the local network are monitored for 5 days to create the habit list of 3 users. We install virtual

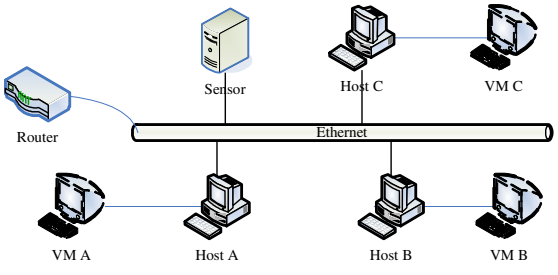


Fig. 1. Topology graph of test environment

machines on each PCs using VMware, which are bridged to the PCs. The access from a virtual machine is regarded the same as the access from PC host.

After having Host A running for a period, we release the Slammer worm in PC host A at a later time. The destination habit of user A is damaged and the number of strange IPs which are not in the habit list of user A is shown in Fig.2.

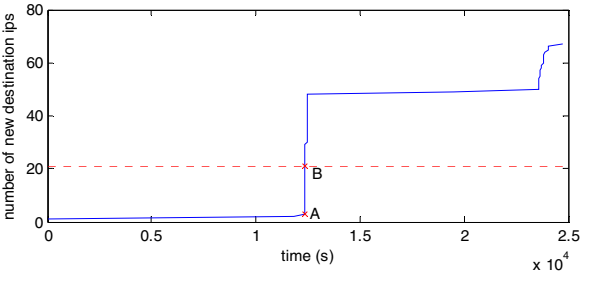


Fig. 2. Sketch map of worm early warning

Since PC A accesses only a few destinations during the period of experiment, the length of its habit list is only 21. With the WCOEF defined as 1, the dashed line in the figure defines the threshold of warning where the number of destination IPs not included in user habit list reaches the upper limit allowed. We release Slammer worm at point A and the system gives out warning at point B after 1.03 seconds. This time includes the delay of network interface. This experiment proves that the system is able to have a quick detection of worm epidemics and thus can be used in the early warning of the worm.

3.2 Experiment B (Detection of Slow Scan Worm)

Using the experiment environment in Experiment A, we analyze the access scene of PC host B. After PC B runs for some time, we execute the Slowping program in the virtual machine of PC B. The Slowping, coded by ourselves, scans the network at the rate of 1host/s and sends an ICMP echo_request message to the active host detected. In this way, we simulate the scan of worm manually. However, the intensity of Slowping is much weaker than the real worm. For example, the CodeRed II worm varies the number of threads according to the operation system: in Chinese version, the worm creates 600 threads whereas in other language version, it creates 300

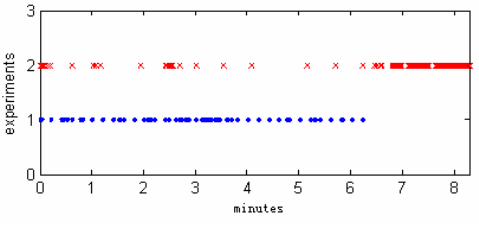


Fig. 3. Detection of the change in single user's access habit

threads. From the result of this experiment, we can see that the slow scan can also be detected effectively as shown in Fig.3.

In Fig.3, each point of Experiment 1 implies a destination IP visited by user B and each cross in Experiment 2 implies a visited destination which is not in the user habit list. At the tail of Experiment 2 is the scene of manual simulation of worm. From the figure we can see that the visits of user are numerous and scattering all around. With the user habit list, regular accesses are recognized as user habit, whereas random accesses not in the user habit emerging in concentrated mode are considered to be the preclusion to a worm epidemic. The length of user B's habit list is 168- a substantially larger size than that of A -due to its activeness. While the Slowping is running in the virtual machine, user B uses the PC as usual. After about 2 minutes, the system gives warning against worm. This experiment indicates that the anomaly detection using user habit can detect the change of user habit very well.

4 Conclusion and Future Work

Each worm comes and goes with heavy loss of money and enormous inconvenience for people. To free people from such miserable experience, we must detect and give warning against the worm as soon as possible. Since to what degree can a defense system give out warning against the outbreak of worm has a significant impact on the ultimate effectiveness of worm treatment, the work of giving early worm warning proves to be a critical link in the overall worm containment project. However, tradition detection based on misuse detection is inefficient to give an in-time warning, therefore, anomaly detection is a must in this scenario. In this paper, we first analyze the user habit and principle of worm propagation. Then utilizing the nature of worm's blind scan, we focus on the influence of worm on user-habit. On the basis of the work above, we construct user-habit model and summarize how the convention of user-habits can be damaged by worm. We propose the approach for early warning of the worm in a large scale network based on the user-habit. By providing a feasible approach, we make constructive exploration in the way to detect the worm effectively and quickly. The experiments in the real network prove that this user-habit based approach can effectively detect worm burst in a large scale network and provide favorable conditions for the next step to quench the epidemic. At the same time, the user-habit model constructed in this paper provides a good platform for the anomaly-based Intrusion Detections.

The trend of worm being more intelligent and shady demands our constant research on the new propagation model and effective warning and containing strategy. Therefore, our future work includes: research on worm's propagation model and the anomaly in network caused by worms, the approach and strategy for early warning of the worm and thinning in user-habit categorization. Furthermore, sophisticated categorization of user-habit also has great significance in the field of IDS.

References

1. Moore, D., Shannon, C., Claffy, K.: Code Red: A case study on the spread and victims of an Internet worm. Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement (2002) 273–284

2. Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., Weaver, N.: Inside the slammer worm. IEEE Magazine of Security and Privacy, Vol.1 (4) (2003) 33–39
3. Berk, V.H., Gray, R.S., Bakos, G.: Using sensor networks and data fusion for early detection of active worms. In Proceedings of the SPIE AeroSense (2003) 92–104
4. Zou, CC., Gao, L., Gong, W., Towsley, D.: Monitoring and Early Warning for Internet Worms. Proceedings of the 10th ACM Conference on Computer and Communication Security (2003) 190–199
5. Kuwatly, I., Sraj, M., Al Masri, Z., Artail, H.: A Dynamic Honeypot Design for Intrusion Detection. IEEE/ACS International Conference on Pervasive Services (2004) 95–104
6. Wen, WP., Qing, SH., Jiang, JC., Wang, YJ.: Research and Development of Internet Worm. Journal of Software, Vol.15 (8) (2004) 1208–1219

A Multi-gigabit Virus Detection Algorithm Using Ternary CAM*

Il-Seop Song, Youngseok Lee, and Taeck-Geun Kwon

Dept. of Computer Engineering, Chungnam National University,
220 Gung-dong, Yuseong-gu, Deajeon 305-764, Korea
{issong95, lee, tgkwon}@cnu.ac.kr

Abstract. During the last few years, the number of Internet worms and viruses has significantly increased. For the fast detection of Internet worms/viruses, the signature-based scheme with TCAM is necessary for the network intrusion detection system (NIDS). However, due to the limit of the TCAM size, all the signatures of Internet worms/viruses cannot be stored. Hence, we propose a two-phase content inspection algorithm which can support a large number of long signatures at TCAM. From the simulation results, it is shown that our algorithm for TCAM provides a fast virus-detection capability at line rate of 10Gbps (OC192).

Keywords: Deep packet inspection, content inspection, pattern matching, TCAM, network security.

1 Introduction

Recently, an epidemic of Internet worms/viruses has motivated a lot of anti-virus solutions at the middle of the network as well as at the end hosts. Network Intrusion Detection Systems (NIDSes), (e.g., Snort[1]) monitor packets in the Internet and inspect the content of packets to detect malicious intrusions. Since the Snort system implements pattern matching algorithms in software, it is hard to keep up with the line rate of the current Internet such as 10Gbps (OC192). In addition, malicious contents in Snort rules do not include worms and virus patterns represented of very long signatures. However, the ClamAV [2] has very long signatures such as 233 bytes signature. Fig. 1 shows length distributions of Snort contents and ClamAV [2] signatures.

To provide multi-gigabit line rates content inspection, hardware-assistant NIDS systems are essential; one possible solution is to use Ternary Content Addressable Memories (TCAMs) which support longest prefix matching as well as exact matching. Fig. 2 illustrates an example of a TCAM search operation with an input pattern of "1000." Since a TCAM finds a matched rule in $O(1)$, it can keep up with the current line rate of high-speed Internet switches or routers.

* This research was supported in part by ITRC program of the Ministry of Information and Communication, Korea.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 220 – 227, 2005.
© Springer-Verlag Berlin Heidelberg 2005

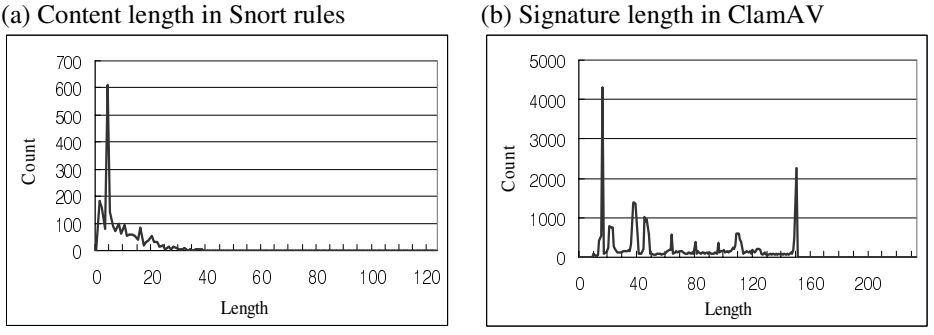


Fig. 1. Distribution of content lengths in Snort rules and virus signature length in ClamAV

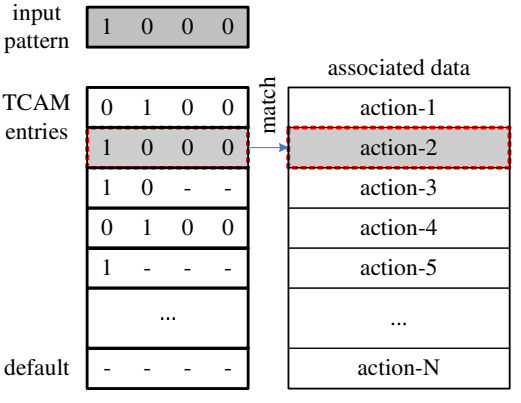


Fig. 2. An example of a TCAM operation

Worm.CodeRed.2

```
ff559c8d855cfefffff50ff55988b40108b08898d58fefffffff55e43d
040400000f94c13d040800000f94c50acd0fb6c9898d54fefffff8b75
08817e309a0200000f84c4000000c746309a020000e80a000000436f
64655265644949008b1c24ff55d8660bc00f958538fefffc78550
```

Worm.CodeRed

```
8bf450ff9590fefffff3bf490434b434b898534fefffffeb2a8bf48b8
d68fefffff518b9534fefffff52ff9570fefffff3bf490434b434b8b8d
4cfefffff89848d8cfefffffeb0f8b9568fefffff83c201899568fefff
f8b8568fefffff0fbe0885c97402ebe28b9568fefffff83c2018995
```

Fig. 3. An example of CodeRed virus signatures

Although a TCAM-based deep packet inspection scheme is suitable for fast pattern matching of malicious intrusion and virus signatures, TCAM has very limited storage capacity and high power consumption. Many researches including [3] and [4] have focused on reducing the size of tables stored in TCAM. Unfortunately, the capacity of current TCAM products – e.g., IDT [5] NSE products – is not enough to store all virus and worm signatures which still grow rapidly every day. Notice that the number of virus signatures in the current ClamAV is more than 37,000 signatures and the average length is more than 67 bytes. Fig. 3 shows an example of the well-known virus, i.e., CodeRed, signatures.

Therefore, in this paper, we propose a fast TCAM-based virus detection algorithm which can support the line rate of 10Gbps for very large virus database and analyze the size of TCAM expected for supporting the current virus signatures in ClamAV.

2 Deep Packet Inspection Using TCAM

The current TCAM of IDT products is capable of fast pattern matching with the performance of 250 Million Searches Per Second (MSPS) which is suitable for not more than 2~3Gbps. In our previous research [6], we have proposed an *M*-times faster pattern matching algorithm using TCAM, where *M* is the TCAM width and it can be up to 72 (576bits) in the current TCAM. Fig. 4 shows two algorithms for Deep Packet Inspection (DPI) one is a naïve algorithm based on sliding window which finds the pattern byte by byte; the other is our *M*-byte jumping window algorithm with all possible patterns.

As shown in Fig. 1, the average length of virus signatures in ClamAV is 67 bytes and the number of signatures is more than 37,000. The total required memory size is more than 2 MB. Since the maximum size of current TCAM products – e.g., the IDT75K product which is used in our development system for DPI – is 1.2Mbytes (exactly, 9Mbits), it is impossible to store the entire signatures in TCAM. For faster DPI, redundant entries require additional memory in the *M*-byte jumping window algorithm.

(a) Sliding window

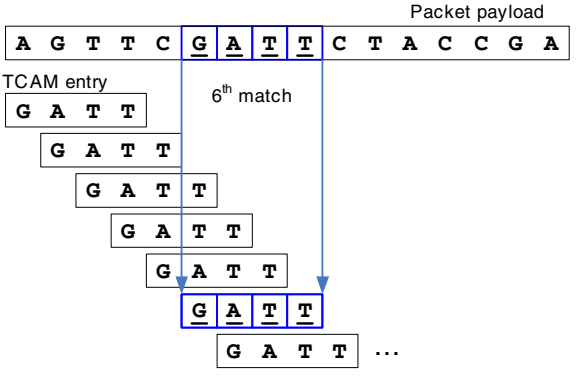


Fig. 4. DPI algorithms using TCAM

(b) *M*-byte jumping window

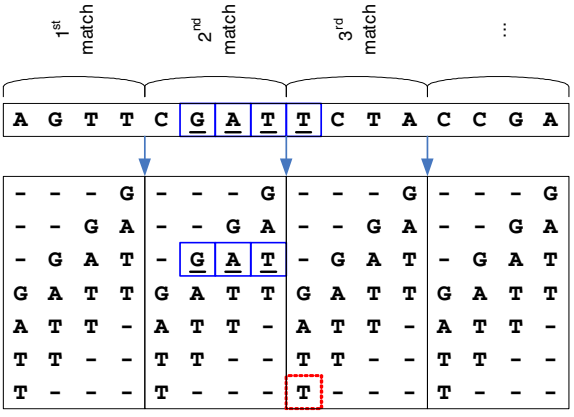


Fig. 4. (*continued*)

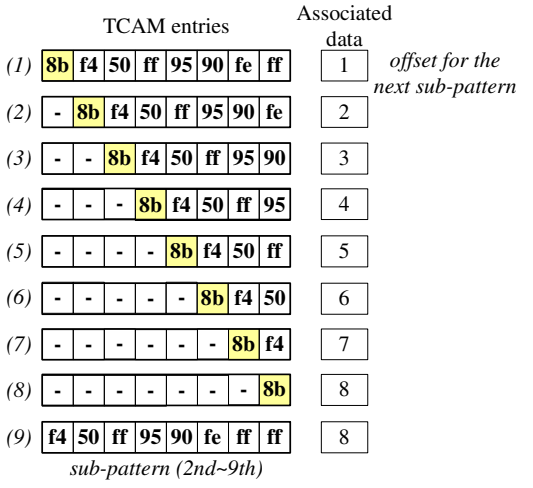
In order to provide faster DPI with the reduced TCAM size, we propose a two-phase content matching algorithm which finds the first *L*-byte signature with TCAM in the first phase and matches remaining bytes of the signature in associated memory, i.e., at the second phase. More details of our fast content inspection algorithm will be explained at the following sections.

3 A Two-Phase Virus Pattern Matching Algorithm with Limited TCAM Size

CodeRed virus signature starts with "`8bf450ff9590feffff`" and it may appear at the arbitrary position in the packet. The sliding window algorithm searches the target pattern byte-by-byte; it significantly decreases the performance of DPI. For our *M*-byte jumping window algorithm, TCAM has to store all possible patterns independent of the position appeared. However, our algorithm can match any pattern within the TCAM window as illustrated in Fig. 5. Excluding the first byte "`0x8b`" in the first 9-byte signature, the last TCAM should match the initial prefix of virus pattern "`8bf450ff9590feffff`." Though the signature length of the CodeRed virus is 109 bytes, the required size of TCAM is only 72 bytes to find the first 1 byte of the total prefix which includes initial prefix and following prefix – e.g., "`0x8b`" and "`0x f450ff9590feffff`." Fig. 5(b) shows the remaining process of pattern matching, in which the remaining pattern is stored in the associated data instead of TCAM. Due to the offset for the next sub-pattern – e.g., following prefix – in the associated data, the following 8-byte prefix will appear at the fixed position.

After the initial prefix has been found, the basic 2-Phase DPI algorithm searches virus patterns *M*-times faster than the naïve DPI algorithm. Yet, a large amount of TCAM for storing the entire virus signatures is required.

(a) CodeRed virus signature representation in TCAM



(b) 2-phase signature detection scheme assuming the initial prefix length is 1

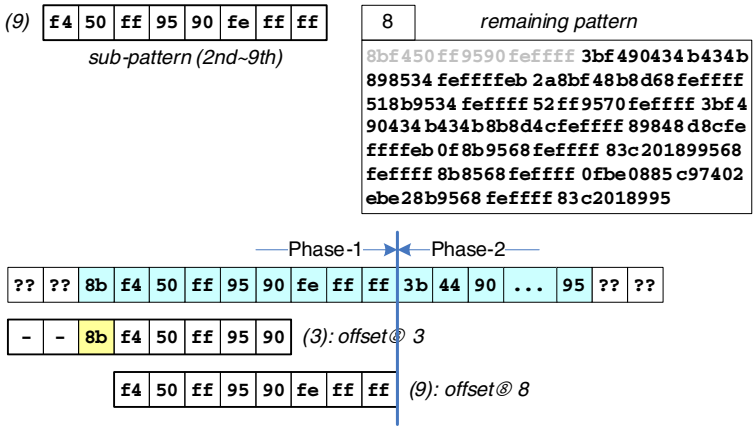


Fig. 5. 2-Phase DIP for virus detection when the TCAM window size (M) is 8 and initial prefix length (L) is 1

One way to further increase the savings is through extending the initial prefix. Instead of initial prefix "0x8b," longer initial prefix, e.g., "8bf450ff95," may be used to find the entire pattern "8bf450ff9590feffff" as shown in Fig. 6. Though the necessary TCAM size becomes small with the increment of L, the signature pattern should appear within the first 4 bytes. Therefore, increasing L will decrease the performance of DPI.

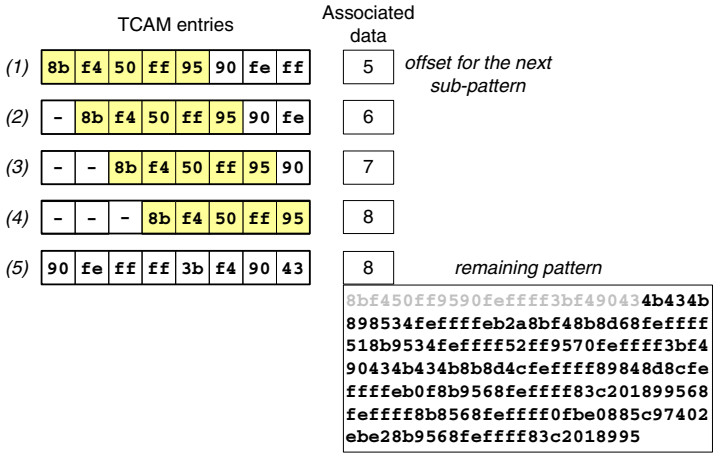


Fig. 6. TCAM representation of virus signature when the initial prefix length (*L*) is 5 and the TCAM window size (*M*) is 8

4 Simulation Results

As mentioned earlier, we have employed the next sub-pattern to be stored in TCAM. Because there are some signatures with the same initial prefix, it makes the remaining pattern match difficult. Fig. 7 shows how many signatures have the prefix which includes initial prefix and following prefix (notice that it shows only top 5 of the signatures sorted). The maximum number of signatures with the same prefix is 42, 31, and 5 when L is 1, 3, and 8, respectively.

Although the TCAM size increases 8 bytes per each signature by the following prefix, the following prefix prevents from increasing duplicated signatures in TCAM. In the worst case, the number of duplicated signatures becomes more than 950,000

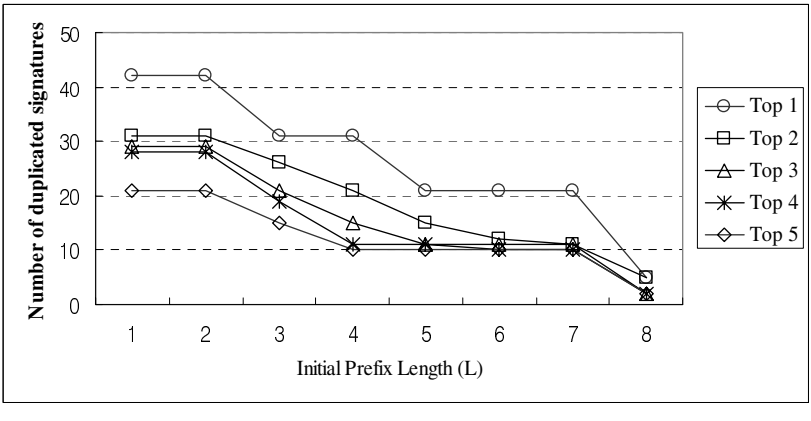


Fig. 7. Top 5 duplicated signatures with the varying the initial prefix length (*L*) when the TCAM window size (*M*) is 8 and total prefix has following prefix

when the following prefix is not used and the initial prefix length is 1. For a 2-Phase DIP, the increasing duplicated signatures should affect more complex processing both at the first phase and at the second phase. However, a very long following prefix increases the TCAM size which needs multiple bytes of *M* per signature.

As the increasing initial prefix length *L*, the size of TCAM decreases. In addition, TCAM window size *M* also has an effect on the TCAM size. Fig. 8 shows the relationship among parameters *L*, *M*, and the TCAM size.

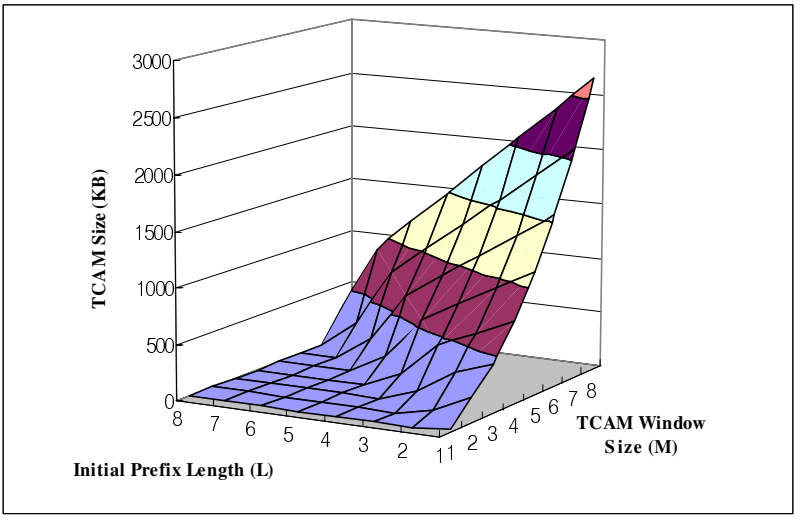


Fig. 8. TCAM size for 37,159 ClamAV signatures according to initial prefix length (*L*) and TCAM window size (*M*), when the total prefix includes the following prefix

Fig. 9 shows the required TCAM performance in MSPS for supporting OC192 – i.e., 9.62 Gbps for total usable bandwidth excluding SONET overhead [8], where 250 MSPS is necessary for the line rate of OC192. The current TCAM with the performance of 250 MSPS supports less than 5 initial prefix lengths in our 2-Phase content inspection algorithm.

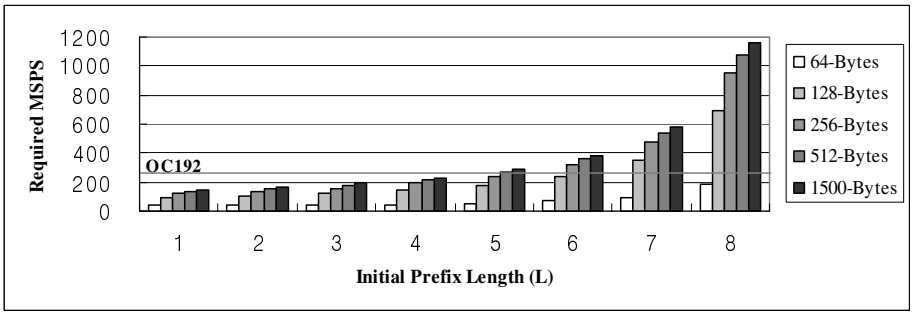


Fig. 9. TCAM performance required for line rate OC192 with varying Initial Prefix Length (*L*) and packet size, when the TCAM window size (*M*) is 8

For the naïve algorithm, the required MSPS is identical to our algorithm when the initial prefix length is 8. Thus, it requires more than 1,000 MSPS for supporting virus detection at OC192. Hence, the naïve algorithm for virus detection with the current TCAM does not keep up with the line rate of the Internet backbone.

5 Conclusion

With TCAM, fast content inspection can be implemented at line rate 10Gbps or OC192. In order to mitigate the limited size of TCAM under a large amount of virus signatures, we have proposed a 2-Phase content inspection algorithm which finds the first L bytes of the signature stored in TCAM. This algorithm decreases TCAM size dramatically, while improving the performance of pattern matching.

From our simulation results, it is shown that a large number of virus signatures can easily fit into the small TCAM of size of 9Mbits. In addition to the reduced TCAM size, our virus detection algorithm can support the line rate of 10Gbps.

References

1. Snort, http://www.snort.org/.
2. Clam Antivirus, http://www.clamav.net/.
3. Liu, H.: Routing Table Compaction in Ternary CAM, IEEE Micro (2002)
4. Ravikumar, V.C., Mahapatra, R.N.: TCAM Architecture for IP Lookup Using Prefix Properties, IEEE Micro (2004)
5. IDT, Network Search Engine (NSE) with QDR™ Interface, http://www1.idt.com/pcms/tempDocs/75K6213452134_DS_80635.pdf
6. Sung, J.S., Kang, S.M., Lee, Y.S., Kwon, T.G., Kim, B.T.: A Multi-gigabit Rate Deep Packet Inspection Algorithm using TCAM, IEEE Globecom (2005, to appear)
7. Yu, F., Kats, R.H., Lakshman, T.V.: Gigabit Rate Packet Pattern-Matching Using TCAM, IEEE International Conference on Network Protocols (2004)
8. Naik, U.R., Chandra, P.R.: Designing High-Performance Networking Applications, Intel Press (2004) 472.

Sampling Distance Analysis of Gigantic Data Mining for Intrusion Detection Systems*

Yong Zeng and Jianfeng Ma

Key Laboratory of Computer Networks and Information Security,
Ministry of EducationXidian University, Xi'an 710071, China
{yongzeng, ejfma}@hotmail.com

Abstract. Real-Time intrusion detection system (IDS) based on traffic analysis is one of the highlighted topics of network security researches. Restricted by computer resources, real-time IDS is computationally infeasible to deal with gigantic operations of data storage and analyzing in real world. As a result, the sampling measurement technique in a high-speed network becomes an important issue in this topic. Sampling distance analysis of gigantic data mining for IDS is shown in this paper. Based on differential equation theory, a quantitative analysis of the effect of IDS on the network traffic is given firstly. Secondly, a minimum delay time of IDS needed to detect some kinds of intrusions is analyzed. Finally, an upper bound of the sampling distance is discussed. Proofs are given to show the efficiency of our approach.

1 Introduction

With the development of the network technology and scale, the network security has already become a global important problem. It is prominent to develop reliable network-based intrusion detection system. Data Mining is a rapidly growing area of research and application that builds on techniques and theories from many fields, including intrusion detection systems [6]. Because the gigantic operations of collection, storing and analysis of every data packages are restricted by computer resources, real world intrusion detection systems (IDS) is not able to deal with the exponential information in high-speed networks in time.

Recent advances showed that some kinds of intrusions can be detected by comparing the network traffic loads with normal ones. Kolmogorov-Smirnov Test was adopted to compare them in [1], while Canonical Correlation Analysis in [2]. The two papers only gave comparing methods of the sampling data and normal data, while how to sample packages from networks was not shown. Authors in [7] gave an empirical measure of processing times of above stages in Snort (a well known network IDS). However, there were no quantitative measures in that paper.

This paper will introduce a quantitative analysis of network traffic loads. Then the affection of IDS on the network traffic will be given based on differential equations.

* This research was partially supported by National Natural Science Foundation of China (No.90204012) and Hi-Tech Research and Development Program of China (No.2002AA143021).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802 , pp. 228 – 235, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Finally, the minimum delay time needed of IDS reacting to some kinds of intrusions is analyzed, which gives a theoretical foundation to analyze the maximum sampling distance of IDS.

2 Related Works

2.1 A Rate Control Model of Internet

Authors in [3] and [4] showed a rate control model of Internet. Consider a network with a set *J* of resources. Let a route *r* be a non-empty subset of *J*, and write *R* for the set of possible routes. Associate a route *r* with a user, and suppose that if a rate $x_r \geq 0$ is allocated to user *r* then this has utility $U_r(x_r(t))$ to the user. Assume that the utility $U_r(x_r(t))$ is an increasing, strictly concave function of x_r over the range $x_r \geq 0$ (following Shenker [5], we call traffic that leads to such a utility function elastic traffic). To simplify the statement of results, assume further that $U_r(x_r(t))$ is continuously differentiable, with $U_r^{'}(x_r) \to \infty$ as $x_r \downarrow 0$ and $U_r^{'}(x_r) \to 0$ as $x_r \uparrow 0$. Suppose that as a resource becomes more heavily loaded the network incurs an increasing cost, perhaps expressed in terms of delay or loss, or in terms of additional resources the network must allocate to assure guarantees on delay and loss: let $C_j(y)$ be the rate at which cost is incurred at resource *j* when the load through it is *y*. Suppose that $C_j(y)$ is differentiable, with

$$\frac{d}{dy} C_j(y) = p_j(y) \tag{1}$$

where the function $p_j(y)$ is assumed throughout to be a positive, continuous, strictly increasing function of *y* bounded above by unity, for $j \in J$. Thus $C_j(y)$ is a strictly convex function.

Next consider the system of differential equations

$$\frac{d}{dt} x_r(t) = k_r \left(w_r(t) - x_r(t) \sum_{j \in r} \mu_j(t) \right) \tag{2}$$

For $r \in R$, where

$$\mu_j(t) = p_j \left(\sum_{s: j \in s} x_s(t) \right) \tag{3}$$

for $j \in J$. We interpret the relations (2)–(3) as follows. We suppose that resource *j* marks a proportion $p_j(y)$ of packets with a feedback signal when the total flow through resource *j* is *y*; and that user *r* views each feedback signal as a congestion indication requiring some reduction in the flow x_r. Then equation (2) corresponds to

a rate control algorithm for user *r* that comprises two components: a steady increase at rate proportional to $w_r(t)$, and a steady decrease at rate proportional to the stream of congestion indication signals received.

2.2 Traffic Model for Network Intrusion Detection

Statistical and dynamic traffic modeling for network intrusion detection were proposed in [1, 2]. [1] discussed that features associated with traffic modeling(volume of traffic in a network, and statistics for the operation of application protocols) are particularly suited for detecting general novel attacks. The motivation for a dynamic modeling of the signals is that, if a baseline traffic model has been identified, if the model is subsequently confronted with traffic data, and if at a certain point in time the model no longer fits the data, some suspicious activity must be on going. By comparing the sampling data with the normal data via canonical correlation analysis and mutual information, [1]'s model and the statistical techniques are utilized to detect UDP flooding attacks. [2] uses the Kolmogorov-Smirnov Test to show that attacks using telnet connections in the DARPR dataset form a population which is statistically different from the normal telnet connections.

3 Effect of IDS on Network Traffic Flow

Consider the variation of traffic of networks with anomaly detection system. Similarly to Section 2.1, define $d(x_r(t))$ as to represent the control of an intrusion detection system to the traffic of *r* then this has utility $D_r(x_r(t))$ to the user. Suppose the time of IDS to sniff, store and analyze traffic loads is τ. Then the judge of IDS to user at time *t* should be deduced according to the data at time $t-\tau$.

Case 1: IDS need a delay time τ to analyze the data sniffed at time $t-\tau$. If they are normal data, IDS does nothing, else IDS will block the intruders.

Following [3, 4], $D_r(x_r(t))$ is differentiable, then we have

$$\begin{aligned} &\frac{\partial}{\partial x_r} D(r,t) = d(x_r(t-\tau)) \\ &\frac{\partial}{\partial t} D(r,t) = d(x_r(t-\tau)) \cdot \frac{d}{dt} x_r(t-\tau) \end{aligned} \tag{4}$$

The anomaly detection system generally blocks the intruders as soon as an intrusion is detected. So we have $d(x_r(t-\tau)) = \begin{cases} 0 & \text{r is not an intruder at time } t-\tau \\ 1 & \text{r is an intruder at time } t-\tau \end{cases}$. Then the variation of traffic of user *r* with networks can be described as (5) and (3):

$$\frac{d}{dt} x_r(t) = k_r \left(w_r(t) - x_r(t) \sum_{j \in r} \mu_j(t) - w_r(t) \cdot d(x_r(t-\tau)) \right) \tag{5}$$

Theorem 1. If $w_r(t) = w_r > 0$ for $r \in R$ then the function

$$U(x) = \sum_{r \in R} w_r \log x_r - \sum_{j \in r} C_j \left(\sum_{s: j \in s} x_s \right) - D(r,t) \tag{6}$$

a) is a Lyapunov function for the system of differential equations (3)–(5). The unique value $x = \frac{w_r}{\sum_{j \in r} \mu_j}$ maximizing $U(x)$ is a stable point of the system, to which all trajectories converge, when there are no intrusions at time $t - \tau$;

b) has unique value $x = 0$ minimizing $U(x)$ for the system of differential equations (3)--(5), when there are intrusions at time $t - \tau$.

Proof: (*Sketch*) $\because \frac{\partial}{\partial x_r} U(x) = \frac{w_r}{x_r} - \sum_{j \in r} p_j \left(\sum_{s: j \in s} x_s \right) - d(x_r(t-\tau))$

and

$$\begin{aligned}
&\frac{d}{dt} U(x(t)) \\
&= \sum_{r \in R} \frac{\partial U}{\partial x_r} \cdot \frac{d}{dt} x_r(t) - d(x_r(t-\tau)) \cdot \frac{d}{dt} x_r(t-\tau) \\
&= \sum_{r \in R} \frac{k_r}{x_r(t)} \left(w_r - x_r(t) \sum_{j \in r} p_j \left(\sum_{s: j \in s} x_s(t) \right) - w_r \cdot d(x_r(t-\tau)) \right) \\
&\quad \times \left(w_r - x_r(t) \sum_{j \in r} p_j \left(\sum_{s: j \in s} x_s(t) \right) - w_r \cdot d(x_r(t)) \right) \\
&\quad - d(x_r(t-2\tau)) \cdot k_r \cdot \left(w_r - x_r(t-2\tau) \sum_{j \in r} p_j \left(\sum_{s: j \in s} x_s(t-\tau) \right) - w_r \cdot d(x_r(t-2\tau)) \right)
\end{aligned} \tag{7}$$

There are no intrusions at time $t - \tau$, so $d(x_r(t-\tau)) = 0$, $d(x_r(t-2\tau)) = 0$, $d(x_r(t)) = 0$ and $U(x) = \sum_{r \in R} w_r \log x_r - \sum_{j \in r} C_j \left(\sum_{s: j \in s} x_s \right)$. According to [3][4] $U(x)$ is strictly concave on the positive orthant with an interior maximum: the maximizing value of x is thus unique, then Further

$$\frac{d}{dt} U(x(t)) = \sum_{r \in R} \frac{k_r}{x_r(t)} \left(w_r - x_r(t) \sum_{j \in r} p_j \left(\sum_{s: j \in s} x_s(t) \right) \right)^2 > 0$$

establishing that $U(x(t))$ is strictly increasing with t, unless $x(t)=x$, the unique x maximizing $U(x)$. Setting above to zero identifies the maximum, we have $x = \frac{w_r}{\sum_{j \in r} \mu_j}$.

The function $U(x)$ is thus a Lyapunov function for the system (3)-(5), so a) Stands.

There are intrusions at time $t-\tau$, so $d(x_r(t-\tau))=1$, $d(x_r(t-2\tau))=0$ $d(x_r(t))=0$, then

$$\frac{d}{dt}U(x(t))$$

$$=\sum_{r\in R}\frac{k_r}{x_r(t)}\left(w_r-x_r(t)\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)-w_r\right)\times\left(w_r-x_r(t)\cdot\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)\right)$$

$$=\sum_{r\in R}\frac{k_r}{x_r(t)}\left(-\left(x_r(t)\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)\right)\times\left(w_r-x_r(t)\cdot\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)\right)\right)$$

$$\because x_r(t)\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)>0, w_r-x_r(t)\sum_{j\in r}p_j\left(\sum_{s:j\in s}x_s(t)\right)>0$$

$$\therefore \frac{d}{dt}U(x(t))<0$$

establishing that $U(x(t))$ is strictly decreasing with t, unless $x(t)=x$, the unique x minimizing $U(x)$. Setting above to zero identifies the minimum, we have $x=0$. So b) stands. ■

Remark 1. Theorem 1.a) tell us that when there are no intrusions in networks or intrusions are not detected at time t, IDS does no affection on network traffic at time $t-\tau$. A user can obtain a maximum and stable service or a maximum traffic flow $x=\frac{w_r}{\sum_{j\in r}\mu_j}$, which has a simple interpretation in terms of a charge per unit flow: the variable μ_j is the shadow price per unit of flow through resource j, and w_r can be viewed as the willingness to pay of user r.

Remark 2. Theorem 1.b) tell us that when intrusions happened at time $t-\tau$ are detected at time t, IDS can affect networks traffic flow. The traffic flow x of intruders will decrease. The shadow price and willingness to pay cannot alter the influence of IDS on networks, but can work on the deceleration of traffic flow.

4 Analysis

The influence of IDS on networks traffic flow is given in Theorem 1. In this section we will discuss the minimum delay time of IDS needed to react to intrusions and the relationship of τ with the sampling distance.

The shadow price $\mu_j(t)=p_j\left(\sum_{s:j\in s}x_s(t)\right)$ can be divided into two parts. One is the charge of user *r*, and the other is that of other users. Suppose the charge varies

with flow at a direct ratio and the charge of other users is δ . Then $\sum_{j\in r}\mu_j(t) = n\cdot\beta_r^{'}\cdot x_r(t)+\delta$, in which $n\cdot\beta_r^{'}\cdot x_r(t)$ is the charge of user *r*. So we have Case 2:

Case 2: (5) can be simplified into $\frac{d}{dt}x_r(t) = k_r\left(w_r - x_r(t)\cdot n\cdot\beta_r^{'}\cdot x_r(t) - x_r(t)\cdot\delta\right)$.

Setting $\lambda_r = n\beta_r^{'}$, notice $x_r(0)=0$, then

$$\begin{cases}\frac{d}{dt}x_r(t)+k_r\lambda_r\cdot x_r^2(t)+k_r\delta\cdot x_r(t)=k_r w_r\\ x_r(0)=0\end{cases} \tag{8}$$

Setting $a=k_r\lambda_r$, $b=k_r\delta$, $c=k_r w_r$, and $\frac{d}{dt}x_r(t)=v$, then we have $a(x_r(t)+\frac{b}{2a})^2 = c+\frac{b^2}{4a}-v$, further $a\cdot u^2 = c+\frac{b^2}{4a}-v$, according to Eq.(8) we have $x_r(t)=u-\frac{b}{2a}$, $\frac{d}{dt}x_r(t)=c+\frac{b^2}{4a}-a\cdot u^2$ and an indefinite integral

$$t=\int\frac{1}{c+\frac{b^2}{4a}-a\cdot u^2}du=\frac{1}{a}\cdot\ln\left|\frac{\frac{\sqrt{b^2+4ac}}{2a}+u}{\frac{\sqrt{b^2+4ac}}{2a}-u}\right|$$

noticing that $x_r(t)=u-\frac{b}{2a}$, then the solution of Eq.(8) follows,

$$x_r(t)=\frac{\frac{\sqrt{b^2+4ac}}{2a}}{1+\frac{2}{e^{at}-1}}-\frac{b}{2a} \tag{9}$$

According to Theorem 1 and Case 2, $x_r=\frac{w_r}{\sum_{j\in r}\mu_j}$, $x_r(\tau)=\frac{\sqrt{4w_r\lambda_r+\delta^2}-\delta}{2\lambda_r}$, and $x_r(0)=0$, so we have

$$\tau=\frac{1}{k_r\lambda_r}\ln\frac{2}{\delta}(\sqrt{\delta^2+4w_r\lambda_r}+\lambda_r-\delta)$$

Noticing that w_r is the willingness to pay per unit flow of user *r*, while $\lambda_r=n\beta_r^{'}$ is the charge per unit flow of user *r*, so w_r is equal to λ_r. Then

$$\tau = \frac{1}{k_r w_r} \ln \frac{2}{\delta} (\sqrt{\delta^2 + 4w_r^2} + w_r - \delta)$$

where k_r is a constant and δ is the charge of flow of other users. Obviously τ decreases when w_r or δ increases.

According to above analysis, we have the following Theorem:

Theorem 2. For the system of differential equations (3)-(5) and Case 1-2, the minimum time needed to finish an intrusion is

$$\tau = \frac{1}{k_r w_r} \ln \frac{2}{\delta} (\sqrt{\delta^2 + 4w_r^2} + w_r - \delta)$$

in which k_r is a constant , w_r is the willingness to pay per unit flow of user r and is the charge of flow of other users. is a decreasing function of w_r and .

Remark 1. It is general that the more a user pays, the higher bandwidth he can obtain. According to Theorem 2 τ is a decreasing function of w_r and w_r is the willingness to pay per unit flow of user r, so we know that the more the willingness to pay per unit flow of user r is, the minimum time of the user r needed to finish an intrusion.

Remark 2. δ is the charge of flow of other users. Generally speaking the more money other users is charged, the less willingness of other users to pay is, as a result the less bandwidth of other users will be. So the networks traffic will be more smoothness and much less blocks will happen, which can avail to decrease the packages losing ratio of user r, thus can decrease the time of the user r needed to finish an intrusion.

Actually we have gotten a bound of sampling distance according to Theorem 2. If the delay time of IDS is longer than τ , an intruder can finish its intrusion before the IDS find it. So the delay time of IDS have to be smaller than τ . If the sampling distance of IDS is longer than τ , it is obviously that the detection rate will fall down. So τ is an upper bound of the sampling distance of IDS.

5 Conclusion

This paper has discussed the effect of IDS on the network traffic, the minimum reacting delay time of IDS to some kind of intrusions, and the upper bound of sampling distance. Differential theory has been employed to develop the model of IDS and network traffic. Proofs are given to show the efficiency of our approach. We leave here future further studies for real-world simulation and lower bound of sampling distance.

References

1. E.Johnckheere, K.Shah and S.Bohacek.: Dynamic Modeling of Internet Traffic for Intrusion Detection, *Proceedings of the American Control Conference* (ACC2002), Anchorage, Alaska, May 08-10, (2002), Session TM06, 2436-2442.

2. Joao B.D.Cabrera, B.Ravichandran and Raman K.Mehra.: Statistical Traffic Modeling for Network Intrusion Detection, *Proceedings of the 8th International Symposium on Modeling*, Analysis and Simulation of Computer and Telecommunication Systems, IEEE press, (2000)466-473
3. Frank Kelly.: Mathematical Modeling of the Internet, Proceedings of the fourth International Congress on Industrial and Applied Mathematics,(1999)685-702
4. F.P.Kelly, A.K.Maulloo and D.K.H.Tan.: Rate Control in Communication Networks: Shadow prices, Proportional Fairness and Stability, *J Opl Res Soc* 46(1998)237-252
5. S.Shenker.: Fundamental issues for the future Internet. *IEEE J. Selected Areas in Commun*. 13(1995)1176-1188.
6. Wenke Lee, Salvatore J. Stolfo, Kui W. Mok.: A Data Mining Framework for Building Intrusion Detection Models, *Proceedings of IEEE Symposium on Security and Privacy*, Oakland, CA, May (1999)
7. Joao B. D. Cabrera, Jaykumar Gosar, Wenke Lee, and Raman K. Mehra.: On the Statistical Distribution of Processing Times in Network Intrusion Detection, Prooceedings of the 43rd IEEE Conference on Decision and Control, Bahamas, December (2004).

Hardware-Software Hybrid Packet Processing for Intrusion Detection Systems

Saraswathi Sachidananda, Srividya Gopalan, and Sridhar Varadarajan

Applied Research Group, Satyam Computer Services Ltd,
Indian Institute of Science, Bangalore, India
{Saraswathi_Sachidananda, Srividya_Gopalan, Sridhar}@Satyam.com

Abstract. Security is a major issue in today's communication networks. Designing Network Intrusion Detection systems (NIDS) calls for high performance circuits in order to keep up with the rising data rates. Offloading software processing to hardware realizations is not an economically viable solution and hence hardware-software based hybrid solutions for the NIDS scenario are discussed in literature. By deploying processing on both hardware and software cores simultaneously by using a novel Intelligent Rule Parsing algorithm, we aim to minimize the number of packets whose waiting time is greater than a predefined threshold. This fairness criterion implicitly ensures in obtaining a higher throughput as depicted by our results.

1 Introduction

With the ever increasing amounts of malicious content within internet traffic, it has become essential to incorporate security features in communication networks. Network security is defined as the protection of networks from unauthorized modification, destruction, or disclosure. Rule checking is one of the popular approaches of Network Intrusion Detection Systems (NIDS). Incoming packets are matched against a rule-set database; the rule-set containing a variety of pre-defined possible malicious contents. The most popular rule-set the Snort rules [1] is an open-source, highly configurable and portable NIDS. Incoming packets are compared with the rules in the Snort database; a match indicates a malicious packet while no match implies that the packet is safe. Some methodologies and implementations of Snort rule checking are based on Content addressable memories [4], regular expression string matching using finite automata [3], and decision tree based rule checking [5]. The most popular string matching algorithms include Boyer Moore [7], Knutt Morris Pratt [8] and Aho Corasick [6].

While incorporating Snort rule checking in IDS systems, it is important to maintain sufficiently high data rates, which can be achieved by reducing processing time. High-performance hardware implementations can be easily implemented using Field Programmable Gate Arrays (FPGAs). While maintaining the performance factor as a design goal, it is equally essential to save logic resources in view of cost and design scalability. At any future date, the Snort rule database can be increased and the system must be capable of incorporating the upgrade with minimum modification. Attig et al [11] has implemented the FPGA based Bloom filter string matching and has shown that 35,475 strings can be processed on the Xilinx VirtexE 2000 FPGA device using 85 %

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 236–243, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of the logic slices, 54 % of look-up tables, and 96 % of BRAMs. With less than 15 % slices and 5 % memory resource, and with gradual increase in the number of rules to be processed, a hardware-software based co-design is a more economical solution. Software execution is advantageous in terms of minimum resources, flexibility and ease. Its main drawback is that the executions are relatively slow compared to the hardware counterparts. The inevitable disadvantage of FPGAs is the consumption of logic resources, which thereby introduces cost expenses.

Attempting to extract the performance gain provided by hardware and the cost and resource utilization advantage offered by software, we propose an algorithm that makes use of a hardware-software hybrid for Snort rule checking. Both the processors work in parallel; both operations being mutually independent. Further, given a hardware-software based co-design, the success and optimized rule processing that reduces the overall packet processing time greatly depends on the optimized utilization of the resources. Addressing the inefficiency issue, we have proposed an Intelligent Rule Parsing (IRP) algorithm which aims at reducing packet processing period by appropriately deploying packets to hardware and software processors, thereby minimizing the number of packets whose waiting time is greater than a predefined system threshold (derived from system resources).

The proposed IRP algorithm is based on the fact that processor intensive tasks require high performance processing and are rightfully deployed to FPGA resources while those rule components estimated as less computationally severe, are sent for software string matching. The IRP algorithm begins with offline computation of hardware and software processing time for rule components. Next is online scheduling of packets based on the calculation of hardware-software processing time for current packets and also the ranking and selection of packets based on predefined system threshold and packet TTL value. Initial results depict the improvement in throughput, thereby implementing the fairness criterion. Section 2 traces the packet flow in the design. Section 3 explains the proposed algorithm in the rule based IDS. Section 4 describes the implementation, while Section 5 is the conclusion.

2 System Overview

Figure 1 depicts an overview of the high-level system containing the various modules and Figure 2 traces the packet flow in the system.

The core of our contribution i.e. the IRP algorithm contains two inter-related processes, offline rule processing and online dynamic scheduling. Offline or design time

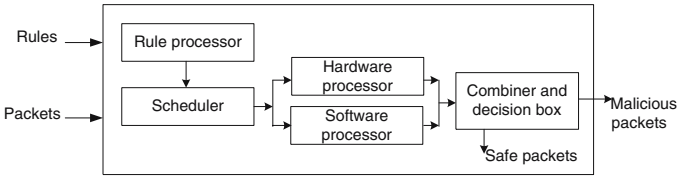


Fig. 1. High-level system overview

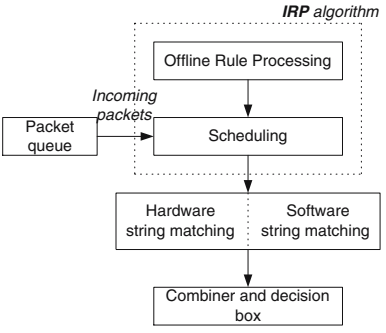


Fig. 2. Packet flow

computation involves calculating the standard parameters that are required for run-time processing. At run-time, the timing parameters for incoming data packets are estimated; based on which the packets are dynamically allocated to hardware or software cores. Data packets are compared against the parsed rules, in parallel in both the processors. While assigning packets to the two processors, it is essential to meet the timing constraint i.e. the total packet processing time, must not exceed the total permissible packet processing period, accounting for the delay associated with the packet queue. To consolidate the results of string matching, a combinational logic and decision box ultimately classify packets as good and bad.

3 Intelligent Rule Parsing (IRP) Algorithm

The proposed algorithm consists of two inter-dependent processes: rule processing and dynamic scheduling is depicted in Figure 3.

Principle: At run-time, the dynamic scheduler allocates the incoming packets to hardware and software string matching based on the output of the offline rule processor which parses the rule and estimates the computation time of each rule component, in both hardware and software processors.

The IRP algorithm is best illustrated by means of an example. Let us consider a single packet *A*, which is to be processed by matching with a single rule *R1*. The rule processor begins by parsing *R1* into its 3 individual components *C1*, *C2* and *C3*. Each component can be processed in either hardware or software processor; the processing time associated with each processing is *Th1*, *Ts1*, *Th2*, *Ts2*, *Th3* and *Ts3* accordingly. The Rule processor generates all possible patterns of *R1* as a combination of *C1*, *C2* and *C3*. There are 2^n possible patterns; each pattern is expressed as *Rk*, where *k:1 to 8*. For a fixed size sample packet, the computation time for software and hardware components of a rule pattern are calculated and tabulated in Tables 1 and 2.

Now, consider the processing of an incoming packet *A*, with packet size *A* bytes and queue delay *Tad*. For every packet, there is a pre-defined total permissible processing time period, denoted as *Ttot*. For the data packet *A*, this value is *Ttota*. Due to the delay *Tad*, the total allowable processing period of packet *A* reduces to,

$$Ta = Ttota - Tad \tag{1}$$

Rule processing: Offline

1. Input a rule *R* to the Rule Processor.
2. Parse the rule *R* into its *Ci* components *i: 1 to N*
3. Generate the 2^N patterns of *Ci, i: 1 to N,* as a combination of hardware and software *Hi* and *Si* components, respectively. Thus, each *Ci* is represented by either *Hi* or *Si.*
4. We now have bb patterns *Rk* with *k : 1 to* 2^N
5. Input a sample packet of known size *M*, say *M=1000 bytes.*
6. For each combination pattern *Rk*, with *k: 1 to* 2^N, compute the total processing time *Tk* for each *Rk* of a sample packet, as a summation of hardware and software processing periods *Thi* and *Tsi* respectively.
   $Tk \quad Thi \quad Tsi$ ... (1)
7. Write to file *File_IRP.*

Dynamic scheduling: Online

1. Read rule - info from file *File_IRP.*
2. Incoming data packet enters the scheduler from the packet queue. Let us consider a single packet *A,* of size *A* bytes and associated with queue delay of *Tad*. Also, input to the scheduler is the total allowable packet processing time, denoted as *Ttot.*
3. Calculate the threshold packet processing time for packet *A.*
   $Ta \quad Ttot \quad Tad$ ... (2)
4. Calculate the relative estimation factor *Tma.*
   $Tma \quad Ta(M)/A$ ... (3)
5. Look into rule-info for a rule pattern *Rk* which has total processing time Tk, which is lower than the actual allowable processing time *Tma* with the constraint that,
   $Max(\min(Tk)) \quad Tma$... (4)
6. Repeat the process for all rules and every new packet entering the queue. Based on the rule processing time and TTL of the packets, select and allocate packets to hardware and software processors.

Fig. 3. IRP algorithm

Table 1. Processing time for every rule component

Rule component Ci	Hardware processing time Thi	Soft processing time Tsi
C1	Th1	Ts1
C2	Th2	Ts2
C3	Th3	Ts3

Table 2. Processing time for every rule pattern

Rule pattern Rk	Processing time period Tk
S1S2S3	T1
S1S2H3	T2
S1H2S3	T3
S1H2H3	T4
H1S2S3	T5
H1S2H3	T6
H1H2S3	T7
H1H2H3	T8

For a data packet *M*, the total processing time can be estimated as,

$$Tma = Ta \times M/A \tag{2}$$

where *A* : size of data packet *A*; *M*: size of sample packet *M*
Tma: Relative estimated processing time for packet *A* with respect to packet *M*
Ta: Actual packet processing time for packet *A*

The value of *Tma* as computed by the scheduler is sent to the rule processor, which searches its rule pattern database for that combination of rule pattern *Rk* which is has total processing time *Tk* lower than total actual allowable processing time *Tma*. The timing constraint imposed is,

$$Max(min(Tk)) \leq Tma \tag{3}$$

The selected rule *Rk*, is sent back to the scheduler. This pattern *Rk* can be used in the deployment of the packet to rule component processing in hardware and software processors. A packet is therefore string matched with the complete rule by selectively deploying rule components to hardware and software processors. The same process is repeated for every rule and matched with every packet.

In summation we can say that, the IRP algorithm aims at minimizing packet processing time by dynamically allocating packets to either hardware or software rule components, based on their estimated computation time. Processor intensive components are associated with hardware, while those rule components which can quickly process packets, are executed in software. Ultimately, the waiting time of any arbitrary packet in the queue is minimized by dynamically dispatching it to either one resource at run-time. The scheduler has yet another important task related to aborting redundant matching. A packet is classified as good, only when it matches all the rules. Hence if a single rule is not matched, the packet cannot be malicious. The scheduler immediately interrupts the co-processor, and aborts the processing of this packet, by removing it from the queue.

4 Implementation

The IRP algorithm is tested for its performance by making the following comparisons. Firstly, we compare the packet processing periods in pure hardware, pure software and hybrid hardware-software string matching implementations. For initial testing, we represent hardware string matching using the better performing Aho Corasick (AC) algorithm [6] and software string matching using the lower performing Wu Manber (WM) algorithm [1]. Second, we compare hybrid hardware-software implementation with respect to allocation in a First Hardware-Next Software (FHNS) approach and allocation accordingly to the proposed IRP algorithm. For initial testing, we compare both the scenarios on the software platform. Table 3 contains sample values of per packet processing time (μs); wherein 10 packets are each matched with 10 rules. String matching is performed on single processors using only AC and WM algorithms. Hybrid hardware-software implementation is represented by the FHNS and IRP algorithms.

The graphs in Figures 4 and 5 depict processing time and cost per processing, for AC, WM and IRP algorithms. We observe that the software implementation takes maximum processing time while the hardware-only and hardware-software show considerable increase in the packet processing rate. From the average packet processing time

Table 3. Packet processing time (μs) for AC, WM, IRP and FHNS algorithms

Packet number	AC	WM	IRP	FHNS
1	6	11	7	7
2	8	12	7	7
3	7	14	7	9
4	9	15	7	9
5	13	23	13	14
6	17	24	14	15
7	16	27	15	17
8	19	31	16	19
9	22	35	20	23
10	25	38	21	24

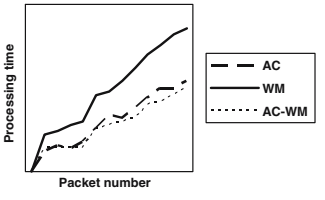


Fig. 4. Packet processing time(μs)

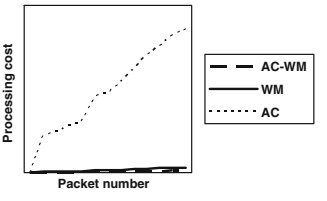


Fig. 5. Packet processing cost

periods, we calculate that the performance gain obtained from software to hardware-software hybrid is 44.28 % and hardware to hardware-software hybrid is 9.8 %. We also note that although hardware solutions provide higher performances, hardware-software hybrid design provides almost close performance at considerably lesser cost expenses.

Next we consider the hardware-software implementation and compare the FHNS approach with the proposed IRP distribution. We compare the IRP and FHNS algorithms for different number of hardware and software processors. We assume 4 parallel processes on the FPGA and 4 parallel threads on the software. We compare the FHNS and IRP, for cases when the hardware:software processor ratios is 3:1, 1:3 and 2:2. Figure 6 considers 2 simultaneous co-processors, one hardware and one software. Figures 7, 8 and 9 represent the per packet processing time (μs) for FHNS and IRP when the hardware-software processor ratio is 2:2, 3:1 and 1:3.

The permissible packet processing period is represented by the *delta* threshold line. In Figures 6, 7, 8 and 9, we observe that fewer packets fall within the *delta* threshold when using the FHNS algorithm, while more packets fall within the limit using IRP algorithm. The performance gain achieved in moving from FHNS to IRP was noted to be 12.64 %. These results indicate that the IRP approach performs best in comparsion to either only hardware or only software processing, and also better than the first hardware next software approach. Hardware-software hybrid implementations therefore, not only bring in area and performance advantages, but provide higher performances when the processes are rightfully deployed to FPGA and software processors.

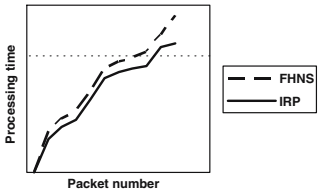


Fig. 6. Processing time for FHNS and IRP algorithms

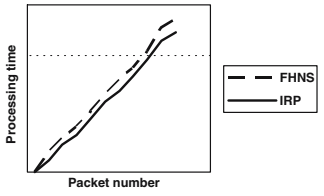


Fig. 7. Processing time for processor ratio 2:2

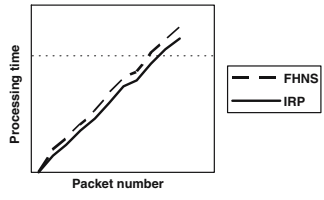


Fig. 8. Processing time for processor ratio 3:1

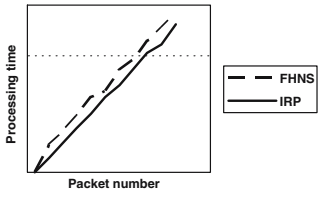


Fig. 9. Processing time for processor ratio 1:3

For the actual implementation, the design is translated to hardware using the VHDL description language. The design is implemented using the Xilinx Synthesis tool Xilinx ISE 7.0 [10] and synthesized for the Virtex FPGA device [9]. The Virtex-II Pro Platform FPGA [9] allows for integrated solutions and can be used to implement both the software and hardware functions on a single platform. Synthesis results can be used to estimate area, performance and power characteristics of the design. Area utilization can be studied with respect to total gate count and percentage of used FPGA resources. Percentage utilization indicates the amount of resources required to implement the design and hence the resources available for scalability of the design. Area synthesis also gives an estimate of the actual chip area when the design is acquired to full-custom ASIC. The design can be optimized to obtain higher operating frequencies. Ongoing work includes area and frequency synthesis in order to estimate area and frequency trade offs.

5 Conclusion and Future Work

The main purpose of offloading functions to hardware is to gain performance. At the same time FPGAs and ASICs are expensive and area consuming when compared to software. Division of functions between hardware and software is thus a trade-off between performance and logic resources. Greater complexity of the function or required speed of operation will bias the design towards software or hardware. This work deals with the development of an algorithm for the rightful and dynamic deployment of packets to hardware and software processors for rule checking in an IDS. A possible avenue for future work is to prioritize the rule components, so that the important ones would get processed first. Another option is to schedule the packets in such a way to maximize the number of packets to get processed within the threshold limit.

The authors thank Prashant Kumar Agrawal of Applied Research Group, Satyam Computer Services Ltd, for analyzing the Snort open source and performing the implementations for the algorithms. We also thank him for the useful ideas and feedback.

References

1. Snort The Open Source Network Intrusion Detection System.: http://www.snort.org
2. M. Fisk, G. Varghese, I.: An analysis of fast string matching applied to content-based forwarding and intrusion detection. In Techical Report CS2001-0670 (updated version), University of California, San Diego 2002
3. R. Sidhu, V.K Prasanna, I.: Fast regular expression matching using FPGAs. In IEEE Symposium on Field-Programmable Custom Computing Machines, CA, USA, April 2001
4. Janardhan Singaraju, Long Bu, John A. Chandy,.: A Signature Match Processor Architecture for Network Intrusion Detection. In FCCM 2005
5. William N. N. Hung, Xiaoyu Song,.: BDD Variable Ordering By Scatter Search. In Proceedings of the International Conference on Computer Design: VLSI in Computers and Processors, ICCD 2001
6. A. Aho, M. Corasick,.: Efficient string matching: An aid to bibliographic search. In Communications of the ACM, vol. 18, **6** (1975) 333–343
7. R. S. Boyer, J. S. Moore,.: A fast string searching algorithm. In Communications of the ACM, vol. 20, **10** (1977) 762–772
8. D. Knuth, J. Morris,V. Pratt,.: Fast pattern matching in strings, In SIAM Journal on Computing, vol. 6, **2** (1977) 323–350
9. Virtex-II Pro and Virtex-II Pro X Platform FPGAs,.:Complete Data Sheet, (v4.3), 2005
10. Xilinx ISE 7.0 In-Depth Tutorial, version 2005
11. Michael Attig, Sarang Dharmapurikar, John Lockwood,.:Implementation Results of Bloom Filters for String Matching, In Proceedings of FCCM 2004

D-S Evidence Theory and Its Data Fusion Application in Intrusion Detection

Junfeng Tian, Weidong Zhao, and Ruizhong Du

Faculty of Mathematics and Computer Science, Hebei University,
Baoding 071002, China
zwdeasenet@163.com

Abstract. Traditional Intrusion Detection System (IDS) focus on low-level attacks or anomalies, and too many alerts are produced in practical application. Based on the D-S Evidence Theory and its data fusion technology, a novel detection data fusion model-IDSDFM is presented. By correlating and merging alerts of different types of IDSs, a set of alerts can be partitioned into different alert tracks such that the alerts in the same alert track may correspond to the same attack. On the base of it, the current security situation of network is estimated by applying the D-S Evidence Theory, and some IDSs in the network are dynamically adjusted to strengthen the detection of the data which relate to the attack attempts. Consequently, the false positive rate and the false negative rate are effectively reduced, and the detection efficiency of IDS is improved.

1 Introduction

The current Intrusion Detection System (IDS) products can be divided into two major categories: Network-based IDS and Host-based IDS. But there are some problems in practical application, for example the false positive rate and the false negative rate are high. To improve the detection accuracy of IDS, some scholars try to apply data fusion technology to IDS in recent years. Tim Bass discusses the multi-sensor data fusion technology for next generation distributed intrusion detection systems [1][2]. The MIRADOR project [3] defines the similarity function which is used for calculating the similarity with two alerts by using expert system method, then merges them, but the performance of real time is low. Burroughs and Wilson consider the intrusion behaviors from an attacker-centered viewpoint, and present a new approach [4] that can track and identify the attacker by using bayesian method. But the false rate reaches about 20%. Moreover, The research project EMERALD [5] of SRI International merges alerts of different types of IDSs by using bayesian methods. But in practical application, it is very difficult to set meta alerts accurately.

The research of applying data fusion technology to IDS in our country is seldom concerned. How to apply data fusion technology to IDS are discussed and some data fusion models are given in the reference 6, 7 and 8. But the concrete fusion methods are rarely presented. Based on the D-S Evidence Theory and its data fusion technology, a novel detection data fusion model-IDSDFM is presented in this paper.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 244 – 251, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Data Fusion Model-IDSDFM

To protect the network security against the attack effectively, we also need to gather quantity of suspicious and intrusive information from Network-based IDS, Host-based IDS and other security products. In order to process them in uniform format, we suppose that alerts processed by IDSDFM accord with the alert instances that were defined in IDMEF [9]. IDSDFM consists of alert correlation module, security situation estimation module and management & control module (see Fig.1).

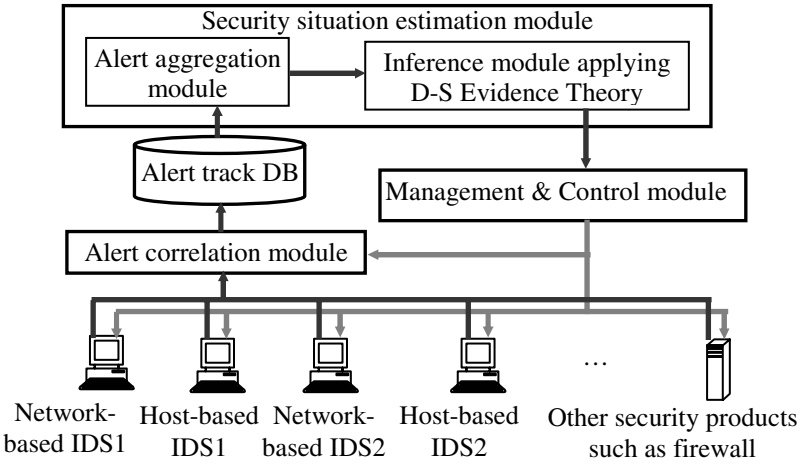


Fig. 1. The logical diagram of IDSDFM

Define1: Alert track describes a set of alerts that relate to an attack event.

By correlating and merging the alerts of the different types of IDSs or other security products, the alert track database possibly has many alert tracks. On the base of it, we can estimate the current security situation by applying the D-S Evidence Theory, and then identify the attack classifications or attack attempts. All these information are conveyed to the management & control module, and some IDSs in the network are adjusted dynamically. On the other hand, according to the current security situation, the weights of some alert attributes can be modified while computing the dissimilarity. As a result, the false positive rate and the false negative rate are effectively reduced, and the detection efficiency of IDS is improved.

2.1 Alert Correlation Module

When an attack is occurring in the network, a great deal of alerts may be produced by the different types of IDSs or other security products. This module processes each alert corresponding to these attack events and there are exactly two possibilities with regard to the alerts produced:

Rule1: The event also triggered other alerts that already belong to an alert track.
Rule2: None of the existing alert tracks is the result of the event that triggered the alert.

These two possibilities are shown in Fig.2. In this figure, Host-based IDS and Network-based IDS1 detect attack event 1 and event 3. Network-based IDS2 detects event 1 and event 2.

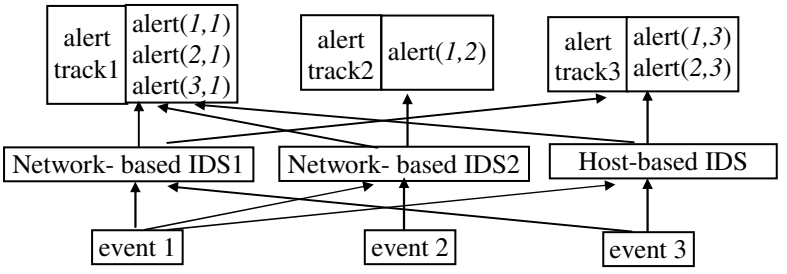


Fig. 2. The schematic diagram of the alert correlation

Define2: Alert (*x,y*) describes that the alert is triggered by an attack event *y*, which number is *x*.

So, alert(*1,1*), alert(*2,1*) and alert(*3,1*) are the result of attack event 1, which number is 1, 2 and 3. Hence they are correlation alerts. Alert(*1,2*) is the only alert triggered by attack event 2. Alert(*1,3*) and alert(*2,3*) are the result of attack event 3, which number is 1 and 2. Hence they are correlation alerts.

Define3: Dissimilarity describes the different degree between two alerts.

The more similar alert *i* and alert *j* are, the more their dissimilarity is close to 0. While a new alert coming, this module searches the alert track database and calculates the dissimilarity with the new alert and each alert of database. The most likely correlation is the pair of alerts with the lowest dissimilarity. When the dissimilarity is low enough, this module decides to assign the new alert to the alert track which the existing alerts belong to. Or we obtain a new alert track, when every dissimilarity which is calculated with the new alert and each alert of database, is higher than the maximum dissimilarity (an expert value). This new alert is considered to be triggered by a new attack event, and there is not an alert track that relate to this attack event.

These attributes of the alert can be described as *p* attributes such as Source IP address, Destination IP address, Protocol type and Create time etc. The alert has some different types of attributes such as numerical variable, boolean variable and enumerated variable. The method of calculating the dissimilarity of different types of attributes is introduced [10] and then the total calculating formula is given.

2.1.1 The Method of Calculating the Dissimilarity of Different Types of Attributes

(1) Numerical variable

For the numerical variable we can adopt Euclidean Distance to describe the different degree of two numerical variables. Here, we suppose two alert sets: (x_{i1}, x_{i2} ,…, x_{ip}) and (x_{j1}, x_{j2}, …, x_{jp}). These alerts are *p*-dimensional variable, and their calculating dissimilarity method is given as formula (1).

$$d(i,j) = \sqrt{w_1 \mid x_{i1} - x_{j1} \mid^2 + w_2 \mid x_{i2} - x_{j2} \mid^2 + \ldots + w_p \mid x_{ip} - x_{jp} \mid^2} \tag{1}$$

Where w_i is the weight of attribute *i* of an alert. x_{ip} is the attribute *p* of the alert *i*.

(2) Boolean variable
We adopt the famous brief match coefficient method to calculate the dissimilarity, and the calculating method is given as formula (2).

$$d(i, j) = \frac{r + s}{q + r + s + t} \tag{2}$$

Where q is the number of the corresponding attribute values which the alert i and alert j are equal to "1". t is the number of the corresponding attribute values which the alert i and alert j are equal to "0". r is the number of the corresponding attribute values which the alert i is "1"and alert j is "0". s is the number of the corresponding attribute values which the alert i is "0"and alert j is "1".

(3) Enumerated variable
Enumerated variable has many values. We may also use the famous brief match coefficient method. The calculating method is given as formula (3).

$$d(i, j) = \frac{p - m}{p} \tag{3}$$

Where m is the number of corresponding attributes which alert i and alert j have the same value. p is the number of all enumerated variables of alert i and alert j.

2.1.2 The Method of the Total Calculating the Dissimilarity

We suppose that alert i and alert j include p different types of attributes. Then the $d(i,j)$ which describes the dissimilarity between alert i and alert j, is defined as formula (4).

$$d(i, j) = \frac{\sum_{f=1}^{p} w^{(f)} d_{ij}^{(f)}}{p} \tag{4}$$

Where p is all the attribute number of alert i and alert j. f is one of the p attributes. $w^{(f)}$ is the weight of attribute f while calculating the dissimilarity. $d_{ij}^{(f)}$ is the dissimilarity of attribute f between alert i and alert j.

2.2 Security Situation Estimation Module

2.2.1 Alert Aggregation Module

After the correlation module processing, an alert track possibly has many alerts which are produced by the different types of IDSs. There are mainly two types of relations. In other words, two types of alert aggregations:

1. Alerts that together make up an attack
2. Alerts that together represent the behavior of a single attacker

These are different types of aggregations, because a single attacker can be involved in multiple attacks and multiple attackers can be involved in a single attack. Why we

did see this becomes clear when the relationships between attackers and attacks are analyzed. This shows which attackers are involved in which attack.

2.2.2 D-S Evidence Theory Introduce [11]

(1) Frame of discernment

The frame of discernment (FOD) Θ consists of all hypotheses for which the information sources can provide the evidence. This set is finite and consists of mutually exclusive propositions that span the hypotheses space.

(2) Basic probability assignment

A basic probability assignment (*bpa*) over a FOD Θ is a function m: $2^{\Theta} \rightarrow$ [0,1] such that

$$\begin{cases} m(\phi) = 0 \\ \sum_{A \subseteq \Theta} m(A) = 1 \end{cases}$$

The elements of 2^{Θ} associated to non-zero values of *m* are called focal elements and their union core.

(3) Belief function and plausibility function

The belief function given by the basic probability assignment is defined as:

$$\mathrm{Bel(A)} = \sum_{B \subseteq A} m(B)$$

The value *Bel*(*A*) quantifies the strength of the belief that event *A* occurs.

(4) Dempster's rule of combination

Given several belief functions on the same frame of discernment based on the different evidences, if they are not entirely conflict, we can calculate a belief function using Dempster's rule of combination. It is called the orthogonal sum of the several belief functions.

The orthogonal sum $Bel_1 \oplus Bel_2$ of two belief functions is a function whose focal elements are all the possible intersections between the combining focal elements and whose *bpa* is given by

$$m(A) = \begin{cases} 0 & when \quad A \neq \phi \\ \dfrac{\sum_{A_i B_j = A} m_1(A_i) m_2(B_j)}{1 - \sum_{A_i B_j = \phi} m_1(A_i) m_2(B_j)} & when \quad A = \phi \end{cases}$$

2.2.3 Inference Module Applying the D-S Evidence Theory

(1) Intrusion detection data fusion based on the D-S Evidence Theory

In the process of IDS data fusion, the targets are propositions that are all the possible estimation of the current security situation (where and when the attack happen). The alerts produced by each IDS result in the measure of the security

situation, which constitute the evidences. Fig.3 illustrates how to estimate the intrusion situation by using the D-S Evidence Theory. In the figure, $m_{s1}(A_j)$, $m_{s2}(A_j)$, …, $m_{sk}(A_j)$, (s=1,2,…,n; j=1,2,…,m)is the *bpa* the s*th* IDS assigns to proposition A_j in the i*th* (i=1, 2, …, k) detection cycle, $m_1(A_j)$, $m_2(A_j)$,…, $m_k(A_j)$ is the conjunctive *bpa* calculated from the aggregation of the n *bpa*s in each of the k detection cycles by using Dempster's rule and $m(A_j)$ is the conjunctive *bpa* calculated from the k *bpa*s.

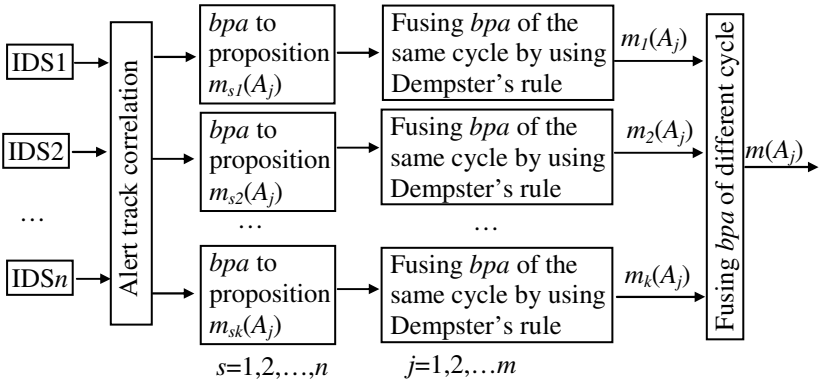


Fig. 3. Data fusion model based on the D-S Evidence Theory

(2) The course of fusion applying the D-S Evidence Theory

In IDSDFM, the fusion process like this: at first, we ought to determinate the FOD, consider all possible kinds of result and list all possible propositions. Then the total conjunctive *bpa* will be calculated by using the Dempster's rule of combination, based on the *bpa* of each proposition obtained according to the evidences that the IDSs provide in the cycle. At last, we get the belief function and inference by a certain rule from these combined results, and estimate the current security situation.

2.3 Management and Control Module

After processing of the security situation estimation module, we can judge the attempts and the threat levels of the attacks, and whether the new attack events happen. A report to the network administrator will be formed. The management & control module adjusts some IDSs of the network dynamically to strengthen the detection of the data which relate to the attack, and add the new attack rules to the IDS feature database. As a result, the detection quality would be improved. On the other hand, according to the current security situation, the weight of some alert attributes can be adjusted by this module so that we pay more attention to some specific alerts.

3 Experiments and Analysis

We configure the LAN with Snort and E-Trust, and attack one of computers in the LAN. IDSs produce the 111 alerts which are obtained as follows: 85 for Snort and 26 for E-Trust. We uniform the format of these alerts and calculate the dissimilarities with them, then get the alert track according to the calculating result (see Table 1).

Table 1. Alert track database after correlation

Alert track	E-Trust	Snort	Description in detail
1	11	31	Port-scan
2	3	18	ICMP Flood
3	4	8	Scan UPNP service discover attempt
4	3	8	WEB-PHP content-disposition
5	4	14	Udp-flood
6	0	4	Web-Iss attack attempt
7	1	2	RPC sadmind Udp ping

At this time, we get the intrusion situation which possibly has 7 attacks as follows: Port-scan, Udp-flood, Scan UPNP service discover attempt, WEB-PHP content-disposition, ICMP Flood, Web-Iss attack attempt, and RPC sadmind Udp ping. The value *m* of *bpa* that the two IDSs assign to the proposition is shown as follow:

Table 2. *m* of *bpa* is assigned by two IDS

Alert track	1	2	3	4	5	6	7
m of *bpa* assigned by Snort	0.268	0.038	0.392	0.029	0.201	0.029	0.043
m of *bpa* assigned by E-Trust	0.341	0.061	0.182	0.036	0.340	0.024	0.016

The total conjunctive *bpa* is calculated by applying the D-S Evidence Theory.

Table 3. The fusion result

Propositions	*m* of total conjunctive *bpa*
Port-scan	0.354
ICMP Flood	0.014
Scan UPNP services discover attempt	0.338
WEB-PHP content-disposition	0.007
Udp-flood	0.278
Web-Iss attack attempt	0.003
RPC sadmind Udp ping	0.006

From the result we can find out the belief of the Port-scan, Scan UPNP service discover attempt and Udp-flood are much greater than that of ICMP Flood, WEB-PHP content-disposition, Web-Iss attack attempt and RPC sadmind Udp ping. So we

can confirm that attackers are attacking the network by Port-scan, Scan UPNP services discover attempt and Udp-flood.

4 Conclusions

Traditional intrusion detection systems (IDS) focus on low-level attacks or anomalies, and too many alerts are produced in practical application. In situations where there are intensive intrusions, not only will actual alerts be mixed with false alerts, but the amount of alerts will also become unmanageable. The D-S Evidence Theory and its data fusion technology is an effective solution to these problems. A novel fusion model-IDSDFM is presented in this paper. It merges these alerts of different IDSs and makes intelligent inference according to those fusion results. Consequently, the number of the raw alerts decreases effectively. The false positive rate and the false negative rate are reduced. But how to identify the very similar alerts which have no logical relation (the similar alerts describe the different attack events) and the dissimilar alerts which have logical relation (the different alerts describe the same attack event) will be studied in our further work.

References

1. Tim Bass.: Intrusion Detection Systems and Multisensor Data Fusion. Communications of the ACM, Vol.43, April (2000) 99-105.
2. Tim Bass, Silk Road.: Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems. IRIS National Symposium Draft (1999) 24-27.
3. Frédéric Cuppens.: Managing Alerts in a Multi-Intrusion Detection Environment. In Proceedings of the 17th Annual Computer Security Applications Conference, December (2001) 22-32.
4. Daniel J. Burroughs, Linda F. Wilson, and George V.: Cybenko. Analysis of Distributed Intrusion Detection Systems Using Bayesian Methods, in Proceedings of IEEE International Performance Computing and Communication Conference, April (2002).
5. Alfonso Valdes, Keith Skinner.: Probabilistic Alert Correlation. In Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (2001) 54-68.
6. Jiaqing Bao, Xianghe Li, Hua Xue.: Intelligent Intrusion Detection Technology. Computer Engineering and Application. (2003) Vol.29 133-135.
7. Jianguo Jiang, Xiaolan Fan.: Cyber IDS-New Generation Intrusion Detection System. Computer Engineering and Application. (2003) Vol.19 176-179.
8. Guangchun Luo, Xianliang Lu, Jun Zhang, Jiong Li.: A Novel IDS Mechanism Based by Data Fusion with Multiple Sensors. Journal of University of Science and Technology of China (2004) Vol. 33 71-74.
9. Intrusion Detection Working Group.: Intrusion detection message exchange format data model and extensible markup language (XML) document type definition. Internet-Draft, Jan (2003) 21-26.
10. Ming Zhu. Data Mining. Hefei: University of Science and Technology of China Press. (2002) 132-136.
11. Jinhui Lan, Baohua Ma, Tian Lan etc.: D-S evidence reasoning and its data fusion application in target recognition. Journal of Tsinghua University (2001) Vol. 41 53-55.

A New Network Anomaly Detection Technique Based on Per-Flow and Per-Service Statistics

Yuji Waizumi[1], Daisuke Kudo[2], Nei Kato[1], and Yoshiaki Nemoto[1]

[1] Graduate School of Information Sciences(GSIS), Tohoku University, Aza Aoba 6-6-05, Aobaku, Sendai, Miyagi 980-8579, Japan
wai@nemoto.ecei.tohoku.ac.jp,
[2] DAI NIPPON PRINTING CO., LTD., 1-1, Ichigaya Kagacho 1-chome, Shinjuku-ku, Tokyo 162-8001, Japan

Abstract. In the present network security management, improvements in the performances of Intrusion Detection Systems(IDSs) are strongly desired. In this paper, we propose a network anomaly detection technique which can learn a state of network traffic based on per-flow and per-service statistics. These statistics consist of service request frequency, characteristics of a flow and code histogram of payloads. In this technique, we achieve an effective definition of the network state by observing the network traffic according to service. Moreover, we conduct a set of experiments to evaluate the performance of the proposed scheme and compare with those of other techniques.

1 Introduction

Today the Internet is expanding to a global scale and serves as a social and economical base. Along with its growth, network crimes and illegal accesses are on the rise. Network Intrusion Detection Systems (NIDSs) have been researched for defending networks against these crimes. The two most common detection techniques adopted by NIDSs are signature based detection and anomaly based detection. The signature based detection technique looks for characteristics of known attacks. Although this technique can precisely detect illegal accesses which are contained in the signature database, it can not detect novel attacks.

The anomaly detection technique which adopts the normal state of the network traffic as a criteria of anomaly, can detect unknown attacks without installation of new database for the attacks. However, it makes a lot of detection errors because of the difficulty to define the normal state of the network traffic precisely. In order to tackle the problem of the detection error, many researches have recently been done in the anomaly detection field [1] [2] [5] [6]. These schemes can detect the definite anomaly traffic because they pay attention to only the packet header fields such as IP address, port number, TCP session state. NETAD[6] has achieved the highest detection accuracy among conventional methods by examination using a data set[4]. However, NETAD might not be a practical system because it detects attacks based on a white list of IP addresses which are observed in learning phase. Thus, NETAD regards an IP address which is not

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 252–259, 2005.
© Springer-Verlag Berlin Heidelberg 2005

included in the white list as anomaly. It cannot detect attacks which are camed out from hosts whose IP addresses are included in the white list.

In this paper, we believe that more detailed traffic information without IP addresses is necessary to build an advanced anomaly detection system and propose a new anomaly detection technique based on per-flow and per-service statistics. The proposed method consists of three statistical models and an alert generation part. The statistical models are defined from three different viewpoints using numerical values, which represent the state of network traffics, extracted from network flows based on Realtime Traffic Flow Measurement (RTFM) which is defined by RFC 2721[3]. The theree statistical models are 1) Service request frequency, 2) Characteristics of packets of a flow unit and 3) Histogram of character code of payload.

The alert generation part employs two-stage distinction algorithm. The algorithm has *suspicious* stage and *danger* stage in order to control incorrect detections and detects anomalies for each flow.

Following this introduction, this paper is organized as follows. Section 2 describes details of the statistical models and distinction algorithm. Section 3 evaluates the detection performance of the proposed method using a data set. Section 4 discusses about the proposed statistical models. Section 5 concludes this paper.

2 System Architecture

Our proposed system is shown in Fig.1. The system consists of three statistical model construction parts and an anomaly discrimination part. The model construction parts extract the state of network traffic as numerical values based on service request frequency, characteristics of a flow unit and histogram of the codes included in the payload of flows. The anomaly discrimination part integrates the three degrees of anomalies, which are calculated from the above three statistics, as a Anomaly Score and generates an alert based on the integrated value.

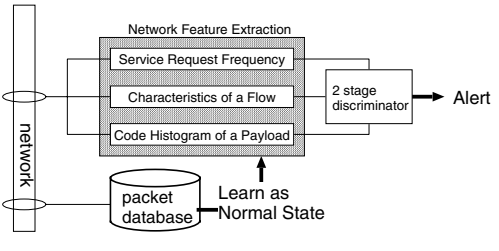


Fig. 1. Proposed System

2.1 Traffic Observation Unit

To extract the characteristics of a flow unit and the histogram of the codes of the payload, we adopt RTFM[3] as an observation method. We adopt a complete 5-tuple (protocol, source and destination IP addresses, source and destination

port numbers). There are two reasons for proposing observation by a flow unit. First, we can extract information about network endpoints independently of time zone. Second, it is possible to acquire quickly the information about the host that causes the network anomaly.

2.2 Statistical Models

Service Request Frequency to a Host in LAN. In a managed network, a host that offers a certain service is considered to experience a slight change. For example, services requested from external networks to a client host are rare and unusual. In order to detect the rare requests, the request frequencies of each services to each hosts are measured.

N denotes the total number of requests of a certain service issued over the whole local area network during a given period of time. A_X is the number of requests of the service to a specific host X. The probability $P(X)$ that host X will operate as a server of a specific service, is computed by (1). This $P(X)$ is the model of a normal state.

$$P(X) = \frac{A_X}{N} \tag{1}$$

In order to quantitatively express the degree of the anomaly, the amount of information is used. If an event rarely occurs, its amount of information gets large value. Accordingly, the degree of anomaly can be evaluated using the probability $P(X)$ given by the model. For a certain service, the degree of anomaly H when host X functions as a server host is given by (2):

$$H = -logP(X) \tag{2}$$

Characteristics of a Flow Unit. To define a state of packets of a flow, 13 different characteristics of flows, which are shown in Table 1, are extracted.

Table 1. The Characteristics of a Flow

01.flow duration time	02.average of packet interval time
03.standard deviation of packet interval time	
04.variation coefficient of packet interval time	
05.total number of packets	06.number of received packets
07.number of sent packets	08.total number of bytes
09.number of received bytes	10.number of sent bytes
11.number of bytes per packet	12.number of received bytes per packet
13.number of sent bytes per packet	

The value Y_i of each characteristic $i(i = 1, 2, ..., 13)$ can undergo a big change. Therefore, we define the probability distribution of a characteristic as the logarithm value of the compensation term of Y_i as is shown in (3):

$$Y_i{}' = \ln(Y_i + 0.00001) \tag{3}$$

Since it can be considered that $Y_i{}'$ values of normal flows center near the average value, we evaluate the degree of anomaly of a flow using the deviation from the average value in formula (4). In every characteristics, Z_i expresses the degree of anomaly as its difference from 0.

$$Z_i = \left| \frac{Y_i{}' - \mu_i}{\sigma_i} \right| \tag{4}$$

where μ_i and ${\sigma_i}^2$ are the average and variance of $Y_i{}'$ respectively, and are calculated from the packet database in Fig.1.

The degree of anomaly, C, of a flow is expressed as the average value of the amount of information of characteristics as in (5).

$$C = \frac{1}{13} \sum_{i=1}^{13} [-\log P_i] \tag{5}$$

where

$$P_i = \Phi(t \geq Z_i) = \int_{Z_i}^{\infty} \frac{1}{\sqrt{2\pi}} exp[-\frac{t^2}{2}]dt \tag{6}$$

The Histogram of the Code of a Payload. Some attacks in accordance with normal protocol show anomalies only in their payloads. Therefore, it is considered that anomaly detection technique should have a statistical model based on the code sequence of the payload of packets.

The payload is regarded as a set of codes and it consists of 1 byte of characters ($2^8 = 256$ types). The appearance probability of each code is used for constructing a normal model using the packet database in Fig.1. The histogram of the codes which has 256 classes for each service are built. Then, this histogram is sorted according to the appearance probability of each code. By comparing it with the frequency-of-appearance distribution of the codes of the normal service, anomalous flows such as illegal accesses, can be detected because they are considered to have extreme deviation in some codes. Based on this consideration, we can say that attacks with illegal payloads can be detected by using the degree of deviation from the normal model which is evaluated by the goodness of fit test.

We use the 20 higher rank classes because the deviation of a character is notably seen only in some classes of higher ranks of histogram. Thus, the degree of deviation of the payload characters is expressed by averaged χ^2 *test* shown as (7).

$$T = \frac{1}{L} \sum_{l=0}^{19} \frac{(O_l - E_l)^2}{E_l} \tag{7}$$

where E_l denotes the expected value of lth class of the histogram, O_l denotes the the observation appearance frequency, and L is the data size of the flow.

2.3 State Discrimination

In above sections, we described three quantities based on three different statistics and the evaluation method of the degree of anomaly. In the proposed

technique, the degree of the anomaly of a flow, referred to as Anomaly Score (AS) throughout this paper, is regarded as the absolute value of $\boldsymbol{AS}$ which is expressed as (8).

$$\boldsymbol{AS} = (H, C, T)^T \tag{8}$$

The anomaly score of a flow is the length of $\boldsymbol{AS}$, $|AS|$.

Moreover, in order to control the number of detection failure, a two stage discrimination which consists of *suspicious* stage and *danger* stage is performed.

AS is calculated for each flow observed. If AS exceeds $Th(P)$, the stage shifts to *danger* stage and generates an alert whose interval time with respect to the previous alert exceeds Dt. If the interval time is shorter than Dt, the danger stage generates no alerts.

When AS is below $Th(P)$, the stage becomes *suspicious* and AS is accumulated for each service(port P) and each host ($Client$ or $Server$). The value of AS is referred to as $Sus.C.Sc(P)$ or $Sus.S.Sc(P)$. When this $Sus.C.Sc(P)$ or $Sus.S.Sc(P)$ exceeds K times the value of $Th(P)$ during a pre-determined period of time S_t[sec],it shifts to *danger* stage. When a timeout occurs to $Sus.C.Sc(P)$ or $Sus.S.Sc(P)$ of a certain host, $Sus.C.Sc(P)$ or $Sus.S.Sc(P)$ of all services provided by the host is initialized.

By establishing *suspicious* stage, it is possible to control the detection failure (false negative:FN). For instance, in case of Denial of Service(DoS), a large number of flows are generated over a short interval of time. Considering only the anomaly score AS of individual flows, the system is likely to fail in detecting the network anomaly as each flow's anomaly score does not exceed the threshold. However, since *suspicious* stage considers the aggregate value of the anomaly score of the whole flows, the system will be able to successfully detect the anomaly.

3 Evaluation of the Proposed Technique

3.1 Experiment Environment

To evaluate the performance of the proposed system, we conducted an experiment using the data set available at [4]. This data set includes five-week network traffic (from Week1 to Week5). Week1 and Week3 constitute the normal state of network traffic where no attack has occurred, and the traffic data of Week2, Week4 and Week5 contain labeled attacks. In this experiment, the data of Week1 and Week3 is used as data for the normal state model construction of the proposed system. By using the normal state model, we detect the known attacks included in the data of Week4 and Week5.

Moreover, since the proposed technique builds models according to every flows, the traffic used as the target for detection is TCP traffic for which the host in LAN operates as a server. 149 attacks using TCP can be found in the Week4 and Week5 data set. As simulation parameter, we set S_t, D_t, K, and $Th(P)$ to 120, 60, 30, and 10, respectively.

Table 2. Comparison with other techniques(1)

IDS	Detection Target	Detection Success	Success Rate [%]
NETAD	185	132	71
Expert-1	169	85	50
Expert-2	173	81	47
DMine	102	41	40
Forensics	27	15	55
PHAD & ALAD	180	70	39
Proposed System	149	98	66

3.2 Results

Table 2 shows the experiment result of the proposed scheme and results of other detection techniques obtained from [5]-[12].

In Table 2, we compare the performance of the proposed algorithm to former techniques based on Detection Success and Detection Success Rate. Compared with other techniques, the proposed technique exhibits good performance. However, this comparison has a limitation of difference in the number of detection targets because each technique targets different attacks. Since the targeted traffic of the proposed technique is set only to TCP traffic, the number of attacks to be detected is 149.

There is a limitation of the maximum number of False Alarm(FA) in case of the performance evaluation using this data set. This is due to the fact that detection rate can increase when there are no limitation for the number of FA. Generally, in the performance evaluation, the number of FA per day is limited to 10. That is, the tolerable number of FA is 100 during 10 days (Week4 and Week5). The number of FA in the proposed technique is 57. This shows another good performance of the proposed scheme in terms of reducing the number of FAs.

4 Consideration

In the proposed technique, the deviation from the normal state model defined by three kinds of statistics is evaluated as anomaly score. The issue to be discussed here is whether we should use the three statistics, to detect network anomaly simultaneously, or only one or two kinds are sufficient to guarantee a good detection system. Moreover, in the conducted experiments, the degree of anomaly of a flow is simply defined as the absolute value of $\boldsymbol{AS}$, which consists of the three statistical elements H, C, and T, as shown in (8). Alternatively, we can assign a weight for each element as:

$$\boldsymbol{AS} = (\alpha H, \beta C, \gamma T)^T \tag{9}$$

where α, β, and γ are the assigned weights of H, C, and T respectively.

Table 3. Change of the Detection Performance by Weighting to the Statistics

Weighting (H:C:T)	Detection Rate	False Positive	Weighting (H:C:T)	Detection Rate	False Positive
1 : 0 : 0	71/149 : 48%	205	1 : 2 : 1	104/149 : 70%	90
0 : 1 : 0	94/149 : 63%	172	1 : 1 : 2	90/149 : 60%	48
0 : 0 : 1	73/149 : 49%	37	1 : 3 : 3	97/149 : 65%	64
0 : 1 : 1	85/149 : 57%	56	1 : 4 : 2	104/149 : 70%	99
1 : 1 : 1	98/149 : 66%	57	1 : 4 : 1	103/149 : 69%	180

H : The Access Frequency to a Server Host
C : The Characteristic of Flow Unit
T : The Histogram of the Character Code of Payload

In the remainder of this section, we investigate the system performance when only one or two statistics are used and when $\boldsymbol{AS}$ is defined as (9). Table 3 shows the experiment result.

From Table 3, we can see that the highest detection rate is 70% if the weights are set adequately and the detection rate 70% comes near the rate of NETAD 71%. The detection rate of NETAD is estimated under the limitation of the number of FA per day to 10, so the total number of FA of NETAD in 10 days is 100. On the other hand, the number of FA of the proposed method is 90 when the parameters are set as $(\alpha, \beta, \gamma) = (1, 2, 1)$. From this result, the proposed method can detect anomalies with lower false positive rate compared to NETAD.

About the effect of parameters, γ, which is the weight for the histogram of the character code of payload, functions to reduce the number of False Alarms. Since the other statistics, H and C, check off the contents of communications, it can be considered that false alarms subject to be generated by occasional errors independently of T.

5 Conclusion

In this paper, a highly precise anomaly detection system was proposed for early detection of network anomalies. The proposed technique groups traffic in accordance with the service type. The scheme adopts three kinds of statistics to detect network anomalies. These statistics help to define the normal state of a network and evaluate the deviation of the network state from its normal state by observing flows. Based on this evaluation, the scheme judges the anomaly of the network. Experiment results demonstrate the good performance of the proposed technique compared with other NIDS(s). By using this technique, it is possible to improve network security and its management.

References

1. Anderson, Debra, Teresa F.Lunt, Harold Javits, Ann Tamaru, Alfonso Valdes, "Detecting unusual program behavior using the statistical component of the Nextgeneration Intrusion Detection Expert System(NIDES)", Computer Science Laboratory SRI-CSL 95-06 May 1995

2. SPADE, Silicon Defense,
http://www.silicondefense.com/software/spice/
3. RTFM, http://www.auckland.ac.nz/net/Internet/rtfm/
4. 1999 DARPA off-line intrusion detection evaluation test set,
http://www.ll.mit.edu/IST/ideval/index.html
5. Matthew V. Mahoney and Philip K. Chan, "Detecting Novel Attacks by Identifying AnomalousNetwork Packet Headers",
Florida Institute of Technology Technical Report CS-2001-2
6. M.Mahoney,"Network Traffic Anomaly Detection Based on Packet Bytes",Proc. ACM-SAC, pp.346-350,2003.
7. Matthew V.Mahoney and Philip K.Chan, "Learning Nonstationary Models of Normal Network Traffic for Detarcting Novel Attacks", SIGKDD '02, July 23-26, 2002, Edmonton, Alberta, Canada
8. P.Neumann and P.Porras, "Experience with EMERALD to DATE", in Proceedings 1st USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, California, April 1999, 73-80,
http://www.sdl.sri.com/projects/emerald/inde.html
9. G.Vigna, S.T.Eckmann, and R.A.Kemmerer,
"The STAT Tool Suite", in Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), IEEE Press, january 2000
10. R. Sekar and P. Uppuluri, "Synthesizing Fast Intrusion Prevention/Detection Systems from High-Level Specifications", in Proceedings 8th Usenix Security Symposium, Washington DC, Aug. 1999, http://rcs-sgi.cs.iastate.edu/sekar/abs/usenixsec99.htm.
11. S.Jajodia, D.Barbara, B.Speegle, and N.Wu, "Audit Data Analysis and Mining (ADAM)",
project described in http://www.isse.gmu.edu/ dbarbara/adam.html, April, 2000.
12. M.Tyson, P.Berry, N.Williams, D.Moran, D.Blei, "DERBI: Diagnosis, Explanation and Recovery from computer Break-Ins",
project described in http://www.ai.sri.com/ derbi/, April. 2000.

SoIDPS: Sensor Objects-Based Intrusion Detection and Prevention System and Its Implementation

SeongJe Cho[1], Hye-Young Chang[1], HongGeun Kim[2], and WoongChul Choi[3,†]

[1] Division of Information and Computer Science, DanKook University
{sjcho, chenil}@dankook.ac.kr
[2] Korea Information Security Agency
hgkim@kisa.or.kr
[3] Department of Computer Science, KwangWoon University
wchoi@daisy.kw.ac.kr

Abstract. In this paper, we propose an intrusion detection and prevention system using sensor objects that are a kind of trap and are accessible only by the programs that are allowed by the system. Any access to the sensor object by disallowed programs or any transmission of the sensor object to outside of the system is regarded as an intrusion. In such case, the proposed system logs the related information on the process as well as the network connections, and terminates the suspicious process to prevent any possible intrusion. By implementing the proposed method as Loadable Kernel Module (LKM) in the Linux, it is impossible for any process to access the sensor objects without permission. In addition, the security policy will be dynamically applied at run time. Experimental results show that the security policy is enforced with negligible overhead, compared to the performance of the unmodified original system.

1 Introduction

Research on intrusion detection and prevention is increasing, mainly because of the inherent limitations of network perimeter defense. Intrusion detection system can be classified into one of the two categories of network-based and host-based in terms of the deployment position [1, 3]. Due to the inherent limitations of network-based intrusion detection system and the increasing use of encryption in communication, intrusion detection must move to the host where the content is visible in the clear.

Generally, the research on the intrusion detection can be divided into misuse detection, anomaly detection and specification-based detection. *Misuse Detection* systems [4, 5] first defined a collection of signatures of known attacks. Activities matching representative patterns are considered attacks. This approach can detect known attacks accurately, but is ineffective against previously unseen attacks, as no signatures are available for such attacks. *Anomaly Detection* systems [6, 7] first characterize a statistical profile of normal behavior usually through the learning phase. In the operation phase, a pattern that deviates significantly from the normal profile is considered an attack. Anomaly detection overcomes the limitation of misuse

† Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 260 – 266, 2005.
© Springer-Verlag Berlin Heidelberg 2005

detection by focusing on the *normal system behaviors,* rather than the attack behaviors. But, anomaly detection often exhibits legitimate but previously unseen behavior, which may produce a high degree of false alarms. Moreover, the performance of the anomaly detection systems largely depends on the amount of the heuristic information and the degree of the accuracy of the learning. In *specification-based* detection [8, 9], the correct behaviors of critical objects are manually abstracted and crafted as security specifications, which are compared to the actual behavior of the objects. Intrusions, which usually cause an object to behavior abnormally, can be detected without the exact knowledge about them. Specification-based detection overcomes the high rate of false alarms caused by the legitimate but unseen behavior in the anomaly detection. But, the development of detailed specifications can be time-consuming.

In this paper, we propose *SoIDPS* (Sensor objects-based Intrusion Detection and Prevention System) that uses policy-based detection and can reduce the possibility of false detections. The sensor objects are installed in critical directories or files to control any access to them that is not permitted. They are also only accessible from inside of the system, and are not transmittable to the outside of the system, which is used to determine whether there is an intrusion or not. Therefore, this method is very effective to unknown intrusion detection. Another advantage of the proposed system is that it is implemented as Loadable Kernel Module (LKM) in the Linux, therefore, it can be regarded as a security model that is secure, dynamical and configurable.

The remainder of the paper is organized as follows. In Section 2, we propose a new intrusion detection model using sensor objects. Section 3 explains the implementation of the proposed system in the Linux kernel and the experimental performance evaluation results. In section 4, we conclude with the future work.

2 Security Policy and Its Usage for Intrusion Detection

The goal of the *SoIDPS* is to detect and prevent any intrusion or anomaly behavior. To achieve such goal, the *SoIDPS* uses '*sensor objects*' that are software objects like a simple sensitive file or a character string. In the proposed system, there are two kinds of sensor objects, *sensor file* and *sensor data*.

A *sensor file* is a special file created by a privileged command. One or more of its instances can be put into a critical directory. Only the privileged command can create the sensor files. That is, other programs should not access the sensor files. The sensor file is effective for detecting an intrusion when an unauthorized program tries to steal out arbitrary set of files in important directories. If unknown security-breaking programs try to access the sensor file, it is sure that there is an attempt of an intrusion to the directory. Because of that, *SoIDPS* can detect any unknown threat as well.

However, using sensor files might not be effective for the case where a malicious program tries to steal only a specific file. For such case, *sensor data* which are a sensitive text string can be used. Sensor data are inserted into every critical file by a privileged command. It is allowed for authorized programs to access the critical files with sensor data only from the internal system or intranet, but should not be sent out to any external host via network. If there is an outgoing network packet including sensor data, the transmission of the sensor data to external network is regarded as an

attack. This determination is effective especially for the case where any malicious program tries to steal out a specific sensitive file selectively

The *SoIDPS* regards the following action as an intrusion or a threat. If there is an attempt to access to a sensor file not by the privileged command, it is an intrusion. If sensor data are transmitted to any external network, it is an intrusion too. Every access to a sensor object is logged in terms of the sensor object, the process that tried to access to the sensor data, and the time of the event, etc., and will be used for security analysis. It can be even useful for tracing back the intrusion attempt.

In order to increase the degree of the security level, it is possible to put multiple sensor objects in a critical directory, and therefore, it is helpful to have a classification of the security levels of the system directories and files like in **Table 1**.

Table 1. Classification of security level

Security level	Directories or files (example)	Number of sensor objects
Secret	Password file, Secret program source directories or files	2 or above
Confidential	Directories or files of system & network administration, Directories or files of system configurations	1 ~ 2
Sensitive	Directories or files of users	0 ~ 1
Unclassified	Sharing directories or files	0

3 Implementation and Experimental Results

In order to verify the validity of the proposed method, we have implemented and evaluated a prototype system in the Linux kernel 2.4.22, IBM PC Pentium IV. To manage sensor objects in the system, several commands are implemented, and some system calls are added and modified using the hooking feature of LKM.

3.1 Sensor Object Management

The proposed system first determines the directories and files where sensor objects are installed, and then the name for sensor objects. We use a string started with *.(dot)* as a file name for a sensor file, and the string of *sensor_data* for the name of the sensor data. Notice that the sensor files seem to be hidden from the users. Once the *SensorFile_Install* command installs sensor files using the configuration file named *sfInstall.cfg* which has a list of the target directory and the number of sensors, a file named *sfInstalled.txt* is consequently created that has a record of the installed sensor files. The *sfInstalled.txt* is used to remove the sensor files. Fig. 1 shows the sequence of installing sensor files named *.sensor_file*, *.passwd*, *.source*.

Fig. 2 shows a configuration file named *sdInstall.cfg* that has a list of the secret files and the name of the sensor data (*sensor_data*). The program named *SensorData_Install* installs sensor data using *sdInstall.cfg*. The *sdInstall.cfg* is also used to remove the sensor data. As sensor data are installed, it is important to guarantee that the normal operation of the files are not bothered by such installation, therefore, the system administrator needs to configure the sensor data carefully. For

example, since a string following '//' in the C source file is interpreted as a comment, sensor data are installed after '//'. In case of the password file, since '*' at the beginning of a line is interpreted as an inactive user, sensor data are inserted there. Such installed sensor data will be used effectively for detecting intrusions.

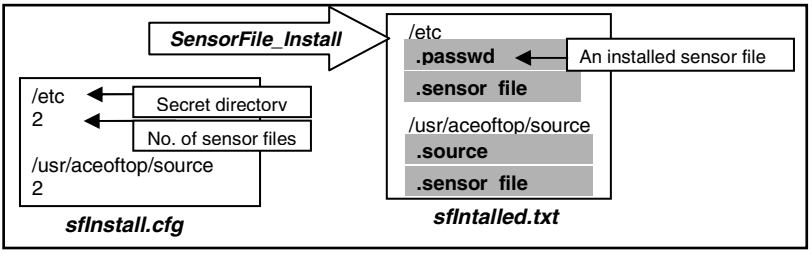


Fig. 1. Sensor File Installation

/home/aceoftop/project.c **2** **// sensor_data**	← *Secret file* ← *Line number to be installed* ← *Name of sensor data*
/etc/passwd **5** ***sensor_data:x:511:511:/sensor_data:/bin/bash**	← *Secret file* ← *Line number to be installed* ← *Name of sensor data*

Fig. 2. An Example of ***sdInstall.cfg*** file

3.2 *SoIDPS* and Its Implementation Module

The proposed system consists of **host module** and **network module**. The **host module** monitors any access to the sensor objects, and saves the log of process ID, the command, user ID, time information, which file is accessed or whether the accessed object is sensor data or sensor file, and then terminates the command.

In order to check if a malicious program accesses sensor objects, we have modified and added several kernel internal routines such as *sys_open, sys_read, sys_close,* and *sys_unlink* which are corresponding to *open, read, close,* and *unlink* system calls, respectively. The *sys_open* examines if the opened file is *'.sensor_file'* at each *open* operation. If so, it records the name of the program of the current task, the PID, the UID, the absolute path of the sensor file, and the current time onto the log file *SIDS_open.log*, kills the task, raises the alarm, and then sends an alarming email to the superuser. As shown in Fig. 3, the *'cat'* command breached the system security policy and touched */etc/.sensor3* with supervisor's security level (UID=0). The host module killed the process and notified it to the super-user. The super-user can suspect the *'cat'* to be an abnormal program and replace it with a secure '*cat*'.

Sys_read examines if *read* call accesses to the sensitive string *'sensor_data'* using the Boyer-Moore algorithm [10] for the exact string matching. If so, it records the program name of the current task, the PID, the current UID, the absolute path of the file with the sensor data, and the current time onto the log file *SIDS_read.log*, and then raises an alarm. Fig. 4 shows an example of *SIDS_read.log***.** To decide if the

access to sensor data is a threat or not, the host module needs more information from its corresponding network module.

The **network module** scans the outgoing packets from the internal to the external networks to check if they contain any sensor objects, and if so, it concludes that there was an intrusion and shuts down the connection between the internal and the external networks. As the network module scans the packets, pinpointing which critical files were tried to be transmitted is not easy, therefore, the collaboration with the host module is very important for successful intrusion detection and prevention. To monitor the network traffics, we have modified *sys_socketcall* and *sys_sendto* functions. Using the functions, the network module inspects if a program transmits a packet with *sensor data* to the external networks. If so, it records the connection information onto the file *SIDS_net.log*(Fig. 5). It then sends an alarming e-mail to the super-user. We have also used the Boyer-Moore algorithm to perform exact string matching between the packet data and the sensor data. The network module has been installed using *libpcap* library [11] which is a network packet analysis package.

```
[root@LocalHost root]# cat SIDS_open.log
[cat] accessed /etc/.sensor3 pid = 686 uid = 0 time = 2005/07/28 thu 23:02:32
[vi] accessed /root/.sensor10 pid = 687 uid = 0 time =  2005/07/28 thu 23:02:40
```

Fig. 3. An example of the log file, *SIDS_open.log*

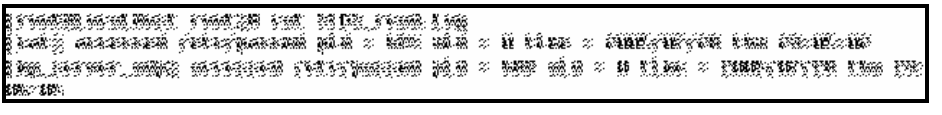


Fig. 4. Log file about the access history related to reading sensor data

```
[root@LocalHost root]# cat SIDS_net.log
my_server_udp => [203.234.83.141] pid = 689 uid = 0 time = 2005/07/28 thu 23:06
:05
```

Fig. 5. Log file about the access history related to transmitting sensor data

To verify if our system can detect an illicit transmission of a secret file with sensor data over the Internet, we have written the program '*my_server_udp*' that tries to transmit a secret file out to an external network. We can see from Fig. 4 and Fig. 5 that */etc/passwd* with the sensor data have been accessed by '*cat'* and '*my_server_udp*'. The access by the '*cat*' seems to have no problem because its access to */etc/passwd* happened inside the system. However, the access by the '*my_server_udp'* is suspicious because it may convey sensor data out to an external network. From Fig. 5, we can see the '*my_server_udp'* process tries to convey a secret file with the sensor data out to the external host with IP address 203.234.83.141. As soon as the outgoing packet with the sensor data is noticed, the system logs the packet information and even kills the corresponding process *my_server_udp* and disconnects the session if the destination IP address is one of the external hosts. From the log files, we can conclude that an intruder tried to steal out the file */etc/passwd* using *my_server_udp* program.

3.3 Performance Analysis

Under the *SoIDPS*, the cost for monitoring a sensor file is very low because the system checks whether the name of the opened file is a sensor file or not. We have measured the access time of a sensor file with varying the number of sensor files. The results are shown in Table 2. In our system with 160 sensor files, the average access time to sensor files is 60μs. Compared to the original system where the average access time to a file is 4μs, the overhead of our system is 56μs.

Table 2. File access time by varying the number of sensor files

Number of sensor files	10	20	40	80	160
Average file access time on the proposed system	23μs	24μs	33μs	57μs	60μs
Average file access time on the original system	4μs	4μs	4μs	4μs	4μs

Table 3. Average time to read a file completely with detecting sensor data

File size	500B	1KB	10KB	100KB	1MB
Average read time on the proposed system	11μs	17μs	100μs	944μs	12,142μs
Average read time on the original system	6μs	11μs	87μs	841μs	11,214μs

The large part of the implementation overhead is caused by executing the Boyer-Moore algorithm to check whether a given string includes sensor data in *sys_read* routine. We have evaluated the overhead by measuring the execution time required to make a string comparison between the sensor data and the fetched data in *sys_read*. The experimental results are shown in Table 3. In our system, the average size of the files with the sensor data is around 10KB. In that case, it needs some additional time by about 13μs to check if a process accesses the file with the sensor data. If the sensor data is put in the beginning part of the file, the search time can be reduced significantly. The overhead in Table 3 is unavoidable at the cost of enhancing the security level. To discriminate between the secret files and the normal files, we put a string "*public*" into the first line of each normal file. We can reduce the string matching overhead by stopping the string comparison in *sys_read* if the string "*public*" was detected. The time to read a normal file on our system is same as the time to read the same sized secret file with the sensor data placed in the first line.

4 Conclusion and Future Work

In this paper, the *SoIDPS*, an intrusion detection and prevention system is presented and implemented. The *SoIDPS* detects any illegal behavior and prevents critical directories and files against security attacks, especially against unknown ones. It can perform such security functions by using sensor files and sensor data, which are installed in critical directories or files to monitor and prevent any unallowable access to them. The *SoIDPS* consists of the host module and the network module. The host module monitors every access to sensor objects, records the access information onto log files, and prevents the system from the attacks of exploiting the important data.

The network module checks if each outgoing packet with sensor data is transmitted to external networks. In that case, it records the information of the process and the packets onto a log file, and kills the process. The performance measurement explicitly shows that the *SoIDPS* achieves the security goal with negligible overhead.

References

1. Marcus J. Ranum, Experiences Benchmarking Intrusion Detection Systems, NFR Security Technical Publications, Dec. 2001 (http://www.nfr.com)
2. Understanding Heuristics: Symantec's Bloodhound Technology. Symantec White Paper Series Volume XXXIV
3. ISO/IEC WD 18043 (SC 27 N 3180): Guidelines for the implementation, operation and management of intrusion detection systems (IDS), 2002-04-26
4. U. Lindqvist and P. Porras, "Detecting Computer and Network Misuse through the Production-Based Expert System Toolset (P-BEST)", Proceedings of the Symposium on Security and Privacy, 1999.
5. S. Kumar and E.H. Spafford. "A Pattern Matching Model for Misuse Intrusion Detection", Proceedings of the 17th National Computer Security Conference, 1994.
6. C. C. Michael, Anup Ghosh, "Simple, state-based approaches to program-based anomaly detection", ACM Transactions on Information and System Security, Vol. 5, Issue 3, 2002.
7. Salvatore J. Stolfo, et al., "Anomaly Detection in Computer Security and an Application to File System Accesses", Proceedings of 15th International Symposium of Foundations of Intelligent Systems, 2005.
8. R. Sekar, et al., "Intrusion Detection: Specification-based anomaly detection: a new approach for detecting network intrusions", Proceedings of the 9th ACM conference on Computer and communications security, 2002.
9. R. Sekar and P. Uppuluri, "Synthesizing Fast Intrusion Prevention/Detection Systems from High-Level Specifications", Proceedings of USENIX Security Symposium, 1999.
10. Charras C. and Lecroq T.: Handbook of Exact String-Matching Algorithms. (http://www-igm.univ-mlv.fr/ ~lecroq/biblio_en.html)
11. ftp://ftp.ee.lbl.gov

A Statistical Model for Detecting Abnormality in Static-Priority Scheduling Networks with Differentiated Services

Ming Li[1] and Wei Zhao[2]

[1] School of Information Science & Technology, East China Normal University, Shanghai 200062, China
mli@ee.ecnu.edu.cn, ming_lihk@yahoo.com
http://www.ee.ecnu.edu.cn/teachers/mli/js_lm(Eng).htm
[2] Department of Computer Science, Texas A&M University, College Station, TX 77843-1112, USA
w-zhao@tamu.edu
http://faculty.cs.tamu.edu/zhao/

Abstract. This paper presents a new statistical model for detecting signs of abnormality in static-priority scheduling networks with differentiated services at connection levels on a class-by-class basis. The formulas in terms of detection probability, miss probability, probabilities of classifications, and detection threshold are proposed.

Keywords: Anomaly detection, real-time systems, traffic constraint, static-priority scheduling networks, differentiated services, time series.

1 Introduction

Anomaly detection has gained applications in computer communication networks, such as network security, see e.g. [1], [2], [3], [4], [5], [6], [7]. This paper considers the abnormality identification of arrival traffic time series (traffic for short) at connection levels, which relates to traffic models.

In traffic engineering, traffic models can be classified into two categories [8]. One is statistically modeling as can be seen from [9], [10], [11]. The other bounded modeling, see e.g. [12], [13], [14], [15].

Though statistically modeling has gained considerable progresses, one thing worth noting is that they are well in agreement with real life data in aggregated case. In general, nevertheless, they are not enough when traffic at connection levels has to be taken into account. In fact, traffic modeling at connection level remains challenging in the field [16]. In the academic area of computer science, a remarkable thing to model traffic at connection level is to study traffic from a view of deterministic queuing theory, which is often called network calculus or bounded modeling. *One of the contributions in this paper is to develop traffic constraint (a kind of deterministically bounded model [13]) into a statistical bound of traffic.*

Recent developments of networking exhibit that there exists an increased interest in differentiated services (DiffServ) [13], [17]. From a view of abnormality detection, instead of detecting abnormality of all connections, we are more interested in

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 267 – 272, 2005.
© Springer-Verlag Berlin Heidelberg 2005

identifying abnormality of some connections in practice. Thus, this paper studies abnormality detection in the environment of DiffServ.

As far as detections were concerned, the current situation is not lacking methods for detections [18] but short of reliable detections as can be seen from the statement like this. "The challenge is to develop a system that detects close to 100 percent of attacks. We are still far from achieving this goal [19]." From a view of statistical detection, however, instead of developing a way to detect close to 100 percent of abnormality, we study how to achieve an accurate detection for a given detection probability. By accurate detection, we mean that a detection model is able to report signs of abnormality for a predetermined detection probability. This presentation proposes an accurate detection model of abnormality in static-priority scheduling networks with DiffServ based on two points: 1) the null hypotheses and 2) averaging traffic constraint in [13]. *A key point in this contribution is to randomize traffic constraint on an interval-by-interval basis so as to utilize the techniques from a view of time series to carry out a statistical traffic bound*, which we shall call average traffic constraint for simplicity. To our best knowledge, this paper is the first attempt to propose average traffic constraint from a view of stochastic processes and moreover apply it to abnormality detection.

The rest of paper is organized as follows. Section 2 introduces an average traffic constraint in static-priority scheduling networks with DiffServ. Section 3 discusses detection probability and detection threshold. Section 4 concludes the paper.

2 Average Traffic Constraint

In this section, we first brief the conventional traffic constraint. Then, randomize it to a statistical constraint of traffic.

The traffic constraint is given by the following definition.

Definition 1: Let $f(t)$ be arrival traffic function. If $f(t+I)-f(t) \le F(I)$ for $t>0$ and $I>0$, then $F(I)$ is called traffic constraint function of $f(t)$ [13]. □

Definition 1 is a general description of traffic constraint, meaning that the increment of traffic $f(t)$ is upper-bounded by $F(I)$. It is actually a bounded traffic model [13]. The practical significance of such model is to model traffic at connection level. Due to this, we write the traffic constraint function of group of flows as follows.

Definition 2: Let $f^i_{p,j,k}(t)$ be all flows of class i with priority p going through server k from input link j. Let $F^i_{p,j,k}(t)$ be the traffic constraint function of $f^i_{p,j,k}(t)$. Then, $F^i_{p,j,k}(t)$ is given by $f^i_{p,j,k}(t+I)- f^i_{p,j,k}(t) \le F^i_{p,j,k}(I)$ for $t>0$ and $I>0$. □

Definition 2 provides a bounded model of traffic in static-priority scheduling networks with DiffServ at connection level. Nevertheless, it is still a deterministic model in the bounded modeling sense. We now present a statistical model from a view of bounded modeling.

Theoretically, the interval length I can be any positively real number. In practice, however, it is usually selected as a finite positive integer in practice. Fix the value of

I and observe $F^i_{p,j,k}(I)$ in the interval $[(n-1)I, nI]$, $n=1,2,...,N$. For each interval, there is a traffic constraint function $F^i_{p,j,k}(I)$, which is also a function of the index n. We denote this function $F^i_{p,j,k}(I, n)$. Usually, $F^i_{p,j,k}(I, n) \neq F^i_{p,j,k}(I, q)$ for $n \neq q$. Therefore, $F^i_{p,j,k}(I, n)$ is a random variable over the index n. Now, divide the interval $[(n-1)I, nI]$ into M non-overlapped segments. Each segment is of L length. For the mth segment, we compute the mean $E[F^i_{p,j,k}(I, n)]_m (m=1,2,...,M)$, where E is the mean operator. Again, $E[F^i_{p,j,k}(I, n)]_l \neq E[F^i_{p,j,k}(I, n)]_m$ for $l \neq m$. Thus, $E[F^i_{p,j,k}(I, n)]_m$ is a random variable too. According to statistics, if $M \geq 10$, $E[F^i_{p,j,k}(I, n)]_m$ quite accurately follows Gaussian distribution [1], [20]. In this case,

$$E[F^i_{p,j,k}(I, n)]_m \sim \frac{1}{\sqrt{2\pi}\sigma_F} \exp[-\frac{\{E[F^i_{p,j,k}(I, n)]_m - F_\mu(M)\}^2}{2\sigma_F^2}], \quad (1)$$

where σ_F^2 is the variance of $E[F^i_{p,j,k}(I, n)]_m$ and $F_\mu(M)$ is its mean. We call $E[F^i_{p,j,k}(I, n)]_m$ *average traffic constraint* of traffic flow $f^i_{p,j,k}(t)$.

3 Detection Probability

In the case of $M \geq 10$, it is easily seen that

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{F_\mu(M) - E[F^i_{p,j,k}(I, n)]_m}{\sigma_F \sqrt{M}} \leq z_{\alpha/2} \right] = 1-\alpha, \quad (2)$$

where $(1-\alpha)$ is called confidence coefficient. Let $C_F(M, \alpha)$ be the confidence interval with $(1-\alpha)$ confidence coefficient. Then,

$$C_F(M, \alpha) = \left[F_\mu(M) - \frac{\sigma_F z_{\alpha/2}}{\sqrt{M}}, F_\mu(M) + \frac{\sigma_F z_{\alpha/2}}{\sqrt{M}} \right]. \quad (3)$$

The above expression exhibits that $F_\mu(M)$ is a template of average traffic constraint. Statistically, we have $(1-\alpha)\%$ confidence to say that $E[F^i_{p,j,k}(I, n)]_m$ takes $F_\mu(M)$ as its approximation with the variation less than or equal to $\frac{\sigma_F z_{\alpha/2}}{\sqrt{M}}$.

Denote $\xi \triangleq E[F^i_{p,j,k}(I, n)]_m$. Then,

$$\text{Prob}\left(\xi > F_\mu(M) + \frac{\sigma_F z_{\alpha/2}}{\sqrt{M}} \right) = \frac{\alpha}{2}. \quad (4)$$

On the other hand,

$$\text{Prob}\left(\xi \leq F_\mu(M) - \frac{\sigma_F z_{\alpha/2}}{\sqrt{M}} \right) = \frac{\alpha}{2}. \quad (5)$$

For facilitating the discussion, two terms are explained as follows. Correctly recognizing an abnormal sign means *detection* and failing to recognize it *miss*. We explain the detection probability as well as miss probability by the following theorem.

Theorem 1 (Detection probability and detection threshold): Let

$$V = F_{\mu}(M) + \frac{\sigma_F z_{\alpha/2}}{\sqrt{M}} \tag{6}$$

be the detection threshold. Let P_{det} and P_{miss} be detection probability and miss probability, respectively. Then,

$$P_{det} = P\{V < \xi < \infty\} = (1-\alpha/2), \tag{7}$$

$$P_{miss} = P\{-\infty < \xi < V\} = \alpha/2. \tag{8}$$

Proof: The probability of $\xi \in C_F(M,\ \alpha)$ is $(1-\alpha)$. According to (2) and (5), the probability of $\xi \le V$ is $(1-\alpha/2)$. Therefore, $\xi > V$ exhibits a sign of abnormality with $(1-\alpha/2)$ probability. Hence, $P_{det} = (1-\alpha/2)$. Since detection probability plus miss one equals 1, $P_{miss} = \alpha/2$. □

From Theorem 1, we can achieve the following statistical classification criterion for a given detection probability by setting the value α.

Corollary 1 (Classification): Let $f^i_{p,j,k}(t)$ be arrival traffic of class i with priority p going through server k from input link j at a protected site. Then,

$$f^i_{p,j,k}(t) \in N \ \text{ if } \ E[F^i_{p,j,k}(I,\ n)]_m \le V \tag{9a}$$

where N implies normal set of traffic flow, and

$$f^i_{p,j,k}(t) \in A \ \text{ if } \ E[F^i_{p,j,k}(I,\ n)]_m > V. \tag{9b}$$

where A implies abnormal set. The proof is straightforward from Theorem 1. □

The diagram of our detection is indicated in Fig. 1.

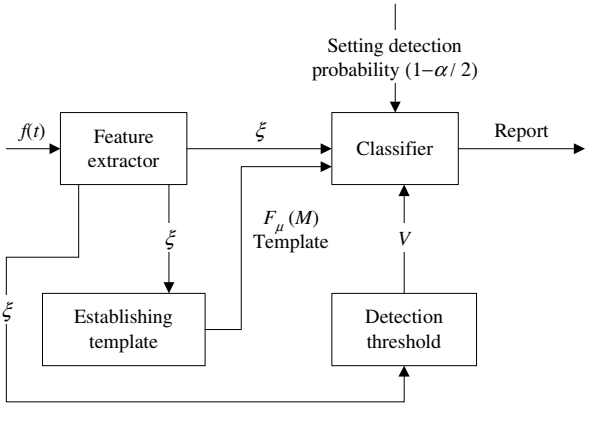


Fig. 1. Diagram of detection model

4 Conclusions

In this paper, we have extended the traffic constraint in [13], which is conventionally a bound function of arrival traffic, to a time series by averaging traffic constraints of flows on an interval-by-interval basis in DiffServ environment. Then, we have derived a statistical traffic constraint to bound traffic. Based on this, we have proposed a statistical model for the purpose of abnormality detection in static-priority scheduling networks with differentiated services at connection level.

With the present model, signs of abnormality can be identified on a class-by-class basis according to a detection probability that is predetermined. The detection probability may be very high and miss probability may be very low if α is set to be very small. The results in the paper suggest that abnormality signs can be detected at early stage that abnormality occurs since identification is done at connection level.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under the project grant number 60573125, by the National Science Foundation under Contracts 0081761, 0324988, 0329181, by the Defense Advanced Research Projects Agency under Contract F30602-99-1-0531, and by Texas A&M University under its Telecommunication and Information Task Force Program. Any opinions, findings, conclusions, and/or recommendations expressed in this material, either expressed or implied, are those of the authors and do not necessarily reflect the views of the sponsors listed above.

References

1. Li, M.: An Approach to Reliably Identifying Signs of DDOS Flood Attacks based on LRD Traffic Pattern Recognition. Computer & Security 23 (2004) 549-558
2. Bettati, R., Zhao, W., Teodor, D.: Real-Time Intrusion Detection and Suppression in ATM Networks. Proc., the 1st USENIX Workshop on Intrusion Detection and Network Monitoring, April 1999, 111-118
3. Schultz, E.: Intrusion Prevention. Computer & Security 23 (2004) 265-266
4. Cho, S.-B., Park, H.-J.: Efficient Anomaly Detection by Modeling Privilege Flows Using Hidden Markov Model. Computer & Security 22 (2003) 45-55
5. Cho, S., Cha, S.: SAD: Web Session Anomaly Detection based on Parameter Estimation. Computer & Security 23 (2004) 312-319
6. Gong, F.: Deciphering Detection Techniques: Part III Denial of Service Detection. White Paper, McAfee Network Security Technologies Group, Jan. 2003
7. Sorensen, S.: Competitive Overview of Statistical Anomaly Detection. White Paper, Juniper Networks Inc., www.juniper.net, 2004
8. Michiel, H., Laevens, K.: Teletraffic Engineering in a Broad-Band Era. Proc. IEEE 85 (1997) 2007-2033
9. Willinger, W., Paxson, V.: Where Mathematics Meets the Internet. Notices of the American Mathematical Society 45 (1998) 961-970

10. Li, M., Zhao, W., and et al.: Modeling Autocorrelation Functions of Self-Similar Teletraffic in Communication Networks based on Optimal Approximation in Hilbert Space. Applied Mathematical Modelling 27 (2003) 155-168
11. Li, M., Lim, SC.: Modeling Network Traffic Using Cauchy Correlation Model with Long-Range Dependence. Modern Physics Letters B 19 (2005) 829-840
12. L.-Boudec, J.-Yves, Patrick, T.: Network Calculus, A Theory of Deterministic Queuing Systems for the Internet. Springer (2001)
13. Wang, S., Xuan, D., Bettati, R., Zhao, W.: Providing Absolute Differentiated Services for Real-Time Applications in Static-Priority Scheduling Networks. IEEE/ACM T. Networking 12 (2004) 326-339
14. Cruz, L.: A Calculus for Network Delay, Part I: Network Elements in Isolation; Part II: Network Analysis. IEEE T. Inform. Theory 37 (1991) 114-131, 132-141
15. Chang, C. S.: On Deterministic Traffic Regulation and Service Guarantees: a Systematic Approach by Filtering. IEEE T. Information Theory 44 (1998) 1097-1109
16. Estan C., Varghese, G.: New Directions in Traffic Measurement and Accounting: Focusing on the Elephants, Ignoring the Mice. ACM T. Computer Systems 21 (2003) 270–313
17. Minei, I.: MPLS DiffServ-Aware Traffic Engineering. White Paper, Juniper Networks Inc., www.juniper.net, 2004
18. Leach, J.: TBSE—An Engineering Approach to The Design of Accurate and Reliable Security Systems. Computer & Security 23 (2004) 265-266
19. Kemmerer, R. A., Vigna, G.: Intrusion Detection: a Brief History and Overview. Supplement to Computer (IEEE Security & Privacy) 35 (2002) 27-30
20. Bendat, J. S., Piersol, A. G.: Random Data: Analysis and Measurement Procedure. 2nd Edition, John Wiley & Sons (1991)

Tamper Detection for Ubiquitous RFID-Enabled Supply Chain

Vidyasagar Potdar, Chen Wu, and Elizabeth Chang

School of Information System, Curtin Business School, Curtin University of Technology, Perth, Western Australia
{Vidyasagar.Potdar, Chen.Wu, Elizabeth.Chang}@cbs.curtin.edu.au
http://www.ceebi.research.cbs.curtin.edu.au/docs/index.php

Abstract. Security and privacy are two primary concerns in RFID adoption. In this paper we focus on security issues in general and data tampering in particular. Here we present a conceptual framework to detect and identify data tampering in RFID tags. The paper surveys the existing literature and proposes to add a tamper detection component in the existing RFID middleware architecture. The tamper detection component is supported by mathematical algorithm to embed and extract secret information which can be employed to detect data tampering.

1 Introduction

RFID means Radio Frequency Identification. A RFID tag is an electronic device that holds data. Typically these tags are attached to an item and contain a serial number or other data associated with that item. RFID technology uses radio waves to automatically identify objects which have such RFID tags attached to it. It is composed of three components; firstly a *tag* that contains the identification number and secondly a *reader* that activates the tag to broadcast its identification number, and finally *RFID Middleware* which is designed to provide the messaging, routing, and connectivity for reliable data integration with existing backend systems such as ERP, SCM, and WMS, etc [1, 2] The standard data structure of a RFID tag is shown in the Table 1.

The widespread adoption of RFID's has been hindered because associated security issues. Considering the *security* aspect of RFID's its worth noting that most of the Class 0 and 1 RFID's are not capable of any secure communication. This is attributed to the fact that this class of RFID's don't have enough computation power and storage capacity to perform any encrypted communication as a result all the data is transmitted in open which leaves doors open for eavesdroppers and attackers. For example Weis et al. (2004) estimate that as few as 500-5000 gates are employed in a typical RFID design, which are normally used for basic logic operation; hence there is no room for extras such as security [3]. Secondly the communication between reader and tag is wireless, which increases the possibility of eavesdropping by third parties.

Considering this insecure communication, data tampering with the RFID's cannot be ruled out. Lukas Grunwald showed how vulnerable RFID's can be when he used a small program called RFDump to show how the tags could be read, altered or even deleted using an inexpensive tag reader [4]. This small software showed how anyone

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 273 – 278, 2005.
© Springer-Verlag Berlin Heidelberg 2005

could tamper the RFID data easily. We realise that the issue of data tampering is not completely addressed by the RFID community and thus we take this opportunity to present a feasible and cost effective solution to detect data tampering in low-cost RFID systems.

Table 1. The data structure in an RFID tag is composed of four sections the Header, EPC Manager, Object Class and the Serial Number

Element	Header	EPC Manager	Object Class	Serial Number
Bits	2	28	24	10
$Value_{10}$	0-3	0-268435455	0-16777215	0-1023
Example		Toyota	Camry	1AUT315

The paper is organized in the following manner. In Section 2 we survey the existing literature. In Section 3 we propose the conceptual framework and in Section 4 we conclude the paper.

2 Related Work

While researchers are just starting to address security questions, privacy advocates and legislators have for some time been attempting to address the privacy issues. A lot of work has been done to address the privacy issues in RFID deployment, however literature doesn't show the existence of any solid full proof work to address the security issues except those in [3, 5].

In [5] a data authentication technique is presented to validate the content of RFID tag. Its termed as relational-check-code" which is created to authenticate the data on bag tag which is part of the solution of the Smart Trakker application. In this technique if the data in the RIFD is changed it can be detected. This solution can be implemented on low-end RFID tags however it doesn't provide a way to ascertain which data is changed and to what extent.

In [3] a solution is presented that assumes cryptographic properties on the tags which prevent unauthorized readers to read the RFID tags unless they are authenticated. Each tag holds the hash value of the authenticated key of a reader. When the reader request for an RFID tag the tag responds by the hash value and the reader responds by the authentication key. The tag then calculates the hash value based on the key and verifies the reader. However this hash value calculation is not feasible in low-end cost effective RFID's [3, 6].

On the other hand some schemes offer a combined privacy and security solution for example the one discussed in [7]. A solution presented in [8] offers location privacy but it is not scalable because it requires a lot of hash value calculations as well as it needs a good random value generator [9]. Here each tag shares an authentication key with the reader. Once the tag is queried the tag generates a pseudo random number (R) and sends the hash of key and the random number along to the reader. The reader then computes the hash using R and finds the appropriate key. Since the random number changes every time it protects from tracking. However this solution also relies on cryptographic properties.

The literature that we surveyed mostly focused on solutions based on next generation RFID's. Such solutions assume some kind of cryptographic capabilities on the RFID tags which is currently very expensive solution. From a deployment perspective using next generation RFID tags is an expensive bet and assuring security on current generation RFID tags is still a major issue. Security issues haven't been completely addressed in literature which gives us motivation to present our work.

3 Proposed Conceptual Framework

The proposed conceptual framework introduces a tamper detection component in the RFID middleware architecture. The tamper detection component is specially introduced to ascertain that no data tampering has happened on the RFID tag. Within the layered hierarchy, the tamper detection component is placed between the *data management* layer and the *application integration* layer. This is done to ascertain that whatever data is being propagated to the higher levels in the RFID middleware is tamper proof. The proposed framework is shown in Fig. 1.

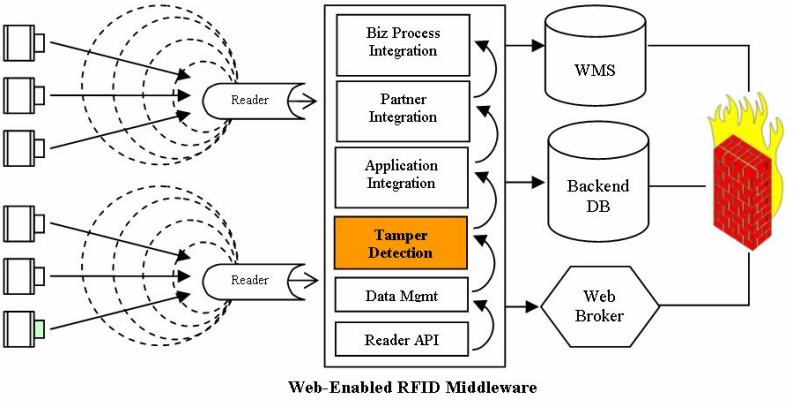


Fig. 1. Web-Enabled RFID Middleware Architecture with Tamper Detection Component

The tamper detection component is composed of an embedding and extraction algorithm which is used to detect data tampering. The proposed algorithm for tamper detection works by embedding secret information with in the RFID tags. In order to embed secret information we have to identify some space within the data which can be modified to represent secret information. In order to identify this space we investigated the RFID data structure. Table 1 describes the basic RFID data structure. On the basis of that investigation we determined that the serial number partition within the RFID tags can offer reasonable amount of space to foster embedding of the secret data. This selection is attributed to the following facts. The *Header* component is fully used for identifying the EAN.UCC key and the partitioning scheme. Hence there is no redundant space so there is no possibility for embedding any secret information. The *EPC Manager* is used to identify the manufacturer uniquely. Hence this partition also doesn't offer any space for embedding. The *Object Class* is used to identify the

product manufactured by the manufacturer. It may follow some product convention taxonomy where the first two digits might represent the classification of that product and so on. Hence modifying any of this data might interfere with the existing industry standard. As a result this partition also does not offer any space for embedding. The *Serial Number* is used to uniquely identify an item which belongs to a particular Object Class. It is orthogonal to first three partitions and can be decided by the manufacturer at will without violating any existing industry standards. Consequently it offers enough space to embed sufficient amount of data. Meanwhile the length of this partition is 38 bits (in EPC96) which offers enough room to accommodate the required amount of secret data. Thus this becomes most appropriate candidate for embedding the secret. Hence in our proposed conceptual framework we focus on information embedding in Serial Number partition in the RFID tags. We now discuss the embedding and extraction algorithm.

3.1 Embedding Algorithm for Tamper Detection

The embedding algorithm begins by selecting a set of one way functions F $\{f_1, f_2, f_3\}$. Each one way function is applied to the values within the RFID tags partition to generate a secret value as shown in Table 2.

Table 2. One Way functions

Partition	Function	Secret Value
EPC Manager (EM)	f_1	A = f(EM)
Object Class (OC)	f_2	B = f (OC)
SN_{org}	f_3	C = f (SN_{org})

This secret value is then embedded at predefined location within the Serial Number partition by appending it to the original Serial Number Value (SN_{org}) to generate the appended Serial Number (SN_{app}).

Table 3. Embedded Pattern for Tamper Detection in RFID Tags

Header	EPC Manager	Object Class	Serial Number (SN_{app})			
10111010	1101010101101	101010101011	A	SN_{org}	B	C

As illustrated in Table 3, the four partitions represent the internal data structure of the RFID tag. The EPC Manager and the Object Class values are hashed as shown in Table 2 and are appended to the SN_{org} in the Serial Number Partition as shown in Table 3. This pattern P (i.e. A, SN_{org}, B, C) can be user defined and is accessible by the RFID middleware's tamper detection component.

3.2 Extraction Algorithm for Tamper Detection

The extraction algorithm is completed in two stages. The first is the extraction stage where the secret values are extracted while the second is the detection stage where the tamper detection is carried out. In the extraction stage the algorithm begins by extracting the following parameters i.e. A, B and C from SNapp using the pattern P. We note

that this pattern is shared by the collaborating trading partners in distributed network. In the detection stage the values of EM, OC and SNorg are hashed using the same one way function set F {f1, f2, f3}, which were used for embedding. This is shared in the tamper detection module. These values are now compared with the extracted parameters to identify any data tampering. If the parameters match then we conclude that there is no tampering happened for EC, OC and SNorg. However, if the extracted parameters do not match then data tampering can be detected. The actual source of data tampering can be identified based on the facts which are illustrated in Table 4.

Table 4. Logic for Tamper Detection

If			Result
$f_1(EM)$	=	A	No Tampering
$f_1(EM)$	!=	A	***EM or*** $\boldsymbol{SN_{app}}$ Tampered
$f_2(OC)$	=	B	No Tampering
$f_2(OC)$	!=	B	***OC or*** $\boldsymbol{SN_{app}}$ Tampered
$f_3(SN_{org})$	=	C	No Tampering
$f_3(SN_{org})$	!=	C	$\boldsymbol{SN_{org}}$ ***or*** $\boldsymbol{SN_{app}}$ Tampered

The tamper detection technique that we presented is useful in identifying whether data tampering has happened and where the data is tampered. We would like to emphasize that this is not a tamper proof solution i.e. we cannot protect the RFID tags from being tampered, however if the RFID tag is tampered we can ascertain (to a certain extent) that tampering has happened. We assume that in normal scenario the most likely location where tampering would happen is the EPC Manager (EM) or the Object Class (OC) partition. This is because we assume that the motivation behind tampering would be to disguise a product against another (for e.g. cheaper shipping cost or other malicious intentions) and this can only be done if the details with EM or OC are modified but not the Serial Number. As a result we embed the secret value in the serial number and hence we can identify whether EM or OC is tampered. However we still consider the last possibility i.e. in case the serial number is tampered, to ascertain that we hash the value of original serial number SN_{org} and hide it in the newly generated serial number SN_{app}. The proposed technique offer a binary result i.e. it can precisely tell that the tampering has happened in EM or OC or SNapp but it is not possible to ascertain whether the tampering was in EM or in SNapp or (OC or SNapp). However the mere fact that there is an inconsistency between 'EM and A' or 'OC and B' is enough to detect tampering.

4 Conclusions

In this paper, we addressed the security issues in RFID deployment; we offered a solution to detect data tampering. We found majority of recent research work in RFID security assumes the ubiquitous deployment of next generation RFID technology with excessive computing capability, however the cost associated with such RFID's is very high. We proposed a new layer in the RFID middleware architecture to detect data tampering. We also gave a detailed description of the tamper detection algorithm which can detect and identify whether and what data is tampered on the RFID tags.

References

1. Molnar, D., Wagner, D.: Privacy and security in library RFID: Issues, practices, and architectures. In: Conference on Computer and Communications Security – CCS, ACM Press, (2004) 210-219
2. Hennig, J. E., Ladkin, P. B., Siker, B.: Privacy Enhancing Technology Concepts for RFID Technology Scrutinized. (2005)
3. Weis, S., A., Sarma, S., E., Rivest, R., L., Engels, D., W.: Security and Privacy Aspects of Low-cost Radio Frequency Identification Systems. In D. Hutter et al. (eds.): Security in Pervasive Computing, Lecture Notes in Computer Science, Vol. 2802. Springer-Verlag, Berlin Heidelberg New York (2003) 201-212.
4. Claburn, T., Hulme, G., V.: RFID's Security Challenge- Security and its high cost appears to be the next hurdle in the widespread adoption of RFID. In InformationWeek, Accessed onNov. 15, 2004 URL: http://www.informationweek.com/story/showArticle.jhtml?articleID =52601030
5. Kevin Chung, "Relational Check Code" Awaiting US Patent.
6. CAST Inc. AES and SHA-1 Crypto-processor Cores http://www.cast-inc.com/index.shtml
7. Gao, X., Xiang, Z., Wang, H., Shen, J., Huang, J., Song, S.: An Approach to Security and Privacy of RFID Systems for Supply Chain. In Proceedings of the IEEE International Conference on E-Commerce Technology for Dynamic E-Business (2003)
8. Weis, S.: Security and Privacy in Radio-Frequency Identification Devices. Massachusetts Institute of Technology (2003)
9. Henrici, D., Müller, P.: Tackling Security and Privacy Issues in Radio Fre-quency Identification Devices. In A. Ferscha and F. Mattern (eds.): PERVASIVE 2004, Lecture Notes in Computer Science, Vol. 3001. Springer-Verlag, Berlin Heidelberg New York (2004) 219-224.

Measuring the Histogram Feature Vector for Anomaly Network Traffic

Wei Yan

McAfee Inc.
wyan@avertlabs.com

Abstract. Recent works have shown that Internet traffics are self-similar over several time scales from microseconds to minutes. On the other hand, the dramatic expansion of Internet applications give rise to a fundamental challenge to the network security. This paper presents a statistical analysis of the Internet traffic Histogram Feature Vector, which can be applied to detect the traffic anomalies. Besides, the Variant Packet Sending-interval Link Padding based on heavy-tail distribution is proposed to defend the traffic analysis attacks in the low or medium speed anonymity system.

1 Introduction

A number of recent measurements and studies demonstrated that real traffic exhibits statistical self-similarity and heavy-tail [PS1]. In the heavy tail distribution, most of the observations are small, but most of the contribution to the sample mean or the variance comes from the few large observations [PW1]. In this paper, we statistically analyze the histograms of ordinary-behaving bursty traffic traces, and apply the Histogram Feature Vector (HFV) to detect and defend the intrusion attacks. Given a traffic trace including some bins, HFV is a vector composed of the histogram frequency of every bin. Since the self-similar traffic traces follow the heavy tail distribution, and the bulk of the values being small but with a few samples having large values, their HFVs present left skewness. On the other hand, based on heavy-tail distribution, we propose Variant Packet Sending-interval Link Padding Method (VPSLP) to defend against the traffic analysis attacks. VPSLP shapes the outgoing packets transmitting within the anonymity network, and makes the packet sending-interval time follow a mixture of multiple heavy tail distributions with different tail index.

The rest of the paper is organized as follows. Section 2 briefly introduces the definition of self-similarity, and Section 3 describes how to generate HFV. HFV-based method to detect self-similar traffic anomalies is introduced in section 4. In section 5, VPSLP is expounded to defend the traffic analysis attacks, and section 6 is the conclusion.

2 Self-similarity Traffic Model

A number of empirical studies [PS1,PW1] have shown that the network traffic is self-similar in nature. For a stationary time series $X(t), t \in \Re$, where $X(t)$

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 279–284, 2005.
© Springer-Verlag Berlin Heidelberg 2005

is interpreted as the traffic at time instance t . The aggregated X^m of $X(t)$ at aggregation level m is defined as $X^m(k) = \frac{1}{m}\sum_{i=km-(m-1)}^{km} X(t)$. That is, $X(t)$ is partitioned into non-overlapping blocks of size m, their values are averaged, and k indexes these blocks. Denote $r^m(k)$ as the auto-covariance function of $X^m(k)$. $X(t)$ is called self-similar with Hurst parameter $H(0.5 < H < 1)$, if for all $k \geq 1$, $r(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})$. Self-similarity has two important features: heavy-tailed distribution and the multi-time scaling nature. A statistical distribution is heavy-tailed, if $P[X > x] \sim x^{-\alpha}$ where $0 < \alpha < 2$. The Pareto distribution is a simple heavy tailed distribution with probability mass function defined as $p(x) = \alpha k^{\alpha} x^{-\alpha-1}, 1 < \alpha < 2, k > 0, x \geq k$ where α is the tail index, and k is the minimum value of x. Multi-time scaling nature means that $X(t)$ and its time scaled version $X(at)$, after normalizing, must follow the same distribution. That is, if $X(t)$ is self-similar with Hurst parameter H, then for any $a > 0, t > 0$, $X(at) =_d a^H X(t)$, where $=_d$ stands for equality of first and second order statistical distributions and a is called the scale factor.

The work in [MB1] showed the aggregation or superimposition of WWW traffics does not result in a smoother traffic pattern. Here we focus on the aggregated bursty traffic on the edge devices of high speed networks, such as edge routers and edge switches. Compared to the traffic nearby the source node, the high-speed aggregated traffic at the edge network devices is relatively more stable. Besides, the aggregated traffic is easy to be extracted from the whole network at those edge devices, without much influence on the whole network's performance.

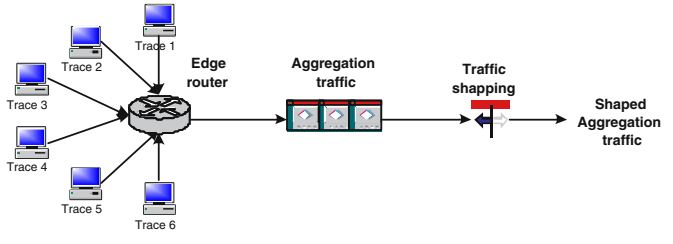


Fig. 1. Self-similar traffic aggregation

In Fig. 1, six incoming traffic traces were generated, and these input traces are aggregated at the edge device. After going through the queuing FIFO buffers, shaped aggregated traffic is forwarded to the high speed network. In our simulation, the input aggregated traffic H is 0.832, which approaches to the maximum H value of the input traces. The output aggregated traffic H is 0.815, which means that simply aggregating self-similar traffic traces at the edge devices will not decrease burstiness sharply. However, after going through the traffic shaping, H decreases to 0.764.

3 Histogram Feature Vector

In this section, we defined HFV of self-similar traffic based on two important features of self-similarity: the multi-time scaling nature and heavy-tailed distri-

bution. A heavy-tailed distribution gives rise to large values with non-negligible probability so that sampling from such a distribution results in the bulk of small values but with a few large values. Multi-time scaling nature means that self-similar trace $X(t)$ and its aggregated traffic at time scale "at",$X(at)$, follow the same first order and second order statistical distribution. Afterwards, $X(t)$ is divided into k bins, and HFVs of these bins and $X_r(t)$ are calculated. The optimum bin number k has an important effect on HFV. If k is too small, it will make the histogram over-smoothed, while an excessively large k will not correctly reflect the traffic changes. Sturges [Sh1] used a binomial distribution to approximate the normal distribution. When the bin number k is large, the normal density can be approximated by a binomial distribution $B(K-1,\ 0.5)$, and the histogram frequency is $\binom{a}{k}$. Then the total number of data is $n = \sum_{i=0}^{k-1} \binom{k-1}{i}$. According to Newton's Binomial Formula:

$$(x+y)^n = \sum_{k=0)}^{n} \binom{n}{k} x^{n-k} y^k \tag{1}$$

Let $x = 1$ and $y = 1$, leading to $n = \sum_{i=0)}^{k-1} \binom{k-1}{i} = 2^{k-1}$. Therefore, the number of bins to choose when constructing a histogram is $k = 1 + \log_2 n$. Doane [DD1] modified Sturges' rule to allow for skewness. The number of the extra bins is:

$$k_e = \log_2(1 + \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^3}{(n-1)\sigma^3}}) \tag{2}$$

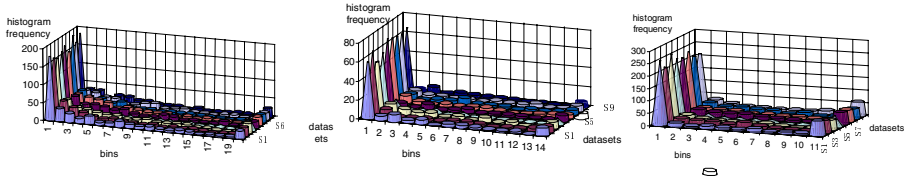
However, because of the normal distribution assumption, Sturges' rule and Doane still do not always provide enough bins to reveal the shape of severely skewed distribution(i,e., self-similarity). Therefore, they often over-smooth the histograms. We modified the Doane's equation to be:

$$k = 1 + \log_2 n + 4H \log_2(1 + \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^3 n}{(n-1)\sigma^3}}) \tag{3}$$

In our simulations, the traffic traces come from two sources: traffic trace 4 mentioned in section 2 and the aug89 trace from [LBL1]. Fig. 1 shows HFVs of these data sets with the left skewness distributions. Due to heavy-tail distribution, a large number of small values locate in first three bins, whereas few large values locate in the rest bins(the relative high histogram frequency in last bin is caused by the large bin size).

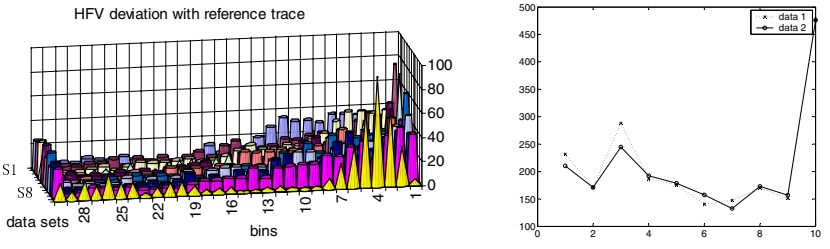
4 Traffic Anomaly Detection

In this section, HFV is applied to detect the self-similar traffic anomalies by comparing the distribution deviation between traffic sets with reference trace. Fig. 2 shows a segment packet inter-arrival time trace $X(t)$ of the input aggregated traffic at edge switch described in section 3, which includes the abnormal



(a) HFV of traffic trace 4 (b) HFV of 1999 DARPA traffic trace (c) HFV of aug89 traffic trace

Fig. 2. HFVs of different datasets



(a) Traffic HFVs with reference trace (b) HFV Deviations

Fig. 3. HFV traffic anomaly detecion

traffic. Total bin number is calculated as 30, and aggregate traffic on time scale $30t$, $X(30t)$, is build up. The reference trace is $X_r(t) = 30^{-H}X(30t)$, and the whole time series is divided into 30 data segments. For each segment, the histogram frequency is calculated, and HFV includes 30 histogram frequencies. We compare the HFV of $X(t)$ with $X_r(t)$ by deviation. Fig. 3 presents the HFV of last 10 bins(21^{th} to 30^{th}) and reference model. The traffic anomaly exists in the 30^{th} bin, with large distinction. Fig. 3(a) shows that the traffic anomaly exists in the 30th data set, which is distinctly different from the reference model.

5 Link Padding

In this section, we propose VPSLP to defend against the traffic analysis attacks. Link encryption is not enough to protect the anonymity of the system. The attacker can eavesdrops the links before and after the ingress anonymizing node, and collects the number of the packets transmitted or the packet inter-arrival time. After comparisons, the attacker can derive the route between the client and the server, and launch attacks on the nodes. The countermeasure against traffic analysis is to use link padding where the cover traffic and the real traffic are mixed so that every link's total traffic looks constant or similar to the attackers.

Consider the link between a client and a server in anonymizing system. On the client side, there exist two kinds of buffers: the traffic buffer and the constant length buffer. The traffic buffer stores the incoming packets, and is large enough without overflow (since the link transmission rate in the anonyminity system is

not very high). The constant length buffer sends the packets exactly according to heavy tail distribution(Pareto distribution). Every value from the heavy tail distribution can be treated as a timer. If the timer does not expires, the constant length buffer will hold the packet, otherwise, the packet is sent right away.

Let the constant length buffer's length be l. If the incoming packet's length is bigger than l, the sending node splits the packet into several segments. We assume the smallest value in the heavy tail distribution is larger than the time to fill up the constant length buffer, and the time to padded the whole constant length buffer can be ignored. When the constant length buffer is filled up, it sends out the packet, and fetches the next value from the heavy tail distribution. The cover traffic is generated under two conditions. First, If the sum of the incoming packet size and the total size of the packets in the buffer is greater than l, then the cover traffic is padded to fill up the constant length buffer. If the timer does not expire, the buffer will hold until the timer expires. Otherwise, the buffer sends the packet out immediately. Second, if the constant length buffer is not padded up and the timer has already expired, the cover traffic is padded to fill up the buffer, and then the packet is sent out.

VPSLP does not generate the cover traffic all the time, but based on the incoming traffic and the generated heavy tail distribution. Links with the same starting node are called the node's link group. Every link group uses the same value of a and k to generate the heavy-tail distribution. If the control unit detects the traffic change on any link within the link group, lets $l = \mu$, and $k = \mu - \gamma$. Initially, the three Pareto distributions ($\alpha = 1.3, 1.5$, and 2.0) are mixed with equal probability. They are modified based on the change of the traffic burstiness degree. Let P_1, P_2, and P_3 are mixed probabilities of the link0, link1 and link2 respectively,

$$P_i = \begin{cases} \frac{H}{A_i}, & i = j \\ 1 - \frac{(H - A_j)}{2}, & i \neq j \end{cases} \; i = 1, 2, 3. \qquad j = \begin{cases} 1, 0.5 \leq H < 0.75 \\ 2, 0.75 \leq H < 0.85 \\ 3, 0.85 \leq H \end{cases}$$

We simulated the scenario as shown in Fig. 4. The links from node A to node B, node A to node C, and node A to node D, are labeled as link0, link1, and link2 respectively. Traffic datasets from [LBL1] are used. With link padding, the traffic throughput patterns of all the links are statistically similar. As shown in Fig. 5, unlike the original traffic without the link padding, the total traffic throughput patterns and HFVs of all the links are statistically similar with VPSLP.

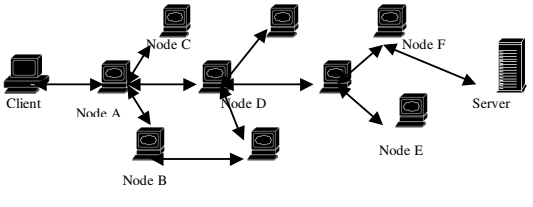
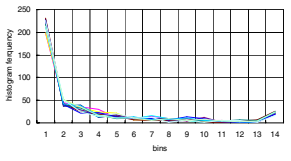


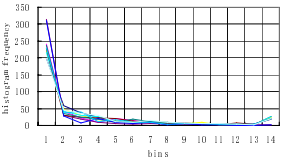
Fig. 4. Link padding in anonymity system

Table 1. Descriptive statistics of link traffic w/o link padding

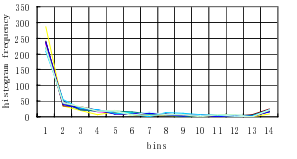
Link	Mean	StDev	Q1	Q3	Entropy
before link padding					
Link 0	492	295	300	650	-
Link 1	379	243	200	550	-
Link 2	259	200	100	350	-
after link padding					
Link 0	423	136	350	500	8.978
Link 1	417	141	350	500	8.926
Link 2	375	132	300	450	8.802



(a) Link0 HFV



(b) Link1 HFV



(c) Link2 HFV

Fig. 5. HFVs of link0, link1, and link2

6 Conclusion

In this paper, we statistically analyze HFV of ordinary-behaving bursty traffic traces, and apply HFV to detect the traffic anomalies in high speed netowrk, and defend the traffic attacks in the media speed in the anonymity system.The simulation results showed that our work can detect the traffic anomalies and defend against traffic analysis attacks efficiently.

References

1. Paxson V., and Floyd S.: Wide-area traffic: the failure of Poisson modeling. Proceedings of ACM Sigcomm (1994) 257–268
2. Park K., and Willinger W.: Self-similar network traffic and performance evaluation. John Wiley & Sons Inc (2000) 17–19
3. Mark E., and Bestavroe A.:, Explaining World Wide Web Traffic Self-Similarity. Technical Report TR-95-015 (1995)
4. Sturges H.: The choice of a class-interval. The American Statistical Association. **21** (1926) 65–66
5. http://ita.ee.lbl.gov/html/traces.html
6. Doane, D.: Aesthetic frequency classification. American Statistician, **30** (1976) 181-183

Efficient Small Face Detection in Surveillance Images Using Major Color Component and LDA Scheme

Kyunghwan Baek, Heejun Jang, Youngjun Han, and Hernsoo Hahn

School of Electronic Engineering, Soongsil University, Dongjak-ku, Seoul, Korea
{khniki, sweetga}@visionlab.ssu.ac.kr,{young, hahn}@ssu.ac.kr

Abstract. Since the surveillance cameras are usually covering wide area, human faces appear quite small. Therefore, their features are not identifiable and their skin color varies easily even under the stationary lighting conditions so that the face region cannot be detected. To emphasize the detection of small faces in the surveillance systems, this paper proposes the three stage algorithm: a head region is estimated by the DCFR(Detection of Candidate for Face Regions) scheme in the first stage, the face region is searched inside the head region using the MCC(major color component) in the second stage and its faceness is tested by the LDA(Linear Discriminant Analysis) scheme in the third stage. The MCC scheme detects the face region using the features of a face region with reference to the brightness and the lighting environment and the LDA scheme considers the statistical features of a global face region. The experimental results have shown that the proposed algorithm shows the performance superior to the other methods' in detection of small faces.

1 Introduction

Many methods have been developed to face detection in surveillance system. They can be classified into four categories: (a) knowledge-based, (b) feature invariant, (c) template matching and (d) appearance-based approaches. Knowledge-based methods are rule-based approaches which try to model intuitive knowledge of facial features. A representative work is the multi-resolution rule-based method[1]. Feature invariant methods aim to find structural features to locate faces even when the pose, the viewpoint, or the lighting conditions change. Space Gray-Level Dependence (SGLD) matrix of face pattern[2] and skin color[3] belong to this class. Template matching methods use several standard patterns to describe the whole face or the facial features separately. The correlations between an input image and the stored patterns are computed for the detection[4]. Appearance-based methods highlight relevant characteristics of the face images by machine learning from a training set. The methods based on the eigenface[5] are typical techniques belonging to this category.

The above approaches have their own advantages and drawbacks depending on the applications. In some applications, their drawbacks are prevalent than the advantages. When a facial image is small, then it is very difficult to extract the enough information on the shape, color, facial features, etc. those are required by the conventional approaches.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 285 – 290, 2005.
© Springer-Verlag Berlin Heidelberg 2005

We propose a new face detection algorithm that has strong in detecting small faces appearing in surveillance images. It searches the possible head region first using the DCFR(Detection of Candidate for Face Regions)[6] and then the MCC(Major Color Component) detection scheme. The MCC region is selected by taking into account the brightness variation depending on the lighting environments as well as the red color component which appears the strongest in the skin regions. The faceness of the MCC region is tested by the LDA(Linear Discriminant Analysis) scheme. If the MCC region does not pass the faceness test of the LDA scheme, then it is fed back to the MCC detection process and its region is searched again. The proposed algorithm is visually summarized in Fig. 1.

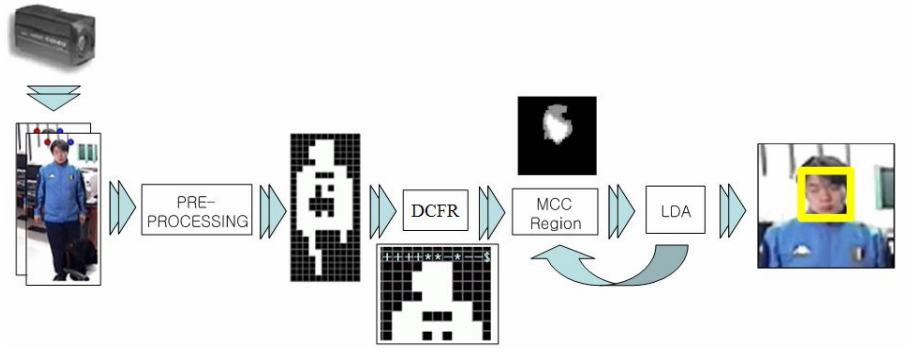


Fig. 1. Flow of the proposed face detection algorithm

2 Detection of the MCC Region in the Head Region

Once the head region is determined, the face region is searched inside it with two schemes based on the MCC and the LDA. Two face regions detected by the two schemes are fused to determine the final face region.

In order to extract the major color component in the head region, three statistical features of a head region are considered here. The first one is that the red component appears in general stronger than the green and blue components in a face when using RGB color space, although there is a slight change in its magnitude depending on the

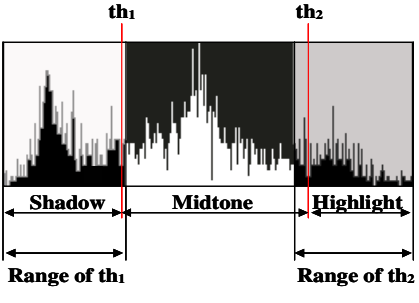
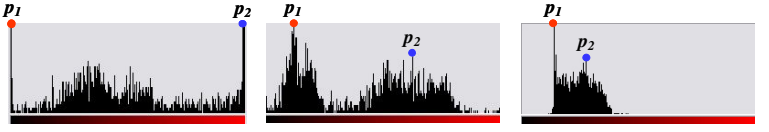


Fig. 2. The histogram of head region

lighting conditions. The second one is that the brightness of a pixel in a head region can be classified by one of the three groups, H(Highlight), M(Midtone), and S(Shadow) as shown in Fig. 2, using the histogram of the head region. The two thresholds dividing the brightness range into three groups are determined by using the Ping-Sung Liao multi-threshold algorithm[7].

The third one is that the brightness of a pixel changes depending on the lighting conditions which can be classified into three cases: N(Normal), B(Backside), and P(Polarized) lighting environments. Depending on the lighting conditions, the thresholds of dividing H, M, and S vary and the relative ratio of red component to the other color components can be changed.

To determine under which lighting environment the head region is obtained, the strength of red component is analyzed. It is confirmed that the red color histogram of a head region has the different distributions depending on the lighting environments. For example, Fig. 3 shows the typical distributions of the red color histogram of the head region, obtained under three lighting environments. Each figure in Fig. 2 is obtained from statistically analyzing 210 histograms of the head region under the same lighting environment.



(a) Normal (b) Backside or Dark (c) Rolarized

Fig. 3. Typical distributions of the red color histograms obtained under three lighting environments

Using the characteristics of the distributions of the three typical histograms, respectively representing three lighting environments, the lighting environment of the head region is determined based on the Median(μ) and by the standard deviation(σ) of the distribution as follows:

$$LightEnvironment = \begin{cases} Normal & , 120 \le \mu \le 140,\ 50 \le \sigma \le 80,\ 70 \le d \le 130 \\ Backside\,or\,Dark & ,\ 50 \le \mu \le 70\ ,\ 10 \le \sigma \le 40,\ 10 \le d \le 50 \\ Polarized & , 125 \le \mu \le 145,\ 70 \le \sigma \le 90,\ 140 \le d \le 210 \end{cases} \quad (1)$$

$$d = \left| p_i - p_j \right|$$

where p_i and p_j represent the two peak points in the histogram.

Before extracting the MCC in the head region based on the above characteristics of the head region, for making the process more efficient, the possible non-face pixels are eliminated from the head region using the red color component once more. Since the red color appears as the strongest component in the skin pixels, only those pixels whose red component is the largest among the RGB colors are selected to be considered further. Therefore, the upper and lower limits of the head region to be processed become the same as those of the selected pixels as above. In the last column of Fig. 6, those pixels where the red color is the strongest component are displayed and the

uppermost and lowermost pixels among them are used to obtain the histogram from which the head region is divided into three groups.

Fig. 4 shows the results of two classification processes. A head region detected in the original image is classified by analyzing the red color image (given in the rightmost column in Fig. 3) into Normal, Backlight or Dark, and Polarized, and once more classified by analyzing the histogram of the pixels into Highlight, Midtone, and Shadow.

Once the face region is segmented using the two classification processes as shown in Fig. 6, then the major color component is selected as follows.

1) In Normal lighting environment, color does not change mostly in a head region. Also, most pixels in a face region are included in M group and Red component is stronger than other colors in a face area. Thus, those pixels included in both the M group and Red component region are merged as the MCC region.
2) In Backside or Dark and Polarized lighting environment, the M region overlapped with the Red component region is smaller than in N environment. In the polarized case, the color of the portion normal to the light source is white but that of the other potions is mostly back. Consequently, the MCC region is determined by merging the M and H regions. In the backside and dark case, with the same reason, the MCC region is determined by merging the M and S regions.

Fig. 4 shows the face regions detected in different lighting environments by using several features: the red color component, the conventional skin color, and the MCC.

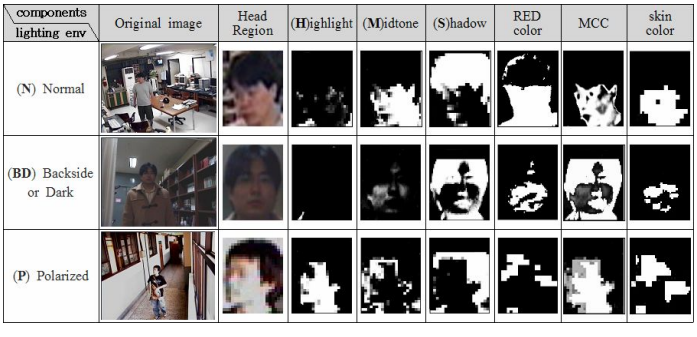


Fig. 4. Face regions detected by various color features in a head region

3 Test of Faceness Using the LDA Scheme

The LDA, which classifies a group of data into the classes where each of them has the minimum within-class variance and the maximum between-class scatter variance[8], is used to find a face inside the head region. It projects an input image with a large dimension into a new vector with a much lower dimension where it can be easily classified. In the training mode, the optimal class coefficient matrix A consisted of q eigen-vectors is obtained from using the training sets of faces and non-faces. All the images in the set of faces are transformed by A to generate the new feature vectors with a reduced dimension, using Eq. (6).

$$y = Ax \tag{6}$$

In here, x is a column vector of the input image normalized to the 12×12 size and y is the new feature vector. The set of the faces are represented by the representative feature vector y_{face} which is the average of the feature vectors of the faces. In the same way, $y_{nonface}$ is obtained. In the testing mode, if a new image is provided, then its class is determined by Eq. (7) in the following:

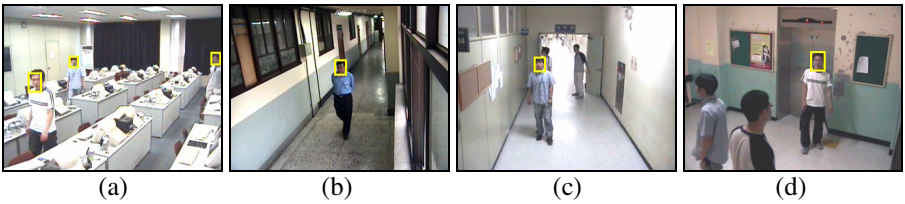
$$x = \begin{cases} face & , \ if \ D_0 > D_1 \\ nonface & , \ if \ D_1 > D_0 \end{cases} \tag{7}$$

Where $D_i = \left|(y - y_i)^2\right|$ and i is face or nonface.

4 Experimental Results

In experiments, the proposed small face detection method was evaluated by using video scenes captured by a camera that acted as a surveillance camera. Video scenes were captured at various places such as rooms, aisles and place in front of elevator. A digital video camera that acted as a surveillance camera was set about 2.0~2.5 meters from a floor for recording the 25 video scenes was about ten seconds. The captured video scenes were sampled of 320× 240 resolutions.

The database consists of the facial images of 300 peoples, which were obtained from various viewpoints using multiple cameras for training in LDA process. The images of small face detection are shown in Fig. 5 The reason of false negatives (Fig. 5(c), (d)) is perhaps due to their appearances from behind or low resolutions in the input image.



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

Fig. 5. Examples of face detection result

In this work, the proposed method was evaluated by using the face detection rate and the reliability of detection. The face detection rate is obtained from dividing a number of detected faces with a number of peoples in video scenes. The reliability of detection is calculated by dividing a correct detection number with a number of detected regions through all frames. The rates of face detection is 82.7%(798/962) for the reliability of detection, respectively. All of no detected faces were due to their small sizes in the images. Since the images for training were ±90° views of a face, the back views were not included.

5 Conclusions

This paper has dealt with the problems which have hindered the successful detection of small faces in surveillance systems: the first one is that their color information changes easily by the illumination conditions and the second one is that it is difficult to extract the facial features. As a solution, two stage approach is suggested. In the first stage, the head region is firstly estimated by using the DCFR scheme which analyzes the boundary shape. In the second stage, the face is searched respectively using the MCC detection scheme and the LDA scheme, and then their results are fused to return the face region. In the MCC scheme, the MCC is selected by analyzing the color distributions and the lighting conditions in the head region and the region with the MCC is considered as the face region. In the experiments, those small faces with less than 20×20 but larger than 12×12 pixels are considered, which are collected in various lighting conditions. The proposed algorithm has shown superior performance to the other's performances in detection of small faces.

References

1. Yang G., and Huang, T.S.: Human Face Detection in Complex Background. Pattern Recognition, Vol. 27. no.1. (1994) 53–63
2. Dai, Y., Nakano, Y.: Face-texture model based on SGLD and its application in face detection in a color scene. Pattern Recognition, Vol. 29. (1996) 1007–1017
3. Chai, D., Bouzerdoum, A.: A Bayesian approach to skin color classification in Ycbcr color space. IEEE region Ten Conference. Kuala Lampur Malaysia, Vol 2. (2000) 421–424
4. Lanitis, A., Taylor, C.J. and cootes, T.F.: An automatic face identification System Using Flexible Appearance Models. Image and Vision Computing, Vol. 13. no.5. (1995) 393–401
5. Vaillant, R., Monrocq, C. and Le Cun, Y.: An original approach for the localization of objects in images. IEEE Proc. Vision. Image and Signals Processing, Vol. 141. (1994) 245–250
6. Soohyun Kim, Sunghyun Lim, Boohyung Lee, Hyungtai Cha, Hernsoo Hahn.: Block Based Face Detection Scheme Using Face Color and Motion Information, Proceedings of SPIE Volume 5297, Real-Time Imaging Ⅷ, p78-88, Jan. 2004.
7. Ping-Sung Liao, Tse-Sheng Chen, Pau-Choo Chung.: A Fast Algorithm for Multilevel Thresholding. Information Science and Engineering, Vol. 17. (2001) 713–727
8. Fukunaga, K.: Introduction to Statistical Pattern Recognition (2nd ed.), Academic Press, 1990.

Fast Motion Detection Based on Accumulative Optical Flow and Double Background Model*

Jin Zheng, Bo Li, Bing Zhou, and Wei Li

Digital Media Lab, School of Computer Science & Engineering, Beihang University, Beijing, China
zhengjin@cse.buaa.edu.cn, boli@buaa.edu.cn

Abstract. Optical flow and background subtraction are important methods for detecting motion in video sequences. This paper integrates the advantages of these two methods. Firstly, proposes a high precise algorithm for optical flow computation with analytic wavelet and M-estimator to solve the optical flow restricted equations. Secondly, introduces the extended accumulative optical flow and also provides its computational strategies, then obtains a robust motion detection algorithm. Furthermore, combines a background subtraction algorithm based on the double background model with the extended accumulative optical flow to give an abnormity alarm in time. All obvious proofs of experiments show that, our algorithm can precisely detect moving objects, no matter slow or little, preferably solve the occlusions as well as give an alarm fast.

1 Introduction

Motion detection has important application in many fields, such as pattern recognition, image understanding, video coding, safety surveillance etc. Especially in the field of safety surveillance, it is the key of abnormity detection and alarming in video sequences. Effective and accurate motion detection can be used to improve the system intelligence and realize unmanned surveillance.

Optical flow[1,2] and background subtraction[3] are two methods for detecting moving objects in video sequences. The method of statistical model based on the background is flexible and fast, but it can not detect slow or little motion, also, it can not track the objects effectively; The method based on optical flow is complex, but it can detect the motion even without knowing the background, and can deal with the tracking easily.

In the paper, an integrated optical flow computation method is represented, which makes full use of the relativity between each wavelet transform coefficient layer to compute the motion vector on the supporting sets of analytic wavelet filters[3,4]. With the use of the robust estimation(M-estimation) to restrain the big error items in the restricted gradient equations, the method minimizes the accumulative error in the optical flow computation from one layer to another. To reduce the probability of false alarm and detect the slow moving or little objects, it introduces the extended accumu-

* This work is supported by the NSFC(63075013), the Huo Ying Dong Education Foundation (81095).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 291 – 296, 2005.
© Springer-Verlag Berlin Heidelberg 2005

lative optical flow, which acquires a pixel's moving distance through adding multi-frame optical flow vectors. Considering the special problems in the accumulative optical flow computation, it puts forward the corresponding computational strategies, finally gets a robust motion detection algorithm.

Moreover, for the real-time requirement, a fast motion detection algorithm integrating background subtraction with optical flow method is provided. It uses background subtraction to detect moving objects in time, while using optical flow to eliminate false alarm.

2 Multi-scale Model and Wavelet Optical Flow Computation

The multi-resolution method adopts different filters to compute an image's optical flow from different frequency bands and different directions. Firstly, it applies the wavelet decomposition to the original image, from the finer layer to the coarser layer, and constructs the multi-resolution and multi-layer image. The low-frequency component on the finer layer is decomposed, which is used to construct the image on the coarser layer, and produces the multi-resolution structure. Secondly, within the neighboring frames, it computes the optical flow on each frequency band of each layer from the coarser layer to the finer one, and the result of the coarser layer computation is treated as the initial value of the finer layer optical flow computation.

In the paper, the multi-resolution computational framework uses multi-filter whose supporting sets are proportionate to 2^{-j}. Through combining the computational results of different scales filters, it solves the magnitude problem of the supporting sets preferably. Bernard has proved that analytic wavelet resolves real wavelet's parameter surge. It combines the character of frequency and phase, and helps to improve the accuracy of optical flow estimation. Let$\psi^n(\Delta)$ denote the analytic wavelet function which is produced via the parallel motion of $(\psi^n)_{n=1\ldots N} \in L_2(R_2)$. Here, I denotes the intensity value, $V=(v_1,v_2)^T$ is the optical flow vector. The optical flow restrain equation is $AV=b$, where

$$A = \begin{bmatrix} I \bullet \dfrac{\partial \psi^1(\Delta)}{\partial x} & I \bullet \dfrac{\partial \psi^1(\Delta)}{\partial y} \\ \vdots & \vdots \\ I \bullet \dfrac{\partial \psi^N(\Delta)}{\partial x} & I \bullet \dfrac{\partial \psi^N(\Delta)}{\partial y} \end{bmatrix}_{N\times 2} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}, \quad b = \begin{bmatrix} \dfrac{\partial}{\partial t} I \bullet \psi^1(\Delta) \\ \vdots \\ \dfrac{\partial}{\partial t} I \bullet \psi^N(\Delta) \end{bmatrix}. \tag{1}$$

Because the minimal square estimation is sensitive to the system error, instead M-estimator is introduced. Unlike the minimal square estimation, M-estimator restrains the weight that leads to the bigger error, and improves the accuracy of optical flow estimation. At the same time, when the coarser layer estimation has bigger error, multi-resolution optical flow computation can use the finer layer to estimate the optical flow accurately, and won't be influenced by the coarser layer.

3 The Extended Accumulative Optical Flow

The difficulty of motion detection is how to eliminate the noise disturbance, and detect the slow or little object motion. In general, the moving direction of a truly signifi-

cant motion primarily holds consistently in a period time, while the direction of the insignificant motion changes constantly, and its motion distance may not exceed a certain scope. Through accumulating multi-frame optical flow vectors, one can not only compute each point's moving distance cursorily, but also needn't set the parameters beforehand, such as the object' size and the intensity etc. The idea of the accumulative optical flow comes from the article [5], but we have better strategies, which avoid the [5]'s blindness in choosing the extended optical flow, and improve the capability of detecting optical flow. The flow of accumulative optical flow computation is shown in Fig 1. The corresponding symbols are explained as follows.

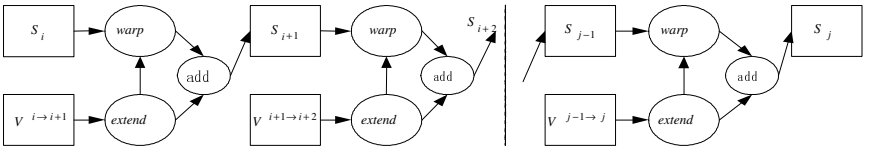


Fig. 1. The computational process of the accumulative optical flow

A warp can be denoted as $warp(I(p,t),V(P))=I(p',t)$, where $p'=p+V(p)$, $I(p,t)$ is the intensity of point p at time t, $V(p)$ is the optical flow vector at each pixel location p. The *warp* of the whole image can be denoted as $I(t)^{t+1}=warp(I(t),V^{t\to t+1})$.

To improve the significant optical flow, the algorithm uses the extended flow $E^{j-1\to j}$ to substitute the practical flow $V^{j-1\to j}$. The extended strategies are similar to those in article [5], except that to avoid the point not on the moving track being regarded as the current point. When the point to be extended gets its maximal accumulative optical flow, its intensity is similar to the current point's intensity, that is

$$\left| I_m(p) - I_m(p + s \cdot V^{j-1\to j}(p)) \right| < T_I \quad . \tag{2}$$

Where I_m is the intensity when the point gets its maximal accumulative optical flow, and T_I is the threshold of the intensity fluctuation, s is a scalar. After the process, gets the extended optical flow: $E^{j-1\to j}=s \bullet V^{j-1\to j}$.

The accumulative optical flow $S^{i\to j}$ can be defined as

$$S^{i\to j} = \begin{cases} V^{i\to i+1} & if \quad j = i+1 \\ V^{j-1\to j} + warp\ (S^{i\to j-1}, V^{j-1\to j}) & if \quad j > i+1 \end{cases} . \tag{3}$$

Because of the existence of the occlusions and the movement of the non-rigid, and the influence of the branch swaying, the computed optical flow always has some errors. Moreover, along with the time goes, the accumulative optical flow can't keep the proper value, the objects can't be tracked accurately. Therefore, substitutes $V^{j-1\to j}$ with the extended accumulative optical flow $E^{j-1\to j}$, this paper uses the add strategy.

(1) If $|V^{j-1\to j}(p)+V^{j\to j-1}(p+V^{j-1\to j}(p))|>k_c$, $V^{j-1\to j}(p)=0$. k_c is a predefined threshold.

(2) If $m_x=M_{j-1,x}(p+E^{j-1\to j}(p))$, where M_j is the maximal accumulative optical flow.

$$M_{j,x}(p) = \begin{cases} S_{j,x}(p), & if \quad sign\ (S_{j,x}(p) = sign\ (m_x) \quad and \quad \left|S_{j,x}(p)\right| > \left|m_x\right| . \\ m_x, & else \end{cases} \tag{4}$$

$$S'_{j,x}(p) = \begin{cases} 0 & if \quad |M_{j,x}(p)| > k_s \quad and \quad |S_{j,x}(p) - M_{j,x}(p)| / M_{j,x}(p) > k_r \\ S_{j,x}(p) & else \end{cases}. \tag{5}$$

(3) If two objects move across, it should avoid regarding the point outside the current moving track as the legal point. If $M_{j,x}$ is updated, I_m should be updated using the point's current intensity too, else it should be kept the former intensity.

$$I_m(p) = \begin{cases} I(p) & \begin{array}{l} if \quad sign(S_{j,x}(p)) = sign(m_x) \quad and \quad |S_{j,x}(p)| > |m_x| \\ or \quad sign(S_{j,y}(p)) = sign(m_y) \quad and \quad |S_{j,y}(p)| > |m_y| \end{array} \\ I_m(p + E^{j-1 \to j}(p)) & else \end{cases}. \tag{6}$$

If occlusion appears, the accumulative optical flow keeps the former value, else it is updated by the current value.

$$S'_{j,x}(p) = \begin{cases} S_{j,x}(p) & if \quad |I_m(p) - I_m(p + E^{j-1 \to j}(p))| < T_I \\ S_{j,x}(p + E^{j-1 \to j}(p)) & else \end{cases}. \tag{7}$$

(4) For hostility cheat, suppose its activity area is larger than the unmeaning motion. When getting $max(S_{j,x})$, where j denotes the j^{th} frame, and x denotes the direction, the coordinate is (x^j_1, x^j_2). Consider this in a reasonable n frames.

$$abnormity = \begin{cases} true & if \quad \max_{i,j \in n} (abs(x_1^j - x_1^i) + abs(x_2^j - x_2^i)) > T \\ false & else \end{cases}. \tag{8}$$

4 Double Background Model and Abnormity Management

For real-time requirement, the paper combines background subtraction with optical flow, and sets different detection qualifications for abnormity alarm and storage. Strict qualification must be made for abnormity alarm to reject boring false alarm, while loose qualification is needed to store enough information for back-check.

The background subtraction algorithm uses two Gaussian background models, that is, quick model and slow model, and the construction of models adopts K-means clustering. Quick model uses selective update to make sure that the moving objects will not be recognized as a part of the background, while slow model uses complete update to achieve different results: it can recognize the just-stopped objects (e.g. a running car stops, then parks for a long time) as a part of the background, or renew the moving objects which has been occluded (e.g. the parking car goes away).

The method uses a two-values (0 or 1) array *Warning* to represent whether the pixel is fit for the background models. In *Warning*, when the number of the elements whose values are equal to 1 exceeds the given threshold T, abnormity alarm will occur.

Similarly, another two-values (0 or 1) array *Recording* is used to represent whether the pixel's accumulative optical flow is bigger than a certain threshold. Moreover, the open operation is used to eliminate the little block, and the number of element whose value is equal to 1 is counted. If the number is bigger than the given threshold, abnormity storage process will be started.

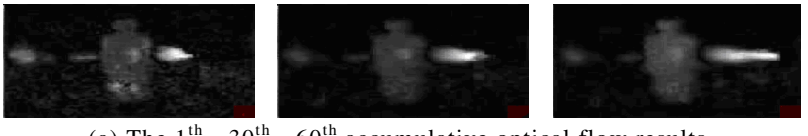
5 The Experiment of Abnormity Detection

The experiments are based on the video sequences provided by the paper [6], which are usually used to evaluate optical flow algorithm. Our method and other representational optical flow methods are compared. The correlative data of other methods comes from the articles of the correlative authors, and the blank items denote that the author didn't provide the correlative data. The targets compared are Average Error and Computation Density, which reflect the optical flow computation quality furthest. Average Error is the average angle difference between the optical flow field computed by the program and the real optical flow field of the test sequence, and Computational Density is the ratio of computational pixels, in which the bigger computation density predicates that it can provide more integrated optical flow field. A group of experimental results are shown in table 1.

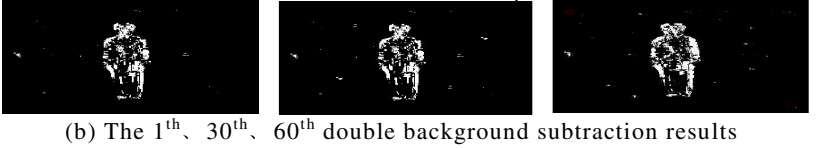
Table 1. The compare of the optical flow methods ①Image②The average error③Computation density④Magarey & Kingsbury⑤Ye-Te Wu ⑥Bernard ⑦Our method

Image①	The average error②				Density(%) ③			
	MK④	Wu⑤	Ber⑥	Our⑦	MK	Wu	Ber	Our
Yosemite	6.20	4.63	6.50	4.83	100		96.5	98.6
Translating Tree	1.32	0.85	0.78	0.67	100		99.3	99.5
Diverging Tree		2.49		2.32				88.2
Sinusoid		0.40		0.36				100

Table 1 shows our method has better results not only on computation density but also on average error, and can get perfect optical flow only using two frames images, so our method has more excellent integrated performance than the existing methods.



(a) The 1th、30th、60th accumulative optical flow results



(b) The 1th、30th、60th double background subtraction results

Fig. 2. The compare of moving objects detection methods in the complex background

Fig 2 denotes the algorithm performance in complex background. There is a person riding a bicycle is going from the left to the right, who is little. When time passes, the accumulative optical flow existing on the contrail of the rider increases. The rider leaves a light line behind his body. The standing people's motion is little, and its accumulative optical flow is little. The swaying of the branch and the grass is the background noise, in No.60 frame, its accumulative optical flow is little or disappears. The

experiments show the motion detection algorithm based on the accumulative optical flow won't be affected by the change of the object's motion speed, even can allow the object to stop for a period of time, and it can deal with the background noise better. Meanwhile, the background subtraction algorithm is fast though it can't detect the little motion. In the paper, the advantages of these two methods are combined, and the algorithm satisfies the practical requirement.

6 Conclusion

The paper integrates the advantages of optical flow and background subtraction, and presents a fast and robust motion detection and abnormity alarm algorithm. Firstly, it introduces M-estimator in the optical flow estimation algorithm to improve the reliability of the estimation, and reduces the computational error. Secondly, represents an improved motion detection algorithm based on the accumulative optical flow, which is supposed to increase the significance of the motion continuously in the moving direction, while the change of background will be restrained in a certain scope. Finally, the combination of the background subtraction algorithm based on the quick and the slow model realizes the fast and flexible motion detection. The experiments prove the algorithm can detect moving objects precisely, including slow or little objects, and solve the occlusions as well as fast give an alarm.

References

1. Iketani, A., Kuno, Y., Shimada, N.: Real time Surveillance System Detecting Persons in Complex Scenes. Proceedings of Image Analysis and Processing (1999) 1112–1115
2. Ong, E.P., Spann, M.: Robust Multi-resolution Computation of Optical Flow. Acoustics, Speech and Signal Proceedings 1996 (ICASSP-96) 1938–1941
3. Suhling, M., Arigovindan, M., Hunziker, P., Unser, M. Multi-resolution moment filters: theory and applications. Image Processing, IEEE Transactions, Vol. 13. Issue. 4. (April 2004) 484–495
4. Bernard, C.P.: Discrete Wavelet Analysis for Fast Optic flow Computation. PhD Dissertation, Ecole polytechnique (1999)
5. Wixson, L.: Detecting Salient Motion by Accumulating Directionally-Consistent Flow. IEEE Transactions on Pattern Analysis And Machine Intelligence, 22(8), (August 2000), 744–780
6. Barron, J., Fleet, D., Beauchemin, S.: Performance of Optical Flow Techniques. International Journal of Computer Vision, 12(1), (1994) 42–77

Reducing Worm Detection Time and False Alarm in Virus Throttling*

Jangbok Kim[1], Jaehong Shim[2,†], Gihyun Jung[3], and Kyunghee Choi[1]

[1] Graduate School of Information and Communication, Ajou University, Suwon 442-749, South Korea
bbok@ajou.ac.kr
khchoi@ajou.ac.kr
[2] Department of Internet Software Engineering, Chosun University, Gwangju 501-759, South Korea
jhshim@chosun.ac.kr
[3] School of Electrics Engineering, Ajou University, Suwon 442-749, South Korea
khchung@ajou.ac.kr

Abstract. One of problems of virus throttling algorithm, a worm early detection technique to reduce the speed of worm spread, is that it is too sensitive to burstiness in the number of connection requests. The algorithm proposed in this paper reduces the sensitivity and false alarm with weighted average queue length that smoothes sudden traffic changes. Based on an observation that normal connection requests passing through a network has a strong locality in destination IP addresses, the proposed algorithm counts the number of connection requests with different destinations, in contrast to simple length of delay queue as in the typical throttling algorithm. The queue length measuring strategy also helps reduce worm detection time and false alarm.

1 Introduction

Internet worms are self-generated and propagate, taking advantage of the vulnerabilities of systems on the Internet [1,2]. One of the typical behaviors of worm virus is that the viruses send a lot of packets to scan vulnerable systems. A system contaminated by a worm scans other vulnerable systems utilizing IP addresses generated in a random fashion. Then the contaminated system propagates the worm to the scanned vulnerable systems. Even though the self-propagated worms do not send out harmful data, they produce abundant traffic during the propagation process and the huge amount data slow down the network.

One of the best ways to minimize the damages to be given to the systems on the network from worm attack is to detect worm propagation as early as possible and to stop the propagation. There are many published *worm early detection techniques* [3,4,5]. The virus throttling [3] which is one of such techniques slows down or stops worm propagation through restricting the number of new connection (or session) requests.

* This research was supported by research funds from National Research Lab program, Korea, and Chosun University, 2005.

† Jaehong Shim is the corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 297 – 302, 2005.
© Springer-Verlag Berlin Heidelberg 2005

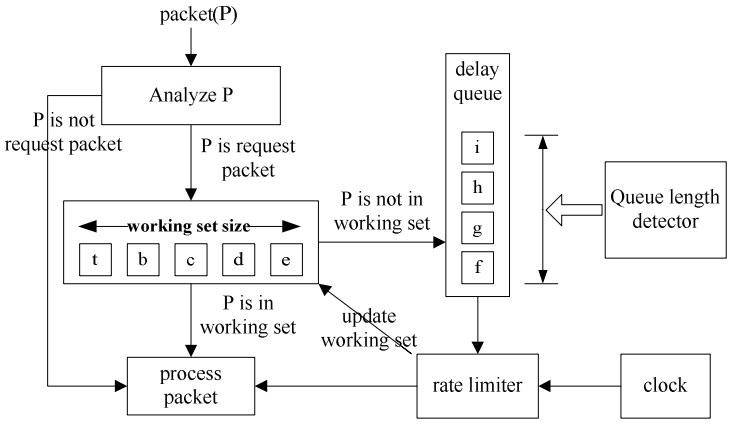


Fig. 1. Virus throttling mechanism

Fig. 1 shows the flow of packets controlled by the virus throttling. When the packet (*P*) for a new connection request arrives, the working set is searched to find an IP address matched to the destination IP address of *P*. If there is such an IP address, *P* is sent to its receiver. Otherwise, *P* is pushed into the *delay queue*. The policy lets a connection request to the host that has made a normal connection recently, pass to the client without any delay. And all other new session requests are kept into the delay queue for a while. The *rate limiter* periodically pops out the oldest packet from the delay queue depending on the period of rate limiter and sends it to the receiver. If there are more than one packet with the same IP address, the packets are also sent to the receiver. Once the packet(s) is delivered to the receiver, the destination IP address is registered in the working set. The *queue length detector* checks the length of delay queue whenever a new packet is stored in the queue. If the length exceeds a predefined threshold, a worm spread is alerted and no more packets can be delivered. This mechanism helps slow down the speed of worm propagation. This paper suggests a mechanism to shorten worm detection time as well as to reduce false alarm.

2 Proposed Worm Detection Algorithm

2.1 Measuring the Length of Delay Queue

Many other studies revealed that normal connection requests on the Internet concentrate on a few specific IP addresses. In contrast, the destination IP addresses of connection requests generated by worm are almost random. The facts mean that it is enough to count the number of different destination IP addresses in the delay queue for measuring the queue length, instead of the number of entries in the queue.

In the typical throttling, two different packets with the same destination IP address are stored in the delay queue unless the IP address is in the working set. In the case, they are counted twice for measuring the length of queue and the length increases by two. However, the increase is not helpful to detect worm since worm usually generated many connection requests with different IP addresses in a short period.

In the proposed algorithm, connection requests with the same IP address increase the length of delay queue by one. For implementing the idea, the connection requests with the same IP are stored in a linked list. To see how helpful the idea is to lower the length of delay queue and thus eventually to reduce false alarm in the algorithm, we performed a simple experiment in a large size university network. In a PC in the network, sixteen web browsers performed web surfing simultaneously. We made the traffic passes through both typical and proposed throttling algorithms and measured both *DQL* (delay queue length) and *NDDA* (the number of different destination IP addresses).

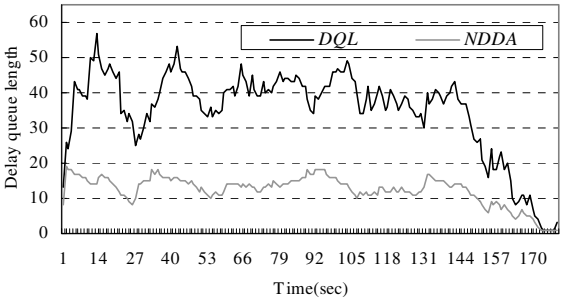


Fig. 2. *DQL* vs. *NDDA* in normal connection requests

Fig. 2 shows the change of delay queue length when using *DQL* and *NDDA*. Obviously, there is a big discrepancy between two values, which mainly comes from the fact that many requests aim to a few IP's, as expected. If the threshold was set to 50, the typical virus throttling using *DQL* alone must have fired a couple of false alarms even though there was no a single worm propagation. However the proposed algorithm using *NDDA* never produces any false alarm. The complexities of implementing *DQL* and *NDDA* are same, $O(n^2)$. Consequently, we can say that the throttling can reduce false alarms with using *NDDA* instead of *DQL* without any penalty to increase the implementation complexity.

2.2 An Algorithm Utilizing Weighted Average Queue Length

The way for the typical virus throttling to decide a worm spread is that whenever a new connection request packet is entered into the delay queue, the virus throttling checks the length of delay queue and if the length exceeds a pre-defined threshold, it decides there is a worm spread.

However, the proposed worm detection algorithm utilizes both the current queue length and the recent trend of queue length, instead of the current queue length alone. To take account of the dynamic behavior of delay queue length, the proposed technique uses *weighted average queue length* (*WAQL*). *WAQL* is calculated with the following well-known exponential average equation.

$$WAQL_{n+1} = \alpha * avgQL_n + (1 - \alpha) * WAQL_n$$

Where $avgQL_n$ is the average length of delay queue between instant n and n+1 and α is a constant weight factor. The constant α controls the contribution of past

($WAQL_n$) and current ($avgQL_n$) traffic trend to $WAQL_{n+1}$. That is, *WAQL* is a moving average of delay queue length reflecting the change in the delay queue length. If the sum of *WAQL* and the delay queue length is greater than a threshold, the proposed algorithm notifies a worm intrusion. The pseudo code looks like

```
IF ((WAQL + NDDA) > threshold) THEN
   Worm is detected.
END IF
```

Note that the typical throttling compares just *DQL* and threshold. A merit of using *WAQL* is that *WAQL* roles as an alarm buffer. In the typical throttling utilizing *DQL* alone, a sudden burstiness in normal connection requests can instantly make the length of delay queue longer than the threshold. Thus, it produces a false alarm. However, *WAQL* can reduce the sensitivity of false alarm to the sudden change in the requests.

But worm detection time will increase unless $\alpha = 0$ (normally true and $\alpha < 0.5$) if the algorithm uses *WAQL* alone since *WAQL* smoothes the change in the requests. Instead of monitoring just *DQL* as in the typical throttling, the proposed algorithm takes account of both *NDDA* and *WAQL* for speeding up worm detection time. In case that a worm intrusion happens, *WAQL+NDDA* reaches to the threshold as nearly same as or faster than *DQL* alone does. Thus, the proposed algorithm can smell worm propagation as fast as the typical throttling does. On the other hand, the slower a worm propagates, the greater *WAQL* becomes (from the equation $WAQL_{n+1}$). Thus, *WAQL+NDDA* grows faster while worms propagate slowly. Consequently, the proposed algorithm using *WAQL+NDDA* can detect worm propagation faster than the typical throttling algorithm.

3 Performance Evaluation

For the empirical study to see the performance, both the typical throttling and the proposed algorithms were implemented in a PC running Linux, which was connected to a campus network through a firewall. The firewall included several rules to protect the systems on network from harmful packets generated by the worm generator.

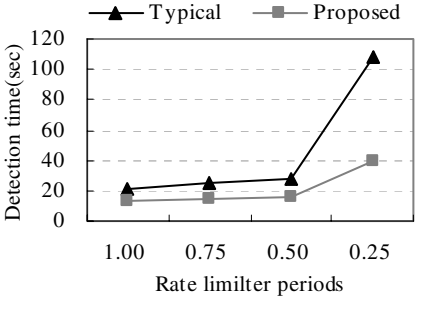


Fig. 3. Worm detection time in different periods of rate limiter

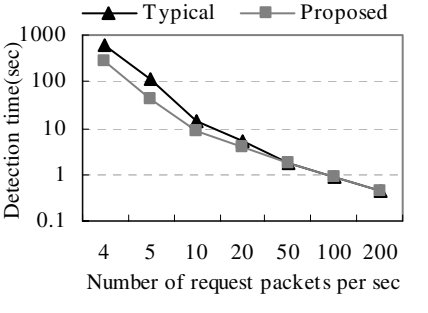


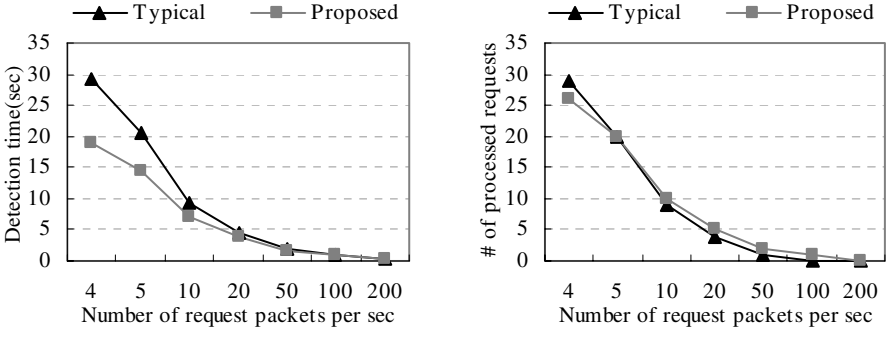
Fig. 4. Worm detection time in different number of request packets per sec

The numbers of packets generated by worms are different depending on virus type or environment. We made the packet generator produce worm packets in different rates. During the study, the sampling time of weighted average queue length is set to one second, $\alpha = 0.2$, threshold = 100 and the working set size was set to 5 as in [3].

Fig. 3 compares worm detection times by the two algorithms. The period of rate limiter varies from 1 to 0.25 sec. The worm generator transmits five connection request packets per second with all different IP addresses. The proposed algorithm detects worm spread faster than the typical throttling at all periods. When the period is shortened from 1 to 0.25 sec, the typical throttling algorithm needs about five times longer detection time but the proposed technique needs about three time longer time. This is because the increase rate of *WAQL* becomes greater and greater as worm propagates for longer time (that is, worm detection time gets longer).

Fig. 4 shows worm detection times by the two algorithms when the number of connection requests (= worm arrival rates in the figure) varies 4, 5, 10, 20, 50, 100, 200 per seconds, respectively. The proposed algorithm detects worm faster in all cases. In lower arrival rates, the difference between the detection times becomes larger. This is because as the arrival rate is lower, worm propagates longer and as worm propagates longer, *WAQL* becomes greater than that in higher arrival rates. Consequently, the proposed algorithm shows better performance than the typical one in lower arrival rates. It indicates that *WAQL* reflects the trend of worm spread very well into worm detection.

In the next study shown in Fig. 5, we measured the worm detection time while making it nearly equal the number of worm packets processed by two algorithms until worm spread is detected. To make the number of worm packets processed by the algorithms nearly same, the periods of rate limiter in the typical and proposed algorithms are set 1 and 0.7 second. Even though the period of proposed algorithm is shorter than that of the typical one, the worm detection time is nearly same due to the same reason we mentioned in Fig. 3. The detection times are depicted in Fig. 5-(a). Actually, in lower worm arrival rates, the proposed algorithm detects worm faster. As shown in Fig. 5-(b), the number of worm packets passed through the proposed throttling and that of the typical one are almost same. That is, Fig. 5 tells us that, by



(a) Worm detection time

(b) # of processed worm packets

Fig. 5. Performance comparison of algorithms with different periods of rate limiter

shortening the period of rate limiter, the proposed algorithm can reduce worm detection time and end-to-end connection delay without increasing the number of worm packets passed through the throttling mechanism.

4 Conclusion

This paper proposed a throttling algorithm for worm propagation decision. The proposed algorithm reduces worm detection time using weighted average queue length which significantly lowers the sensitivity to burstiness in normal Internet traffic, reducing false alarm. Another way for the proposed algorithm to reduce false alarm is in counting the length of delay queue. Instead of counting the number of entries as in the typical throttling, the proposed algorithm counts the number of different destination IP's in the delay queue, based on an observation that normal traffic has a strong locality in destination IP's. Experiments proved that the proposed algorithm detects worm spread faster than the typical algorithm without increasing the number of worm packets passed through the throttling.

References

1. CERT.: CERT Advisory CA-2003-04 MS-SQL Server Worm, Jan. (2003). http://www.cert.org/advisories/CA-2003-04.html
2. CERT.: CERT Advisory CA-2001-09 Code Red II Another Worm Exploiting Buffer Overflow in IIS Indexing Service DLL, Aug. 2001. http://www.cert.org/incident_notes/IN-2001-09.html
3. Matthew M. Williamson.: Throttling Viruses: Restricting propagation to defeat malicious mobile code, Proc. of the 18th Annual Computer Security Applications Conference, Dec. (2002).
4. J. Jung, S. E. Schechter, and A. W. Berger.: Fast Detection of Scanning Worm Infections, Proc. of 7th International Symposium on Recent Advances in Intrusion Detection (RAID), Sophia Antipolis, French Riviera, France, Sept. (2004).
5. X. Qin, D. Dagon, G. Gu, and W. Lee.:Worm detection using local networks, Technical report, College of Computing, Georgia Tech., Feb. (2004).

Protection Against Format String Attacks by Binary Rewriting

Jin Ho You[1], Seong Chae Seo[1], Young Dae Kim[1], Jun Yong Choi[2], Sang Jun Lee[3], and Byung Ki Kim[1]

[1] Department of Computer Science, Chonnam National University, 300 Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
{jhyou, scseo, utan, bgkim}@chonnam.ac.kr
[2] School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu 702-701, Korea
jychoi@ee.knu.ac.kr
[3] Department of Internet Information Communication, Shingyeong University, 1485 Namyang-dong, Hwaseong-si, Gyeonggi-do 445-852, Korea
aura88@empal.com

Abstract. We propose a binary rewriting system called Kimchi that modifies binary programs to protect them from format string attacks in runtime. Kimchi replaces the machine code calling conventional `printf` with code calling a safer version of `printf`, `safe_printf`, that prevents its format string from accessing arguments exceeding the stack frame of the parent function. With the proposed static analysis and binary rewriting method, it can protect binary programs even if they do not use the frame pointer register or link the `printf` code statically. In addition, it replaces the `printf` calls without extra format arguments like `printf(buffer)` with the safe code `printf("%s", buffer)`, which are not vulnerable, and reduces the performance overhead of the patched program by not modifying the calls to `printf` with the format string argument located in the read-only memory segment, which are not vulnerable to the format string attack.

1 Introduction

A majority of distributed binary programs are still built without any security protection mechanism. Although the static analysis of binary programs is more difficult compared with source programs, the security protection of a binary program without source code information is expedient when we can neither rebuild it from the patched source code nor obtain the patched binary program from the vendor in a timely manner; or when a malicious developer might introduce security holes deliberately in the binary program [1].

In this paper, we focus on the protection of a binary program—whose source code is not available—from format string attacks [2].

There are limitations to the previous binary program level protection tools against format string attacks. Libformat [3] and libsafe [4] can treat only the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 303–308, 2005.
© Springer-Verlag Berlin Heidelberg 2005

binary programs to which the shared C library libc.so is dynamically linked, and libsafe requires the program to be compiled to use a frame pointer register. TaintCheck [5] slows the traced program execution by a factor 1.5 to 40, because it runs a binary program in *traced mode* similar to a debugger monitoring all running binary code while tracing the propagation paths of user input data—incurring a significant amount of overhead.

We propose a tool called Kimchi for the UNIX/IA32 platform that modifies binary programs—even if they are statically linked to the libc library, or they do not use a frame pointer register—to prevent format string attacks at runtime.

2 Format String Attack

Figure 1 shows the stack memory layout while the function call `printf("%d%d%d%100$n%101$n", 1, 2)` is running; the function arguments are pushed onto the stack. The `printf` function reads the arguments corresponding to each `%` directive on the stack. In the example shown in Fig. 1, the first two `%d`s' accesses to the `printf`'s actual parameters $arg_1(1)$ and $arg_2(2)$ respectively are valid; while the `%100$n`'s access to arg_{100}—which is not a parameter of `printf`—is not valid. However, previous implementations of `printf` permit such invalid accesses. `Printf` stores the total number of characters written so far into the integer indicated by the `int` $*$ (or variant) pointer argument corresponding to the `%n` directive. In Fig. 1, arg_{100} located in the manipulated user input has 0xbeebebee, the location of the return address of `printf`. Thus, `printf` will overwrite and change its return address processing the `%100$n` directive. It will interrupt the control flow of the program; the attacker can execute arbitrary binary code under the program's privilege. There are many ways to change the program's control flow by overwriting critical memory.

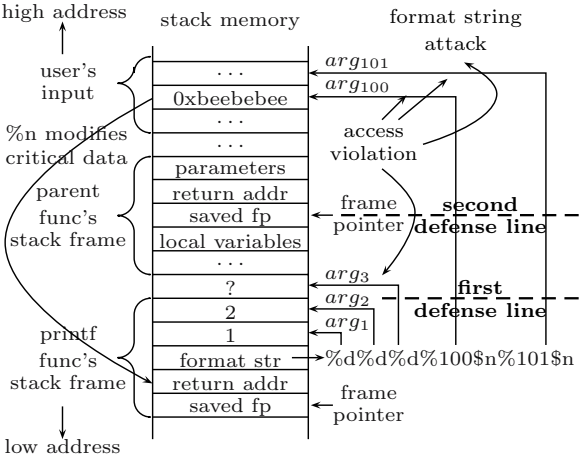


Fig. 1. `printf` call and format string attack

3 Binary Rewriting Defense Against Format String Attacks

In this section, we describe how Kimchi modifies binary programs so that previous calls to `printf` are redirected to the safe version, `safe_printf` and how `safe_printf` defends against format string attacks in runtime.

3.1 Read-Only Format String

A `printf` call with a constant format string argument located in a read-only memory region is not affected by format string attacks, because the attack is possible only when it is modifiable. Therefore, `printf` calls with constant format strings do not need to be protected. Kimchi skips the `printf` calling codes of pattern: a `pushl` *`$address`* instruction directly followed by `call printf`, where the *address* is located in a read-only memory region. We can get read-only memory regions from the section attribute information of the binary program file [6].

3.2 `Printf` Calls Without Extra Format Arguments

A `printf` call without extra format arguments like `printf(buffer)` can be changed to be not vulnerable to format string attacks by replacement with `printf("%s", buffer)`. Conventional C compilers generate the binary code of the `printf` call without extra format arguments typically as shown in Fig. 2, though it can be changed by optimization process. Kimchi finds such a code pattern in CFG, and replaces it with the code to call `safe_printf_noarg` which executes `printf("%s", buffer)`. Most of the known format string vulnerable codes are `printf` calls without extra arguments; therefore, the proposed binary rewriting should be very effective.

```
main:
 ...
 subl  $12, %esp ; for 16 byte alignment
 movl  $12(%ebp), %eax
 addl  $4, %eax  ; %eax = &argv[1]
 pushl (%eax)    ; format string arg.
 call  printf    ; printf(argv[1])
 addl  $16, %esp ; remove arguments
 ...
```

(a) The original binary code

```
main:
 ...
 subl  $12, %esp
 movl  $12(%ebp), %eax
 addl  $4, %eax
 pushl (%eax)
 call  safe_printf_noarg
 addl  $16, %esp
 ...
safe_printf_noarg:    ; INSERTED CODES
 movl $4(%esp), %eax
 subl $4, %esp
 pushl %eax     ; format_str arg.
 pushl $.FMT    ; "%s"
 call  printf   ; printf("%s",format_str)
 addl  $12, %esp
 ret
.FMT: .string "%s"
```

(b) The rewritten binary code

Fig. 2. An example of the modification of a call to printf without extra format arguments

3.3 `Printf` Calls with Extra Format Arguments

The Detection of Format String Attacks in Runtime. The defense against format string attacks is to prevent `%` directives from accessing arguments which are not real parameters passed to `printf`. An adequate solution is to modify `printf` so that it counts arguments and checks the range of argument accesses of the directives for preventing access beyond "the first defense line" as shown in Fig. 1. However, it is not easy to analyze the types of stack memory usage of the optimized or human written binary code. Kimchi protects from accessing arguments beyond "the second defense line"—i.e. the stack frame address of the parent function of `printf`. The improved version of `printf`, `safe_printf` is called with the length of extra format arguments as an additional argument, and checks the existence of the argument access violation of % directives while parsing the format string. And then, if all of them are safe, `safe_printf` calls the real `printf`, otherwise, regards the access violation as a format string attack and runs the reaction procedure of attack detection. The reaction procedure optionally logs the attack detection information through `syslog`, and terminates the process completely or returns -1 immediately without calling the real `printf`.

This same defense method is applied to other functions in the `printf` family: `fprintf`, `sprintf`, `snprintf`, `syslog`, `warn`, and `err`.

`Printf` Calls in the Function Using Frame Pointer Register. Our detection method needs to know the stack frame address of the parent function of `safe_printf`; its relative distance to the stack frame address of `safe_printf` is passed to `safe_printf`.

If the parent function uses the frame pointer register storing the base address of the stack frame, `safe_printf` can get the parent's stack frame address easily by reading the frame pointer register.

We can determine whether a function uses the stack frame pointer register by checking the presence of prologue code, `pushl %ebp` followed by `movl %esp, %ebp` which sets up the frame pointer register `%ebp` as shown in Fig. 3.

In the proposed binary rewriting method of Kimchi, `printf` calls in the function using the stack frame pointer are replaced with calls to `safe_printf_fp` as in the example shown in Fig. 3.

`Printf` Calls in the Function Not Using Frame Pointer Register. `Printf` calls in the function not using the stack frame pointer are replaced with calls to `safe_printf_`n as shown in Fig. 4, where n is the stack frame depth of the current function at the time it calls `printf`: this value is previously calculated by static analysis of the change in the stack pointer at the function during Kimchi's binary rewriting stage.

The stack frame depth at any given node of the machine code control flow graph is defined as the sum of changes in the stack pointer at each node on the execution path reachable from function entry. The static analysis calculates the stack frame depth at the node calling `printf`, and determines whether this value is constant over all reachable execution paths to the node. The problem is a

```
.FMT: .string "%d%d%d%100$n"
foo:
  pushl %ebp        ; setup frame pointer
  movl  %esp, %ebp  ;
  subl  $24, %esp   ; alloc local var mem
  subl  $4, %esp    ; typical pattern of
  pushl $2          ; function call
  pushl $1          ;
  pushl $.FMT       ; printf(.L0,1,2);
  call  printf      ;
  addl  $16, %esp   ;
  leave             ; reset frame pointer
  ret               ; return
```

(a) The original binary code

```
.FMT: .string "%d%d%d%100$n"
foo:
  pushl %ebp
  movl  %esp, %ebp
  subl  $24, %esp
  subl  $4, %esp
  pushl $2
  pushl $1
  pushl $.FMT
  call safe_printf_fp
  addl  $16, %esp
  leave
  ret
safe_printf_fp:     ;INSERTED CODES
  movl  %ebp, %eax
  subl  %esp, %eax
  subl  $8, %eax
  pushl %eax        ;call
  call  safe_printf ;safe_printf(%eax,
  addl  $4, %esp    ;retaddr,format,...)
  ret
safe_printf:
  ...
```

(b) The rewritten binary code

Fig. 3. An example of the modification of a call to printf in a function using the frame pointer register

```
.FMT: .string "%d%d%d%100$n"
foo:
  subl  $12, %esp
  subl  $4, %esp
  pushl $2
  pushl $1
  pushl $.FMT
  call  printf
  addl  $16, %esp
  addl  $12, %esp
  ret
```

(a) The original binary code

```
.FMT: .string "%d%d%d%100$n"
foo:                ; STACK CHANGE (  0)
  subl  $12, %esp   ;        %esp = -12
  subl  $4, %esp    ;             = -16
  pushl $2          ;             = -20
  pushl $1          ; stack depth = -24
  pushl $.FMT
  call  safe_printf_sp_24
  addl  $16, %esp
  addl  $12, %esp
  ret
safe_printf_sp_24:  ; INSERTED CODES
  pushl  $24        ; stack depth = 24
  call   safe_printf
  addl   $4, %esp
  ret
safe_printf:
  ...
```

(b) The rewritten binary code

Fig. 4. An example of the modification of a call to printf in a function not using the frame pointer register

kind of data flow analysis of *constant propagation* [7]; we use Kildall's algorithm giving *maximal fixed point(MFP)* solution.

3.4 Searching the `printf` Function Address

In case `libc` library is dynamically linked to the binary program, Kimchi can get the address of the `printf` function from the dynamic relocation symbol table in the binary program [6]. Otherwise, Kimchi searches the address of the `printf` code block in the binary program by a pattern matching method using the signature of binary codes [8].

4 Performance Testing

We implemented a draft version of proposed tool Kimchi, which is still under development. We measured the marginal overhead of Kimchi protection on `printf` calls. The experiment was done under single-user mode in Linux/x86 with kernel-2.6.8, Intel Pentium III 1GHz CPU and 256MB RAM.

Experiments shows that `safe_sprintf` and `safe_fprintf` have more 29.5 % marginal overhead than the original `sprintf` and `fprintf`. `Safe_printf` has more 2.2 % marginal overhead than `printf` due to its heavy cost of terminal I/O operation. The overall performance overhead of the patched program is much smaller, because general programs have just a few `printf` calls with non-constant format strings. Kimchi increases the size of binary programs by the sum of the following: memories for `safe_printf` code, `safe_printf_noarg` code, `safe_printf_fp` code, and `safe_printf_`n codes of the number of `printf` call patches in the function not using the frame pointer register.

5 Conclusions

We proposed a mechanism that protects binary programs that are vulnerable to format string attacks by static binary translation. The proposed Kimchi can treat the binary programs not using the frame pointer register as well as the ones statically linked to the standard C library; moreover, the patched program has a very small amount of performance overhead.

We are currently researching static analysis of the range of `printf` call's parameters and a format string defense mechanism applicable to the `vprintf` family functions.

References

1. Prasad, M., Chiueh, T.C.: A binary rewriting defense against stack-based buffer overflow attacks. In: the Proceedings of USENIX 2003 Annual Technical Conference, USENIX (2003) 211–224
2. Lhee, K.S., Chapin, S.J.: Buffer overflow and format string overflow vulnerabilities. Software: Practice and Experience **33** (2003) 423–460
3. Robbins, T.J.: libformat (2000) http://www.securityfocus.com/data/tools/libformat-1.0pre5.tar.gz.
4. Singh, N., Tsai, T.: Libsafe 2.0: Detection of format string vulnerability exploits (2001) http://www.research.avayalabs.com/project/libsafe/doc/whitepaper-20.ps.
5. Newsome, J., Song, D.: Dynamic taint analysis for automatic detection, analysis, and signature gerneration of exploits on commodity software. In: Proceedings of the 12^{th} Annual Network and Distributed System Security Symposium (2005)
6. Tool Interface Standard (TIS) Committee: Executable and linking format (ELF) specification, version 1.2 (1995)
7. Kildall, G.A.: A unified approach to global program optimization. In: ACM Symposium on Principles of Programming Languages (1973) 194–206
8. Emmerik, M.V.: Signatures for library functions in executable files. Technical Report FIT-TR-1994-02 (1994)

Masquerade Detection System Based on Principal Component Analysis and Radial Basics Function

Zhanchun Li, Zhitang Li, Yao Li, and Bin Liu

Network and Computer Center,
Huazhong University of Science and Technology, Wuhan, China
lzc@mail.hust.edu.cn

Abstract. This article presents a masquerade detection system based on principal component analysis (PCA) and radial basics function (RBF) neural network. The system first creates a profile defining a normal user's behavior, and then compares the similarity of a current behavior with the created profile to decide whether the input instance is valid user or masquerader. In order to avoid overfitting and reduce the computational burden, user behavior principal features are extracted by the PCA method. RBF neural network is used to distinguish valid user or masquerader after training procedure has been completed by unsupervised learning and supervised learning. In the experiments for performance evaluation the system achieved a correct detection rate equal to 74.6% and a false detection rate equal to 2.9%, which is consistent with the best results reports in the literature for the same data set and testing paradigm.

1 Introduction

Masquerade Detection System is a system for monitoring and detecting data traffic or user behaviors to distinguish intruders who impersonate valid users. It first creates a profile defining a normal user's behavior, and then compares the similarity of a current behavior with the created profile to decide whether the input instance is normal or not.

The interest in systems for the automatic detection of masquerade attacks has recently increased. Many methods have been used for masquerade detection. Maxion [1-2] used Naive Bayes classification algorithm based on the data set of truncated command lines or the data set of Enriched Command Lines. Schonlau et al.[3] described an anomaly detection technique based on unpopular and uniquely used commands to detect masqueraders. They made available UNIX command-line data from 50 users collected over a number of months. Schonlau et al[4] summarizes several approaches based on pattern uniqueness, Bayes one-step Markov model, hybrid multistep Markov model, text compression, Incremental Probabilistic Action Modeling (IPAM), and sequence matching.

Yung [5] used self-consistent naive-Bayes to detect masquerades. Kim et al.[6,7] applied Support Vector Machine (SVM) technique to detect masqueraders. Seleznyov et al. [8] used continuous user authentication to detect masqueraders. Okamoto et al. [9] used immunity approach for detecting masqueraders and evaluated this approach by utilizing multiple profiles for detecting masqueraders in UNIX-like systems.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 309 – 314, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we describe the method based on Principal Component Analysis and Radial Basics Function (PCA-RBF) to detect masquerader. PCA-RBF originated in the technique of recognizing humans facial images The PCA-RBF method has three main components: (1) modeling of the dynamic features of a sequence; (2) extraction of the principal features of the resulting model by using PCA; and (3) recognizing masquerader by using RBF neural network.

The remaining of this paper is organized as follows. Section 2 describes the methodology of the system proposed, presenting a brief description of the cooccurrence method, of the PCA technique and of RBF neural networks. Section 3 describes the experiments that have been performed. The results are then shown in section 4, which is followed by the conclusions in section 5.

2 Methodology

The coocurrence method considers the causal relationship between events in terms of a distance between two events [10]. The closer the distance between two events is and the more frequent an event-pair appears, the stronger their causal relationship becomes. Strength of correlation represents the occurrence of an event-pair in the event sequence within a scope size, namely *s*. Scope size defines what extent the causal relationship of an event-pair should be considered. Fig.1 show that the scope size in the example is 6, and then two events of *emacs* and *ls* correlation strength is 4. When strength of correlation for all events was been calculated, the matrix named coocurrence matrix will be created. On the assumption that all unique commands of the all user sequences is *m* , then the dimension of coocurrence matrix is *m*m*.

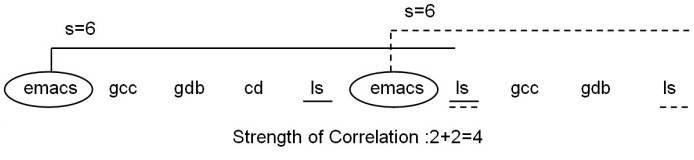


Fig. 1. Correlation between *emacs* and *ls* for User

With the number of user increased, the dimension of coocurrence matrix is increased. In order to detect masquerader rapidly, dimension of coocurrence matrix must be reduced. PCA is a typical statistical tool for dimension reduction, which just uses second-order statistics information to extract the components. An cooccurrence matrix having $n = m*m$ elements can be seen as a point in an n dimensional space. PCA identifies the orthonormal base of the subspace which concentrates most of the data variance.

An cooccurrence matrix is represented by a *n*1* row vector obtained by concatenating the rows of each matrix of size $n = m*m$ elements. Given a sample with N coocurrence matrixs, a training set such as $\{X_k \mid k=1,2,\ldots N\}$ can be built. The base for the subspace is derived from the solution of the eigenvector/eigenvalue problem:

$$U^T W U = \Lambda \tag{1}$$

where U is the $N*N$ eigenvectors matrix of W, Λ is the corresponding diagonal matrix of eigenvalues, W is the sample covariance matrix that is defined as

$$W = \frac{1}{N}\sum_{k=1}^{N}(X_k - \overline{X})(X_k - \overline{X})^T \tag{2}$$

$\overline{X}$ is the mean vector of training set. To reduce the dimensionality, only the p ($p < N$) eigenvectors associated with the p greatest eigenvalues are considered. The representation y_{PCA} in the p dimensional subspace of X_k can be obtained by:

$$y_{PCA} = W_{eig}(X_k - \overline{X})^T \tag{3}$$

where X_k is the row vector, $\overline{X}$ is the corresponding sample mean and W_{eig} is the $N*p$ matrix formed by the p eigenvectors with the greatest eigenvalues. It can be proved that this approach leads to the p-dimensional subspace for which the average reconstruction error of the examples in the training set is minimal.

The RBF neural network is based on the simple idea that an arbitrary function $o(y)$ can be approximated as the linear superposition of a set of radial basis functions $g(y)$, leading to a structure very similar to the multi-layer perceptron. The RBF is basically composed of three different layers: the input layer, which basically distributes the input data; one hidden layer, with a radial symmetric activation function; and one output layer, with linear activation function. For most applications, the Gaussian format is chosen as the activation function $g(y)$ of the hidden neurons.

$$g_j(Y_{PCA}) = \exp\left[\frac{-\left\|Y_{PCA} - c_j\right\|^2}{2\sigma_j^2}\right] \qquad j = 1,2,...,h \tag{4}$$

Where g_j is output of j^{th} hidden node; $Y=(y_{PCA1}, y_{PCA2}\ldots, y_{PCAn})^T$ is the input data; c_j is the center vector of j^{th} radial basic function unit ; $\|\ \|$ indicates the Euclidean norm on the input space; σ_j is the width of the j^{th} radial basic function unit; h is the number of hidden node.

The k^{th} output O_k of an RBF neural network is

$$O_k = \sum_{j=1}^{h} w_{kj} g_j(Y_{PCA}) \qquad (k = 1,2,\cdots,m) \tag{5}$$

Where w_{kj} is the weight between the j^{th} hidden node and the k^{th} output node; h is the number of hidden nodes.

The training procedure of RBF is divided into two phases. At the first, k-means clustering method is used to determine the parameters of RBF neural network, such as the number of hidden nodes, the center and the covariance matrices of these nodes. After the parameters of hidden nodes are decided, the weights between the hidden layer and output layer can be calculated by minimizing.

3 Experiments

The data used in the present work is the same as the data used in the several Schonlau et al [4].The data can be download at http://www.schonlau.net. Masquerade detection studies commonly called *SEA* dataset. As described by Schonlau et al., user commands were captured from the UNIX auditing mechanism and built a truncated command dataset. In *SEA* dataset, 15,000 sequential commands for each of 70 users are recorded. Among 70 users, 50 are randomly chosen as victims and the remaining 20 as intruders. Each victim has 5,000 clean commands and 10,000 potentially contaminated commands.

In our experiment, the first 5,000 commands for each user are taken as the training data and the remaining 10,000 commands form the testing data. Because of the data grouped into blocks of 100 commands, the length of command sequences in cooccurrence method is 100 *(l=100)* and each user has 50 sequences.

The procedure of create the profile for each user include three steps: creating initialization behavior matrix for each user, dimensionality reduction using PCA, and training RBF neural network.

Because the number of unique commands in the experiment accounted is 100, the dimension of cooccurrence matrix for each user is 100*100. The element of the matrix is calculated by using cooccurrence method where the scope size is six (*s=6*). The rows of each cooccurrence matrix are concatenated in order to form line vectors which having 10000 elements.

The dataset is divided in two subsets: one used for neural network training and the other one used for test. At th first, an average vector of the training subset is calculated and subtracted from all vectors, as required by the PCA algorithm. Then the covariance matrix is calculated from the training set. The p eigenvectors, correspondents to the p largest eigenvalues of this matrix, are used to build the transformation matrix W_{eig}. In this experiment p is fifty. By applying the transformation matrix on each vector, its projection on the attribute space is then obtained, which contains only p ($p=50$) elements instead of the initial number of 10000. Thus each user's profile is composed of the feature vectors y_{PCA} which number of elements is 50. At the last, y_{PCA} is the input vector of RBF neural network and RBF neural network will be trained.

When a sequence of one user was to be tested, the procedure was very similarly with the procedure of create the profile. At the first, initialization behavior matrix for the sequence was created. Then the obtained the matrix was converted to the row vector X_{test}. The representation y_{test} in the p dimensional *(p=50)* subspace of X_{test} can be obtained by:

$$y_{test} = W_{eig}(X_{test} - \overline{X})^T \tag{6}$$

where X_{test} is the row vector, $\overline{X}$ is the corresponding sample mean and W_{eig} is the N*p matrix formed by the *p(p=50)* eigenvectors with the greatest eigenvalues.

Then y_{test} is the input vector of RBF neural network, the output vector o_{test} will be calculated by RBF neural network and legitimate user or masquerader will be recognized.

4 Results

The results of a masquerade detector assessed are the tradeoff between correct detections and false detections These are often depicted on a receiver operating characteristic curve (called an ROC curve) where the percentages of correct detection and false detection are shown on the y-axis and the x-axis respectively. The ROC curve for the experiment is shown in Fig.2. Each point on the curve indicates a particular tradeoff between correct detection and false detection. Points nearer to the upper left corner of the graph are the most desirable; as they indicate high correct detection rates and correspondingly low false detection rates. As a result, the PCA-RBF method achieved a 74.6% correct detection rate with a 2.9% false detection rate.

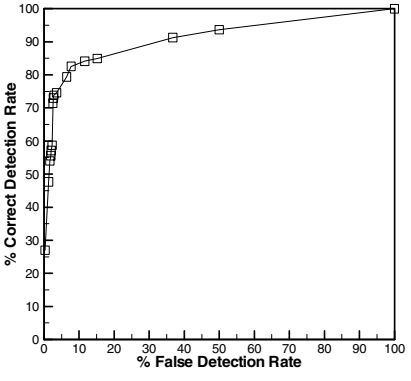


Fig. 2. ROC curve for experiments

Schonlau et al[4] summarizes several approaches based on pattern uniqueness, Bayes one-step Markov model, hybrid multistep Markov model, text compression, Incremental Probabilistic Action Modeling (IPAM), and sequence matching. The results from previous approaches to masquerade detection on same dataset are showed in table 1. So the PCA-RBF method achieved the best results.

Table 1. Results from previous approaches masquerade detection

Approaches	False Rate (%)	Correct Rate (%)
Uniqueness	1.4	39.4
Bayes one-step Markov	6.7	69.3
Hybrid Multistep Markov	3.2	49.3
Compression	5.0	34.2
IPAM	2.7	41.4
Sequence Matching	3.7	36.8

5 Conclusions

In this paper, we outline our approach for building a masquerade detection system based on PCA and RBF neural network. We first describe architecture for building valid user behavior. The cooccurrence method ensures that the system can model the dynamic natures of users embedded in their event sequences. Second, PCA can discover principal patterns of statistical dominance. Finally, Well-train RBF neural network contained the normal behavior knowledge and can be used to detect masquerader.

Experiments on masquerade detection by using SEA dataset show that the PCA-RBF method is successful with a higher correct detection rate and a lower false detection rate (74.6% correct detection rate with 2.9% false detection rate). It also shows PCA can extract the principal features from the obtained model of a user behavior, and that RBF neural networks can lead to the best detection results.

References

1. Maxion, R.A. and T.N.: Townsend. Masquerade detection using truncated command lines. in Proceedings of the 2002 International Conference on Dependable Systems and Networks DNS 2002, Jun 23-26 2002. (2002). Washington, DC, United States: IEEE Computer Society.
2. Maxion, R.A.: Masquerade Detection Using Enriched Command Lines. in 2003 International Conference on Dependable Systems and Networks, Jun 22-25 (2003). 2003. San Francisco, CA, United States: Institute of Electrical and Electronics Engineers Computer Society.
3. Schonlau, M. and M. Theus.: Detecting masquerades in intrusion detection based on unpopular commands. Information Processing Letters,76(1-2) (2000)33-38.
4. Schonlau, M., et al.: Computer Intrusion: Detecting Masquerades. Statistical Science, 16(1) (2001)58-74.
5. Yung, K.H.: Using self-consistent naive-Bayes to detect masquerades. in 8th Pacific-Asia Conference, PAKDD 2004, May 26-28 2004. (2004). Sydney, Australia: Springer Verlag, Heidelberg, Germany.
6. Kim, H.-S. and S.-D. Cha.: Efficient masquerade detection using SVM based on common command frequency in sliding windows. IEICE Transactions on Information and Systems, E87-D(11) (2004)2446-2452.
7. Kim, H.-S. and S.-D. Cha, Empirical evaluation of SVM-based masquerade detection using UNIX commands. Computers and Security, 2005. 24(2): p. 160-168.
8. Seleznyov, A. and S. Puuronen, Using continuous user authentication to detect masqueraders. Information Management and Computer Security, 2003. 11(2-3): p. 139-145.
9. Okamoto, T., T. Watanabe, and Y. Ishida. Towards an immunity-based system for detecting masqueraders. in 7th International Conference, KES 2003, Sep 3-5 2003. 2003. Oxford, United Kingdom: Springer Verlag, Heidelberg, D-69121, Germany.
10. Oka, M., et al. Anomaly Detection Using Layered Networks Based on Eigen Cooccurrence Matrix. in RAID. 2004.

Anomaly Detection Method Based on HMMs Using System Call and Call Stack Information*

Cheng Zhang and Qinke Peng

State Key Laboratory for Manufacturing Systems
and School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an 710049, China
cheng_zhang@stu.xjtu.edu.cn, qkpeng@xjtu.edu.cn

Abstract. Anomaly detection has emerged as an important approach to computer security. In this paper, a new anomaly detection method based on Hidden Markov Models (HMMs) is proposed to detect intrusions. Both system calls and return addresses from the call stack of the program are extracted dynamically to train and test HMMs. The states of the models are associated with the system calls and the observation symbols are associated with the sequences of return addresses from the call stack. Because the states of HMMs are observable, the models can be trained with a simple method which requires less computation time than the classical Baum-Welch method. Experiments show that our method reveals better detection performance than traditional HMMs based approaches.

1 Introduction

Intrusion detection is very important in the defense-in-depth network security framework and has been an active topic in computer security in recent years. There are two general approaches to intrusion detection: misuse detection and anomaly detection. Misuse detection looks for signatures of known attacks, and any matched events are considered as attacks. Misuse detection can detect known attacks efficiently. However, it performs poorly with unknown attacks. Anomaly detection, on the other hand, creates a profile that describes normal behaviors. Any events that significantly deviate from this profile are considered to be anomalous.

The key problem of anomaly detection is how to effectively characterize the normal behaviors of a program. System calls provide a rich resource of information about the behaviors of a program. In 1996, Forrest et al. introduced a simple anomaly detection method based on monitoring the system calls issued by active, privileged processes. Each process is represented by the ordered list of system calls it used. The *n-gram* method builds a profile of normal behaviors by

* Supported by National Natural Science Foundation under Grant No.60373107 and National High-Tech Research and Development Plan of China under Grant No.2003AA142060.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 315–321, 2005.
© Springer-Verlag Berlin Heidelberg 2005

enumerating all unique, contiguous sequences of a predetermined, fixed length n that occur in the training data [1]. This work was extended by applying various machine learning algorithms to learn the profile. These methods include neural networks [2], data mining [3], variable length approach [4], Markov Chain Models [5, 6] and Hidden Markov Models [7,8,9], etc.

Recently, researchers have shown that the call stack can be very useful to build the normal profile for detecting intrusions. Sekar et al. used the program counter in the call stack and the system call number to create a finite state automaton [10]. Feng et al. first utilized return address information extracted from the call stack to build the normal profile for a program [11]. The experimental results showed that the detection capability of their method was better than that of system call based methods.

In this paper, we propose an anomaly detection method based on HMMs to model the normal behaviors of a program. The method uses system call and call stack information together to build hidden Markov models. The system calls in a trace are used as hidden states and return addresses from the call stack are used as observation symbols. If the probability of a given observation sequence is below a certain threshold, the sequence is than flagged as a mismatch. If the ratio between the mismatches and all the sequences in a trace exceeds another threshold in advance, the trace is considered as an intrusion.

The rest of the paper is organized as follows. Section 2 gives a brief introduction to call stack data and how HMMs are applied to model the normal program behaviors. In section 3, the anomaly detection method is presented. Then experimental results on wu-ftpd are summarized and discussed in section 4. Conclusion is given in section 5.

2 Building HMMs for Anomaly Detection

2.1 Audit Data

The biggest challenge for anomaly detection is to choose appropriate features that best characterize the program or user usage patterns so that normal behaviors would not be classified as anomalous. Anomaly detection based on monitoring of system calls has been shown to be an effective method for unknown attacks. However, system call is only one aspect of program behaviors, the call stack of the program provides an additional rich source of information that can be used for detecting intrusions [11].

As each system call is invoked, we extract both system call and all the return addresses from the call stack. In this paper, a sequence of return addresses is defined as a function chain, which denotes the overall execution path when a system call is made. For example, assume that a function $f()$ is called within the function $main()$. Then the function chain is $A = \{a_1, a_2\}$ when a system call is invoked in function $f()$. a_1 is the return address of $main()$ and a_2 is the return address of the function $f()$.

Using these two types of data can be more useful and precise than using only system calls to model program behaviors. Therefore, we apply system calls and function chains to establish HMMs for anomaly detection.

2.2 Establishing Hidden Markov Models

A Hidden Markov Model (HMM) is a doubly stochastic process. A HMM contains a finite number of states. Transitions among the states are governed by a stochastic process to form a Markov Model. In a particular state, an outcome or observation can be generated according to a separate probability distribution associated with the state. The Markov model used for the hidden layer is a first-order Markov model, which means that the probability of being in a particular state depends only on the previous state.

In traditional method based on HMMs, system calls are the observations of the model and the states are unobservable [7]. The method chooses a number of states roughly corresponding to the number of unique system calls used by the program. In our method, the states of the HMM are associated with system calls and the observations are associated with function chains (sequences of return addresses from the call stack). The HMM is characterized by the following:

(1) N, the number of states in the model. We denote the states as $S = \{S_1, S_2, \cdots, S_N\}$, which is the set of unique system calls of the process. The state at time t is q_t. Here, system calls are applied as the states of the model. So the number of states is the number of unique system calls.
(2) M, the number of possible observations. We denote the observations as $V = \{V_1, V_2, \cdots, V_M\}$, which is the set of unique function chains of the process. The observation at time t is O_t. Here, function chains are applied as the observation symbols of the model. A function chain is associated with a system call when the system call is invoked. The number of observations is the number of unique function chains used by the program.
(3) The state transition probability distribution $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, $1 \leq i, j \leq N$. The states are fully connected; transitions are allowed from any system call to any other system call in a single step.
(4) The observation symbol distribution $B = \{b_j(k)\}$, where $b_j(k) = P(O_t = V_k | q_t = S_j)$, $1 \leq j \leq N$, $1 \leq k \leq M$. When a system call is invoked, a function chain associated with it can be observed.
(5) The initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i)$, $1 \leq i \leq N$.

We use the compact notation $\lambda = (A, B, \pi)$ to indicate the complete parameter set of the model. Given the form of the hidden Markov model, the training of the HMM is the key problem. The Baum-Welch method is an iterative algorithm that uses the forward and backward probabilities to solve the problem of training by parameter estimation. But the Baum-Welch method is locally maximized and requires high computational complexity. In our method, system calls are states and function chains are observation symbols. Both of them are observable so that we use a simple method to compute parameters of the HMM instead of the classical Baum-Welch method.

Given observation sequence $O = \{O_1, O_2, \cdots, O_T\}$ and state sequence $Q = \{q_1, q_2, \cdots q_T\}$ of the process, where O_t is the function chain when system call q_t is made at time t, $t = 1, 2, \cdots, T$. The initial state distribution, the state transition probability distribution and the observation symbol distribution of the HMM $\lambda = (A, B, \pi)$ can be computed as follows:

$$a_{ij} = \frac{N_{ij}}{N_i*}, \pi_i = \frac{N_i}{N_{\text{total}}}, b_j(k) = \frac{M_{jk}}{M_j*}. \tag{1}$$

Here, N_{ij} is the number of state pairs q_t and q_{t+1} with q_t in state S_i and q_{t+1} in S_j; N_i* is the number of state pairs q_t and q_{t+1} with q_t in S_i and q_{t+1} in any one of the state $S_1, \cdots, S_N$; N_i is the number of q_t in state S_i; N_{total} is the total number of states in state sequence; M_{jk} is the number of pairs q_t and O_t with q_t in state S_j and O_t in V_k; M_{j*} is the total number of O_t in any one of the observation $V_1, \cdots, V_M$ and q_t in state S_j.

3 Anomaly Detection

Given a testing trace of system calls and function chains (length T) that were generated during the execution of a process, we use a sliding window of length L, and a sliding step 1 to pass training trace and get $T - L + 1$ short sequences of function chain X_i $(1 \leq i \leq T - L + 1)$ and $T - L + 1$ short sequences of system call $Y_i (1 \leq i \leq T - L + 1)$.

Using the normal model $\lambda = (A, B, \pi)$ which was constructed by our training method described above, a given function chain sequence $X = \{O_{t-L+1}, \cdots, O_t\}$ with corresponding system call sequence $Y = \{q_{t-L+1}, \cdots, q_t\}$ can be computed as follows:

$$P(X|\lambda) = \pi_{q_{t-L+1}} b_{q_{t-L+1}}(O_{t-L+1}) \prod_{i=t-L+1}^{t-1} a_{q_i q_{i+1}} b_{q_{i+1}}(O_{i+1}) \tag{2}$$

Ideally, a well trained HMM can maintain sufficiently high likelihood only for sequences that correspond to normal behaviors. On the other hand, sequences which correspond to abnormal behaviors should give a significantly lower likelihood values. By comparing the probability of short observation sequence, we can determine whether it is anomalous.

In the procedure of actual anomaly detection, there will be some function chains or system calls that have not appeared in the training set. It might be an anomaly in the execution of the process, so we define these function chains as an anomalous observation symbol V_{M+1}. Also we define those system calls as an anomalous state S_{N+1}. Because they are not included in the training set, we define their parameters of the model by the following: $\pi_{S_{N+1}} = a_{S_{N+1}S_j} = a_{S_i S_{N+1}} = 10^{-6}$, $b_j(V_{M+1}) = 10^{-6}$, $1 \leq i \leq N$, $1 \leq j \leq M$.

Given a predetermined threshold ε_1, with which we compare the probability of a sequence X in a testing trace, if the probability is below the threshold, the sequence is flagged as a mismatch. We sum up the mismatches and define the

anomaly index as the ratio between the numbers of the mismatches and all the sequences in the trace.

$$\text{anomaly index} = \frac{\text{number of the mismatches}}{\text{number of all the sequences in a trace}} \leq \varepsilon_2 \tag{3}$$

If the anomaly index exceeds the threshold ε_2 that we choose in advance, the trace (process) is considered as a possible intrusion.

4 Experiments

4.1 Training and Testing Trace Data

On a Red Hat Linux system, we use the wu-ftpd as an anomaly detection application, because wu-ftpd is a widely deployed software package to provide File Transfer Protocol (FTP) services on Unix and Linux systems, and exists many intrusion scripts on the Internet. Using kernel-level mechanism, we extract both system call and call stack information during execution of the process. All of the training data are normal trace data, while the testing data include both the normal traces and intrusion trace data. The description of the datasets in the experiments is shown in Table 1.

Table 1. Description of the data sets used in the experiments

Data	Number of trace	Number of call
Training data(Normal data)	82	823,123
Testing data(Normal data)	114	1,023,628
Testing data(Intrusion data)	43	83,234

4.2 Normal Profile Training

Each line in the trace data contains the pid, system call name, followed by its function chain. Each trace is the list of system calls and function chains issued by a single process from the beginning of its execution to the end. Data are preprocessed before training. Each different system call name is assigned with its corresponding system call number in Linux. Each different function chain is also assigned with a different number. Then, system call and function chain in a trace are represented as two sequences of numbers.

The process of learning wu-ftpd normal profiles is to construct the initial state distribution, the state transition probability distribution and the observation distribution for the HMM from wu-ftpd training trace data by (1). After training, the total number of states is 45 and the total number of observations is 765.

4.3 Experimental Results and Discussion

In the experiments, the sliding window size is set as 6 and the anomaly index is used as classification rule for the proposed method and the traditional method [7]. Table 2 shows the detection performance of these two methods. Here, the false positive rate is the minimal value of the false positive rate when the true positive rate obtains 100%. From Table2, we can conclude that:

Table 2. Comparisons of the detection performance

	Traditional method	The proposed method
True positive rate	100%	100%
False positive rate	3.51%	1.75%
Average anomaly index of testing data	2.398%	4.11%
Average anomaly index of intrusion data	16.21%	29.69%
Computation time(s)	79835.4	2154.2

(1) The proposed method has a better detection performance. A desirable intrusion detection system must show a high true positive rate with a low false positive rate. The true positive rate is the ratio of the number of correctly identified intrusive events to the total number of intrusive events in a set of testing data. The false positive rate is the ratio of the number of normal events identified as intrusive events to the total number of normal events in a set of testing data. From Table 2, we can see that the false positive rate of the proposed method is lower than traditional method, while both of the methods can detect all the abnormal traces.
(2) In the proposed method, the difference between the average anomaly index of the intrusion data and that of the normal data is higher than the difference of the traditional method. This means that the detection accuracy of the proposed method is better than the traditional method based on HMMs.
(3) During training of traditional method, the probabilities were iteratively adjusted to increase the likelihood that the automaton would produce the traces in the training set. Many passes through the training data were required. The proposed method only needs one pass through the training data. Table 2 shows the computation time occupied for training process. The experimental results indicate that the proposed method need much less time than traditional method to estimate the parameters of the HMM.

5 Conclusions

In this paper, a new anomaly detection method based on HMMs has been proposed. We use system calls as states and use sequences of return addresses as observation symbols of the HMMs. The normal program behaviors are modeled by using HMMs and any significant deviation from the model is considered as possible intrusion. Because the states are observable in the hidden Markov Model, we can use a simple method to compute parameters of the HMM instead of the Baum-Welch method.

The experimental results on wu-ftpd clearly demonstrate that the proposed method is more efficient in terms of detection performance compared to traditional HMMs method. For future work, we will extend our model to other applications and anomaly detection techniques.

References

1. S. Forrest, S. A. Hofmery, A. Somayaji: A Sense of Self For Unix Processes. Proceedings of the 1996 IEEE Symposium on Security and Privacy, Oakland, California, (1996) 120-128
2. A. K. Ghosh, A. Schwartzbard: Learning program behavior profiles for intrusion detection. In Proceedings: 1st USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, California, (1999) 51-62
3. W. Lee, S. J. Stolfo: Data Mining Approaches for intrusion detection. Proceedings of the 7th USENIX Security Symposium, pages 79-94, San Antonio, Texas, (1998) 26-29
4. A. Wespi, M. Dacier, H. Debar: Intrusion Detection using Variable-length audit trail patterns. Proceedings of the 3rd International Workshop on the Recent Advances in Intrusion Detection (RAID' 2000), (2000) No.1907 in LNCS
5. N. Ye. A Markov chain model of temporal behavior for anomaly detection: Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop [C]. Oakland : IEEE, (2000) 166-16
6. N. Ye, X. Li, Q. Chen, S. M. Emran, M. Xu: Probabilistic Techniques for Intrusion Detection Based on Computer Audit Data. IEEE Trans. SMC-A, (2001) Vol.31, No.4, 266-274
7. C. Warrender, S. Forrest, B. Pearlmutter: Detecting intrusions using system calls: alternative data models. Proceedings of the 1999 IEEE Symposium on Security and Privacy, May 9-12, (1999) 133-145
8. Y. Qiao, X. W. Xin, Y. Bin, S. Ge. Anomaly intrusion detection method based on HMM. Electronics Letters, (2002) Vol. 38, No.13, pages 663-664
9. W. Wei, G. X. Hong, Z. X. Liang. Modeling program behaviors by hidden Markov models for Intrusion Detection: Proceedings of 3rd International Conference on Machine Learning and Cybernetics, August 26-29, (2004) 2830-2835
10. R. Sekar, M. Bendre, D. Dhurjati, P. Bollineni: A fast autiomation-based method for detection anomalous program behaviors. Proceedings of IEEE symposium on Security and Privacy, Oakland, California, (2001) 144-155
11. H. H. Femg, O. M. Kolesnikov, P. Fogla, W. Lee, W. Gong: Anomaly detection using call stack information: Proceedings of IEEE symposium on Security and Privacy, Berkeley, California, (2003)

Parallel Optimization Technology for Backbone Network Intrusion Detection System

Xiaojuan Sun, Xinliang Zhou, Ninghui Sun, and Mingyu Chen

National Research Center for Intelligent Computing Systems,
Institute of Computing Technology, Chinese Academy of Sciences,
P.O. Box 2704, Beijing 100080, China
{sxj, zxl, snh, cmy}@ncic.ac.cn

Abstract. Network intrusion detection system (NIDS) is an active field of research. With the rapidly increasing network speed, the capability of the NIDS sensors limits the ability of the system. The problem is more serious for the backbone network intrusion detection system (BNIDS). In this paper, we apply parallel optimization technologies to BNIDS using 4-way SMP server as the target system. After analyzing and testing the defects of the existed system in common use, the optimization policies of using fine-grained schedule mechanism at connection level and avoiding lock operations in thread synchronization are issued for the improved system. Through performance evaluation, the improved system shows more than 25 percent improvement in CPU utilization rate compared with the existed system, and good scalability.

1 Introduction

NIDS discovers the actions offending the security policies and the traces of network attacking, and gives some reactions to them. It performs security analysis on the packets eavesdropped on a network link. Existing NIDS can barely keep up with the bandwidth of a few hundred Mbps, which falls behind the requirement of the ever-increasing network traffic.

BNIDS is a NIDS for the interface of backbone Internet. It has much higher network traffic throughput compared with the common NIDS used in enterprises and universities. The rapidly increasing rate of the backbone network bandwidth greatly challenges the existing BNIDS. With the development of the computer manufacturing industry, the 4-way and 8-way SMP computers are becoming cheap, and the dual-core processors have been announced and widely used in high performance computers. Affordable high performance computers are available to BNIDS.

In this paper, after analyzing and verifying the defects of the existed multithreading NIDS implemented by our lab three years before, we present parallel optimization technologies for SMP sensors. The improved system acquired about 30 percent improvement of CPU utilization rate on 2-way system, and 25 percent improvement on 4-way system compared with the existed system. The improved system gained good scalability on SMP systems in experiments. It achieved an average of 56 percent enhancement of CPU utilization rate on 4-way system compared with on 2-way system.

The rest of the paper is organized as followed: After a discussion of related work in Section 2, Section 3 introduces the existed system and its shortcomings in thread

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 322 – 327, 2005.
© Springer-Verlag Berlin Heidelberg 2005

scheduling and synchronizing, gives our optimization technologies and shows the obvious improvement by tests. Section 4 is the performance analysis; we give three conclusions from the experimental results. Section 5 is our conclusion and future works.

2 Related Works

The related works of this field can be classified into the research of load balancer in distributed NIDS and the research of higher throughput sensors.

Concluded from the distribution strategies in [1], [2] and [3], the agreed requirement for load balancer is to scatter the network traffic non-overloading to the sensors, to guarantee all the relative packets in a malicious attack to be handled, and to minimize the introduced vulnerability through the load balancing algorithm.

A number of vendors, such as Juniper Network [4], ISS [5] and Top Layer Networks [6], claim to have sensors that can operate on high-speed ATM or Gigabit Ethernet links. But there are some actual difficulties of performing NIDS at high-speed links [1]. These products are not suitable for the BNIDS in two reasons. First, the BNIDS is required to cover several gigabits of network traffic, and the traffic is usually much higher than the real deployed traffic of NIDS products. The off-the-shelf hardware and software of high performance computers are a good choice for BNIDS. Second, compared with the high performance computers, the dedicated system has more difficulty to upgrade and operate and less flexibility to use. Besides these products, there are many novel algorithm researches on protocol reassembling [7] and pattern matching [8] [9] for higher sensors. They can gain limited increase of the efficiency of the analysis applications on sensors.

3 Parallel Optimization Technologies

3.1 Defects of the Existed System

The existed system is a multithreaded program for SMP computers. The packets are dispatched to eight slave threads that are created by the master thread. The rules of dispatch ensure that all the packets of the same connection can not be dispatched to different threads by hashing of the source and destination internet addresses. The work flow of the existed system, described in Fig. 1, goes through three main stages: packet capturing, protocol reassembling and content filtering.

The existed system meets two intractable problems in the real network execution. First, the multithreading model cannot acquire good parallelism and the resources of the system are consumed seriously. Second, the pending packets are accumulated by the best-effort mechanism and forced to discard for the healthy execution.

These problems are resulted from the parallel mechanism of the system. The packets are dispatched to the slave threads at the very beginning of the protocol reassembling before the IP fragment process stage. It results that the packets belonging to different connections are still mingled in the buffer of the slave threads. The concurrence of different connections is hard to implement. Moreover, the slave threads handle the packets on their own packet queue, the synchronization with the master thread uses *Pthread* lock operations.

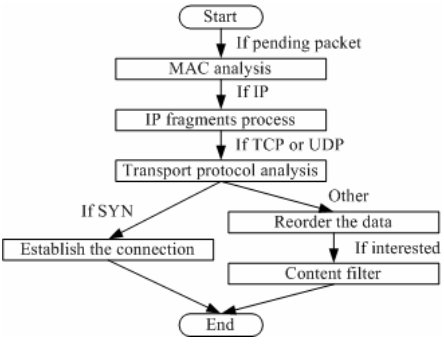


Fig. 1. The work flow of the system

Table 1. The comparison of the wait time in the existed system and the improved system

Thread	Existed System(microsecond)		Improved System(microsecond)	
	Time	Wait time	Time	Wait time
0	2,544,559	1,137,660	1,295,123	70
1	17	0	1,805	210
2	267	0	885	453
3	1	0	197	0
4	5	0	515	105
5	5,015,763	2,161,582	4,725	0
6	2	0	282	0
7	903	164	230	0
8	125	0	10,248	1070

These defects were verified by experiments. We used Vtune Performance Analyzer [10] to record the wait time of all the threads. The network traffic between a FTP server and a client went through the system. The results were in Table 1. We concluded that Thread5 had an extremely long runtime. According to the hashing algorithm, the packets were dispatched to the same thread. We also discovered the great wait time of the synchronization between threads by Oprofile[11]. The executions of the *pthread_mutex_lock* and *pthread_mutex_unlock* functions, which could put the thread from running to wait and back to runnable, took a relatively long time. They were less than the packet checksumming operations and nearly the same as the packet copying operations, the buffer allocating and initializing operations.

3.2 Optimization Policies

Paralleling at Connection Level and Dispatching Evenly. To maximize the parallelism, we multithread the system as soon as the connections are distinguished after the transport protocol reassembling. We repeated the tests in Section 3.1, and found out that the self time of slave threads didn't concentrate on one thread and the wait time between the master and the slave threads decreased greatly. As shown in Table 1, the wait time of Thread0 and Thread5 fell to no more than 1.07 milliseconds, while it was high to about one and two seconds in existed system.

In addition, we dispatch the tasks in round robin fashion to all slave threads according to connections. In this way, the pending packets are not accumulated and the speed and accuracy of responses are improved.

Decreasing the Synchronization Overhead. Because the proportion of thread synchronization is high according to the tests in Section 3.1, we replace many lock operations with queue operations. The repeated test showed the proportion decreased greatly. In addition, we reduce the slave threads to three, and hope one thread for one processor. The tests showed that the context switches of 4-way system with eight slave threads increased 9.5% compared with that of three slave threads in 200 Mbps workload.

4 Performance Analysis

In order to simulate a heavy workload of network, we select SPECWeb99 [12] and Httperf [13] to generate HTTP requests. We use three types of workloads in the experiments as shown in Table 2. The SPECWeb99 load has much more newly established connections per second than Httperf loads. We can see the drop trend of the following curves because of less per-connection overhead.

The experiments are implemented on the 2-way system with 2.0GHz Intel XEON™ processors and the 4-way system with 2.2GHz AMD Opteron™ processors. We also configured the mirror source port and monitor port at Cisco 3500XL switch.

Table 2. The detail of three workloads

	SPECWeb99 200Mbps	Httperf 370Mbps	Httperf 520Mbps
Conns/s	600	52	0.6
Packets/s	28,000	50,000	64,000

Scalability. Fig. 2 shows that the improved system scales well in 4-way system. Compared with the 2-way system, the CPU utilization rate of 4-way system was enhanced at an average of 56%. The CPU cycles for user instructions were improved at an average of 51%, and for system instructions they were improved at least 5 times.

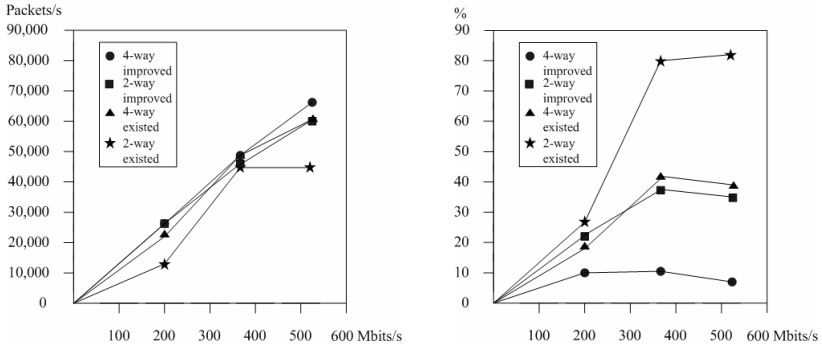


Fig. 2. Comparison of throughput and CPU utilization on 2-way and 4-way

Optimization Effects. In Fig. 2, we can see that the improved system is better than the existed system in any case. In the first workload of the 2-way system, the dropped packet rate of the existed system was high to 50%, while it was not obvious the improved system had dropped packets in any load. The CPU idle rates of the improved system on the 2-way and the 4-way system were about 60% and 90% respectively. Despite the high dropped packet rate of the existed system, the improved system had lower CPU utilization.

Load Balance. As Fig. 3 illustrated, the threads' loads of the improved system is much more balanced than the existed system. In the experiment, the total packets, the HTTP requests and the HTTP handling bytes processed by each slave thread of the improved system were also even. On the contrary, the thread loads of the existed system were not balanced. Some threads' load was much lighter than others, or even zero.

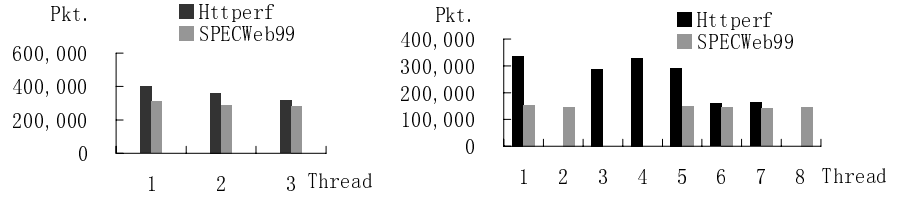


Fig. 3. Monitored packets of the slave threads in improved system (left) and existed system (right)

5 Conclusion and Future Works

The improved system with new parallel optimization technologies shows good performance. This proves SMP high performance computers with the parallel optimized multithreaded applications are helpful for BNIDS. The improved system achieves an average of 56 percent enhancement on 4-way system than on 2-way system. Compared with the existed system, the improved system acquires about 30 percent improvement of CPU utilization rate on 2-way system, and 25 percent improvement on 4-way system.

In the future, with the trend of developing 8-way or higher SMP servers, the parallelism of program is more important. To acquire good scalability on higher SMP servers continues to be a challenging work. In addition, we will pay much attention at the master thread optimization, and apply more fair mechanism for task dispatch.

References

1. Christopher, K., Fredrik, V., Giovanni, V., Richard, K.: Stateful Intrusion Detection for High-Speed Networks. Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos, Calif. (2002)
2. Simon, E.: Vulnerabilities of Network Intrusion Detection Systems: Realizing and Overcomign the Risks -- The Case for Flow Mirroring. Top Layer Networks. (2002)

3. Lambert, S., Kyle, W., Curt, F.: SPANIDS: A Scalable Network Intrusion Detection Loadbalancer. Conf. Computing Frontiers. Ischia, Italy (2005) 315-322
4. IDP 1000/1100. Juniper Networks. http://www.juniper.net/products/
5. RealSecure Network Gigabit. ISS. http://www.iss.net/products_services/enterprise_protection/rsnetwork/gigabitsensor.php
6. Top Layer Networks. http://www.toplayer.com/
7. Xiaoling, Z., Jizhou, S., Shishi, L., Zunce, W.: A Parallel Algorithm for Protocol Reassembling. IEEE Canadian Conference on Electrical and Computer Engineering. Montreal (2003)
8. Spyros, A., Kostas, G.A., Evangelos, P.M., Michalis, P.: Performance Analysis of Content Matching Intrusion Detection Systems. Proceedings of the IEEE/IPSJ Symposium on Applications and the Internet. Tokyo (2004)
9. Rong-Tai, L., Nen-Fu, H., Chin-Hao, C., Chia-Nan, K.: A Fast String-matching Algorithm for Network Processor-based Intrusion Detection System. ACM Transactions on Embedded Computing System, Volume 3, Issue 3 (2004)
10. Vtune Performance Analyzer. http://www.intel.com/software/products/
11. OProfile. http://sourceforge.net/projects/oprofile/
12. SPECWeb99 Benchmark. http://www.spec.org/osg/web99/
13. David, M., Tai, J.: Httperf: A Tool for Measuring Web Server Performance. Proceedings of the SIGMETRICS Workshop on Internet Server Performance. Madison (1998)

Attack Scenario Construction Based on Rule and Fuzzy Clustering

Linru Ma[1], Lin Yang[2], and Jianxin Wang[2]

[1] School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, Hunan, China
[2] Institute of China Electronic System Engineering, Beijing 100039, China

Abstract. Correlation of intrusion alerts is a major technique in attack detection to build attack scenario. Rule-based and data mining methods have been used in some previous proposals to perform correlation. In this paper we integrate two complementary methods and introduce fuzzy clustering in the data mining method. To determine the fuzzy similarity coefficients, we introduce a hierarchy measurement and use weighted average to compute total similarity. This mechanism can measure the semantic distance of intrusion alerts with finer granularity than the common similarity measurement . The experimental results in this paper show that using fuzzy clustering method can reconstruct attack scenario which are wrecked by missed attacks.

1 Introduction

A great deal of research activity has been witnessed dealing with IDS alerts correlation and fusion. The rule-based approach and the data mining approach are representative. In existing correlation approaches, there are different strengths and limitations. All of them neither clearly dominate the others, nor solve the problem absolutely. In this paper, we combine the rule-based approach and the data mining approach, which are two complementary correlation methods, to improve the ability of construct attack scenario. Our result show that using fuzzy clustering method to complement the rule-base correlation is feasible for reconstructing the attack scenario separated by missed attack.

The remainder of this paper is organized as follows. Section 2 reviews related work. In section 3, we integrate rule based method and fuzzy clustering to reconstruct attack scenario and give a new algorithm to compute alert attribute similarity. Section 4 gives experimental evidence in support of our approach, and in Section 5 we summarize this work and draw conclusions.

2 Related Work

In [1], the approach models attacks by specifying their preconditions (prerequisites) and post-conditions (consequences). This technique has the potential to discover novel attack scenarios. However, specifying preconditions and post-conditions of attacks requires knowledge of individual attacks, and is time-consuming and error-prone. The

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 328 – 333, 2005.
© Springer-Verlag Berlin Heidelberg 2005

notable shortage of these approaches is that they all depend on underlying abilities of IDSs for alerts. If the IDSs miss the critical attacks, they can't get the true attack scenario, thus provide misleading information.

The approach in [2] uses data mining method with probabilistic approach to perform alert correlation. Conceptual clustering is used to generalize alert attributes [3]. Fuzzy clustering method is applied to building intrusion detection model [4]. Yet, these fuzzy approaches using simple measure to compute the fuzzy distance or similarity between objects are not fine enough to measure the semantic distance. In contrast, we present a new method to compute the fuzzy similarity coefficient.

Our approach is based on such prior work, and extends prior work through integrating rule-based correlation method and fuzzy clustering method. We also introduce a new method to compute fuzzy similarity coefficients.

3 Integrating Fuzzy Clustering to Reconstruct Correlation Graphs

3.1 Application of Rule-Based Correlation Method and Problem Statement

In all rule-based approaches, we are particularly interested in a representative correlation method based on prerequisites and consequences of attacks [1]. This method relies on a knowledge base for prior knowledge about different types of alerts as well as implication relationships between predicates. The result of the rule-based method is alert correlation graphs, and the nodes of the graph represent the hyper-alert type [1].

Rule-based correlation method has the good performance in multi-step attack recognition. But its intrinsic shortage is that it depends on underlying IDSs for attack detection. Thus the results of alert correlation are limited to the abilities of the IDSs. If the IDSs miss critical attacks in the multi-step intrusion, the results of correlation will not reflect the true attack scenario. Consequently, the integrated scenario may be truncated and alerts from the same attack scenario may be split across several dispersive correlation graphs.

3.2 Using Fuzzy Clustering Method as Complementarity

In order to solve the above problem, we integrate the rule-based and data mining based methods. This approach uses the results from the clustering methods to combine correlation graphs generated by the technique introduced in [1]. The clustering method is potential to identify the common features shared by these alerts [4], and therefore to reconstruct the relevant correlation graphs together.

In this paper, we integrate two correlation graphs which both contain alerts from a common cluster. According to the similarity among the alerts, clustering method organizes objects into groups. Since clusters can formally be seen as subsets of the data set, one possible classification method can be whether the subsets are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory, and it requires an object that either does or does not belong to a cluster. Fuzzy clustering methods (FCM) allow objects to belong several clusters simultaneously with different degrees of membership. In many real situations, such as IDSs alerts classification, fuzzy clustering is more natural than hard clustering.

The alerts to be classified are the objects to study. A set of n objects is denoted by $X = \{x_1, \cdots, x_n\}$, and each object consists of m measured variables, grouped into an m-dimensional column vector $x_i = (x_{i1}, \cdots, x_{im})(i = 1, \cdots, n)$. In order to easy our analysis, we restrict the domain of the alerts to be studied to six kinds of attributes, that is source IP address and port, destination IP address and port, attack type and timestamp. Thus, m=6.

3.3 Calculation of Fuzzy Correlation Coefficient

After determining the objects to be classified and restricting the domain, the most important step of fuzzy clustering is establish of fuzzy similarity relationship, that is calculating the correlation coefficient. Using $r_{ij} \in [0,1]$ to represent the degree of the similarity relationship between two objects x_i and x_j, suppose $\tilde{R}$ is the fuzzy similarity matrix, where $\tilde{R} = (r_{ij})_{n \times n}$ satisfying: $r_{ij} = r_{ji}, r_{ii} = 1(i, j = 1, \cdots, n)$.

To determine the value of r_{ij}, the methods mostly used are similar coefficient method, distance method, minimum arithmetic mean, etc [5]. These common methods are applicable to the objects with numeric attributes. The attribute fields of IDS alerts which we study on are almost all categorical attributes, such as IP address, attack type, etc. Since means only exist for numeric attributes, the traditional similarity algorithm is not applicable to domains with non-numeric attributes.

The common measurement used for dissimilarity of categorical attributes is denoted as follows. x_i, x_j are objects with categorical attributes, and can be represented as an m-dimensional vector $x_i = (x_{i1}, \cdots x_{im})$, where m is the dimensionality of the categorical attributes. The similarity is defined as $d(x_i, x_j) = \sum_{l=1}^{m} \delta(x_{il}, x_{jl})$, where $\delta(x_{il}, x_{jl}) = \begin{cases} 1 & if\ x_{il} = x_{jl} \\ 0 & otherwise \end{cases}$. We observe that this similarity measurement is not fine enough to measure the semantic distance of two objects, because they have only two values 0 and 1 to present the attribute equal or not. In this paper, our approach is, for each attribute we define an appropriate similarity function. The overall similarity is a weighted average of each similarity function. To the similarity function of IP address and port, we propose the hierarchy measurement. In the conceptual clustering, the generalization hierarchy graph is used to generalize the attribute and form the more abstract conception. Different from the front work, we use hierarchy measurement and the algorithmic method in graph theory to compute the similarity of the attributes. We describe this approach in detail as follows.

We define the total similarity between the two alerts as follows:

$$r_{ij} = SIM(x_i, x_j) = \frac{\sum_k \omega_k SIM(x_{ik}, x_{jk})}{\sum_k \omega_k}, k = 1, \cdots, m \tag{1}$$

Where ω_k is the weight of similarity function of the k-th attribute.

The first categorical attribute is IP address, which include source IP address and destination IP address. The hierarchy graph is a tree-structured graph that shows how attribute is organized and which objects are more similar. Figure 1(a) shows a sample attribute hierarchy graph of IP addresses.

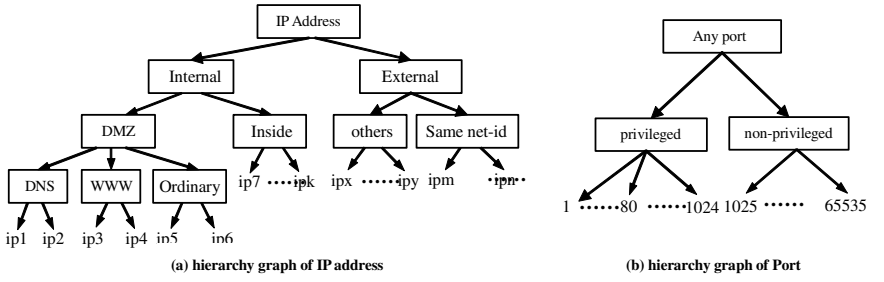


Fig. 1. Hierarchy graphs

The construction of the hierarchy graphs depends on the experts' knowledge and the actual network topology. In order to compute the hierarchy measurement, we give the following definition.

Definition 3.1. An attribute hierarchy graph $G=(N,E)$ is a single-rooted and connected acyclic graph, where the subset $N'(N' \subset N)$ consists of terminal nodes is a set of the value of certain attribute Ai, and for each pair of nodes $n_1, n_2 \in N'$, there is one and only walk W_{n_1,n_2} from n_1 to n_2 in E.

We use $L(n_1,n_2)$ represent the length of the walk W_{n_1,n_2}. The similarity of the two nodes is defined as follows:

$$SIM(x_{n_1}, x_{n_2}) = \frac{L(n_1,n_2)}{\max(L(n_i,n_j))} \quad ,n_1,n_2,n_i,n_j \in N' \tag{2}$$

The denominator is the max of the length of any two nodes in the subset N'. We use formula (2) to compute the similarity of IP address between two alerts.

Computation of the similarity of the port attribute is analogous to the IP address attribute. Figure 1(b) shows a sample attribute hierarchy graph of port.

To compute the similarity of attack type attribute and timestamp attribute, we use the definition 3.1. It is the Hamming distance [5] of two objects with categorical attributes. Different from the standard definition 3.1, we use the range of the time as condition when we compute the similarity of timestamp. We extend the definition as follows: Given the time range ε, $\delta(x_{il}, x_{jl}) = \begin{cases} 1 & if \left| x_{il} = x_{jl} \right| \le \varepsilon \\ 0 & otherwise \end{cases}$. The correlation coefficient can be calculated as above. Fuzzy similar matrix $\tilde{R}$ may not be transferable. Through the square method in [5], we get the transitive closure of $\tilde{R}$, which is the fuzzy equivalent matrix R^*, then, a dynamic cluster result can be obtained.

At last, we use the clustering result to integrate related correlation graphs. We integrate two correlation graphs when they both contain alerts from a common cluster. To simplify the problem, we used a strait-forward approach which uses the alert timestamp to hypothesize about the possible causal relationships between alerts in different correlation graphs.

4 Experiment and Result

We have developed an off-line alert correlation system based on [1] and our fuzzy clustering algorithm, and performed several experiments using the DARPA 2000 intrusion detection evaluation datasets.

In the experiments, we performed our tests on the inside network traffic of LLDOS 1.0. To test the ability of our method, we design three experiments and drop different alerts that Real Secure Network Sensor detected as table 1 shows. Thus, the correlation graph is split into multiple parts.

Table 1. Experiment results

Test Data	alerts	Hyper alerts	cluster number	reconstruct result	missing alert	
					name	NUM
original data	922	44	null	null	null	null
experiment 1	908	28	6	succeed	Sadmind_Amslverify_ overflow	14
experiment 2	905	37	8	succeed	Rsh	17
experiment 3	916	27	5	succeed	Mstream_Zombie	6

As mentioned in section 3, we build the hierarchy graph of IP through the description of the network topology. The hierarchy graphs of IP and port are just like figure 1 and figure 2. We try to use the value on average as an example in formula (1), so that neither attribute is given prominence to others. If the fuzzy similarity coefficient of field i and j, r_{ij} is larger than 0.5, there is a strong correlation between filed i and j, and can put them into one cluster. We use the weave net method [5] to compute and get the final clusters.

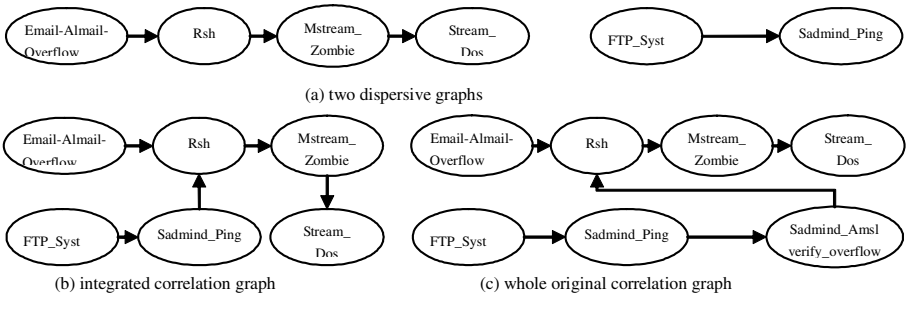


Fig. 2. Reconstruction of attack scenario

In experiment 1, we identify the alert Sadmind_Ping936 to be integrated. Through observing the clustering result, Sadmind_Ping936 and Rsh928 belong to the same cluster, the coefficient r (20,28) is 0.83. Thus, the two correlation graph should be integrated together as figure 3. We compute the fuzzy similarity equivalent matrix R^* which is not showed here due to space limit.

In order to evaluate the result of integration of the two graphs, we use the original network traffic of LLDOS 1.0 and get the whole correlation graph in figure 3. From this figure, we can see that Sadmind_Ping and Rsh have causal relationship, and integrating them is correct. The graphs of experiment 2 and 3 are not shown here due to space limit.

5 Conclusion and Future Work

This paper integrates rule-based approach and data-mining approach to correlate alerts in IDSs. The rule-based approach depended on IDSs output can not reconstruct the whole attack scenario, if the critical attack is missed. In our method, fuzzy clustering which can catch the intrinsical relationship of the alerts is applied to reconstruct the whole attack scenario entirely. In the process of clustering, a novel method - attribute hierarchy measurement is applied to compute the alert attribute similarity. At last, weighted average method is used to obtain the whole attribute similarity.

However, we still lack valid evaluation of the proposed methods in this paper. We combine the separate correlation graphs directly without reasoning the missed attacks. We will try to settle the problems in the following research.

References

1. P. Ning, Y. CUI, and D. S. Reeves. Constructing attack scenarios through correlation of intrusion alerts. In Proceedings of the 9th ACM Conference on Computer and Communications Security, Washington, D.C., (2002)245–254.
2. A. Valdes and K. Skinner. Probabilistic alert correlation. In Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection (RAID 2001). (2001)54–68.
3. K. Julisch. Clustering intrusion detection alarms to support root cause analysis. ACM Transactions on Information and System Security. 6, 4 (Nov.), (2003)443–471.
4. H. Jin, Jianhua Sun. A Fuzzy Data Mining Based Intrusion Detection Model. Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems (FTDCS'04). 2004.
5. P. Y. Liu and M. D. Wu. Fuzzy theory and its applications. National University of Defense Technology Press. China, 1998.

A CBR Engine Adapting to IDS

Lingjuan Li[1], Wenyu Tang[1], and Ruchuan Wang[1,2]

[1] Nanjing University of Posts and Telecomm., Dept. of Computer Science and Technology, Nanjing 210003, China
lilj@njupt.edu.cn
[2] Nanjing University, State Key Laboratory for Novel Software Technology, Nanjing 210093, China

Abstract. CBR is one of the most important artificial intelligence methods. In this paper, it is introduced to detect the variation of known attacks and to reduce the false negative rate in rule based IDS. After briefly describes the basic process of CBR and the methods of describing case and constructing case base by rules of IDS, this paper focuses on the CBR engine. A new CBR engine adapting to IDS is designed because the common CBR engines cannot deal with the specialties of intrusion cases in IDS. The structure of the new engine is described by class graph, and the core class as well as the similarity algorithm adopted by it is analyzed. At last, the results of testing the new engine on Snort are shown, and the validity of the engine is substantiated.

1 Introduction

Intrusion Detection System (IDS) is one of the most important technologies of information security. In rule based IDS, the attributes of intrusion behavior are elicited to compose detection rules; exact match is made between the captured data packages and rules; if they are matched, the intrusion behavior is confirmed, and an intrusion alarm is output. If rules are too common or special, there will be many wrong or missing reports. That will reduce the accuracy of intrusion detection.

Case Based Reasoning (CBR), one of the most important artificial intelligence methods, does similar matching between old case and new sample. It focuses on best match rather than exact match. So by using CBR to detect intrusion, attacks escaping from IDS detection rules can be detected and the false negative rate can be reduced.

However, intrusion cases in IDS have some specialties, for example, cases corresponding to different attacks usually include different attributes, and the same attribute often has different importance (i.e. weight) in different cases. That means the common CBR engine used for case base with fixed attributes and weights is not suitable for IDS. So we design a new CBR engine adapting to IDS (simply denoted as new engine).

In this paper, firstly, the basic process of CBR and the methods of describing case and constructing case base by rules of IDS are briefly described. Then, the structure of the new engine is depicted with class graph, and the core class as well as the similarity algorithm it adopts is analyzed. At last, the results of testing the engine on Snort, a rule based IDS, are shown.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 334 – 339, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 CBR in IDS

The case is the summary of answers to former questions. Right cases are saved in case base. When a new sample (we call it target case) is handled, CBR computes the similarities between target case and each source case in case base, the case which is most similar to the target case will be found out, and it will be modified to match the target case. The modified case will be added in the case base as a new one[1].

2.1 The Basic Process of CBR

The basic process of CBR is as follows:

(1) Pretreatment and filtration: In rule based IDS, many rules have set thresholds for determinant attributes. If one user's behavior makes the determinant attributes' values of one case larger than or equal to their thresholds (i.e. the target case exactly matches the detection rule), it will be confirmed to be attack. It means that reasoning process doing on those that can be detected by rules will increase system cost. Therefore we suggest that pretreatment and filtration should be carried out before reasoning. Pretreatment recalculates the values of time related attributes based on the intervals defined by rules. Filtration discriminates the target case that exactly matches a certain detection rule. After pretreatment and filtration, the target case that cannot match any rule will be handled by next step. In a word, we try to make good use of the detection ability of rule based IDS, and enhance it by using CBR.
(2) Describing target case: This step mainly analyzes target case and extracts attributes in order to search source cases in case base.
(3) Searching: This step mainly searches case base for suitable cases. Similarity computing is of great importance to this step, and the result of this step is a similarity list. Cases founded should be similar to target case, and fewer cases should be better.
(4) Drawing conclusion: In this step, detection results will be ascertained, and a similarity threshold is referenced. If no similarity in step 3 is larger than or equal to the similarity threshold, the behavior related to the target case is confirmed not to be an attack. Otherwise, next step should be taken.
(5) Correcting case and saving it: This step mainly modifies the founded cases in order to match the target case, and adds modified cases into case base.
(6) Outputting attack name.

2.2 Describing Case and Constructing Case Base

Describing Case is coding knowledge into a group of data structure which can be accepted by computer. Adequate description can make problem easier to solute effectively, whereas inadequate description will result in troubles and lower effect.

It's a new idea to apply CBR to intrusion detection, there is no mature case set to refer to, and constructing a new case base is a burdensome task that cannot be all-around. So we propose to construct the case base by the rule set of rule based IDS, and attributes must be evaluated with proper weight for each case in order to improve the accuracy of reasoning because intrusion cases have specialties (see section 1).

3 A CBR Engine Adapting to IDS

The third step mentioned in section 2.1 plays an important role in reasoning. Many functions such as weight processing needed by IDS and similar matching are all in it. We design a CBR engine to implement these functions based on an existing searching engine [see reference 2]. We use the Nearest Neighbor Algorithm (i.e. NN algorithm) to do similar matching, and improve it for implementing and case specialty in IDS.

3.1 The Structure of the New Engine

The structure of the new engine is shown in Figure 1[2], in which:

IDS_CBREngine is the core class, which computes the similarity of each source case to target one. ComputeSimilarity( ) is main method of IDS_CBREngine. It takes IDSItems, IDS_CBRCriteria and DataBaseWeight as its arguments, and returns collection IDS_CBRItems that ranks the similarity values in descending order.

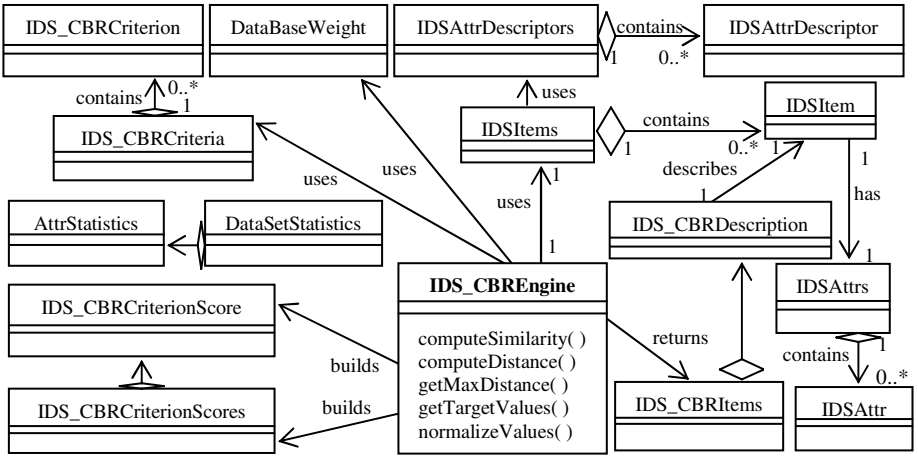


Fig. 1. The structure of the CBR Engine Adapting to IDS

IDSItems is a collection of IDSItem objects. An IDSItem is an intrusion case.

IDSAttrs is a collection of IDSAttr objects. An IDSAttr object is one attribute of the intrusion case. It is just a name/value pair, i.e. attribute name/ attribute value.

IDSAttrDescriptors is a collection of IDSAttrDescriptor objects. The object plays the meta-data role. It is also a name/value pair, i.e. attribute name/ attribute data type. It recognizes integers, floating point numbers, strings and booleans.

IDS_CBRCriteria is a collection of IDS_CBRCriterion objects. IDS_CBRCriterion describes the relationship made up of an attribute, an operator and a value. It can determine the max and min values for each of the IDSItem's attributes.

It's deserved to emphasize that the main difference between the new engine and the existing engine is the storage and passing of the weights. In IDS, each intrusion case has a particular group of weights stored in the database with the intrusion case. So in the

engine shown in Figure 1, before passed to engine, every group of weights is read from the database and stored in an array by object DataBaseWeight. Thus each intrusion case uses a group of weights suitable for it to calculate its weighted distance to target case. That meets the special needs of weight process in IDS.

3.2 Core Class Analysis

Core class IDS_CBREngine mainly makes similar matching between source case and the target case. NN algorithm is the most used algorithm for matching search in case study. It views attribute vector of the case as a point in the multi-dimensional space, and searches for the nearest point to the point of target sample. The distance between the tested point and the sample is determined by the weighted Euclidean distance.

IDS_CBREngine adopts NN algorithm, but improves it for implementing and case specialty in IDS. The process of core class IDS_CBREngine is as follows:

Suppose that, there is a target case $\boldsymbol{O}$ which has n attributes, $\boldsymbol{O^i}$ means the $\boldsymbol{i}$th attribute. There are $\boldsymbol{m}$ cases in case base, $\boldsymbol{C_k}$ means the $\boldsymbol{k}$th case, $\boldsymbol{C_k^{\,i}}$ means the $\boldsymbol{i}$th attribute of the $\boldsymbol{k}$th case, $\boldsymbol{W_k^{\,i}}$ means the $\boldsymbol{i}$th attribute's weight of the $\boldsymbol{k}$th case.

Step 1: Get the data setting information, i.e. read all the IDSItem objects and get MAX[i], MIN[i] and RANGE[i] of every numeric attribute.

$$MAX[i] = \max(C_1^i, C_2^i, C_3^i, ..., C_k^i, ..., C_{m-1}^i, C_m^i)$$

$$MIN[i] = \min(C_1^i, C_2^i, C_3^i, ..., C_k^i, ..., C_{m-1}^i, C_m^i)$$

$$RANGE[i] = MAX[i] - MIN[i]$$

Step 2: Normalize each attribute value of the source case and that of the target case. Then, acquire the weighted normalization value.

Normalized value of source case : $N_Cvalue[C_k^i] = (C_k^i - MIN[i]) / RANGE[i]$

Normalized value of target case : $N_Ovalue[O^i] = (O^i - MIN[i]) / RANGE[i]$

Weighted normalized value of source case:

$$W_N_Cvalue[C_k^i] = N_Cvalue[C_k^i] * W_k^i$$

Weighted normalized value of target case:

$$W_N_Ovalue[O^i] = N_Ovalue[O^i] * W_k^i$$

Step 3: Get the max distance of the $\boldsymbol{k}$th case.

Max distance of the $\boldsymbol{k}$th case: $\max D = \sqrt{\sum_{i=1}^{n} (W_k^i)^2}$

Step 4: Calculate the weighted Euclidean distance between case $\boldsymbol{k}$ and case $\boldsymbol{O}$.

$$D(O, C_k) = \sqrt{\sum_{i=1}^{n} (W_N_Cvalue[C_k^i] - W_N_Ovalue[O^i])^2}$$

Step 5: Compute the similarity between the $\boldsymbol{k}$th case and target case $\boldsymbol{O}$.

$$Similarity(O, C_k) = 1 - D(O, C_k) / \max D$$

Repeat the above process, compute the similarity between target case and each case in case base, rank the similarity values in descending order and output them.

4 Testing the CBR Engine on Snort

Snort is a rule based network IDS[3]. It cannot avoid omitting some attacks, especially those deliberately escape from detection rules. In order to test the effect of the new engine, several experiments were carried out on Snort.

Figure 2 shows several intrusion cases used in the experiments. They are derived from Snort rules, and their attributes are weighted properly. The attribute with weight "0" is not the component of the case. The first case means: In 60 seconds, if one host establishes connections with 20 ports of 5 hosts, it could be thought as the Port Scan Attack and the system should raise the alarm.

Pro	w1	SH	w2	DH	w3	DP	w4	TS	w5	SC	w6	CC	w7	Http	w8	MB	w9	AN
TCP	3	1	0	5	4	20	10	60	4	0	0	0	0	0	0	0	0	Port Scan Attack
ICMP	5	1	0	1	0	0	0	10	4	24	10	0	0	0	0	0	0	ICMP Ping Attack
TCP	3	1	0	50	4	1	1	20	4	254	10	0	0	0	0	0	0	Red Code Attack
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

Pro: Protocol
SH: Source Host number
DH: Destination Host number
DP: Destination Port number
TS: Time Space
SC: Same Character number
CC: Continuous Character number
HTTP: HTTP request
MB: Memory Buffer
AN: Attack Name

Fig. 2. Attack cases in case base

In one of the experiments, we intruded Snort through scanning 18 ports, 3 hosts in 60 seconds. Obviously it is an attack deliberately escaping from detection rules.

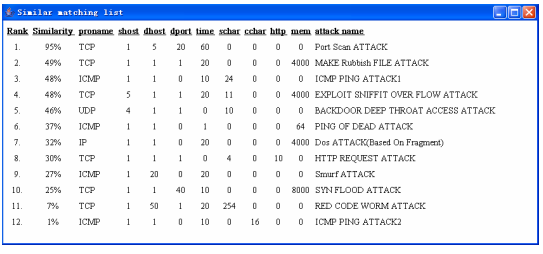
In such case, Snort failed to detect this attack because it doesn't exactly match any rules of Snort. The detection result produced by running the existing engine is as Figure 3. The detection result of running the new engine is displayed in Figure 4.

Similar matching list

Rank	Similarity	proname	shost	dhost	dport	time	schar	cchar	http	mem	attack name
1.	97%	TCP	1	5	20	60	0	0	0	0	Port Scan ATTACK
2.	68%	TCP	1	1	1	20	0	0	0	4000	MAKE Rubbish FILE ATTACK
3.	67%	TCP	5	1	1	20	11	0	0	4000	EXPLOIT SNIFFIT OVER FLOW ATTACK
4.	67%	ICMP	1	20	0	20	0	0	0	0	Smurf ATTACK
5.	65%	ICMP	1	1	0	10	24	0	0	0	ICMP PING ATTACK1
6.	63%	IP	1	1	0	20	0	0	0	4000	Dos ATTACK(Based On Fragment)
7.	63%	ICMP	1	1	0	1	0	0	0	64	PING OF DEAD ATTACK
8.	62%	UDP	4	1	1	0	10	0	0	0	BACKDOOR DEEP THROAT ACCESS ATTACK
9.	52%	TCP	1	1	1	0	4	0	10	0	HTTP REQUEST ATTACK
10.	52%	TCP	1	1	40	10	0	0	0	8000	SYN FLOOD ATTACK
11.	44%	ICMP	1	1	0	10	0	16	0	0	ICMP PING ATTACK2
12.	43%	TCP	1	50	1	20	254	0	0	0	RED CODE WORM ATTACK

Fig. 3. The detection result of the existing engine

It is clearly shown in Figure 3 and Figure 4 that CBR method can effectively catch the attack deliberately escaping from Snort detection rules. However, the existing engine outputs many cases with high similarity to the target case due to all cases using the same group of weights, it is obviously unpractical to IDS. On the contrary, there were proper weights for each intrusion case in the new engine, and as a result of that, only the case of Port Scan Attack was 97% similar to the target case, the similarities of other cases were all less than 50%, the attack was detected accurately.


Similar matching list

Rank	Similarity	proname	shost	dhost	dport	time	schar	cchar	http	mem	attack name
1.	95%	TCP	1	5	20	60	0	0	0	0	Port Scan ATTACK
2.	49%	TCP	1	1	1	20	0	0	0	4000	MAKE Rubbish FILE ATTACK
3.	48%	ICMP	1	1	0	10	24	0	0	0	ICMP PING ATTACK1
4.	48%	TCP	5	1	1	20	11	0	0	4000	EXPLOIT SNIFFIT OVER FLOW ATTACK
5.	46%	UDP	4	1	1	0	10	0	0	0	BACKDOOR DEEP THROAT ACCESS ATTACK
6.	37%	ICMP	1	1	0	1	0	0	0	64	PING OF DEAD ATTACK
7.	32%	IP	1	1	0	20	0	0	0	4000	Dos ATTACK(Based On Fragment)
8.	30%	TCP	1	1	1	0	4	0	10	0	HTTP REQUEST ATTACK
9.	27%	ICMP	1	20	0	20	0	0	0	0	Smurf ATTACK
10.	25%	TCP	1	1	40	10	0	0	0	8000	SYN FLOOD ATTACK
11.	7%	TCP	1	50	1	20	254	0	0	0	RED CODE WORM ATTACK
12.	1%	ICMP	1	1	0	10	0	16	0	0	ICMP PING ATTACK2

Fig. 4. The detection result of CBR engine adapting to IDS

5 Conclusion

As exact match used by rule based IDS may omit reporting intrusion attack, this paper applies CBR method to rule based IDS. As common CBR engine cannot adapt to the specialties of cases in IDS, this paper designs and implements a new CBR engine adapting to IDS, which improves traditional NN algorithm for implementing and case specialty in IDS. By testing it on Snort, the rationality and validity of the new engine are proved. In a word, this paper has made efforts to use CBR to enhance the ability of IDS and provided a base for further research on applying CBR to IDS.

Acknowledgment

We would like to thank the author of reference 2. We design the CBR engine adapting to IDS based on what he has done. This research is sponsored by the National Natural Science Foundation of China (No.60173037 & 70271050), the Natural Science Foundation of Jiangsu Province (No.BK2005146, No.BK2004218), and other Foundations (No.BG2004004, No. kjs05001).

References

1. Kolodner, J.: Case-based Reasoning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA(1993)
2. Baylor Wetzel: Implementing A Search Engine With Case Based Reasoning. http://ihatebaylor.com/technical/computer/ai/selection_engine/CBR…, 9/7/02
3. Brian Caswell**,** Jay Beale**,** James C.Foster**,** Jeffrey Posluns**.:** Snort 2.0 Intrusion Detection. National Defence Industry Press, Beijing(2004)

Application of Fuzzy Logic for Distributed Intrusion Detection*

Hee Suk Seo[1] and Tae Ho Cho[2]

[1] School of Internet Media Engineering, Korea University of Technology and Education, Gajunri 307, Chunan 330-708, South Korea
histone@kut.ac.kr

[2] School of Information and Communications Engineering, Modeling & Simulation Lab., Sungkyunkwan University, Suwon 440-746, South Korea
taecho@ece.skku.ac.kr

Abstract. Application of agent technology in Intrusion Detection Systems (IDSs) has been developed. Intrusion Detection (ID) agent technology can bring IDS flexibility and enhanced distributed detection capability. However, the security of the ID agent and methods of collaboration among ID agents are important problems noted by many researchers. In this paper, coordination among the intrusion detection agents by BlackBoard Architecture (BBA), which transcends into the field of distributed artificial intelligence, is introduced. A system using BBA for information sharing can easily be expanded by adding new agents and increasing the number of BlackBoard (BB) levels. Moreover the subdivided BB levels enhance the sensitivity of ID. This paper applies fuzzy logic to reduce the false positives that represent one of the core problems of IDS. ID is a complicated decision-making process, generally involving enormous factors regarding the monitored system. A fuzzy logic evaluation component, which represents a decision agent model of in distributed IDSs, considers various factors based on fuzzy logic when an intrusion behavior is analyzed. The performance obtained from the coordination of an ID agent with fuzzy logic is compared with the corresponding non-fuzzy type ID agent.

1 Introduction

At present, it is impossible to completely eliminate the occurrence of security events, all the security faculty can do is to discover intrusions and intrusion attempts at a specific level of risk, depending on the situation, so as to take effective measures to patch the vulnerabilities and restore the system, in the event of a breach of security. IDSs have greatly evolved over the past few years. Artificial intelligence can be applied to the field of ID research. Dickerson proposed the development an IDS based on fuzzy logic, with a core technique of substituting fuzzy rules for ordinary rules so as to more accurately map knowledge represented in natural languages to that represented in computer languages [1]. Siraj argued that a Fuzzy Congnitive Map (FCM)

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 340 – 347, 2005.
© Springer-Verlag Berlin Heidelberg 2005

could be used to support the decision making of intelligent IDSs [3]. This kind of graph reflects the fuzzy cause and effect relationship between events, and is used to calculate the degree of confidence for events, enabling the ID engine to make more accurate decisions. Christopher proposed employing artificial intelligent methods in IDSs in order to recognize attackers' plans.

2 Intrusion Detection and Response

2.1 ID Using BBA

The BB hierarchy is set according to Joseph Barrus & Neil C. Rowe as shown in Fig. 1. They proposed Danger values to be divided into five different levels. These five BB levels are Minimal, Cautionary, Noticeable, Serious and Catastrophic. We classified BB levels into five levels for both host and network attacks based on these divisions. Each agent communicates using two types of messages. The first are control messages, the second are data messages. Control messages are used to communicate between the agents and controller, the data messages are required to send data between the agents and BB.

A network attack is defined where hosts in an external network attack the host network. In this case the attacked hosts insert related intrusion information in the Network-Attack area of the blackboard. During a host attack, the blackboard level is transmitted to the Host-Attack area of the BB. When the BB state is at the Host-Attack level and any other host is attacked, the BB state changes to Network-Attack. The Minimal, Cautionary, Noticeable, Serious and Catastrophic levels of Network-Attack represent the case when multiple hosts are attacked. For example, the Cautionary level of Network-Attack is defined in as when at least two hosts are at the Cautionary level of the Host-Attack area. A host at the Cautionary level of the Host-Attack and host, which detects the beginning of an attack, transmits a Minimal level to a Cautionary. Then the whole network is at the Cautionary level of the Network-Attack area. The message transmission mechanism of the Network-Attack area on the BB is basically similar to that of the Host-Attack area. When the BB level is at a Noticeable level of the Network-Attack area in the composed simulation environment, then an attacker's packets coming from the attack point are blocked to protect the network. An attacker continuing the attack when the BB level is at the Serious level of the Network-Attack area, all packets coming from the external network are prevented from the damaging the host network. Fig. 1 shows the communication within IDS and BB architecture during the jolt attack.

This case is for the detection of attacks to a single host within the network presented. In this case the attacked host inserts intrusion related information to the Host-Attack area of the blackboard. Each agent must request permission by transmitting a BB_update_request message to the controller, in order to manage consistency and contention problems. The controller transmits an BB_update_permit message to the agent capable of handling the current problem. The agent receiving this message writes (BB_update_action) the intrusion related information to the BB. After updating is completed, the agent transmits the BB_update_completion message to the

controller. Controller transmits a BB_broadcasting_of_action_request message for reporting this event to other IDSs. IDSs, receiving the necessary information from the BB, transmit the BB_information_acquisition_completion message to the controller. The BB levels are transmitted according to these steps. When the BB level is at the Serious level, the agent adds the source IP address to the blacklist of the Firewall model using the network management policy, then all subsequent packets from these sources are blocked.

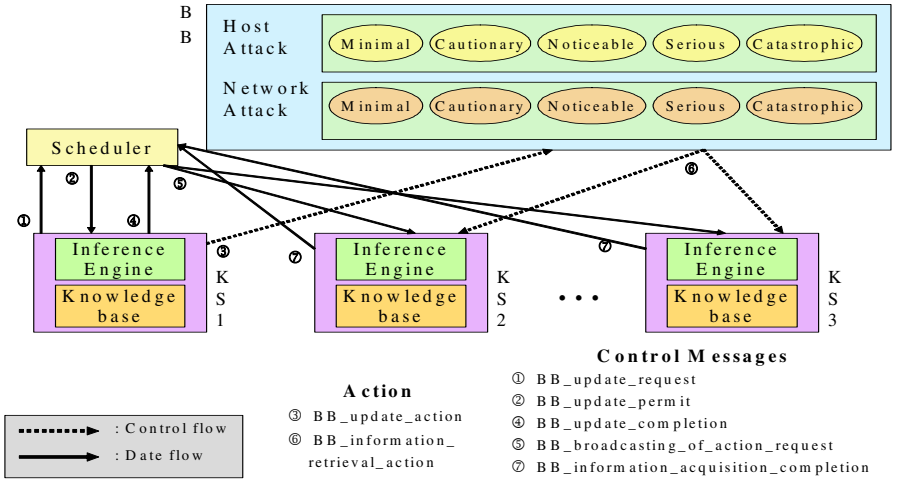


Fig. 1. Messages of IDS and BBA

2.2 Intrusion Evaluation Using Fuzzy Logic

The fuzzy evaluation component, which is a model of a decision-making agent in distributed IDSs, considers various factors for fuzzy logic calculation when an intrusion behavior is analyzed. The BB level selection in the previous system is performed according to the agent with the highest BB level among the participating agents. Namely, if an ID agent A is at the Noticeable level and other agents are at the Cautionary level, then the level of the overall system is decided as the Noticeable level, representing the highest level. When the threshold value of Cautionary level is between 31 and 40 and that of Noticeable level is between 41 and 50, it is paramount that the threshold changes. The threshold change from 38 to 39 is not very important, but from 40 to 41 is very important. Since a difference of 1 changes the level of the agent. Therefore, fuzzy logic has been applied for finding a solution to the problem. The proposed fuzzy evaluation model using the fuzzy logic consists of four steps. First, the degree of the input three basic steps and conditions of the fuzzy rules are calculated. Second, the rule's conclusion based on its matching degree is calculated. Third, the conclusion inferred by all fuzzy rules is combined into a final conclusion and finally the crisp output is calculated.

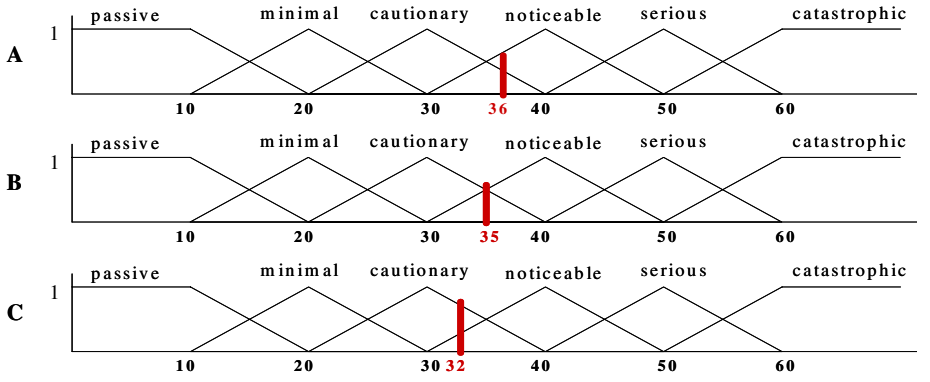


Fig. 2. Membership function of each agent

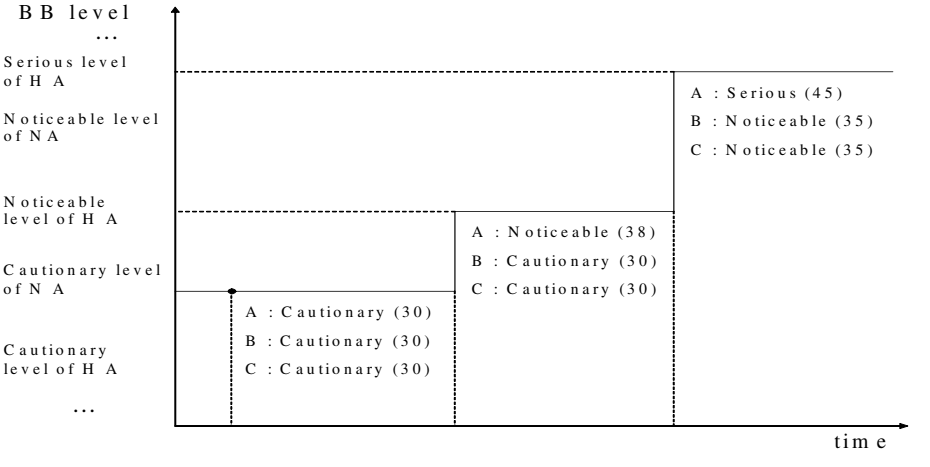


Fig. 3. The BB level transition of non-fuzzy system

Table 1. The level decision in the non-fuzzy system

BB level	level selection
Cautionary level of Network Attack	MAX(Cautionary(A) , Cautionary(B) , Cautionary(C))
Noticeable level of Host Attack	MAX(Noticeable(A) , Cautionary(B) , Cautionary(C))
Serious level of Host Attack	MAX (Serious(A) , Noticeable(B) , Noticeable(C))

Fig. 2 presents the membership function of each agent. Fuzzy inference is used in scaling for fast computation. The Standard Additive Model (SAM) has been applied for combining fuzzy logic. The structure of fuzzy rules in the SAM is identical to that of the Mandani model. This is achieved by analyzing the outputs of the individual

fuzzy rules and then defuzzifying the combined outcome to obtain a crisp value is the standard procedure. The defuzzification value is obtained as below.

$$CoA(C) = \frac{Area(C_1)*CoA(C_1)+ Area(C_2)*CoA(C_2)+ Area(C_3)*CoA(C_3)}{Area(C_1) + Area(C_2) + Area(C_3)}$$

In the two systems shown as Fig.3, one is a non-fuzzy system and the other is a proposed fuzzy system, presented by the level transition of the BB. These are ID agent A, ID agent B, and ID agent C in the current network. The level of all ID agents is assumed at the Cautionary level as Fig. 3. If an attacker continuously attacks ID agent A in the non-fuzzy system, then the threshold of ID agent A is changed to 38. The BB level is transmitted as the Noticeable level of Host Attack. If ID agent A is continuously attacked, the threshold value reaches the Serious level. The other two ID agents are transmitted as the Noticeable level using the ID agent A.

The level transition is presented in the proposed fuzzy system as Fig 4. The threshold value of all ID agents is assumed to be 30 (the Cautionary threshold value of the non-fuzzy system). Though the threshold of ID agent A is modified at 38 (the Noticeable threshold value of the non-fuzzy system) by the attack, the blackboard level, decided by the fuzzy logic calculation, is still the Cautionary level of Network Attack. If ID agent A is transmitted as the Serious level (threshold value is at 45), the other two agents are transmit the Noticeable level (threshold value is at 35). In this case, the BB level is transmitted as the Serious level of Host Attack in the non-fuzzy system, however, the BB level is transmitted as the Noticeable level of Network Attack in the fuzzy system. If the threshold value of ID agent A reaches 53, B reaches 42, and C reaches 44, then the BB level is transmitted as the Serious level of Host Attack.

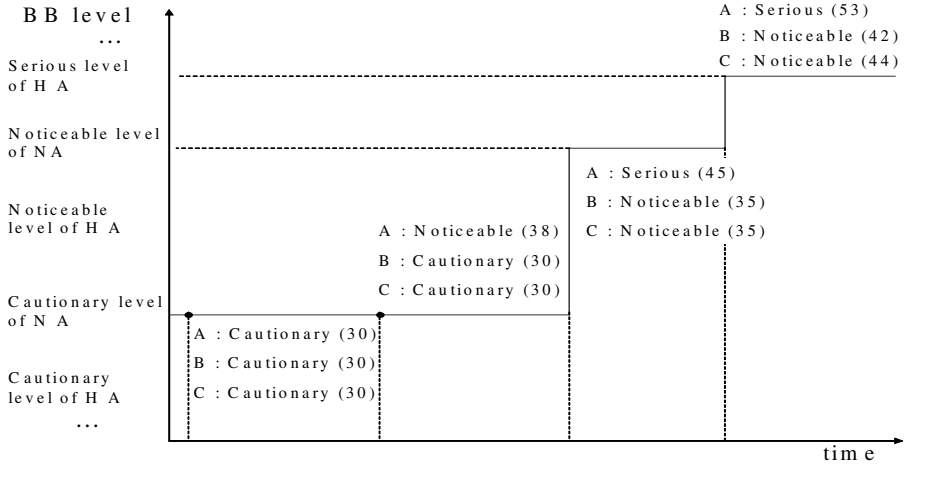


Fig. 4. The BB level transition of fuzzy system

Table 2. The level decision in the fuzzy system

BB Level	Fuzzy matching Rule (A ID:B ID:C ID)	Combing	Defuzzi-fication
Cautionary level of Network Attack	(A:38, B:30, C:30)	$\frac{2*30+8*30}{2+8}$	30
	(N,C,C)		
Noticeable level of Network Attack	(A:45, B:35, C:35)	$\frac{5*30+5*40+5*40+5*40}{5+5+5+5} + \frac{5*30+5*40+5*40+5*40}{5+5+5+5}$	38
	(N,C,C) (S,C,C) (N,C,N) (S,C,N) (N,N,C) (S,N,C) (N,N,N) (S,N,N)		
Serious level of Host Attack	(A:53, B:42, C:44)	$\frac{6*40+4*50+2*50+2*50}{6+4+2+2} + \frac{3*40+3*50+2*50+2*50}{3+35+2+2}$	46
	(S,N,N) (Ca,N,N) (S,N,S) (Ca,N,S) (S,S,N) (Ca,S,N) (S,S,S) (Ca,S,S)		

3 Simulation Result

Simulation is demonstrated for two cases. The first case shows when only the BB detects the intrusion, the second uses the BB and fuzzy logic to detect the intrusion. The jolt and mailbomb attacks are used for simulation in both cases.

The ID time, False Positive Error Ratio (FPER) and False Negative Error Ratio (FNER) are measured as performance indexes in the simulation. This is presented in previous studies where the ID agents using the BB for coordination are superior in ID to those not using a blackboard.

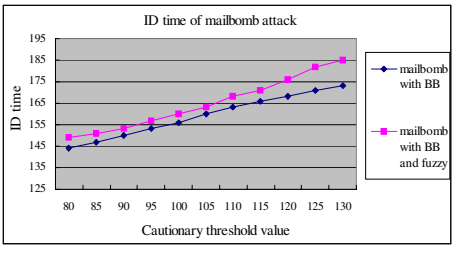


Fig. 5. ID time of mailbomb attack

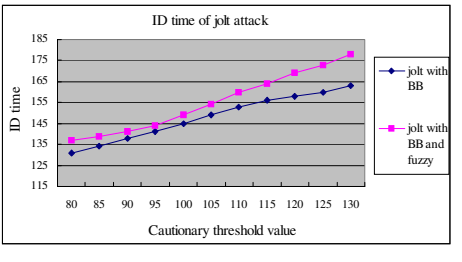


Fig. 6. ID time of jolt attack

Fig. 5,6 present the ID time for the mailbomb and jolt attacks. The selected BB levels for the simulation are at the Cautionary level. The other threshold values are modified according to the same ratio of change in this Cautionary level's values. The system only using the BB detects the intrusion faster than the system using the both BB and fuzzy logic for all the threshold values. The faster the intrusion is detected, the earlier administrators can correspond to the intrusion. When the security level is weakened, by increasing the Cautionary threshold value in the proposed system, the difference in the ID time between the proposed system and the previous system increases. The stronger the sensitivity becomes because of the information sharing among IDSs, resulting from the lower security level. This phenomenon related to the sensitivity applies to all other simulated results. The simulation results of ID time shows that the system using fuzzy logic does not rapidly transmit its level. In the case of using only the BB, the system level is decided by one ID agent that first transmits a different level. However, in the case of using fuzzy logic, the system level is decided by considering the status of all ID agents. As a result, the system using fuzzy logic does not detect the intrusion faster than the system only using the BB, but more accurately detects the intrusion using overall network information.

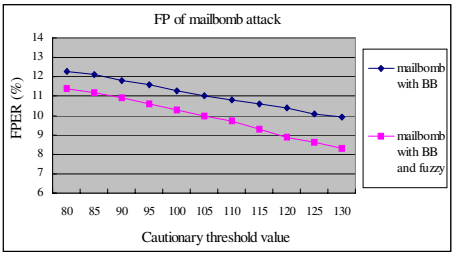


Fig. 7. FPER of mailbomb attack

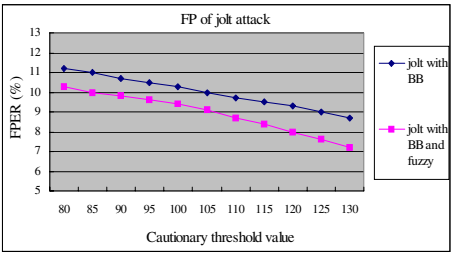


Fig. 8. FPER of jolt attack

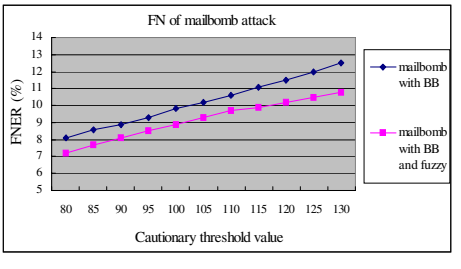


Fig. 9. FNER of mailbomb attack

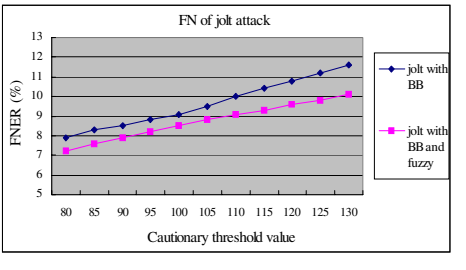


Fig. 10. FNER of jolt attack

Fig. 7,8 presents the false positive error ratio of the system using the BB levels and the system using the BB and fuzzy logic for the mailbomb and jolt attack. A false positive represents an incorrect alarm from acceptable behavior. Fig. 7,8 present an increasing false positive error ratio by strengthening the security level. This increase in the error ratio is due to the fact that the higher the security level, the more error IDSs make in both cases. The FPER of the system using fuzzy logic is lower than the

system using only the BB. The simulation results present the ID agent using fuzzy logic more accurately detects the intrusion using overall network information. Nowadays one of the main problems of an IDS is the high false positive rate. The number of alerts that IDS launches is clearly higher than the number of real attacks. The proposed system lessens the false positive error ratio by using fuzzy logic.

Fig. 9,10 presents the false negative error ratio of the system using BB levels and the system using BB and fuzzy logic for the mailbomb and jolt attacks. A false negative represents a missed alarm condition. Fig. 9,10 presents a decrease in the false positive error ratio as the security level is increased. For all cases, the error ratio of the proposed system is lower than that of the previous system, since intrusions that are detected are based on shared information. The fuzzy logic evaluation component, which is a model of a decision agent in the distributed IDS, considers various factors based on fuzzy logic when intrusion behavior is judged. The performance obtained from the coordination of intrusion detection agent with fuzzy logic is compared with the corresponding intrusion detection agent. The results of these comparisons demonstrate a relevant improvement when fuzzy logic is involved.

4 Conclusion and Future Work

Prevention, detection and response are critical for comprehensive network protection. If threats are prevented then detection and response are not necessary. ID is the art of detecting and responding to computer threats. If multiple intrusion detection agents share intrusion related information with one another, detection capability can be greatly enhanced. A system using BB architecture for information sharing can easily be expanded by adding new agents, and by increasing the number of BB levels. Collaboration between the firewall component and the IDS will provide the added efficiency in protecting the network. Policy-based network management provides a method by which the administration process can be simplified and largely automated. The proposed system has this advantage in terms of ID management. The administrator can easily apply policies to network components (network devices and security systems) with the policy-based system.

References

1. S. Northcutt, Network Intrusion Detection - An Analyst's Handbook, New Riders Publishing, 1999.
2. B.P. Zeigler, H. Praehofer, and T.G. Kim. Theory of Modeling and Simulation 2ed. Academic Press, 1999.
3. J.E. Dickerson, J. Juslin, O. Koukousoula, J.A. Dickerson, "Fuzzy intrusion detection," In IFSA World Congress and 20th NAFIPS International Conference, pp. 1506-1510, 2001.
4. R. Bace, Intrusion Detection, Macmillan Technical Publishing, 2000.
5. H.S. Seo, T.H. Cho, "Simulation Model Design of Security System based on Policy-Based Framework," Simulation Transactions of The Society for Modeling and Simulation International, vol. 79, no. 9, pp. 515-527, Sep. 2003.

Dynamic Access Control for Pervasive Grid Applications*

Syed Naqvi and Michel Riguidel

Computer Sciences and Networks Department, Graduate School of Telecommunications,
46 Rue Barrault, Paris 75013, France
{naqvi, riguidel}@enst.fr

Abstract. The current grid security research efforts focus on static scenarios where access depends on the identity of the subject. They do not address access control issues for pervasive grid applications where the access privileges of a subject not only depend on their identity but also on their current context (i.e. current time, location, system resources, network state, etc). Our approach complements current authorization mechanisms by dynamically granting permission to users based on their current context. The underlying dynamic and context aware access control model extends the classic role based access control, while retaining its advantages (i.e. ability to define and manage complex security policies). The major strength of our proposed model is its ability to make access control decision dynamically according to the context information. Its dynamic property is particularly useful for pervasive grid applications.

1 Introduction

The grid infrastructure presents many challenges due to its inherent heterogeneity, multi-domain characteristic, and highly dynamic nature. One critical challenge is providing authentication, authorization and access control guarantees. Although lots of researches have been done on different aspects of security issues for grid computing, these efforts focus on relatively static scenarios where access depends on identity of the subject. They do not address access control issues for pervasive grid applications where the access privileges of a subject not only depend on its identity but also on its current context (i.e. current time, location, system resources, network state, etc).

In this paper, we present a dynamic and context aware access control mechanism for pervasive grid applications. Our model complements current authorization mechanisms to dynamically grant and adapt permissions to users based on their current context. The underling dynamic and context aware access control model extends the classic role based access control, while retaining its advantages (i.e. ability to define and manage complex security policies). The model dynamically adjusts role assignments and permission assignments based on context information. In our proposed model, each subject is assigned a role subset from the entire role set by the authority service. Similarly, each object has permission subsets for each role that will access it.

* This research is supported by the European Commission funded project SEINIT (Security Expert Initiative) under reference number IST-2002-001929-SEINIT.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 348 – 355, 2005.
© Springer-Verlag Berlin Heidelberg 2005

During a secure interaction, state machines are maintained by delegated access control agents at the subject to navigate the role subset, and the object to navigate the permission subset for each active role. These state machines navigate the role/permission subsets to react to changes in context and define the currently active role at the subject and its assigned permissions at the object.

Besides simulations, a prototype of this model has been implemented. The feasibility, performance and overheads of this scheme are also evaluated. This paper is organized as follows. Section 2 gives an overview of the pervasive grids. Our proposed dynamic and context aware access control model is presented in section 3. A description of the prototype implementation and experimental evaluations are given in section 4. Finally, some conclusions are drawn in section 5 along with the future directions of our present work.

2 Pervasive Grids

The term *Grid* refers to systems and applications that integrate and manage resources and services distributed across multiple control domains [1]. Pioneered in an e-science context, grid technologies are also generating interest in industry, as a result of their apparent relevance to commercial distributed applications [2].

Pervasive Grids augment the list of resources available to include peripherals (or sensors) like displays, cameras, microphones, pointer devices, GPS receivers, and even network interfaces like 3G [3]. Devices on a pervasive grid are not only mobile but also nomad – i.e. shifting across institutional boundaries. The pervasive grid extends the sharing potential of computational grids to mobile, nomadic, or fixed-location devices temporarily connected via ad hoc wireless networks. Pervasive grids present an opportunity to leverage available resources by enabling sharing between wireless and non-wireless resources. These wireless devices bring new resources to distributed computing. Pervasive grids offer a wide variety of possible applications. They can reach both geographic locations and social settings that computers have not traditionally penetrated.

However, the dynamic case of unknown devices creates special challenges. Besides inheriting all the typical problems of mobile and nomadic computing, the incorporation of these devices into the computational grids requires overhauling of its access control mechanism. The authorization and access control mechanisms in a wired grid environment focus on relatively static scenarios where access depends on identity of the subject. They do not address access control issues for pervasive grid applications where the access capabilities and privileges of a subject not only depend on its identity but also on its current context (i.e. current time, location, system resources, network state, etc). In this paper, we have proposed a mechanism to dynamically determine the access rights of the subject.

3 Dynamic and Context Aware Access Control Model

One key challenge in pervasive grid applications is managing security and access control. Access Control List (ACL) is inadequate for pervasive applications as it does

not consider context information. In a pervasive grid environment, users are mobile and typically access grid resources by using their mobile devices. As a result the context of a user is highly dynamic, and granting a user access without taking the user's current context into account can compromise security as the user's access privileges not only depend on *who the user is* but also on *where the user is* and *what is the user's state and the state of the user's environment*. Traditional access control mechanisms such as access control list break down in such environments and a fine-grained access control mechanism that changes the privilege of a user dynamically based on context information is required.

In our proposed access control model, a Central Authority (CA) maintains the overall role hierarchy for each domain. When the subject logs into the system, based on his credential and capability, a subset of the role hierarchy is assigned to his for the session. The CA then sets up and delegates (using GSI) a local context agent for the subject. This agent monitors the context for the subject (using services provided by the grid middleware) and dynamically adapts the active role. Similarly every subject maintains a set of permission hierarchies for each potential role that will access the resource. A delegated local context agent at the subject resource uses environment and state information to dynamically adjust the permissions for each role. We assign each user a role subset from the entire role set. Similarly each resource assigns a permission subset from the entire permission set to each role that has privileges to access the resource.

3.1 Some Definitions

3.1.1 Subject

A *subject* is an active entity that can initiate a request for resources and utilize these resources to complete some task. These subjects are basically the grid actors.

3.1.2 Object

An *object* is a passive repository that is used to store information.

3.1.3 Operation

An *operation* is an action that describes how a subject is allowed to use or access an object. These operations are performed according to the roles/privileges of the various subjects.

3.2 Security Objectives

3.2.1 Availability

Availability of a requested service is an important performance parameter. The credit for efficient service deliveries usually goes to the effective resource brokering agents, which should be smart enough to assure the seamless availability of computing resources throughout the executions. The availability factor is important for all the grid applications but for the critical pervasive grid applications, it is one of the most desirable properties. The availability of resources will be absolutely critical throughout the processes to get the information in good time.

3.2.2 Confidentiality
Confidentiality is one of the most important aspects of a security system. This is the property that information not reach unauthorized individuals, entities, or processes. It is achievable by a mechanism for ensuring that only those entitled to see information or data have access to that information.

3.2.3 Integrity
Integrity is the assurance that information can only be accessed or modified by those authorized to do so. Measures taken to ensure integrity include controlling the physical environment of networked terminals and servers, restricting access to data, and maintaining rigorous authentication practices. Data integrity can also be threatened by environmental hazards, such as heat, dust, and electrical surges.

3.3 Assumptions
The initial assumptions of the proposed model are:

3.3.1 Active Users
Active Users constitute a small community made up of its *permanent members*. These active users, with their intelligent device(s) e.g. Personal Digital Assistant (PDA), form a nomadic pervasive grid of personal smart devices. This grid has its own security model, where trust and confidence is established by reputation and proximity. This grid will be assumed as a *trusted* grid.

3.3.2 Public Users
Public Users is composed of unknown subjects. Such users will gain limited grid access by sending their formal registration request along with the traces of some biometric identity through the associated gadget and that pattern will be used for their subsequent authentications.

3.3.3 Technology Updates
The pervasive technology continues to emerge; newer, better, and more powerful devices and gadgets are surfacing regularly. Today's latest technology will be obsolete tomorrow and hence the security architecture will be periodically revised to accommodate these upcoming technologies without altering the core architecture.

3.3.4 Physical Protection
It is very significant in the pervasive grid environment as the physical theft of the small devices is much easier than stealing the fix computers of the computing centres. These devices and gadgets have no physical perimeter security mechanism.

3.4 Security Functions

3.4.1 Auditability
Keeping track of the various uses of medical records is vital for auditability. The pervasive grid applications should be prepared to answer *who*, *what*, *when*, etc. if required. They should ensure that all the activities are logged (by its actors or by the implanted smart devices). Logging features and log analysis will create user and resource audit trails.

3.4.2 Authentication

Authentication is the process of determining whether someone or something is, in fact, who or what it is declared to be. The authenticated users (actors) are given access to various services/applications according to their roles/privileges defined in the authorization table.

3.4.3 Authorization

It is dynamically determined and a matrix of *subjects* versus the *objects* is constantly updated.

3.4.4 Traceability

Traceability of applications data is a policy issue. The stake holders determine the lifecycle of various data items. Traceability also plays an important role in the auditability and the accountability of the actors.

4 Performance Evaluations

We have carried out a number of simulations and have developed a prototype of pervasive grid that uses our proposed security architecture. An example scenario is:

All the teachers and students of our department are supposed to use their PDAs to gain access to the pedagogic resources. Wireless access points are provided in every room of the department. These access points are also used to determine the context of the users. In the library, students can read e-books but can not read their examination paper; whereas in the exam hall, from 9 am to noon, the students can read the examination paper, write the answers file, but can not read books. The teachers can read and write the examination paper from both library and from the exam hall.

A PDA is placed in the quarantine zone if its user:

1. tries more than three unsuccessful log-in attempts as student or more than two unsuccessful log-in attempts as teacher, as he/she may be a potential intruder;
2. is using too much bandwidth, as he/she may be trying to cause the Denial of Service (DoS) attack;
3. is seeking unauthorized privileges.

Placement in a quarantine zone implies that:

1. other users are informed of his/her presence, as a troublemaker;
2. he/she is asked to *behave normally* otherwise he/she will be expelled;
3. after some time Δt it is evaluated whether to clear him/her out the quarantine zone or disconnect him/her from the system. This decision will be based on the close observation of his/her activities during the quarantine period Δt.

As shown in figure 1, two different Wi-Fi access points at our department building are used to model *library* and *exam hall*. PDAs with embedded Wi-Fi card are used to model students (S), teacher (T) and potential attacker (encircled). One PC is used (to be connected from the third Wi-Fi access point) to act as the CA. The overall happening of the system is displayed on its screen including the log of the various actions taken by these PDAs and the time taken by each operation.

4.1 Simulations

Before developing a prototype, we carried out a comprehensive set of simulations to study the effects of scalability and dynamic distribution of trust among the various pervasive grid nodes. Initial results are published in [4].

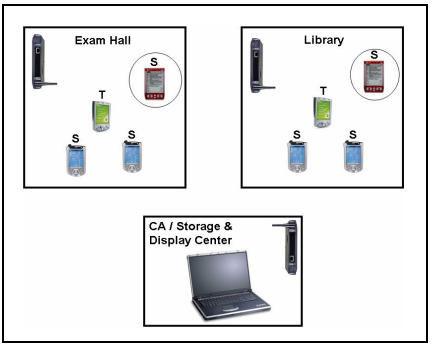


Fig. 1. Experimental Setup

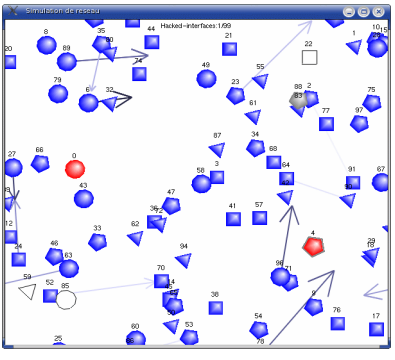


Fig. 2. Simulator graphics display

We consider a bunch of heterogeneous nodes containing some malicious nodes. These blue nodes are mutually trusted nodes until an attack is detected. A malicious node regularly tries to attack the other nodes. Each attack has a probability p of success. This probability depends on the target node type. A successful attack turns the victim node into a new attacking node for the others. However, in the contrary case the attacker is blocked in its firewall and an alert concerning this node is transmitted in the system.

Each node has *xy*-coordinates (figure 2) and its class is determined by its shape (e.g. a triangular shape corresponds to a PDA, a round shape corresponds to a PC, etc.). The color coding used in this scheme is as follows: A node is gray if it does not know about the presence of the malicious node, blue if it is informed of malicious node, and white if it knows all the malicious nodes in the system.

A red halo around a node indicates that it is a victim node (which has become a malicious node itself), blue if the attack was foiled by the security architecture and yellow if the attack failed due to some other reason.

The triangles in the display shows the attack propagation whereas the arrows correspond to the distribution of trust among the nodes. The calculation of the distribution of trust is based on a *trust table*. A trust table is shown in figure 3. The left entry A is the node that evaluates the entry of the node B from the top side. A color code is employed to quickly determine if there remains a danger of attack in the system: green, if A relies on B, and that A and B are indeed trustworthy; red, if A relies on B, and that B belongs to the attacker or is an attack victim; blue, if A does not rely on B and that B is indeed untrustworthy due to the reasons described in the previous case; white, if A's confidence in B has no importance.

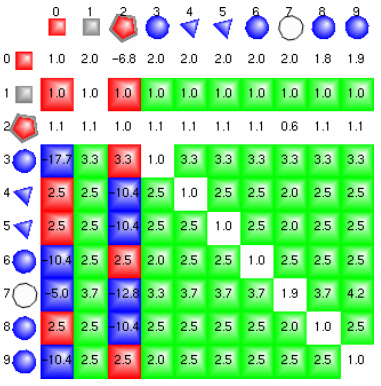


Fig. 3. Ten nodes trust table

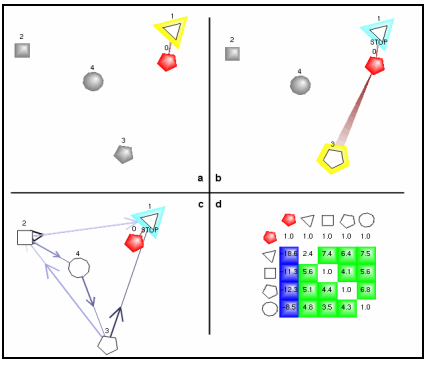


Fig. 4. Failed attack paradigm

Figure 4 presents the collective defense behavior of the nodes with the described infrastructure of confidence. If the attacker fails in its first attempt, it will be difficult for it to take control of the other nodes. Here node 0 escapes an attack from node 1 and blocks its transmissions. The other nodes are promptly informed of the threat so that they do not remain confident in node 0; and hence the overall system is protected (cf. corresponding values in the trust table).

But if the node 0 fell prey to the attack of node 1 (figure 5) and then manages to take control of node 3 all the other nodes will soon be affected resulting in the successful endeavor of the attacker.

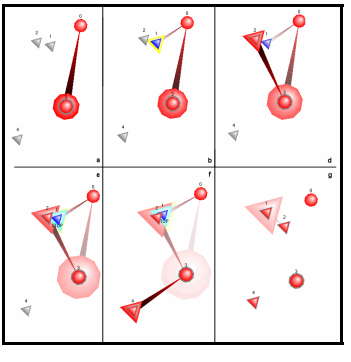


Fig. 5. Successful attack paradigm

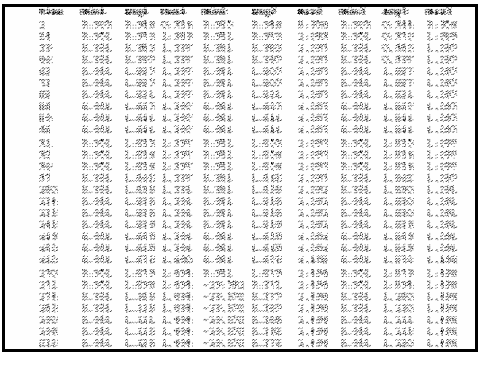


Fig. 6. Contents of a log file

A log file is maintained for keeping the traces of the various actions of the different nodes (figure 6). Its contents include: minimum, average and maximum confidence allotted to a node by the nodes that are not controlled by the attacker.

4.2 Prototype Implementation

The prototype experiments were conducted on one PC using PIII-600MHZ processors, running RedHat Linux 7.2 and four PDAs HP IPAQ HX4700 using 624 MHz

Intel XScale PXA270 with WMMX, running Windows Mobile operating system 2003 second edition. The machines were connected by Ethernet switch. The following factors affect overhead of the proposed model.

- The number of roles assigned to the object.
- The frequency of the events (generated by the context agent at the object) that trigger transitions in the role state machine.
- The number of permissions assigned to each role.
- The frequency of the events that trigger transitions in the permission state machine.

The results show that the overheads of the dynamic and context aware access control model implementation are reasonable. The primary overheads were due to the event generated by the context agent – the higher the frequency, the larger was the overhead. The context agent can be implemented as an independent thread and as a result, the transition overheads at the object and subject are not significant.

5 Conclusions and Future Directions

In this paper, we presented our proposed dynamic and context aware access control mechanism for pervasive grid applications. Our model complements current authorization mechanisms to dynamically grant and adapt permissions to users based on their current context. This approach simply extends the classic role based access control model. A comprehensive set of simulations followed by a prototype implementation for experimental evaluation are carried out to determine the feasibility, performance and overheads of our proposition. The results show that the overheads of the model are reasonable and the model can be effectively used for dynamic and context aware access control for pervasive grid applications.

Dynamic and context aware access control is an important factor for a security architecture of the pervasive grid applications. However, this authorization mechanism should be combined with other security mechanisms (such as authentication) to secure the pervasive grid applications in the real world. We will continue work to include other security mechanisms into our proposed model so that a comprehensive security architecture for the pervasive grid applications could be yielded.

References

1. Foster I., Kesselman C., *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, (1999). ISBN 1558604758
2. Foster I., Kesselman C., Nick J., Tuecke S., *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*, Globus Project, (2002).
3. McKnight L., Howison J., and Bradner S., *Wireless grids: Distributed resource sharing by mobile, nomadic, and fixed devices* IEEE Internet Computing, Jul-Aug (2004) 24-31
4. Naqvi S., Riguidel M., *Performance Measurements of the VIPSEC Model*, High Performance Computing Symposium (HPC 2005), San Diego, California - USA, April 3-7, (2005).

On the Security of the Canetti-Krawczyk Model*

Xinghua Li[1], Jianfeng Ma[1,2], and SangJae Moon[3]

[1] Key Laboratory of Computer Networks and Information Security(Ministry of Education), Xidian University, Xi'an 710071, China
flame168@sohu.com
[2] School of Computing and Automatization, Tianjin Polytechnic University, Tianjin 300160, China
jfma@mail.xidian.edu.cn
[3] Mobile Network Security Technology Research Center, Kyungpook National University, Sankyuk-dong, Buk-ku, Daegu 702-701, Korea
sjmoon@ee.knu.ac.kr

Abstract. The Canetti-Krawczyk (CK) model is a formal method to design and analyze of key agreement protocols, and these protocols should have some desirable security attributes. In this paper, the relationship between the CK model and the desirable security attributes for a key agreement protocol is analyzed. The conclusions indicate that: (1) protocols designed and proved secure by the CK model offer almost all the security attributes, such as perfect forward secrecy (PFS), loss of information, known-key security, key-compromise impersonation and unknown key-share, but the attribute of key control; (2) loss of information and key-compromise impersonation can be guaranteed by the first requirement of the security definition (SK-security) in the CK model, while PFS and known-key security by the second requirement, and unknown key-share can be ensured by either the requirement. Thereafter, the advantages and disadvantages of the CK model are presented.

1 Introduction

The design and analysis of key-agreement protocols continues to be the most challenging areas of security research, so a systematic research of them is necessary. In this field, the Canetti-Krawczyk (CK) model is the most popular methodology now [1,2].

In the past twenty years, researchers have made a lot of efforts in designing and analyzing key-exchange protocols [3,4,5,6,7,8,9,10], they realize that the potential impact of compromise of various types of keying material in a key agreement protocol should be considered, even if such compromise is not normally expected [5]. So some desirable security properties that a key agreement protocol should have are identified. Such security properties include PFS, loss of information, known-key security, key-compromise impersonation, unknown-key share, key control and so on.

* Research supported by the National Natural Science Foundation of China (Grant No. 90204012), the National "863" High-tech Project of China (Grant No. 2002AA143021), the Excellent Young Teachers Program of Chinese Ministry of Education, the Key Project of Chinese Ministry of Education, and the University IT Research Center Project of Korea.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 356 – 363, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The main goal of the CK model is to design and analyze key agreement protocols. Then does a SK-secure key agreement protocol have the desirable security attributes? And what is the relationship between the SK-security and the security attributes? This is the main motivation of this paper. At the same time, the advantages and disadvantages of the CK model are analyzed.

The rest of this paper is organized as follows. In Section 2, the CK model is briefly reviewed. In section 3, the desirable security properties of key-agreement protocols are outlined. In section 4, the relationship between the CK model and desirable security attributes for a key agreement protocol is analyzed. Section 5 presents the advantages and disadvantages of the CK model. We conclude this paper in Section 6.

2 The CK Model

The CK model includes three main components: unauthenticated-link adversarial model (UM), authenticated-link adversarial model (AM) and authenticator [1].

2.1 Definition of Session-Key Security

Definition 1: Session-key security. A key-agreement protocol π is called Session-Key secure (or SK-secure) if the following properties hold for adversary μ in the UM.

1. Protocol π satisfies the property that if two uncorrupted parties complete matching sessions then they both output the same key; and
2. the probability that μ can distinguish the session key from a random value is no more than 1/2 plus a negligible fraction in the security parameter.

3 Desirable Security Attributes of Key Agreement Protocols

A number of desirable security attributes of key agreement protocols have been identified [12].

1. *(perfect) forward secrecy.* If long-term private keys of one or more entities are compromised, the secrecy of previous session keys established by honest entities is not affected [5].
2. *loss of information*. Compromise of other information that would not ordinarily be available to an adversary does not affect the security of the protocol. For example, in Diffie-Hellman type protocols [6], security is not comprised by loss of $\alpha^{s_i s_j}$ (where S_i represents entity i's long-term secret value) [11].
3. *known-key security*. A protocol is said to be vulnerable to a known-key attack if compromise of past session keys allows either a passive adversary to compromise future session keys, or impersonation by an active adversary in the future [12].
4. *key-compromise impersonation.* Suppose A's long-term private key is disclosed. Clearly an adversary that knows this value can now impersonate A, since it is precisely this value that identifies A. However, it may be desirable that this loss does not enable an adversary to impersonate other entities to A [12].

5. *unknown key-share.* Entity A cannot be coerced into sharing a key with entity B without A's knowledge, i.e. when A believes the key is shared with some entity $C \neq B$, and B (correctly) believes the key is shared with A [12].
6. *key control.* Neither entity should be able to force the session key to a preselected value [12].

4 The Relationship Between the CK Model and the Desirable Security Attributes

Firstly, a key-agreement protocol is assumed to be SK-secure, i.e. the protocol is proved secure by the CK model. Suppose that the two parties that participate in the protocol are I and J. According to the definition of SK-security, at the end of the agreement run, both parties will establish a same session key K, and the attacker $\mathcal{A}$ cannot distinguish K from a random number with a non-negligible advantage. In the following part, this SK-secure key-agreement protocol is analyzed on whether it has the desirable security properties.

4.1 The Relationship Between a SK-Secure Key-Agreement Protocol and the Desirable Security Attributes

Lemma 1. A key-agreement protocol which is proved secure by the CK model provides PFS.

Proof. The CK model allows matching sessions not to expire simultaneously. When I and J complete matching sessions, the attacker $\mathcal{A}$ can make the session within I expire first, then he can corrupt I and get his private key. If the protocol cannot provide PFS, $\mathcal{A}$ can get the session key within I. And because the session within I has expired, the attacker cannot get any information specific to this session except the session key. But the matching session within J has not expired, therefore the attacker can choose this session as the test session and performs the test-session query. Because he has got the session key, and from the prior assumption we know that I and J have got the same session key, $\mathcal{A}$ can completely distinguish K from a random number, which contradicts the second requirement of Definition 1. So the protocol is not SK-secure, and this is in contradiction with the prior assumption. Therefore the protocol which is proved secure by the CK model provides PFS. #

Lemma 2. A key agreement protocol that is proved secure by the CK model is secure in the case of loss of information.

Proof. If the protocol cannot offer the security property of loss of information, $\mathcal{A}$ can impersonate I or J to send spurious messages when he gets some secret information (other than the private key of I or J). Then I and J cannot get a same session key at the end of the protocol run**,** e.g. the protocol 1, 2 and 3 in [11], which contradicts the consistency requirement of Definition 1. So the protocol is not SK-secure, as is in contradiction with the prior assumption. So the protocol proved secure by the CK model is secure in the case of loss of information. #

Lemma 3. A key-agreement protocol that is proved secure by the CK model provides known-key security.

Proof: According to [13], known-key attacks can be divided into passive attacks and active attacks. And active attacks are also divided into non-impersonation attacks and impersonation attacks. Here we classify the impersonation attack as the loss of information attack, because it is a form of loss of information attacks. We just focus on passive attacks and active non-impersonation attacks, whose common character is that $\mathcal{A}$ can get present session key from the past ones. It is assumed that I and J have ever agreed on at least one session key, and $\mathcal{A}$ may have gotten them through session-key query before the corresponding sessions expired. If the protocol cannot provide known-key security, then $\mathcal{A}$ can get the session key of the test-session, thus can distinguish K from a random number with a non-negligible advantage. So the protocol is not SK-secure, which contradicts the prior assumption. Therefore the protocol proved secure by the CK model can provide known-key security. #

Lemma 4. A key-agreement protocol that is proved secure by the CK model resists key-compromise impersonation attacks.

Proof: It is assumed that the protocol proved secure by the CK model cannot resist key-compromise impersonation attacks [11] and the private key of I is compromised. Then $\mathcal{A}$ can impersonate J to send messages to I during the negotiation of the session key. In such case, I and J cannot get a same session key at the end of the protocol run, which does not satisfy the first requirement of Definition 1. Therefore this protocol is not SK-secure, which contradicts the prior assumption. So the protocol proved secure by the CK model can resist key-compromise impersonation attacks. #

Lemma 5. A key-agreement protocol that is proved secure by the CK model resists unknown key-share attacks.

Proof. It is assumed that the protocol cannot offer the security property of unknown key-share attacks. Then $\mathcal{A}$ can apply the following strategy: (1) initiate a session (I, s, J) at I where s is a session-id; (2) activate a session (J, s, *Eve*) at J as responder with the start message from (I, s, J) where *Eve* is a corrupted party; (3) deliver the response message produced by J to I. As a result, I believes (correctly) that he shares a session key with J and J believes that he shares a key with *Eve* [14]. In addition, I and J get a same session key. An example of this attack can be found in [7]. In such case, $\mathcal{A}$ can choose the session at I as the test session and expose (J, s, *Eve*) via a session-state reveal attack to get the session key within J which is same as that of the test-session. As a result, $\mathcal{A}$ can completely get the session key of the test session, which contradicts the second requirement of Definition 1. Therefore this protocol is not SK-secure, which is in contradiction with the prior assumption. So the protocol which is proved secure by the CK model can resist unknown key-share attacks. #

Lemma 6. A key-agreement protocol which is proved secure by the CK model cannot offer the security property of key control.

Proof: Suppose that I and J both contribute random inputs to the computation of the session key $f(x,y)$, but that I first sends x to J. Then it is possible for J to compute 2^l variants of y and the corresponding $f(x,y)$ before sending y to I. In this way, entity J

can determine approximately l bits of the joint key [15]. Even though such case happens during the key agreement, $\mathcal{A}$ cannot distinguish the session key from a random number, because this is a normal implementation of the protocol and the session key also comes from the distribution of keys generated by the protocol. Therefore according to the prior assumption, at the end of the protocol, I and J get a same session key and $\mathcal{A}$ cannot distinguish the key from a random number with a non-negligible advantage. So even though key control happens, the protocol can still be proved secure by the CK model. As it is noted in [16], the responder in a protocol almost always has an unfair advantage in controlling the value of the established session key. This can be avoided by the use of commitments, although this intrinsically requires an extra round. #

From Lemma 1 to Lemma 6 presented above, we can obtain Theorem 1 naturally.

Theorem 1: A key-agreement protocol designed and proved secure by the CK model offers almost all the desirable security properties except key control**.** #

4.2 The Relationship Between the Security Attributes and the Two Requirements of SK-Security

We also find that some security attributes can be ensured by the first requirement of SK-security, while others by the second requirement. In the following, theorem 2 and theorem 3 are presented for a detailed explanation.

Theorem 2: The first requirement of SK-security guarantees a protocol to resist impersonation attacks and unknown key-share attacks.

Proof: If there are impersonation attacks in the key agreement protocol, $\mathcal{A}$ can impersonate I (or J) to send messages to J (or I), then at the end of the protocol run, they will not get a same session key, which contradicts the first requirement of SK-security. If there are unknown key-share attacks, then I and J will not complete matching sessions, which contradicts the first requirement of SK-security too.

So from the above analysis, we can get that the first requirement of SK-security can guarantee a protocol to resist impersonation attacks and unknown key-share attacks. Key-compromise impersonation attacks and loss of information attacks belong to impersonation attacks, so these two attacks can be resisted by this requirement. #

Theorem 3: The second requirement of SK-security guarantees a protocol to offer PFS, known-key security and unknown key-share attacks.

Proof: From Theorem 2, we know that known-key security and PFS cannot be guaranteed by the first requirement. From Theorem 1 we know the CK model can ensure these security properties, thus it is the second requirement of SK-security that guarantees PFS and known-key security. #

From Theorem 1 and Theorem 2, we find that unknown key-share can be guaranteed by either requirement of SK-security. In addition, it should be noticed that the first requirement is the precondition of SK-security. Only under the consistency condition, does it make sense to investigate the security properties of PFS and known-key security.

5 The Advantages and Disadvantages of the CK Model

5.1 The Advantages of the CK Model

Why is the CK model applicable for designing and analyzing key-agreement protocols?

First, the indistinguishability between the session key and a random number is used to achieve the SK-security of a key-agreement protocol in the AM. If an attacker can distinguish the session key from a random number with a non-negligible advantage, a mathematics hard problem will be resolved. According to the reduction to absurdity, a conclusion can be gotten: no matter what methods are used by the attacker (except party corruption, session state reveal and session key query [1]), he cannot distinguish the session key from a random number with a non-negligible advantage. So the protocol designed and proved secure by the CK model can resist known and even unknown attacks.

Second, the CK model employs authenticators to achieve the indistinguishability between the protocol in the AM and the corresponding one in the UM. Through this method, the consistency requirement of SK-security is satisfied.

From the above analysis, it can be seen that this model is a modular approach to provably secure protocols. With this model, we can easily get a provably secure protocol which can offer almost all the desirable security attributes. And the CK model has the composable characteristic and can be used as an engineering approach [17,18]. Therefore, it is possible to use this approach without a detailed knowledge of the formal models and proofs, and is very efficient and suitable for applications by practitioners.

5.2 The Disadvantages of the CK Model

Though the CK model is suitable for the design and analysis of key-agreement protocols, it still has some weaknesses as follows:

1. The CK model cannot detect security weaknesses that exist in key-agreement protocols, however some other formal methods have this ability, such as the method based on logic [19] and the method based on state machines [20]. But the CK model can confirm the known attacks, i.e. this model can prove that a protocol that has been found flaws is not SK-secure.
2. In the aspect of the forward secrecy, the CK model cannot guarantee that a key-agreement protocol offers forward secrecy with respect to compromise of both parties' private keys; it can only guarantee the forward secrecy of a protocol with respect to one party. In addition, in ID-based systems this model lacks the ability to guarantee the key generation center (KGC) forward secrecy because it does not fully consider the attacker's capabilities [21].
3. From lemma 6, we know that protocols which are designed and proved secure by the CK model cannot resist key control, which is not fully consistent with the definition of key agreement [12].
4. A key-agreement protocol designed and proved secure by the CK model cannot be guaranteed to resist Denial-of-Service (DoS) attacks. However

DoS attacks have become a common threat in the present Internet, which have brought researchers' attention [22,23].

5. Some proofs of the protocols with the CK model are not very credible because of the subtleness of this model. For example, the Bellare-Rogaway three-party key distribution (3PKD) protocol [24] claimed proofs of security is subsequently found flaws [25].

We know that a protocol designed and proved secure by the CK model can offer almost all the security attributes, and this model has the modular and composable characteristics, so it is very practical and efficient for the design of a key-agreement protocol. But this model still has weaknesses. So when the CK model is employed to design a key-agreement protocol, we should pay attention to the possible flaws in the protocol that may result from the weaknesses of CK model.

6 Conclusion

In this paper, the relationship between the CK model and the desirable security attributes for key agreement protocols is studied. The conclusions indicate that: (1) a SK-secure key-agreement protocol can offer almost all the desirable security properties except the key control; (2) key-compromise impersonation and loss of information can be guaranteed by the first requirement of SK-security, while PFS and known-key secure by the second requirement, and unknown key-share can be ensured by either the requirement. Thereafter, some advantages and disadvantages of the CK model are analyzed, from which we can get that this model is suitable and efficient for the design of provably secure key agreement protocols, but attention should also be paid to the possible flaws resulting from the disadvantages of this model.

References

1. Canetti, R., Krawczyk, H., Advances in Cryptology Eurocrypt 2001, Analysis of Key-Exchange Protocols and Their Use for Building Secure Channels. Lecture Notes in Computer Science, Springer-Verlag,Vol 2045 (2001) 453-474.
2. Colin Boyd, Wenbo Mao and Kenny Paterson: Key Agreement using Statically Keyed Authenticators, Applied Cryptography and Network Security: Second International Conference, ACNS 2004, Lecture Notes in Computer Science, Springer-verlag, volume 3089,(2004)248-262.
3. Bellare, M., Rogaway, P., Entity authentication and key distribution, Advances in Cryptology,- CRYPTO'93, Lecture Notes in Computer Science , Springer-Verlag, Vol 773 (1994) 232-249
4. Bellare, M., Canetti, R., Krawczyk, H., A modular approach to the design and analysis of authentication and key-exchange protocols, 30th STOC, (1998)419-428
5. A. Menezes, P. vanOorschot, and S. Vanstone, Handbook of Applied Cryptography, chapter 12, CRC Press, 1996.
6. W. Diffie, M. Hellman: New directions in cryptography, IEEE Trans. Info. Theory IT-22, November (1976)644-654
7. W. Diffie, P. van Oorschot, M. Wiener. Authentication and authenticated key exchanges, Designs, Codes and Cryptography, 2, (1992)107-125

8. H. Krawczyk. SKEME: A Versatile Secure Key Exchange Mechanism for Internet, Proceeding of the 1996 Internet Society Symposium on Network and Distributed System Security, Feb. (1996)114-127
9. H. Krawczyk, SKEME: A Versatile Secure Key Exchange Mechanism for Internet, Proceeding of the 1996 Internet Society Symposium on Network and Distributed System Security, (1996) 114-127
10. V. Shoup, On Formal Models for Secure Key Exchange, Theory of Cryptography Library, 1999. http://philby.ucsd,edu/cryptolib/1999/99-12.html
11. S. Blake-Wilson, D. Johnson, A. Menezes, Key Agreement Protocols and Their Security Analysis, Proceedings of the sixth IMA international Conference on Cryptography and Coding, 1997
12. L. law, A. Menezes, M.Qu, et.al, An Efficient Protocol for Authenticated Key Agreement. Tech. Rep. CORR 98-05, Department of C&O, University of Waterloo.
13. K. Shim, Cryptanalysis of Al-Riyami-Paterson's Authenticated Three Party Key Agreement Protocols, Cryptology ePrint Archive, Report 2003/122, 2003. http://eprint.iacr.org/2003/122.
14. R. Canetti, H. Krawczyk, Security Analysis of IKE's Signature-based Key-Exchange Protocol, Proc. of the Crypto conference, (2002).
15. Günther Horn, Keith M. Martin, Chris J. Mitchell, Authentication Protocols for Mobile Network Environment Value-Added Services, IEEE Transaction on Vehicular Technology, Vol 51 (2002)383-392
16. C. J. Mitchell, M. Ward, P. Wilson, Key control in key agreement protocols, Electronics Letters, Vol 34 (1998) 980-981
17. Tin, Y.S.T., Boyd, C. Nieto, J.G., Provably Secure Key Exchange: An Engineering Approach. Australasian Information Security Workshop 2003(AISW 2003) (2003) 97-104
18. R. Canetti, H. Krawczyk. Universally Composable Notions of Key Exchange and Secure Channels. Advances in Cryptology-EUROCRYPT 2002, Lecture Notes in Computer Science, Springer-verlag, volume 2332 (2002) 337--351.
19. Burrows, M., Abadi, M., Needham, R.M, A logic of Authentication, ACM Transactions on Computer Systems, vol.8, No,1 (1990) 122-133,
20. Meadows C. Formal verification of cryptographic protocols: A survey. In: Advances in Cryptology, Asiacrypt'96 Proceedings. Lecture Notes in Computer Science, Springer-Verlag, Vol 1163, (1996) 135~150.
21. Li Xinghua, Ma Jianfeng, SangJae Moon, Security Extension for the Canetti-Krawczyk Model in Identity-based Systems, Science in China, Vol 34, (2004)
22. E. Bresson, O. Chevassut, D. Pointcheval. New Security Results on Encrypted Key Exchange, the 7th International Workshop on Theory and Practice in Public Key Cryptography-PKC 2004, Springer-Verlag, LNCS (2947) 145-158-
23. W. Aiello, S.M. Bellovin, M. Blaze. Efficient, DoS-Resistant, Secure Key Exchange for Internet Protocols. Proceedings of the 9th ACM conference on Computer and communications security (2002) 45-58
24. M. Bellare, P. Rogaway. Provably Secure Session Key Distribution: The Three Party Case. The 27th ACM Symposium on the Theory of Computing – STOC(1995) 57-66. ACM Press,1995.
25. K.K.R Choo, Y.Hitchcock. Security requirement for key establishment proof models: revisiting bellare-rogaway and Jeong-Katz-Lee Protocols. Proceedings of the 10th Australasian conference on information security and privacy-ACISP (2005).

A Novel Architecture for Detecting and Defending Against Flooding-Based DDoS Attacks*

Yi Shi and Xinyu Yang

Dept. of Computer Science and Technology, Xi'an Jiaotong University,
Xi'an 710049, China
yxyphd@mail.xjtu.edu.cn

Abstract. Flooding-based distributed denial-of-service (DDoS) attack presents a very serious threat to the stability of the Internet. In this paper, we propose a novel global defense architecture to protect the entire Internet from DDoS attacks. This architecture includes all the three parts of defense during the DDoS attack: detection, filtering and traceback, and we use different agents distributed in routers or hosts to fulfill these tasks. The superiority of the architecture that makes it more effective includes: (*i*) the attack detection algorithm as well as attack filtering and traceback algorithm are both network traffic-based algorithms; (*ii*) our traceback algorithm itself also can mitigate the effects of the attacks. Our proposed scheme is implemented through simulations of detecting and defending SYN Flooding attack, which is an example of DDoS attack. The results show that such architecture is much effective because the performance of detection algorithm and traceback algorithm are both better.

1 Introduction

Flooding-based distributed DoS attack, or simply DDoS attack, have already become a major threat to the stability of the Internet [1]. Much work has been done toward protect against or mitigate the effects of DDoS attacks and they concentrate on three aspects: attack detection, attack filtering and attack source traceback.

DDoS attack detection is responsible for identifying DDoS attacks or attack packets. Approaches for DoS/DDoS detection are categorized into anomaly detection, which belongs to intrusion detecting techniques. Frequently used approaches for anomaly detection include statistics, immunology, neural-network, data mining, machine learning, finite-state automation (hidden Markov approach), and so on. Traffic-based approach is a new direction in the evolvement of attack detection. There have been some developments based on traffic to detect attacks. Haining Wang et al proposed a detecting approach aiming at SYN Flooding attacks based on the protocol behavior of TCP SYN-FIN pair [2]. Statistical models are widely used; they combine the expectation, variance or other statistical values with hypothesis tests for detecting attacks [3]. John E. Dickerson et al used data mining to search the fuzzy rules of attacking traffic [4]. Garcia, R. C. et al introduced wavelet analysis into the attack

* This work is supported by the NSFC (National Natural Science Foundation of China -- under Grant 60403028), NSFS (Natural Science Foundation of Shaanxi -- under Grant 2004F43), and Natural Science Foundation of Electronic and Information Engineering School, Xi'an jiaotong university.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 364 – 374, 2005.
© Springer-Verlag Berlin Heidelberg 2005

detection, they judged the occurrences of attacks by relevant variables after applying wavelet analysis to traffic data [5]. There are also many researches in self-similarity features of traffic, which often have linkages with wavelet analysis approaches. David A. et al proposed to detect attacks according to the variety of the Hurst parameter in discrete wavelet transmission [6]. In addition, some modeling approaches for other network parameters are proposed, for example, establishing dynamic model for the occupied percentage of the link bandwidth [7], but the acquirement of those parameter data are still based on the measurement of traffic. In the proposed architecture, a novel method for attack detection is used. It is based on measuring, analyzing, and modeling of network traffic, focused on the features of Flooding-based DDoS attacks ---- traffic burst and remaining of comparative smooth for some time.

After identifying attack packet flows or attack packets, attack filtering is responsible for classifying those packets and then dropping them. Ingress filtering [8] configures the internal router interface to block spoofed packets whose source IP addresses do not belong to the stub network. This limits the ability to flood spoofed packets from that stub network since the attacker would only be able to generate bogus traffic with internal addresses. Given the reach ability constraints imposed by the routing and network topology, the route-based distributed packet filtering (DPF) [9] exploits routing information to determine if a packet arriving at the router is valid with respect to its inscribed source/destination addresses. The experimental results reported in [9] show that a significant fraction of spoofed packets are filtered out and the spoofed packets that escaped the filtering can be localized into five candidate sites which are easy to trace back. In our current architecture, because both detection and traceback algorithms are traffic-based, our filtering scheme does not parse the packets and just base on traffic rate-limiting.

Attack source traceback is used to locking the attacker. Since the source addresses of flooding packets are faked, various traceback techniques [10-14] have been proposed to find out the origin of a flooding source. There are generally two approaches to the traceback problem. One is for routers to record information about packets they have seen for later traceback requests [12]. Another is for routers to send additional information about the packets they have seen to the packets' destinations via either the packets [11] or another channel, such as ICMP messages. All of these IP traceback technologies rely on additional information. And they are not practical and less effective because they are usually the after-the-fact response to a DDoS attack, and just helpful in identifying the attacker. In this paper, we propose a novel traceback algorithm based on countering SYNs and FINS and traffic filtering. The experiment result shows that, it's more practical because it is quickly enough to abort an ongoing attack with the traced attack source before the attack stops.

Most of current detect-and-filter approaches or traceback methods, as mentioned above, are used solely for security purpose. They cannot possibly achieve effective attack detection and filtering as well as effective attack source traceback. In response to this shortcoming, we propose to deploy a global defense architecture to protect the entire Internet from DDoS attacks.

This paper is organized as follows. In Section 2, we introduce the proposed architecture, especially its detection algorithm and traceback algorithm. In Section 3, this architecture is implemented through simulations of detecting and defending SYN Flooding attack. Finally, conclusions are made based on the simulations and the future work is also discussed.

2 Traffic-Based DDOS Detecting and Defending Architecture

2.1 Architecture Description

As shown in Figure 1, in this architecture, three kinds of agents are considered: Detection Agents, Traceback Agents and Coordination Agents. Detection agents in hosts do the attack detection to find any dubious attack. Traceback Agents in routers are responsible for router traffic filtering and trace back to the source of the attacks (usually slaves, not the genuine attacker hosts). Coordination Agents in hosts play the role of coordinating and communicating with detection agents and traceback agents.

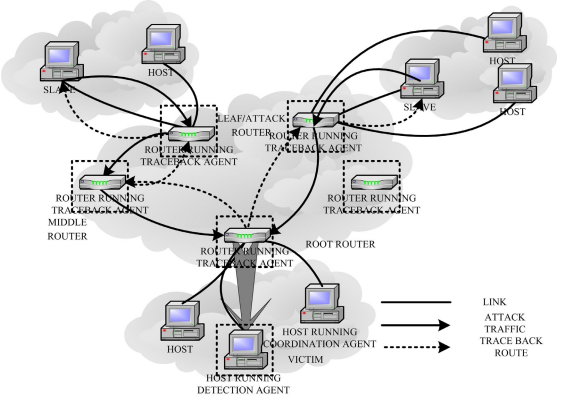


Fig. 1. Traffic-based detecting & defending DDoS architecture

The following are a few terms we will use to describe the problem and our proposed solutions in this paper:

- VICTIM: Under the DDoS attack model, the network bandwidth toward this attacked host will be eventually taken away by the collection of DDoS packets.
- ATTACKER: It's not the genuine attacker host but a slave, which sends out lots of malicious packets toward the victim after receiving the attack commands from its master.
- ROOT ROUTER: The router that directly connected with the VICTIM.
- LEAF ROUTER: The edge router that directly connected with slaves or the remotest router on the attack path to the VICTIM that deployed with our traceback agent.
- ATTACK ROUTER: It's a special one of LEAF ROUTER. In the ideal situation, it would be the leaf router that directly connected with ATTACKER. In practical terms, it is the leaf router that forwards the attack traffic.
- MIDDLE ROUTER: The routers located between ROOT ROUTER and LEAF ROUTER on an attack route.

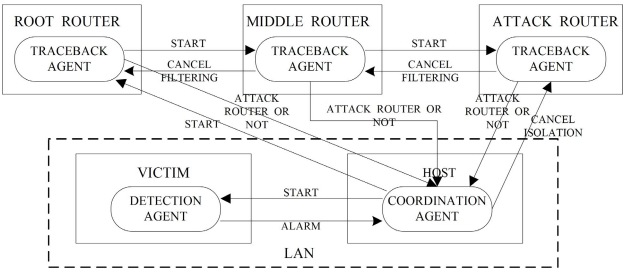


Fig. 2. Communication and cooperation of agents

Figure 2 is the sketch map of communication and cooperation of agents. The approach to the problem of detecting and defending attacks is simply as follows:

1) Detection agent is started by coordination agent to inspect whether there is any dubious attack at the victim, with the traffic-based recursive detection algorithm;
2) When detection agent catches the attack, it will send alarm message to inform coordination agent of starting traceback agent in the root router. The traceback agent of the root router use traceback algorithm to find out which upstream node/nodes (usually middle router) forward the attack traffic to it, and discard the attack traffic in the root router. Then, repeat the process in the upstream attack routers with their traceback agents, to trace their upstream attack node/nodes as well as filter the attack traffic in their level. Step by step, it will finally trace to the attacker router and isolate the slave (or its upstream router forwards the attack traffic to it) from it. Meanwhile, coordination agent collects all the messages come from the traceback agents and finally constructs the attack paths, and puts all attack routers into a queue (called waiting-recovery-queue) waiting to be recovered;
3) After Δt interval, for any attack router in the queue, coordination agent will startup recovery process: it first startup the detection agent at the victim to examine whether new attack occurs. If there is new attack, repeat 2) to trace back; otherwise, it sends cancel message to the traceback agent in the attack router to cancel the isolation, and startup detection agent to work again. If attack detected, it means this attack router is still forwarding attack traffic, so send isolation message to its traceback agent to isolate it and put it into the waiting-recovery-queue once again. If not, we can predicate that this attack router is harmless now. Then, repeat 3) with the next node in the queue. Until the waiting-recovery-queue is empty, all attack routers are recuperated.

The core of this architecture includes three aspects: detection algorithm, filtering & traceback algorithm, and traffic recovery. We will expatiate on them in the following sections.

2.2 Detection Algorithm

Detection agents deployed in hosts use this traffic-based real-time detection algorithm to find dubious attack. Our detection algorithm is composed of the basic recursive algorithm and the adjusted detection algorithm with short-term traffic prediction.

According to the flat-burst feature of the attack traffic, a traffic-based recursive algorithm for detecting DDoS attacks is proposed as follows:

A. Calculation of the Statistical Values

In order to meet the real-time demands of the algorithm, every statistical value is calculated by a recursive way.

Suppose the original traffic sequence is C, the global mean value (the mean value of all data before the current one) sequence is cu, the difference sequence of the traffic is z, and the difference variance sequence (the variance of differenced data before current) is d_var. The mean value of difference sequence can be considered as 0. The calculations of those statistical values are shown as follows:

$$cu(t) = \frac{1}{t} \cdot ((t-1) \cdot cu(t-1) + C(t)) \tag{1}$$

$$z(t) = C(t) - C(t-1) \ (t>1) \tag{2}$$

$$\mathrm{d_}var(t) = \frac{1}{t-1}\sum_{i=2}^{t}(z(i)-u(t))^2 \approx \frac{1}{t-1}((t-2)\cdot \mathrm{d_}var(t-1) + (z(t)-u(t))^2) = \frac{1}{t-1}((t-2)\cdot \mathrm{d_}var(t-1) + z(t)^2) \tag{3}$$

In order to estimate the volume of the traffic at time t, a function uG(t) is defined as:

$$u_G(t) = \begin{cases} 0, C(t) < lt \times cu(t) \\ \dfrac{lt \times C(t)}{(ht-lt) \times cu(t)} - \dfrac{lt}{(ht-lt)}, lt \times cu(t) \le C(t) \le ht \times cu(t) \\ 1, C(t) > ht \times cu(t) \end{cases} \tag{4}$$

This definition of uG(t) borrows the concept of "membership function" in Fuzzy Arithmetic. lt*cu(t) means the lower limit of "great" traffic, that is to say, if the traffic is lower than lt*cu(t), it can be considered to be "not great", namely the weight of "great" is 0. ht*cu(t) means the upper limit of bearable traffic. If the traffic volume is ht times higher than the global mean value, it can be considered to be "very great", namely its weight of "great" is 1. In real applications, the parameters lt and ht should be assigned according to the performance of networks, and the definition of lt may have more influences on the results.

B. Judging Attacks

According to the definition of $u_G(t)$, if $u_G(t)=0$, it is regarded that no attack occurs; if at the time t $u_G(t)>0$, there may be a beginning of an attack, the judging process is started immediately to verify the attack.

The following is the description of the judging algorithm:

```
if (uG(t)>0 && d_var(t)>d_var(t-1))
{  attack_count=1;
   for (i=1; attack_acount<=expected_steps; i++)
   {     if (uG(t+1)>0)
         {      attack_count++;
                if (d_var(t+i)<d_var(t+i-1))
                       A(i)=max(uG(t+i-1), uG(t+i));
```

```
                    else A(i)= uG(t+i);
                     if (attack_count>=expected_steps)
                           if (A(i)>=A(i-1))
                           {    alarm();
                                 t=t+i;
                                 break; }
                            else attack_acount--; }}}
    else
    {  attack_count=0;
       t=t+i;
       break; }
```

C. Joining Short-Term Traffic Prediction into DDoS Attacks Detection

Considering the traffic sequence as a time series, we can establish an adaptive AR model on it and predict the values based on the model. The approach of predicting we used is Error-adjusted LMS (EaLMS), which is shown in another paper of ours[15]. The intent of joining short-term traffic prediction into DDoS detection is to obtain a sequence longer than the current one, thus to accelerate the speed of detecting and to reduce the prediction error.

After obtained the difference of current traffic (presented by z(t)), the single step prediction z_predict(t) for z(t) is calculated. The z_predict(t) that is regarded as z(t+1) is used for calculating the difference variance d_var(t+1) and $u_G(t+1)$ of the time t+1.

$$\begin{aligned} \mathrm{d}_var(t+1) &= \frac{1}{t}\sum_{i=2}^{t+1}(z(i)-\mu(t))^2 \approx \frac{1}{t}((t-1)\cdot \mathrm{d}_var(t) + (z(t+1)-\mu(t))^2) \\ &= \frac{1}{t}((t-1)\cdot \mathrm{d}_var(t) + z_predict(t)^2) \end{aligned} \tag{5}$$

$$u_G(t+1) = \begin{cases} 0, C(t)+z_predict(t) < lt \times cu(t) \\ \dfrac{lt\times(C(t)+z_predict(t))}{(ht-lt)\times cu(t)} - \dfrac{lt}{(ht-lt)}, lt\times cu(t) \le C(t)+z_predict(t) \le ht\times cu(t) \\ 1, C(t)+z_predict(t) > ht\times cu(t) \end{cases} \tag{6}$$

2.3 Filtering and Traceback Algorithm

First, we should indicate that although our architecture is designed for DDoS detecting and defending, current traceback algorithm is just for SYN Flooding attack. The filtering & traceback algorithm is used by traceback agents located on routers, and it is composed of two different parts as we will explain in the following:

A. Traceback Algorithm

When received the start message from coordination agent or from a downstream router, traceback agent of the local router is started to run traceback algorithm. The algorithm is described as following:

1) Received the start message and acquired three or four parameters: α represents the filtering proportion during tracing and $0<\alpha\le 1$; β represents the filtering proportion for defending after located the attacker and $0<\beta\le 1$; ID is an attack serial number to represent a tracing request; and R_d represents the request downstream router;

2) Find all upstream routers to the local router;
3) Execute attack-router-judging algorithm to discern which upstream node/nodes forwards the attack packets, then drop all the packets from it/them;
4) Send the result of its upstream attack node/nodes to coordination agent;
5) If the local router is a leaf router, it is the attack router we defined above. That means the attack source (a slave or a router not in our architecture) is locked. Inform all routers in this attack path of recovering traffic, i.e. β=1; and the algorithm stops;
6) If not, notify traceback agent/agents of its upstream attack node/nodes to work with this algorithm;

B. Attack-Router-Judging Algorithm

Suppose R represents the router runing this algorithm, $T_1 \ldots\ldots T_n$ are the traffic of all upstream nodes (routers or hosts) to R, $\Delta P = P_{request} - P_{response}$ ($P_{request}$ is the sum traffic of request, and $P_{request}$ is the sum traffic of respons), $T = SUM\{T_{1\ldots\ldots} T_n\}$ represents the sum traffic of $T_1 \ldots\ldots T_n$, $T_{max} = MAX\{T_{1\ldots\ldots} T_n\}$ represents the max traffic of $T_1 \ldots\ldots T_n$.

The steps of the algorithm are shown below:

1) Calculate T;
2) Calculate T_{max};
3) Let $T_i = T_i * \alpha$, ($0<\alpha<1$, i = 1,2,…,n, and i<>max);
4) Calculate $P_{request}$, $P_{response}$ and ΔP;
5) Sum $P'_{request}$ and $P'_{response}$ after changing the traffic of Ti, and calculate $\Delta P'$($\Delta P' = P'_{request} - P'_{response}$);
6) Calculate $A = \Delta P' / \Delta P$, and then judge:
   i) If A=1, it indicate that Ti are all normal traffic except T_{max}. Drop the traffic T_{max} and stop this algorithm;
   ii) If $\alpha<A<1$, it indicate that there still are other attack traffic besides T_{max}. Drop the traffic T_{max};
   iii) If $A=\alpha$, then show that T_{max} is not the attack traffic;
7) T=T- Tmax, if T is not empty, jump to 2).

2.4 Traffic Recovery

During tracing back, attack traffic is filtered completely in every router on the attack path step by step, and finally the same is done to the attacker router. However, there is still a problem of how to recover the traffic of attack router securely.

After Δt interval from the attack source locked, for any attack router in the waiting-recovery-queue, coordination agent will startup recovery process:

1) Startup the detection agent at the victim to examine whether new attack occurs during Δt;
2) If there is new attack, repeat traceback algorithm; otherwise, send cancel message to the traceback agent in the attack router to recover it's traffic completely;

3) Startup detection agent to work again. If attack detected, it means this attack router is still forwarding attack traffic; thus, put it into the queue once again and isolate it for a longer time $\Delta t' = \Delta\lambda * \Delta t$ (λ=1,2,…,n according to the times it is put into the waiting-recovery-queue);
4) If no attack detected, we can predicate that this attack router is harmless now. Then, jump to 1) to process the next node in the queue till the queue is empty.

3 Experiments

SYN Flooding attack, an example of Flooding-based DDoS attacks, is used in our simulations. For evaluating our proposal, we construct a network simulation test-bed (based on NS-2 simulator [16]) with 15 routers and 11 end-hosts as shown in Figure 3. Nodes 15~25 are end-host nodes and the rest nodes are routers. Node 20 is supposed as the only victim. The bandwidth of the link is 10Mbps. Agents are distributed as: node 20 is deployed with detecting agent, node 25 is with coordination agent, and every router is disposed with traceback agent. Δt is assigned with 10 seconds. We perform a set of experiments with different attack strategies:

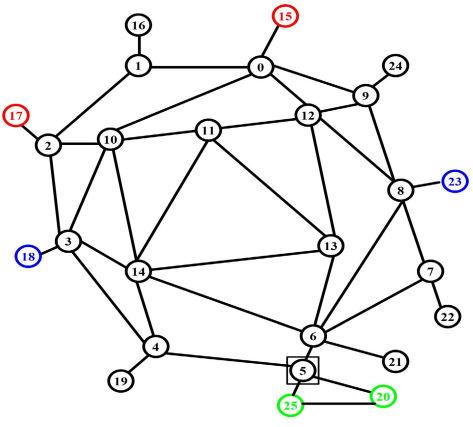


Fig. 3. Test-bed topology

1. Single-attacker versus single-victim: single node assigned with background traffic
In this case, node 15 is the only attacker and node 23 sends background traffic to node 20. The steps of the experiments are:

- At second 0, nodes 15 and 23 start to request normal TCP connections, and the coordination agent starts the detection agent to work.
- At second 4, node 15 launches SYN Flooding attack to node 5 by requesting TCP connections with node 16 as the spoofing source address.
- At second 18, node 15 stops attacking and recover to send normal packets.
- At second 50 the simulation stops.

Results:

- At second 4.41075, detection agent detects the attack and sends alarm message to coordination agent. It takes only 0.41075 seconds from the attack starts to it was caught (Detection Time is 0.41075).
- When detection agent sends alarm message to coordination agent, traceback agent is started at the same time. At 15.472, traceback agent locks the attacker 15 and immediately isolation the attack traffic. Thus, the cost of accomplishing tracing is 11.06125 seconds (Traceback Time is 11.06125).
- At 26.27375, detection agent report no attack occurs after 15 is recovered. The recovery time from finding the attack source 15 is 10.80175 (Recovery Time).
- At second 15.491, coordination agent also calculates the result attack path 15→0→12→13→6→5→20.

The results above can be shown as in table 1 (unit: s):

Table 1. Single-attacker versus single-victim results (unit: s)

Slave Nodes	Detection Time	Traceback Time	Recovery Time
15	0.41075	11.06125	10.80175

2. Double-attackers versus single-victim: one node assigned with background traffic

In this case, node 15 and 17 are attackers and node 23 sends background traffic to node 20. The steps of the experiments are:

- At second 0, nodes 15, 17 and 23 start to request normal TCP connections, and the coordination agent starts the detection agent to work.
- At second 4, node 15 and 17 launch SYN Flooding attacks to node 5 by requesting TCP connections with node 16 as the spoofing source address.
- At second 20 and 14, node 15 and 17 respectively stops attacking and recover to send normal packets.
- At second 100 the simulation stops.

Results:

- Detection Time and Traceback Time are shown in table 2 (unit: s):
- Two attack paths of attacker 15 and 17 are all calculated: 15→0→10→14→6→5→20 and 17→2→3→4→5→20.

Table 2. Double-attackers versus single-victim with single node assigned with background traffic results(unit: s)

Slave Nodes	Detection Time	Traceback Time	Recovery Time
15	0.5175	11.0611	43.3276
17		10.051	10.979

3. Double-attackers versus single-victim: two nodes assigned with background traffic

In this case, node 15 and 17 are attackers. Nodes 18 and 23 respectively send background traffic to node 23 and 20. The steps of the experiments are:

- At second 0, nodes 15, 17, 18 and 23 start to request normal TCP connections, and the coordination agent starts the detection agent to work.
- At second 4, node 15 and 17 launch SYN Flooding attacks to node 5 by requesting TCP connections with node 16 as the spoofing source address.
- At second 20 and 14, node 15 and 17 respectively stops attacking and recover to send normal packets.
- At second 100 the simulation stops.
  Results:
- Detection Time and Traceback Time are shown in table 3 (unit: s):
- Two attack paths of attacker 15 and 17 are all calculated: 15→0→10→14→6→5→20 and 17→2→3→4→5→20.

Table 3. Double-attackers versus single-victim with two nodes assigned with background traffic results (unit: s)

Slave Nodes	Detection Time	Traceback Time	Recovery Time
15	0.41075	14.051	40.9685
17		11.06113	10.8427

From the experiments results, we conclude that:

- The proposed scheme behaves much better in detecting distributed SYN Flooding attacks at the victim (of course it's as well applicable to other kinds of DDoS attacks for the detection algorithm is traffic-based), because the time cost of detecting attacks is just around half a second;
- Under background traffic, it also can effectively trace back to the distributed attack sources and calculate attack paths correctly. With the traceback time is about 10 seconds; it implies that we can identify the sources very quickly after the attack starts. Accordingly, it helps us to prevent the attack as soon as possible before it stops;
- Network can be successfully recovered after a bearable time, and the recovery time has much relevance to the value of Δt.

4 Conclusions and Future Work

In this paper, we study the problem of DDoS attack detecting, tracing and defending, and a novel global defense architecture is proposed to protect the entire Internet from DDoS attacks. In particular, we analyze the feature of the DDoS attack and give a real-time traffic-based detection algorithm, and also present a novel traceback algorithm combining tracing with filtering to defending during tracing back. Simulations are performed under different strategies for detecting and defending SYN Flooding attack. From the results we obtained, we believe that the detecting and tracing performance is much better.

One immediate next step we are considering is to study other traceback methods in order to extend the applicability of our architecture, and not be limited in tracing SYN Flooding attack. Furthermore, current filtering design is the weakest point of our

architecture, because simply dropping packets may discard normal packets as well as malicious ones; accordingly, we will enhance the research of this aspect. Finally, the security of communication between different agents will also be considered before long.

References

1. Comp. Emergency Response Team: Results of the Distributed-Systems Intruder Tools Workshop. http://www.cert.org/reports/dsit_workshop-final.html(1999)
2. Haining, W., Danlu Z., Kang G. S.: Detecting SYN flooding attacks. Proc. IEEE INFOCOM. Volume: 3. IEEE Computer Society Press, Los Alamitos (2002) 1530-1539
3. Caberera, J.B.D., Ravichandran, B., Mehra, R.K: Statistical traffic modeling for network intrusion detection. Proc. IEEE International Symposium on Modeling, Analysis and Simulation of Computer Telecommunication Systems, (2000) 466-473
4. John, E. D., Jukka, J., Ourania, K., Julie, A. D.: Fuzzy intrusion detection. IFSA World Congress and 20th NAFIPS International Conference, Volume 3, (2001) 1506-1510
5. Garcia, R.C., Sadiku, M.N.O., Cannady, J.D.: WAID: wavelet analysis intrusion detection. Proc. IEEE MWSCAS'02, Volume 3, (2002) III-688 - III-691
6. Nash D. A., Daniel J. R.: Simulation of Self-Similarity in Network Utilization Patterns as a Precursor to Automated Testing of Intrusion Detection Systems. IEEE Transactions on systems, Volume 31, No. 4, (2001) 327-331
7. Jonckheere, E., Shah, K., Bohacek, S.: Dynamic modeling of Internet traffic for intrusion detection. Proc. American Control Conferenc, Volume: 3, (2002) 2436-2442
8. Ferguson, P. and Senie, D.: Network Ingress Filtering: Defeating Denial of Service Attacks which Employ IP Source Address Spoofing. RFC 2267, Jan. 1998
9. Park, K., Heejo, L.: On the Effectiveness of Route-Based Packet Filtering for Distributed DoS Attack Prevention in Power-Law Internets. Proc. ACM SIGCOMM (2001) 15-26
10. Bellovin, S. M.: ICMP Traceback Messages. IETF, Internet Draft: draftbellovin-itrace-00.txt, Mar. 2000
11. Stefan, S., David, W., Anna, K., Tom, A..: Practical Network Support for IP Traceback. Proc. ACM SIGCOMM (2000) 295-308
12. Snoeren, A. C., Partridge, C., Sanchez, L. A., Jones, C. E., Tchakountio, F., Kent, S. T., & Strayer, W. T.: Hash-Based IP Traceback. Proc. ACM SIGCOMM (2001) 3-14
13. Song., D. X., and Perrig, A.: Advanced and authenticated marking schemes for IP traceback. Proc. IEEE INFOCOM. IEEE Computer Society Press, Los Alamitos (2001) 878-886
14. Robert, S.: CenterTrack: An IP Overlay Network for Tracking DoS Floods. Proc. Ninth USENIX Security Symp (2000) 199-212
15. Xinyu, Y., Ming, Z., Rui, Z., Yi, S.:A Novel LMS Method for Real-time Network Traffic Prediction. Lecture Notes in Computer Science, Volume 3046. Springer-Verlag Heidelberg (2004) 127-136
16. The Network Simulator - ns-2. http://www.isi.edu/nsnam/ns/

A Variant of Poly1305 MAC and Its Security Proof*

Dayin Wang, Dongdai Lin, and Wenling Wu

SKLOIS Lab, Institute of Software,
Chinese Academy of Science, Beijing 100080, China
{wdy, ddlin, wwl}@is.iscas.ac.cn

Abstract. We give a variant of Poly1305 MAC and prove its security viewing this MAC as a Carter-Wegman MAC. The proposed variant not only keeps all the good properties of the Poly1305, but also makes Poly1305 deterministic.

1 Introduction

BACKGROUND. A Message Authentication Code (MAC) provides a way to detect whether a message has been tampered with during transmission. The usual model for authentication includes three participants: a transmitter, a receiver and an adversary. The transmitter sends a message over an insecure channel, where the adversary can introduce new messages as well as alter existing ones. The adversary's goal is to deceive the receiver into believing that the new message is authentic.

Generally, a message authentication scheme includes a MAC generation algorithm which, on input of a message M, a key Key, and a nonce often, generates a short tag. Normally we think of this tag and nonce as being transmitted along with the message M, as our way of authentication M. A message authentication scheme also includes a MAC verification algorithm which, on input of a message, a key, a nonce, and a purported tag, answers 0 or 1, with 1 indicating that the message M is deemed authentic, and 0 indicating that it is not.

Poly1305 MAC [1], computes a 16-byte authenticator of a variable-length message, using a 16-byte block-cipher key, a 16-byte additional key, and a 16-byte nonce. It has several useful features, such as provable security, extremely high speed, parallelizability and incrementality. The variant, proposed here, called DPoly1305, only needs a 16-byte block-cipher key, a 16-byte additional key. Thus DPoly1305 is deterministic and its security is very close to that of Poly1305. In addition, DPloy1305 keeps all the good properties of Poly1305.

Our Contribution: We give a variant of Poly1305-MAC and make it deterministic, i.e., a nonce is not needed any more. Thus it will decrease the weight of the communication and increase the speed of computing.

* Supported by the chinese national 973 Project (2004CB318004), 863 Project (2003AA144030), NSFC90204016 and NSFC60373047.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 375–380, 2005.
© Springer-Verlag Berlin Heidelberg 2005

On the other hand, we view it as an instance of the Carter-Wegman paradigm: with DPoly1305, one is enciphering the output of a universal hash function. Under this viewpoint the proving security is to analyze the collision probability of the universal hash function. This is a simple, information-theoretic question. Thus this approach effectively eliminates the complexity associated with the adversary's adaptivity. Using this method, we can get a simpler proof for DPoly1305 than for Ploy1305[1, 2].

2 Definition of DPoly1305

Some of the content is the same as the content in the Ploy1305[1]. We include here for completeness.

Messages
A message is any sequence of bytes $m[0]$, $m[1]$, $\cdots$, $m[l-1]$; a byte is any element of $\{0, 1, \cdots, 255\}$. The length l can be any nonnegative integer, and can vary from one message to another.

Keys
DPoly1305 authenticates messages using a 32-byte secret key shared by the message sender and the message receiver. The key has two parts: first, a 16-byte block cipher key k; second, a 16-byte string $r[0], r[1], \cdots, r[15]$. The second part of the key represents a 128-bit integer r in unsigned little-endian form: i.e.,$r = r[0] + 2^8 r[1] + \cdots + 2^{120} r[15]$.

Certain bits of r are required to be 0: $r[3], r[7], r[11], r[15]$ are required to have their top four bits clear(i.e., to be in $\{0, 1, \cdots, 15\}$), and $r[4], r[8], r[12]$ are required to have their bottom two bits clear(i.e., to be in $\{0, 4, 8, \cdots, 252\}$). Thus there are 2^{106} possibilities for r. In other words, r is required to have the form $r_0 + r_1 + r_2 + r_3$ where $r_0 \in \{0, 1, 2, 3, \cdots, 2^{28} - 1\}$, $r_1/2^{32} \in \{0, 4, 8, 12, \cdots, 2^{28} - 4\}$, $r_2/2^{64} \in \{0, 4, 8, 12, \cdots, 2^{28} - 4\}$, and $r_3/2^{96} \in \{0, 4, 8, 12, \cdots, 2^{28} - 4\}$.

Block Cipher
DPloy1305 requires a block cipher $E : \mathrm{K} \times \{0,1\}^{128} \rightarrow \{0,1\}^{128}$. Any block cipher that is pseudo-random permutation can be used here.

Conversion and Padding
Let $m[0], m[1], \cdots, m[l-1]$ be a message. Write $q = \lceil l/16 \rceil$, Define integers $c_1, c_2, \cdots, c_q \in \{1, 2, 3, \cdots, 2^{129}\}$ as follows: if $1 \le i \le \lfloor l/16 \rfloor$ then

$$c_i = m[16i-16] + 2^8 m[16i-15] + 2^{16} m[16i-14] + \cdots + 2^{120} m[16i-1] + 2^{128}$$

if l is not a multiple of 16 then

$$c_q = m[16q-16] + 2^8 m[16q-15] + \cdots + 2^{8(l \bmod 16)-8} m[l-1] + 2^{8(l \bmod 16)}$$

In other words: Pad each 16-byte chunk of a message to 17 bytes by appending a 1. If the message has a final chunk between 1 and 15 bytes, append 1 to the

chunk, and then zero-pad the chunk to 17 bytes. Either way, teat the resulting 17-byte chunk as an unsigned little-endian integer.

Authenticators
The DPoly1305 authenticator of a message m under secret key (k, r) is defined as the 16-byte string

$$E_k(((c_1 r^q + c_2 r^{q-1} + \cdots + c_q r^1) \ mod \ 2^{130} - 5) \ mod \ 2^{128})$$

Here $c_1, c_2, \cdots, c_q$ and block cipher E are defined above.

3 Carter-Wegman MAC

Carter-Wegman MAC is based on the universal hashing paradigm introduced by Carter and Wegman [3, 4]. They proposed to hash a given message with a randomly chosen function from a strongly universal family of hash functions, whereafter the output is encrypted with a one-time-pad (OTP) in order to obtain the MAC tag. In the original paper, Wegman and Carter [4] use perfect encryption to produce their MAC. Subsequently Black [5] give several other variants of these methods for producing a MAC given a universal hash family. In this section, we will briefly introduce universal hash families and a variant of Carter-Wegman type constructions that will be used later.

3.1 Universal Hash Families

There are many different variants of Universal Hash Families, and we now present three of them that will be used later.

In the following discussion and throughout the paper it is assumed that the domain and range of universal hash functions are finite sets of binary strings and that the range is smaller than the domain.

Definition 1. ***[Carter and Wegman,1979]*** *Fix a domain D and range R. A finite multiset of hash functions $H = \{h : D \longrightarrow R\}$ is said to be* ***Universal*** *if for every $x, y \in D$ where $x \neq y$, $\Pr_{h \in H}[h(x) = h(y)] = 1/|R|$.*

If we slightly relax the requirement of the collision probability to be some $\epsilon \geq 1/|R|$, we will get the notion of Almost Universal Hash Families.

Definition 2. *[3, 6] Let $\epsilon \in R^+$ be a positive number. Fix a domain D and a range R. A finite multiset of hash functions $H = \{h : D \longrightarrow R\}$ is said to be ϵ-Almost Universal(ϵ-AU) if for every $x, y \in D$ where $x \neq y$, $\Pr_{h \in H}[h(x) = h(y)] \leq \epsilon$.*

Definition 3. *[7, 8] Let $(B, +)$ be an Abelian group. A family H of hash functions that maps from a set A to the set B is said to be ϵ-almost Δ-universal (ϵ-AΔU)w.r.t.$(B, +)$, if for any distinct elements $x, y \in A$ and for all $g \in B$:*

$$\Pr_{h \in H}[h(x) - h(y) = g] \leq \epsilon$$

H is ΔU if $\epsilon = 1/|B|$.

Note that for the classes of hash function families defined in definitions 1-3, the latter are contained in the former, i.e. an ϵ-$A\Delta U$ family is also an ϵ-AU family.

3.2 From Hash to MAC

In [5], Black gives another approach to build a MAC in the Carter-Wegman style. This approach applies a PRF(Pseudo-random Function) directly to the output of an almost universal hash family. We now give a description of it. Suppose the shared key between a transmitter and a receiver is (h, ρ) where $h \in H = \{h : D \to \{0,1\}^n\}$ is a ϵ-AU and ρ is a randomly-choose function from $Rand(n, n)$. Given a message M, the output of MAC is $\rho(h(M))$, then $\rho(h(M))$ is not distinguishable with a random function family for any polynomial time adversary. Formally, there exists the following theorem.

Theorem 1 (PRF(Hash) as a PRF). *Fix $n \geq 1$. Let $H = \{h : Msg \to \{0,1\}^n\}$ be a family of hash functions. Let $\delta(m)$ be a function such that for all distinct pairs $M, M' \in Msg$,with M and M' at most mn bits long, $\Pr_{h \xleftarrow{R} H}[h(M) = h(M')] \leq \delta(m)$. Let A be an adversary that asks q queries from Msg, each query of length at most mn bits. Then*

$$Pr[h \xleftarrow{R} H; \rho \xleftarrow{R} Rand(n,n) : A^{\rho(h(\cdot))} = 1] - Pr[R \xleftarrow{R} Rand(Msg, n) : A^{R(\cdot)} = 1] \leq \frac{q^2}{2}\delta(m)$$

This theorem's proof is given in Lemma 6.3.6 in [5].

4 Security

The security proof of DPoly1305 is done by viewing it as a Carter-Wegman MAC. Firstly, we divide DPoly1305 into two parts naturally, one is a universal hash function, and the other is a PRF. Then we use Theorem 1 to complete the proof. Before giving the definition of the universal hash function in DPoly1305, we give a few lemmas that will be used in the proof.

It is often convenient to replace random permutations with random functions, or vice versa, in security proof. The following lemma lets us easily do this. For a proof see Proposition 2.5 in [9].

Lemma 1 (PRF/PRP Switching). *Fix $n \geq 1$. Let A be an adversary that asks at most q queries. Then*

$$|\Pr[\pi \xleftarrow{R} Perm(n) : A^{\pi(\cdot)} = 1] - \Pr[\rho \xleftarrow{R} Rand(n) : A^{\rho(\cdot)} = 1]| \leq q(q-1)/2^{n+1}$$

As is customary, we will show the security of our MACs by showing that their information-theoretic versions approximate random functions. As is standard, this will be enough to pass to the complexity-theoretic scenario. Part of the proof is Propositon 2.7 of [9].

Lemma 2 (Inf.Th.PRF$\Longrightarrow$ Comp.Th.PRF). *Fix $n \geq 1$. Let $CONS$ be a construction such that $CONS_\rho(\cdot) : \{0,1\}^* \to \{0,1\}^n$ for any $\rho \in Rand(n,n)$. Suppose that if $|M| \leq \mu$ then $CONS_\rho(M)$ depends on the values of ρ on at most p points. Let $E : Key \times \{0,1\}^n \to \{0,1\}^n$ be a family of functions. Then*

$$Adv^{prf}_{CONS[E]}(t,q,\mu) \leq Adv^{prf}_{CONS[Perm(n)]}(q,\mu) + \cdot Adv^{prf}_{E}(t^{'},q), \text{ and}$$
$$Adv^{mac}_{CONS[E]}(t,q,\mu) \leq Adv^{prf}_{CONS[Perm(n)]}(q,\mu) + \cdot Adv^{prf}_{E}(t^{'},q) + \tfrac{1}{2^n}$$

where $t^{'} = t + O(pn)$.

4.1 Universal Hash Function in DPoly1305

Here we describe a hash family that we will call "Poly hash." Let the hash domain $D = \{0,1\}^*$. Let Poly hash $\mathrm{H} = \{H_r : D \to \{0,1\}^{128}\}$ be the family of functions where members are selected by the random choice of some r. For any message m, write $\overline{m}$ for the polynomial $c_1 x^q + c_2 x^{q-1} + \cdots + c_q x^1$, where $q, c_1, c_2, \cdots, c_q$ are defined as in Section 2. Define $H_r(m)$, a member of Poly hash, as the 16-byte unsigned little-endian representation of $(\overline{m}(r) \bmod 2^{130} - 5) \bmod 2^{128}$; then the authenticator $\mathrm{DPoly1305}_{E,\mathrm{H}}(m)$ is equal to $E_k(H_r(m))$ under secret key (k,r).

The crucial property of H_r described in [1] is that it has small differential probabilities: if g is a 16-byte string, and m, m' are distinct messages of length at most L, then $H_r(m) = H_r(m') + g$ with probability at most $8\lceil L/16 \rceil / 2^{106}$.

Theorem 2. *Suppose all distinct pairs m,$m'$are at most L bytes long, the* Poly hash *on domain $D(\{0,1\}^*)$ is $8\lceil L/16 \rceil / 2^{106}$-$AU$.*

Proof. Theorem 3.3 in [1] has proved that the probability of $H_r(m) = H_r(m^{'}) + g$ is at most $8\lceil L/16 \rceil / 2^{106}$ under condition that g is a 16-byte string, and m, m' are distinct messages of length at most L. From the definition above, we know H is a $8\lceil L/16 \rceil / 2^{106}$-$A\Delta U$ hash family. Thus we say H is a $8\lceil L/16 \rceil / 2^{106}$-$AU$ hash family. □

4.2 Security Proof of DPoly1305

In the following we prove the security of the $\mathrm{DPoly1305}_{\pi,\mathrm{H}}$ construction.

Theorem 3 (DPoly1305$\approx$ Rand). *Fix $n \geq 1$ and let $n = 128$. Let A be an adversary which asks at most q queries, each of which is at most L bits, then*

$$Pr[H_r \xleftarrow{R} \mathrm{H},\ \pi \xleftarrow{R} Perm(n) : A^{DPoly1305_{\pi,H_r}(\cdot)} = 1] -$$
$$Pr[\ R \xleftarrow{R} Rand(\{0,1\}^*, n) : A^{R(\cdot)} = 1] \leq \frac{q^2}{2} \cdot \frac{8\lceil L/16 \rceil}{2^{106}} + \frac{q^2}{2^{n+1}}$$

Proof. We first compute a related probability where the random permutation π is replaced by a random function:

$$\Pr[H_r \xleftarrow{R} \mathrm{H}, \rho \xleftarrow{R} Rand(n,n) : A^{\mathrm{DPoly1305}_{\rho,H_r}(\cdot)} = 1] -$$
$$\Pr[R \xleftarrow{R} Rand(\{0,1\}^*, n) : A^{R(\cdot)} = 1]$$

Theorem 2 tells us that H is $8\lceil L/16\rceil/2^{106}$-AU, so we know $\Pr_{H_r \xleftarrow{R} \mathrm{H}}[H_r(m) = H_r(m')] \leq \frac{8\lceil L/16\rceil}{2^{106}}$. Therefore we can conclude from Theorem 1 that

$$\Pr[H_r \xleftarrow{R} \mathrm{H}, \rho \xleftarrow{R} Rand(n,n) : A^{\rho(H_r(\cdot))} = 1] - \Pr[R \xleftarrow{R} Rand(\{0,1\}^*, n) : A^{R(\cdot)} = 1] \leq \frac{q^2}{2} \cdot \frac{8\lceil L/16\rceil}{2^{106}}$$

Finally we replace the PRF ρ on the left-hand with a PRP to properly realize the DPoly1305 construction. From Lemma 1, this costs us an extra $\frac{q^2}{2^{n+1}}$, so our final bound is

$$\frac{q^2}{2} \cdot \frac{8\lceil L/16\rceil}{2^{106}} + \frac{q^2}{2^{n+1}}$$

□

From Lemma 2 it is standard to pass from information-theoretic to a complexity-theoretic.

It is a standard result that being secure in the sense of a PRF implies an inability to forge with good probability see [9, 10].

5 Conclusions

In this paper we propose a MAC called DPloy1305, a variant of Ploy1305 MAC, which keeps all the good properties of Ploy1305 MAC and no nonce is needed any more. Thus DPloy1305 becomes deterministic. Viewing it as a Carter-Wegman MAC, we effectively eliminates the complexity associated with the adversary's adaptivity and get a simple proof for DPloy1305.

References

1. D. J. Bernstein.: The Poly1305-AES message-authentication code. in Proceedings of FSE 2005.
2. D. J. Bernstein.: Stronger security bounds for Wegman-Carter-Shoup authenticators. in Advances in Cryptology:EUROCRYPT 2005, (2005)164-180.
3. J. Carter and M. Wegman.: Universal classes of hash functions. in Journal of Computer and System Sciences, vol.18, (1979)143-154.
4. M. Wegman and J. Carter.: New hash functions and their use in authentication and set equality. in Journal of Computer and System Sciences, vol.22, (1981)265-279.
5. J. Black.: Message Authentication Codes. http://www.cs.colorado.edu/~jrblack.
6. D. Stinson.: Universal hashing and authentication codes. in Crypto'91, LNCS 576(1992)74-85.
7. H. Krawczyk.: LFSR-based hashing and authentication. in Crypto'94, LNCS 839(1994)129-139.
8. D. Stinson.: On the connection between universal hashing, combinatorial designs and error-correcting codes. In Proc. Congressus Numerantium 114(1996)7-27.
9. M. Bellare, J. Kilian and P. Rogaway.: The security of the cipher block chaining message authentication code. in http://www.cs.ucdavis.edu/~rogaway.
10. O. Goldreich, S. Goldwasser and S. Micali.: How to construct random functions. in Journal of the ACM, vol.33, no.4, (1986) 210-217.

Covert Channel Identification Founded on Information Flow Analysis

Jianjun Shen, Sihan Qing, Qingni Shen, and Liping Li

Institute of Software, Chinese Academy of Sciences, 100080 Beijing, China
Graduate University of Chinese Academy of Sciences, 100080 Beijing, China
weishen_cu@yahoo.com.cn

Abstract. This paper focuses on covert channel identification in a nondiscretionary secure system. The properties of covert channels are analyzed by channel types. Information flow characteristics are utilized to optimize channel identification with the Share Resource Matrix method adopted for demonstration, and a general framework for channel identification founded on information flow analysis is presented. At last, timing channels are also discussed.

1 Introduction

The notion of covert channel was first introduced by Lampson [1] in 1973. As a significant threat to system securities, covert channels are required to be analyzed in a secure information system. Channel identification is a particularly difficult task in covert channel analysis, which requires vast manual efforts.

Tsai [2] proposed to categorize information flows and covert channels for the advantages of analysis. However, he failed to give precise definitions and verifications, and thus deduced a series of inexact conclusions. In this paper, covert channels are classified according to their information flow characteristics with explicit definitions. On this basis, the optimization of channel identification is explored.

2 Information Flow and Information Flow Sequence

When analyzing a Trusted Computing Base (TCB) for covert channels, information flows between shared TCB variables and through the TCB interface are considered.

2.1 Information Flow Sequence

Viewed from the TCB interface, an information flow is evoked by a series of TCB primitive operations. Each operation transits the flow to its next medium variable. Thus, it is reasonable to partition information flows with primitive operations. Each transitive step as the basic flow unit can be represented in the form of a quadruple:

$$(s, op, vs_1, vs_2) \in S \times OP \times VS \times VS \tag{1}$$

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 381 – 387, 2005.
© Springer-Verlag Berlin Heidelberg 2005

where S is the set of all subjects, OP is the TCB primitive set, and VS is the shared variable set. It denotes the invocation of primitive op by subject s evokes an information flow from variable vs_1 to vs_2.

If let $s(op,vs)$ represent subject s invokes primitive op causing the flow transits to variable vs, an information flow can be represented by:

$$vs_0 \; s_1(op_1, vs_1) \cdots s_n(op_n, vs_n) \tag{2}$$

which represents op_1 by s_1, …, and op_n by s_n evoke a flow: $vs_0 \rightarrow vs_1 \rightarrow \ldots \rightarrow vs_n$.

2.2 Information Flow Types

Definition 1. **Direct Flow:** Information flow evoked by a single primitive operation. That is, a direct flow takes only one transitive step.

Definition 2. **Indirect Flow:** Information flow evoked by multiple operations.

Definition 3. **Internal Flow:** Information flow between internal TCB variables.

Definition 4. **Process Flow:** Information flow that traverses the TCB interface, and returns to the TCB outside.

We introduce two variables: *In* and *Out* to represent the inputs and outputs of primitives. A process flow can then be expressed as:

$$v_0 \; s_1(op_1, vs_1) \cdots s_n(op_n, Out), v_0 \in VS \bigcup \{In\} \tag{3}$$

3 Information Flow and Covert Channel

A covert channel implies an information flow that violates the security policy. In this paper we adopt a simplified multilevel secure policy model for discussion convenience. Two processes: P_S and P_R, $SL(P_S) \not\leq SL(P_R)$, are introduced to represent subject sender and receiver respectively. According to the model's security policy, any communication from P_S to P_R is covert communication.

3.1 Covert Information Flow

A covert channel can be expressed in the form of an information flow sequence. Such a flow is called a covert information flow, which has following features:

1. Should be an indirect process flow.
2. Is constituted by concatenation of the primitive operation sequence invoked by the sender and the operation sequence invoked by the receiver.
3. May lack an evident initial variable. Or the initial variable is not essential to the existence of the covert channel.

According to the above arguments, a covert channel can be generally expressed as:

$$P_S(op_1, vs_1) \cdots P_S(op_m, vs_m) \; P_R(op_{m+1}, vs_{m+1}) \cdots P_R(op_n, Out) \tag{4}$$

We can find out some characteristics of a covert information flow that are utilizable in optimizing channel identification (see section 4):

a.1: The flow sequence transits to *Out*, and then terminated.
a.2: There is only one subject switch in the flow sequence which occurs at the joint point of the sending sequence and receiving sequence.
a.3: The security levels of the subjects before and after the switching point satisfy: $SL(P_S) \not\leq SL(P_R)$.

Definition 5. **Direct Channel:** Covert channel that is composed of a single sending operation and a single receiving operation.

Definition 6. **Indirect Channel:** Covert channel whose sending sequence or receiving sequence consists of more than one operation.

3.2 Single Channel and Plural Channel

Tsai [2] deemed direct channels to be the components of all covert channels. He thereby proposed that all covert channels could be properly handled by handling just direct channels. In order to clarify the problem, we introduce several notions first:

Definition 7. **Sub-channel:** c and c_1 are two different covert channels. Omitting the destination variable (must be *Out*), if the flow sequence of c_1 is a sub-sequence of c, then c_1 is a sub-channel of c, represented by: $c_1 \subset c$.

That is, if the flow sequence of c appears like: $P_S(op_1,vs_1)\ldots P_S(op_m,vs_m)\ P_R(op_{m+1}, vs_{m+1})\ldots P_R(op_n,Out)$, then c_1 shall be like: $P_S(op_i,vs_i)\ldots P_S(op_m,vs_m)\ P_R(op_{m+1}, vs_{m+1})\ldots P_R(op_j,Out)$, $1\leq i\leq m \wedge m<j\leq n \wedge c_1\neq c$.

Definition 8. **Single Channel:** Covert channel that has no sub-channel.

Definition 9. **Plural Channel:** Covert channel that has sub-channels. Part of the flow sequence of a plural channel already forms a covert channel.

Tsai's viewpoint is just that a single channel must be a direct channel. However, it is true only when the constituent primitives of a channel have no restriction on subjects that invoke them.

Let us consider an instance in practical system – the "File Node Number" channel. In many file systems, each file is identified by a node number which is incrementally assigned. This mechanism can be utilized in covert communication: (1) The sender decides whether or not to invoke primitive *Create* to create a new file depending on the value of the bit to transmit; (2) The receiver invokes *Create* to create a new file; (3) and then invokes primitive *Stat* to obtain the node number of the new file; (4) By comparing this number with the previous one, the receiver is able to know the value of the received bit. This channel can be represented by: $\underline{P_S(Create,LastInode)\ P_R(Create, File.I_No)\ P_R(Stat,Out)}$. It is a single channel, if we only allow a process to observe a file's node number via *Stat* when its security level dominates that of the file.

It is not hard to know that the sending sequence of a single channel consists of only one sending operation. A single channel can be expressed as:

$$scc : P_S(op_1, vs_1)\ P_R(op_2, vs_2) \cdots P_R(op_n, Out), \neg\exists(cc)\left(cc \in CC \wedge cc \subset scc\right) \quad (5)$$

Sequence (5) presents following characteristics:

b.1: The information flow is unable to find an outlet before it transits to the last operation. That is, none of the intermediate variable vs_i ($1 \leq i < n$) is visible to P_R via op_{i+1}.

The flow sequence of any covert channel contains the operation sequence of at least one single channel, and a plural channel must have at least one sub-channel that is a single channel. Therefore all covert channels can be eliminated through eliminating single channels (interdicting the invocation of a flow sequence as a whole). However, this is not equally true for the bandwidth-limitation policy. Tsai argued that the bandwidth of a plural channel is not higher than that of its sub-channel. So if the bandwidths of all its sub-channels are limited to a threshold, the bandwidth of the plural channel should be under the threshold too. Actually, it is not always right.

In conclusion, to attain the completeness of channel handling, at least all single channels shall be identified. The actual identification solution depends on practical completeness requirements, and handling policies should be taken into account.

4 Optimization of Covert Channel Identification

4.1 Identification Framework Founded on Information Flow Analysis

Information flow analysis is a basic means for identifying covert channels; almost all identification methods are based on the identification of illegal flows. As first introduced by Denning [3], original information flow analysis techniques judge the legality of a flow by checking the security-level relationships between its medium variables. This methodology may leas to the generation of large amounts of false illegal flows for two reasons: (1) It is hard or impossible to determine the security level of an individual variable. (2) A generated illegal flow may not be exploitable in real illegal communication.

In fact, a covert channel can be seemed as an information flow between two subjects - the sender and the receiver. So we believe a more practical methodology is to check flow legalities by examining the security levels of flow subjects, while the security restrictions on subjects and objects can help to verify the validity of flows.

Tsai and He et al. [2, 4, 5] attempted to improve the information flow analysis for source code, but they neglect indirect channels. As another class of methods with extensive applications, the SRM [6] and the CFT [7] methods are also grounded on information flow analysis, but they are not complete identification methods.

We divide the whole process of channel identification into three phases: (1) Analyze TCB primitives to find out the direct flows evoked by each primitive. (2) Combine direct flows to construct indirect flows following specific rules. (3) Analyze result doubt flows and identify covert channels. The SRM and other similar methods are just techniques to express and combine information flows, and are independent of the first and third phases. The work by Tsai et al. proposes a practical approach to search direct flows, but skips over the second phase. Therefore, it is possible to inte-

grate these two classes of methods into a complete framework. As for the SRM method, the results of direct flow analysis can be used to initialize the SRM matrix.

4.2 Shared Resource Matrix (SRM) Method

The SRM method adopts a matrix structure to express information flows. Each column of an initialized SRM matrix contains (implicitly) the direct information flows evoked by the corresponding primitive. The SRM transitive closure is actually the cause of combining direct flows to construct indirect flows.

The SRM method is a "lazy" method. It dose not require initial direct flows to be explicitly distinguished when the SRM matrix is constructed, and takes no effort to control the transits of flows during the transitive closure cause. All analysis work is left to last. McHugh [8] made three extensions to the SRM method, which refine the capability of the SRM in describing flows. Nevertheless, the method still has some drawbacks:

1. The SRM method can not record information flow sequences.
2. The transitive cause in the SRM method lacks directivity.
   During the transitive closure cause, each flow transits forward from the second step of its flow sequence. It is impossible to judge whether a flow can reach an outlet (transit to a visible variable) at last. Lots of transits do not contribute to finding a channel.
3. The transits of flows are uncontrolled in the SRM method. Large amounts of false illegal flows may be generated, which increases analysis work unnecessarily.

4.3 Optimizing Shared Resource Matrix and Covert Flow Tree (CFT) Methods

We revise the SRM method against its shortages. First, it is possible to enable the SRM matrix to record flow paths through following two extensions:

1. When a flow transits to a new variable during the transitive closure cause, not only mark the corresponding matrix grid, but also record the previous primitive and variable sequence.
2. Carrying out the transitive closure by rounds. At each round, attempt to extend flows generated in last round. Flows extend step by step in order. A flow advances just one step per round to connect a next direct flow.

Against shortage 2, we propose a Reversed SRM method. In the original positive SRM method, flows transit forward from their sources, and the indirect reference relationships between variables are discovered in this course. By contrast, the flows in the reversed SRM method trace back to their sources from the destination, and the indirect modification relationships are discovered. The reversed transitive closure starts its first round from visible variables, so invalid internal transits are avoided.

Against shortage 3, we adopted a series of control mechanisms:

1. Appending the subject restrictions upon primitives to each matrix column for judging the validity of flow transits.

The occurrences of direct flows often need to pass specific access checks (mainly mandatory access control (MAC) restrictions considered here), which determine most information flows are not exploitable in actual illegal communication. It is not hard to extract the MAC checks upon primitives from ether TCB source code or top-level specifications. They can be effectively applied to inspect the "doubt" flows generated during the transitive closure.

2. Introducing transitive constraints. The transits of flows must satisfy specific transitive constraints, which come from the information flow characteristics of channels.

   We can conclude the positive and reversed transitive constraints for each channel type from the analysis results in section 3.

3. Analyzing newly extended flows after every transitive round. Excluding false and invalid flows as soon as possible.

   The CFT method is virtually a transformation of the SRM method. It can replace the SRM method to fulfill the second-phase mission of combining information flows. Due to its tree structure, the CFT method is capable of recording flow paths. But all other drawbacks of the SRM method still exist in CFT. Our revisions to the SRM method are also applicable to the CFT method.

5 Timing Channel

He [5] divided timing channels into Tightly Coupled and Loosely Coupled Timing Channels. In a tightly coupled channel, the sender modifies the system state in such a way that the execution of the receiver is affected; while in a loosely coupled channel, the sender affects the time at which the receiver starts to execute. We introduce a new variable: *TRef* to represent time reference and a primitive: *Time* which reads the value of *TRef*. Then the two types of timing channels can be expressed respectively as:

$$P_S(op_1, vs_1)\cdots P_S(op_m, vs_m)\ P_R(op_{m+1}, vs_{m+1})\cdots P_R(op_{n-1}, T\,\mathrm{Re}\,f)P_R(Time, Out) \tag{6}$$

$$P_S(op_1, vs_1)\cdots P_S(op_{n-1}, T\,\mathrm{Re}\,f)P_R(Time, Out) \tag{7}$$

6 Conclusion

We think the classification of covert channels benefits channel analysis, and can help evaluate the thoroughness of analysis work. To attain the completeness of covert channel handling, we suggest that at least all single channels shall be identified, and handling polices should be considered when determining an identification solution. We showed how information flow characteristics could be applied to optimize channel identification. Instead of analyzing the security levels of individual variables, we believe a more feasible methodology for identifying illegal flows is to inspect the flows between subjects that actually violate the security policy to some extent.

References

1. B. Lampson. A Note on the Confinement Problem. *Communications of the ACM*, 16:10, Oct. (1973) 613-615.
2. C. Tsai. Covert Channel Analysis in Secure Computer Systems. PhD Dissertation, University of Maryland, (1987).
3. D. Denning. A Lattice Model of Secure Information Flow. *Communications of the ACM*, 19:5, May (1976) 236-243.
4. C. Tsai, V. Gligor and C. Chandersekaran. A Formal Method for the Identification of Covert Storage Channels in Source Code. In *Proc. of IEEE Symposium on Security and Privacy*, Oakland, CA, Apr. (1987) 74-86.
5. J. He and V. Gligor. Formal Methods and Automated Tool for Timing-Channel Identification in TCB Source Code. In *Proc. of 2nd European Symposium on Research in Computer Security*, Nov. (1992) 57–75.
6. R. Kemmerer. Shared Resource Matrix Methodology: An Approach to Identifying Storage and Timing Channels. *ACM Trans. on Computer Systems*, 1:3, Aug. (1983) 256-277.
7. R. Kemmerer. Covert Flow Trees. A Visual Approach to Analyzing Covert Storage Channels. *IEEE Trans. on Software Engineering*, 17:11, Nov. (1991) 1166-1185.
8. J. McHugh. Handbook for the Computer Security Certification of Trusted Systems - Covert Channel Analysis. Technical Report, Naval Research Laboratory, Feb. (1996).
9. J. He and V. Gligor. Information Flow Analysis for Covert-Channel Identification in Multilevel Secure Operating Systems. In *Proc. of the 3rd IEEE Workshop on Computer Security Foundations*, Franconia, NH, Jun. (1990) 139-148.

Real-Time Risk Assessment with Network Sensors and Intrusion Detection Systems

André Årnes, Karin Sallhammar, Kjetil Haslum, Tønnes Brekne, Marie Elisabeth Gaup Moe, and Svein Johan Knapskog

Centre for Quantifiable Quality of Service in Communication Systems,*
Norwegian University of Science and Technology,
O.S. Bragstads plass 2E, N-7491 Trondheim, Norway
{andrearn, sallhamm, haslum, tonnes, marieeli, knapskog}@q2s.ntnu.no

Abstract. This paper considers a real-time risk assessment method for information systems and networks based on observations from networks sensors such as intrusion detection systems. The system risk is dynamically evaluated using hidden Markov models, providing a mechanism for handling data from sensors with different trustworthiness in terms of false positives and negatives. The method provides a higher level of abstraction for monitoring network security, suitable for risk management and intrusion response applications.

1 Introduction

Risk assessment is a central issue in management of large-scale networks. However, current risk assessment methodologies focus on manual risk analysis of networks during system design or through periodic reviews. Techniques for real-time risk assessment are scarce, and network monitoring systems and intrusion detection systems (IDS) are the typical approaches. In this paper, we present a real-time risk assessment method for large scale networks that build upon existing network monitoring and intrusion detection systems. An additional level of abstraction is added to the network monitoring process, focusing on risk rather than individual warnings and alerts. The method enables the assessment of risk both on a system-wide level, as well as for individual objects.

The main benefit of our approach is the ability to aggregate data from different sensors with different weighting according to the trustworthiness of the sensors. This focus on an aggregate risk level is deemed more suitable for network management and automated response than individual intrusion detection alerts. By using hidden Markov models (HMM), we can find the most likely state probability distribution of monitored objects, considering the trustworthiness of the IDS. We do not make any assumptions on the types of sensors used in our monitoring architecture, other than that they are capable of providing standardized output as required by the model parameters presented in this paper.

* The "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" is appointed by The Research Council of Norway, and funded by the Research Council, NTNU and UNINETT.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 388–397, 2005.
© Springer-Verlag Berlin Heidelberg 2005

1.1 Target Network Architecture

The target of the risk assessment described in this paper is a generic network consisting of computers, network components, services, users, etc. The network can be arbitrarily complex, with wireless ad-hoc devices as well as ubiquitous services. The network consists of entities that are either *subjects* or *objects*. Subjects are capable of performing actions on the objects. A subject can be either users or programs, whereas objects are the targets of the risk assessment. An asset may be considered an object. The unknown factors in such a network may represent vulnerabilities that can be exploited by a malicious attacker or computer program and result in unwanted incidents. The potential exploitation of a vulnerability is described as threats to assets. The *risk* of a system can be identified through the evaluation of the probability and consequence of unwanted incidents.

1.2 Monitoring and Assessment Architecture

We assume a multiagent system architecture consisting of agents that observe objects in a network using sensors. The architecture of a multiagent risk assessment system per se is not the focus of this paper, but a description is included as a context.

An *agent* is a computer program capable of a certain degree of autonomous actions. In a multiagent system, agents are capable of communicating and cooperating with other agents. In this paper, an agent is responsible for collecting and aggregating sensor data from a set of sensors that monitor a set of objects. The main task of the agent is to perform real-time risk assessment based on these data. A multiagent architecture has been chosen for its flexibility and scalability, and in order to support distributed automated response.

A *sensor* can be any information-gathering program or device, including network sniffers (using sampling or filtering), different types of intrusion detection systems (IDS), logging systems, virus detectors, honeypots, etc. The main task of the sensors is to gather information regarding the security state of objects. The assumed monitoring architecture is hybrid in the sense that it supports any type of sensor. However, it is assumed that the sensors are able to classify and send standardized observations according to the risk assessment model described in this paper.

1.3 Related Work

Risk assessment has traditionally been a manual analysis process based on a standardized framework, such as [1]. A notable example of real-time risk assessment is presented in [2], which introduces a formal model for the real time characterization of risk faced by a host. *Distributed intrusion detection systems* have been demonstrated in several prototypes and research papers, such as [3, 4]. Multiagent systems for intrusion detection, as proposed in [5] and demonstrated

in e.g. [6] (an IDS prototype based on lightweight mobile agents) are of particular relevance for this paper. An important development in distributed intrusion detection is the recent IDMEF (Intrusion Detection Message Exchange Format) IETF Internet draft [7]. *Hidden Markov models* have recently been used in IDS architectures to detect multi-stage attacks [8], and as a tool to detect misuse based on operating system calls [9]. *Intrusion tolerance* is a recent research field in information security related to the field of fault tolerance in networks. The research project SITAR [10] presents a generic state transition model, similar to the model used in this paper, to describe the dynamics of intrusion tolerant systems. Probabilistic validation of intrusion tolerant systems is presented in [11].

2 Risk Assessment Model

In order to be able to perform dynamic risk assessment of a system, we formalize the distributed network sensor architecture described in the previous section. Let $O = \{o_1, o_2, \ldots\}$ be the set of objects that are monitored by an agent. This set of objects represents the part of the network that the agent is responsible for. To describe the security state of each object, we use discrete-time Markov chains. Assume that each object consisting of N states, denoted $S = \{s_1, s_2, \ldots, s_N\}$.

As the security state of an object changes over time, it will move between the states in S. The sequence of states that an object visits is denoted $X = x_1, x_2, \ldots, x_T$, where $x_t \in S$ is the state visited at time t. For the purpose of this paper, we assume that the state space can be represented by a general model consisting of three states: Good (G), Attacked (A) and Compromised (C), i.e. $S = \{G, A, C\}$. State G means that the object is up and running securely and that it is not subject to any kind of attack actions. In contrast to [10], we assume that objects always are vulnerable to attacks, even in state G. As an attack against an object is initiated, it will move to security state A. An object in state A is subject to an ongoing attack, possibly affecting its behavior with regard to security. Finally, an object enters state C if it has been successfully compromised by an attacker. An object in state C is assumed to be completely at the mercy of an attacker and subject to any kind of confidentiality, integrity and/or availability breaches.

The security observations are provided by the sensors that monitor the objects. These *observation messages* are processed by agents, and it is assumed that the messages are received or collected at *discrete time intervals*. An observation message can consist of any of the symbols $V = \{v_1, v_2, \ldots, v_M\}$. These symbols may be used to represent different types of alarms, suspect traffic patterns, entries in log data files, input from network administrators, and so on. The *sequence* of observed messages that an agent receives is denoted $Y = y_1, y_2, \ldots, y_T$, where $y_t \in V$ is the observation message received at time t. Based on the sequence of observation messages, the agent performs dynamic risk assessment. The agent will often receive observation messages from more than one sensor, and these sensors may provide different types of data, or even inconsistent data.

All sensors will not be able to register all kinds of attacks, so we cannot assume that an agent is able to resolve the correct state of the monitored objects at all times. The observation symbols are therefore probabilistic functions of the object's Markov chain, the object's true security state will be *hidden* from the agent. This is consistent with the basic idea of HMM [12].

2.1 Modeling Objects as Hidden Markov Models

Each monitored object can be represented by a HMM, defined by $\lambda = \{\mathbf{P}, \mathbf{Q}, \pi\}$.

$\mathbf{P} = \{p_{ij}\}$ is the state transition probability distribution matrix for object o, where $p_{ij} = P(x_{t+1} = s_j | x_t = s_i), 1 \leq i, j \leq N$. Hence, p_{ij} represents the probability that object o will transfer into state s_j next, given that its current state is s_i. To be able to estimate $\mathbf{P}$ for real-life objects, one may use either statistical attack data from production or experimental systems or the subjective opinion of experts. Learning algorithms may be employed in order to provide a better estimate of $\mathbf{P}$ over time.

$\mathbf{Q} = \{q_j(l)\}$ is the observation symbol probability distribution matrix for object o in state s_j, whose elements are $q_j(l) = P(y_t = v_l | x_t = s_j), 1 \leq j \leq N, 1 \leq l \leq M$. In our model, the element $q_j(l)$ in $\mathbf{Q}$ represents the probability that a sensor will send the observation symbol v_l at time t, given that the object is in state s_j at time t. $\mathbf{Q}$ therefore indicates the sensor's false-positive and false-negative effect on the agents risk assessments.

$\pi = \{\pi_i\}$ is the initial state distribution for the object. Hence, $\pi_i = P(x_1 = s_i)$ is the probability that s_i was the initial state of the object.

2.2 Quantitative Risk Assessment

Following the terminology in [1], risk is measured in terms of *consequences* and *likelihood*. A consequence is the (qualitative or quantitative) outcome of an event and the likelihood is a description of the probability of the event. To perform dynamic risk assessment, we need a mapping: $\mathcal{C} : S \rightarrow \mathbb{R}$, describing the expected cost (due to loss of confidentiality, integrity and availability) for each object. The total risk $\mathcal{R}_t$ for an object at time t is

$$\mathcal{R}_t = \sum_{i=1}^{N} \mathcal{R}_t(i) = \sum_{i=1}^{N} \gamma_t(i) \mathcal{C}(i) \tag{1}$$

where $\gamma_t(i)$ is the probability that the object is in security state s_i at time t, and $\mathcal{C}(i)$ is the cost value associated with state s_i.

In order to perform real-time risk assessment for an object, an agent has to dynamically update the object's state probability $\gamma_t = \{\gamma_t(i)\}$. Given an observation y_t, and the HMM λ, the agent can update the state probability γ_t of an object using Algorithm 1. The complexity of the algorithm is $O(N^2)$. For further details, see the Appendix.

Algorithm 1. Update state probability distribution

IN: y_t, λ {the observation at time t, the hidden Markov model}
OUT: γ_t {the security state probability at time t}
  if $t = 1$ **then**
    for $i = 1$ to N **do**
      $\alpha_1(i) \leftarrow q_i(y_1)\pi_i$
      $\gamma_1(i) \leftarrow \frac{q_i(y_1)\pi_i}{\sum_{j=1}^{N} q_j(y_1)\pi_j}$
    end for
  else
    for $i = 1$ to N **do**
      $\alpha_t(i) \leftarrow q_i(y_t) \sum_{j=1}^{N} \alpha_{t-1}(j)p_{ji}$
      $\gamma_t(i) = \frac{\alpha_t(i)}{\sum_{j=1}^{N} \alpha_t(j)}$
    end for
  end if
  return γ_t

3 Case – Real-Time Risk Assessment for a Home Office

To illustrate the theory, we perform real-time risk assessment of a typical home office network, consisting of an Internet router/WLAN access point, a stationary computer with disk and printer sharing, a laptop using WLAN, and a cell phone connected to the laptop using Bluetooth. Each of the objects (hosts) in the home office network has a sensor that processes log files and checks system integrity (a host IDS). In addition, the access point has a network monitoring sensor that is capable of monitoring traffic between the outside network and the internal hosts (a network IDS).

For all objects, we use the state set $S = \{G, A, C\}$. The sensors provide observations in a standardized message format, such as IDMEF, and they are capable of classifying observations as indications of the object state. Each sensor is equipped with a database of signatures of potential attacks. For the purpose of this example, each signature is associated with a particular state in S. We define the observation symbols set as $V = \{g, a, c\}$, where the symbol g is an indication of state G and so forth. Note that we have to preserve the discrete-time property of the HMM by sampling sensor data periodically. If there are multiple observations during a period, we sample one at random. If there are no observations, we assume the observation symbol to be g. In order to use multiple sensors for a single object, a round-robin sampling is used to process only one observation for each period. This is demonstrated in example 3.

The home network is monitored by an agent that regularly receives observation symbols from the sensors. For each new symbol, the agent uses Algorithm 1 to update the objects' security state probability, and (1) to compute its corresponding risk value. Estimating the matrices $\mathbf{P}$ and $\mathbf{Q}$, as well as the cost $\mathcal{C}$ associated with the different states, for the objects in this network is a non-trivial task that is out of scope for this paper.

The parameter values in these examples are therefore chosen for illustration purposes only. Also, we only demonstrate how to perform dynamic risk assessment of the laptop.

3.1 Example 1: Laptop Risk Assessment by HIDS Observations

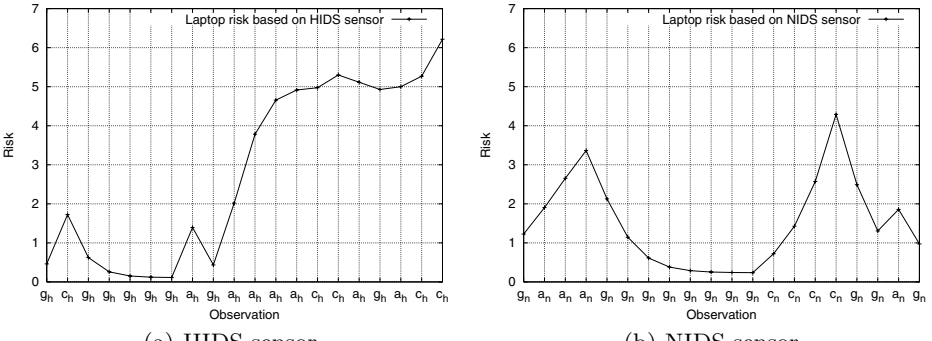
First, we assess the risk of the laptop, based on an observation sequence Y_{HIDS-L}, containing 20 samples collected from the laptop HIDS. We use the HMM $\lambda_L = \{\mathbf{P}_L, \mathbf{Q}_{HIDS-L}, \pi_L\}$, where

$$\mathbf{P}_L = \begin{pmatrix} p_{GG}\ p_{GA}\ p_{GC} \\ p_{AG}\ p_{AA}\ p_{AC} \\ p_{CG}\ p_{CA}\ p_{CC} \end{pmatrix} = \begin{pmatrix} 0.995\ 0.004\ 0.001 \\ 0.060\ 0.900\ 0.040 \\ 0.008\ 0.002\ 0.990 \end{pmatrix}, \tag{2}$$

$$\mathbf{Q}_{HIDS-L} = \begin{pmatrix} q_G(g)\ q_G(a)\ q_G(c) \\ q_A(g)\ q_A(a)\ q_A(c) \\ q_C(g)\ q_C(a)\ q_C(c) \end{pmatrix} = \begin{pmatrix} 0.70\ 0.15\ 0.15 \\ 0.15\ 0.70\ 0.15 \\ 0.20\ 0.20\ 0.60 \end{pmatrix}, \tag{3}$$

$$\pi_L = (\pi_G, \pi_A, \pi_C) = (0.8, 0.1, 0.1). \tag{4}$$

Since the HIDS is assumed to have low false-positive and false-negative rates, both $q_G(a), q_G(c), q_A(c) \ll 1$ and $q_A(g), q_C(g), q_C(a) \ll 1$ in $\mathbf{Q}_{HIDS-L}$. The dynamic risk in Figure 1(a) is computed based on the observation sequence Y (as shown on the x-axis of the figure) and a security state cost estimate measured as $\mathcal{C}_L = (0, 5, 10)$.



(a) HIDS sensor (b) NIDS sensor

Fig. 1. Laptop risk assessment

3.2 Example 2: Laptop Risk Assessment by NIDS Observations

Now, we let the risk assessment process of the laptop be based on another observation sequence, Y_{NIDS-L}, collected from the NIDS. A new observation symbol probability distribution is created for the NIDS

$$\mathbf{Q}_{NIDS-L} = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}. \tag{5}$$

One can see that the NIDS has higher false-positive and false-negative rates, compared to the HIDS. Figure 1(b) shows the laptop risk when using the HMM $\lambda_L = \{\mathbf{P}_L, \mathbf{Q}_{NIDS-L}, \pi_L\}$. Note that the observation sequence is not identical to the one in example 1, as the two sensors are not necessarily consistent.

3.3 Example 3: Aggregating HIDS and NIDS Observations

The agent now aggregates the observations from the HIDS and NIDS sensors by sampling from the observation sequences Y_{HIDS-L} and Y_{NIDS-L} in a round-robin fashion. To update the current state probability γ_t, the agent therefore chooses the observation symbol probability distribution corresponding to the sampled sensor, i.e the HMM will be

$$\lambda_L = \{\mathbf{P}_L, \mathbf{Q}^*, \pi_L\}, \text{where } \mathbf{Q}^* = \begin{cases} \mathbf{Q}_{HIDS-L} \text{ if } y_t \in Y_{HIDS} \\ \mathbf{Q}_{NIDS-L} \text{ if } y_t \in Y_{NIDS} \end{cases}. \tag{6}$$

The calculated risk is illustrated in Figure 2. The graph shows that some properties of the individual observation sequences are retained.

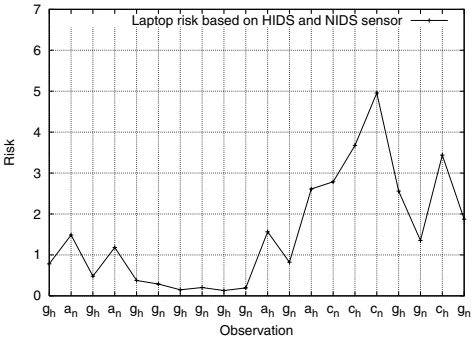


Fig. 2. Laptop risk assessment based on two sensors (HIDS and NIDS)

4 Managing Risk with Automated Response

In order to achieve effective incident response, it must be possible to effectively initiate defensive measures, for example by reconfiguring the security services and mechanisms in order to mitigate risk. Such measures may be manual or automatic. An information system or network can be automatically reconfigured in order to reduce an identified risk, or the system can act as a support system for system and network administrators by providing relevant information and

recommending specific actions. To facilitate such an approach, it is necessary to provide a mechanism that relates a detected security incidence to an appropriate response, based on the underlying risk model. Such a mechanism should include a policy for what reactions should be taken in the case of a particular incident, as well as information on who has the authority to initiate or authorize the response. Examples of distributed intrusion detection and response systems have been published in [13, 14].

The dynamic risk-assessment method described in this paper can provide a basis for automated response. If the risk reaches a certain level, an agent may initiate an automated response in order to control the risk level. Such a response may be performed both for individual objects (e.g. a compromised host) or on a network-wide level (if the network risk level is to high). Examples of a local response may be firewall reconfigurations for a host, changing logging granularity, or shutting down a system. Examples of a global response may be the revocation of a user certificate, the reconfiguration of central access control configurations, or firewall reconfigurations. Other examples include traffic rerouting or manipulation, and honeypot technologies. Note that such adaptive measures has to be supervised by human intelligence, as they necessarily introduce a risk in their own right. A firewall reconfiguration mechanism can, for example, be exploited as part of a denial-of-service attack.

5 Conclusion

We present a real-time risk-assessment method using HMM. The method provides a mechanism for aggregating data from multiple sensors, with different weightings according to sensor trustworthiness. The proposed discrete-time model relies on periodic messages from sensors, which implies the use of sampling of alert data. For the purpose of real-life applications, we propose further development using continuous-time models in order to be able to handle highly variable alert rates from multiple sensors. We also give an indication as to how this work can be extended into a multiagent system with automated response, where agents are responsible for assessing and responding to the risk for a number of objects.

References

1. Standards Australia and Standards New Zealand: AS/NZS 4360: 2004 risk management (2004)
2. Gehani, A., Kedem, G.: Rheostat: Real-time risk management. In: Recent Advances in Intrusion Detection: 7th International Symposium, RAID 2004, Sophia Antipolis, France, September 15-17, 2004. Proceedings, Springer (2004) 296–314
3. Staniford-Chen, S., Cheung, S., Crawford, R., Dilger, M., Frank, J., Hoagland, J., Levitt, K., Wee, C., Yip, R., Zerkle, D.: GrIDS – A graph-based intrusion detection system for large networks. In: Proceedings of the 19th National Information Systems Security Conference. (1996)

4. Snapp, S.R., Brentano, J., Dias, G.V., Goan, T.L., Heberlein, L.T., lin Ho, C., Levitt, K.N., Mukherjee, B., Smaha, S.E., Grance, T., Teal, D.M., Mansur, D.: DIDS (distributed intrusion detection system) - motivation, architecture, and an early prototype. In: Proceedings of the 14th National Computer Security Conference, Washington, DC (1991) 167–176
5. Balasubramaniyan, J.S., Garcia-Fernandez, J.O., Isacoff, D., Spafford, E., Zamboni, D.: An architecture for intrusion detection using autonomous agents. In: Proceedings of the 14th Annual Computer Security Applications Conference, IEEE Computer Society (1998) 13
6. Helmer, G., Wong, J.S.K., Honavar, V.G., Miller, L., Wang, Y.: Lightweight agents for intrusion detection. J. Syst. Softw. **67** (2003) 109–122
7. Debar, H., Curry, D., Feinstein, B.: Intrusion detection message exchange format (IDMEF) – Internet-Draft (2005)
8. Ourston, D., Matzner, S., Stump, W., Hopkins, B.: Applications of hidden markov models to detecting multi-stage network attacks. In: Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS). (2003)
9. Warrender, C., Forrest, S., Pearlmutter, B.: Detecting intrusions using system calls: Alternative data models. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy. (1999)
10. Gong, F., Goseva-Popstojanova, K., Wang, F., Wang, R., Vaidyanathan, K., Trivedi, K., Muthusamy, B.: Characterizing intrusion tolerant systems using a state transition model. In: DARPA Information Survivability Conference and Exposition (DISCEX II). Volume 2. (2001)
11. Singh, S., Cukier, M., Sanders, W.: Probabilistic validation of an intrusion-tolerant replication system. In de Bakker, J.W., de Roever, W.-P., Rozenberg, G., eds.: International Conference on Dependable Systems and Networks (DSN'03). (2003)
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Readings in speech recognition (1990) 267–296
13. Carver Jr., C.A., Hill, J.M., Surdu, J.R., Pooch, U.W.: A methodology for using intelligent agents to provide automated intrusion response. In: Proceedings of the IEEE Workshop on Information Assurance and Security. (2000)
14. Porras, P.A., Neumann, P.G.: EMERALD: Event monitoring enabling responses to anomalous live disturbances. In: Proc. 20th NIST-NCSC National Information Systems Security Conference. (1997) 353–365

Appendix: On Algorithm 1

Given the first observation y_1 and the hidden Markov model λ, the initial state distribution $\gamma_1(i)$ can be calculated as

$$\gamma_1(i) = P(x_1 = s_i|y_1, \lambda) = \frac{P(y_1, x_1 = s_i|\lambda)}{P(y_1|\lambda)} = \frac{P(y_1|x_1 = s_i, \lambda)P(x_1 = s_i|\lambda)}{P(y_1|\lambda)}. \quad (7)$$

To find the denominator, one can condition on the first visited state and sum over all possible states

$$P(y_1|\lambda) = \sum_{j=1}^{N} P(y_1|x_1 = s_j, \lambda)P(x_1 = s_j|\lambda) = \sum_{j=1}^{N} q_j(y_1)\pi_j. \quad (8)$$

Hence, by combining (7) and (8)

$$\gamma_1(i) = \frac{q_i(y_1)\pi_i}{\sum_{j=1}^{N} q_j(y_1)\pi_j}, \tag{9}$$

where $q_j(y_1)$ is the probability of observing symbol y_1 in state s_j, and π is the initial state probability. To simplify the calculation of the state distribution after t observations we use the *forward-variable* $\alpha_t(i) = P(y_1 y_2 \cdots y_t, x_t = s_i|\lambda)$, as defined in [12]. By using recursion, this variable can be calculated in an efficient way as

$$\alpha_t(i) = q_i(y_t) \sum_{j=1}^{N} \alpha_{t-1}(j) p_{ji}, \qquad t > 1. \tag{10}$$

From (7) and (9) we find the initial forward variable

$$\alpha_1(i) = q_i(y_1)\pi_i, \qquad t = 1. \tag{11}$$

In the derivation of $\alpha_t(i)$ we assumed that y_t only depend on x_t and that the Markov property holds.

Now we can use the forward variable $\alpha_t(i)$ to update the state probability distribution by new observations. This is done by

$$\begin{aligned} \gamma_t(i) &= P(x_t = s_i|y_1 y_2 \cdots y_t, \lambda) = \frac{P(y_1 y_2 \cdots y_t, x_t = s_i|\lambda)}{P(y_1 y_2 \cdots y_t|\lambda)} \\ &= \frac{P(y_1 y_2 \cdots y_t, x_t = s_i|\lambda)}{\sum_{j=1}^{N} P(y_1 y_2 \cdots y_t, x_t = s_j|\lambda)} = \frac{\alpha_t(i)}{\sum_{j=1}^{N} \alpha_t(j)}. \end{aligned} \tag{12}$$

Note that (12) is similar to Eq. 27 in [12], with the exception that we do not account for observations that occur after t, as our main interest is to calculate the object's state distribution after a number of observations.

Design and Implementation of a Parallel Crypto Server

Xiaofeng Rong[1,3], Xiaojuan Gao[2], Ruidan Su[3], and Lihua Zhou[3]

[1] School of Computer Science and Engineering, Xi'an Institute of Technology, Xi'an 710032, China
rongxf@mti.xidian.edu.cn
[2] School of Computer, Xi'an University of Engineer Science & Technology, Xi'an 710048, China
gaoxiaojuan@126.com
[3] Key Laboratory of Computer Network and Information Security of the Ministry of Education, Xidian University, Xi'an 710071, China
{surd, zhoulh}@mti.xidian.edu.cn

Abstract. As demands for secure communication bandwidth grow, efficient processing of cryptographic server at the host has become a constraint that prevents the achievement of acceptable secure services at large e-commerce and e-governments. To overcome this limitation, this paper proposes an innovative design in cryptographic server architecture, which based on the hardware of high performance and programmable secure crypto module. The architecture provides a well scalability framework by using a general device API, as well as obtains high performance by carrying cryptography computations in parallel between crypto chips in crypto modules. The system is implemented on an IBM Services345 and hardware of crypto modules. Preliminary measurements are also performed to study the trade-off between numbers of crypto modules parallel computing and performance of generate 1024-bit RSA digital signature. Results indicate that the system implemented by the architecture with high performance and scalability.

1 Introduction

Cryptographic module is the set of hardware, software, and/or firmware that implements approved security functions (including cryptographic algorithms and key generation) and is contained within the cryptographic boundary. It always integrates kinds of crypto-algorithms chips, which implements secure cryptographic functions within a secure computing environment. Functions of cryptographic modules and system resource management are combined with an appropriate key management mechanism within cryptographic server to provide a perfect set of cryptographic service. The growth of the Internet as a vehicle for secure communication and electronic commerce has brought cryptographic processing performance (including system throughput, signature speed, and supporting multi cryptographic algorithms) to the forefront of cryptographic server design. This trend will be further underscored with the widespread adoption of secure protocols such as IPSec and virtual private networks (VPNs). In this paper, we proposed an innovative architecture of high performance cryptographic server. The architecture supports parallel crypto processing among

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 398 – 406, 2005.
© Springer-Verlag Berlin Heidelberg 2005

multi crypto modules with multi crypto algorithms. By combining crypto modules and crypto chips into parallel configurations within crypto server, the overall performance of server is improved remarkably.

We detail the design and implementation of the crypto server and analyze its parallel performance. In Section 2, we present the motivations and background of research. The goals of design is given and discussed in Section 3. The architecture of parallel crypto server is described in Section 4. In Section 5, we present the design of software architecture in crypto module layer, describe how the architecture can be used for parallel crypto processing. Experimental environment and evaluation of performance overheads are presented in Section 6. In the final sections, we conclude the paper, and make suggestions for future work.

2 Research Background

Traditional cryptographic system has been implemented by security toolkits. The capabilities of toolkits are wrapped up in its own set of crypto functions, which encapsulated by crypto application program interface (API). The programming specification of API is complex and with poor security. In addition to these disadvantages, its crypto processing depends throughout on the computing resources of server's. The server will cannot handle other tasks because the limited resources of on-board CPU be tied up by complex crypto algorithm calculate. The advantages of such software-only method for cryptosystem are easier and cheaper to be implemented. Nevertheless, its only appropriate to small applications or personal user.

It is even more important today as customers' aggregate applications onto a smaller number of higher-performance servers. Each server must have the ability to process more operations and transactions per second than earlier generations. Thus rises the need to implement the cryptographic functions in a hardware crypto module with crypto-algorithm chips, which can execute these functions independently while the server CPU is free to perform other functions.

Meanwhile, researchers and engineers addressed on the design of crypto server architecture, presented some specifications and criterions about crypto functions, algorithm, API etc. Such as CDSA, PKCS#11, CCA, GCS-API and GSS-API (RFC 2078), these are all criteria adopted by different organizations or influential companies. The Common Data Security Architecture (CDSA) specification provides a layered security services infrastructure which allows system administrators to "add-in" security modules. The Common Cryptographic Architecture (CCA) is an architected set of cryptographic functions and APIs that provide both general-purpose functions and a broad set of functions designed specifically to secure financial transactions. RSA's Public Key Cryptography Standard (PKCS) #11, also known as Cryptoki, is an API to cryptography devices, and defines a single interface which applications and crypto modules must conform to. The CDSA defines application interfaces to several different security services, whereas Cryptoki was designed as an interface to (solely) cryptographic functionality. Thus, CDSA necessarily includes more auxiliary services to manage its more complex architecture. The CDSA provides a general architecture design at a rather abstract level and defined mostly in terms of the API used to access it.

Nowadays, the performance of generate 1024-bit RSA digital signature by a mature modular math engine is 66kbps in China, while the IBM-developed crypto-chip (Otello chip) in the PCIXCC crypto module is of 3300kbps. The approach of multi-chips and multi-modules to parallel computing is a good solution to provide high performance with low-end crypto chips. Based on the idea, this paper presented the parallel architecture of cryptographic co-server which built on the hardware of programmable secure crypto modules. By using crypto algorithm-agents to abstract computation resource of crypto chips in modules, the crypto-manager can schedule system resources to support parallel data processing among crypto chips in crypto modules.

3 Design Goals of Crypto Server Architecture

The design goals for the architecture are to build a secure, scalable and parallel computing framework. These design principles are summarized below:

- Layered design. The architecture represents a true multilayer design, with each layer of functionality built on its predecessor. The purpose of each layer is to provide certain services to the layer above it, shielding that layer from the details of how the service is. An interface which allows data and control information to pass across in a controlled manner. In this way each layer provides a set of well-defined and understood functions which both minimize the amount of information which flows from one layer to another, and makes it easy to replace the implementation into secure hardware), because all that a new layer implementation requires is that it offer the same service interface as the one it replaces.
- Crypto module independence. Each crypto module is responsible for managing its own resource requirements such as memory allocation. The interface to the management-layer above it is handled in a module-independent manner. As a security kernel, internal software is used to orchestrate the cryptographic operations and to handle parameter checking, data formatting, and part of the access control within crypto module.
- Parallelism design. To meet high-performance cryptographic service, the architecture needs to provide the ability of supporting parallel computing among tasks and crypto-chips. It is important to improve overall system performance on workloads with inter-session and inter-chip parallelism.
- Full isolation of architecture internal from external code. The implementation of management-layer resides in its own address space without the application being aware of this. The crypto module internals are fully decoupled from access by external code, so that it very clearly defines the boundaries of the trusted computing.
- Scalability design. The architecture must be highly scalability so that new crypto module can be added when higher performance be required. The system should also be flexible and programmable so that new crypto algorithm functions can be integrated in the crypto server system.

In addition to the goals mentioned above, the architecture should have built-in error checking for both software and hardware. For example, it should to provide function checking at stated periods, to check the hardware (crypto chip) algorithm results. It means that there must archive the standard input and output parameters of every crypto algorithm which is used in the system.

4 The Architecture of Parallel Crypto Server

In pursuing our goals, the parallel crypto server is designed logically in 4-layer structure, which consists of crypto-module, communication, crypto device scheduler and service provider layers. The authors focus on the high-level description of the architecture and how the environments are used by applications in this section. Fig.1 details the high-level architecture, and shows the main components and their interrelationships.

Crypto-module layer is the origins of cryptographic functions of the whole architecture. It provides implementations of various crypto algorithms and crypto protocols by the means of hardware and software. It also provides tamper-resistant storage to protect sensitive data from disclosure. The cryptographic and computing resources are abstracted by device application programming interface (crypto device API), in order to be managed by scheduler layer.

Crypto-device communication interface layer is the data channel of the distributed system. Meanwhile, it is also the uniform interface for different crypto modules when they are integrated in the system. This layer implements the secure call between modules and manager layer in the case of distributed environment.

Crypto scheduler and manager is the core of the layered structure. The scheduler is responsible for the management of crypto server's system resource. It consists of the following function components:

- Security kernel. Based on the operating system (Linux), the kernel provides security functions of access control, storage management and key management.
- Protocol interpreter. The component is the pre-processing function unit of manager. The main process include: checking format of secure protocol packets and deriving payload from packets.
- Resource manager. The manager abstracts the algorithm function units of software or hardware from low level crypto modules, and describes the resource information of each unit by special data structure. The resources will be registered in an info-table of the manager.
- Task scheduling manager. For required service, through the info-table, the scheduling manager looks for the appropriate crypto modules which algorithm unit resource is unoccupied. On the basis of schedule policy, the manager calls the selected crypto-module to process the task and return the results.

The Crypto service provider layer provides the unified cryptographic service interface to the high layer secure applications. The three main types of services are: confidential service, trust service and certificate service. This layer makes the special service for users from crypto server (e.g. signature and verify service), not isolated crypto function. This means the implementation detail of crypto services are transparent to consumer.

The security kernel of parallel crypto server system is implemented in two parts. For system level, it is implemented in the crypto scheduler and manager. For module level, it is implemented in each crypto-module independently. Any access to the system resources will be monitored by the two-level security kernel. Crypto function unit can be implemented flexibly by several ways. Such as *USB-eKey*, *smartcard*, *crypto-module* and even crypto function library etc.

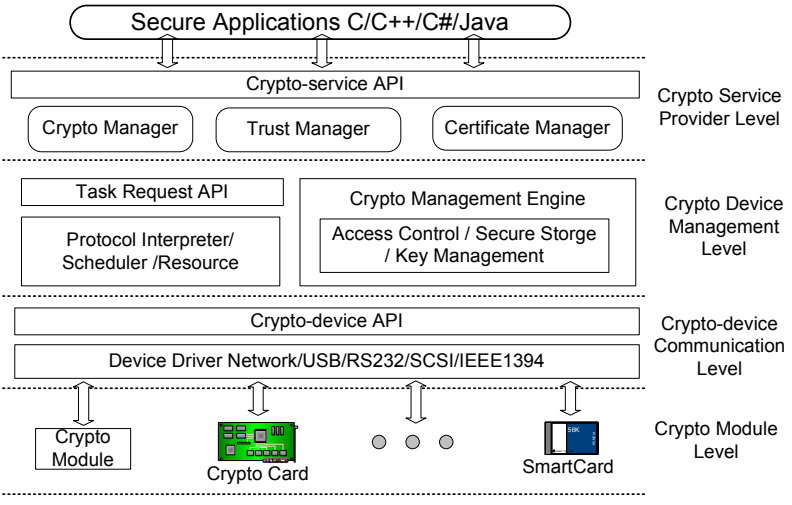


Fig. 1. Security architecture of cryptographic service system

5 Crypto-Module Software Architecture Design

The software architecture of crypto-module internal consists of three layers, which are secure bootstrap software, operating system and crypto-module manager software, crypto-algorithm agent software layers.

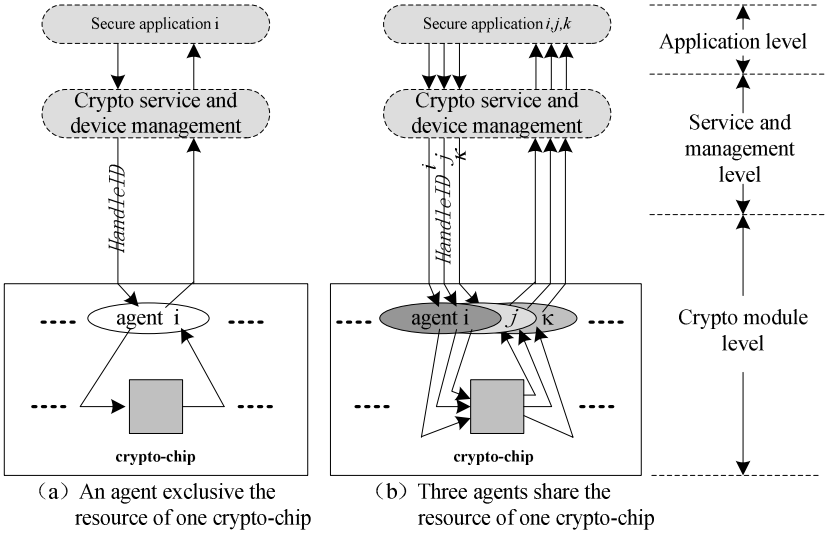
Secure bootstrap software resides in the ROM, and runs at boot-time firstly. The bootstrap boots and controls devices configuration, which ensure the integrity and security of software components loaded.

Crypto-module manager software runs on Linux as forms of background process program. It consists of crypto-function manager and key management controller. The controller provides secure procedures for handling the generation, storage, distribution, deletion, archiving and application of keys in accordance with security policy. In addition, when startup, it collects crypto modules device information and registers the data into local resource registration table. These information will be transmitted to the resource-manager of crypto scheduler and manager during system startup. Depends on different crypto-functions required externals, the crypto-function manager calls corresponding low-level code (such as, multiplying of large integers, DER coding, etc.) and implements the integrated crypto-algorithm function.

According to the information of local resource registration table, crypto-algorithm agents are established in the forms of crypto-service threads. Before it works, agent registers itself at the resource manager in the manager-machine-side. A unique *HandleID* is assigned to the agent to identify the session connection (socket connection) between them. Manager-machine maintains the *HandleID* and the information of corresponding resource into *HandleID* Table.

Fig. 2 shows two different approaches for resources representations. Fig.2(a) is the exclusive way, which means one algorithm-agent (such as, agent i) represents the computation resource of one crypto chip. By this way, only through agent i, can the

application to call the crypto-chip. Fig.2(b) is the share way, computation resource of crypto-chip can be accessed by applications through any of the agents (*i, j, k*). That means agents (*i, j, k*) only represents the crypto-processing capability of this crypto-chip, but not the chip itself.



（a）An agent exclusive the resource of one crypto-chip

（b）Three agents share the resource of one crypto-chip

Fig. 2. Through algorithm agents access computation resources of crypto module

For the exclusive representation, if crypto-computing resource is required by an application through crypto-service API, the chip is unoccupied during the transmission of request data and result. It is waiting for the next task request. By using multiple agents, the processing of main CPU and crypto-chips are overlapping. Therefore, the share representation for resource can make the most of the capability of crypto-chips. It can improve the efficiency and performance of crypto server.

6 Performance Measurements and Analysis

Speed of digital signature and throughput of symmetric encryption are two key performance issues of cryptographic server system. In this paper, the authors measured the performance of generating 1024-bit RSA digital signature on a real parallel crypto-server. The implementation and measurement environment of crypto-server is shown in Fig.3.

Multiple threads for requesting signature service running on the top of the Client-side's operating system and crypto-service API software. Based on the TCP/IP protocol and socket API, application threads access crypto-service. The Manager host-side implements the functions of Crypto Service Provider and Crypto Device Management Levels in layered architecture which is illustrated in Fig.1.

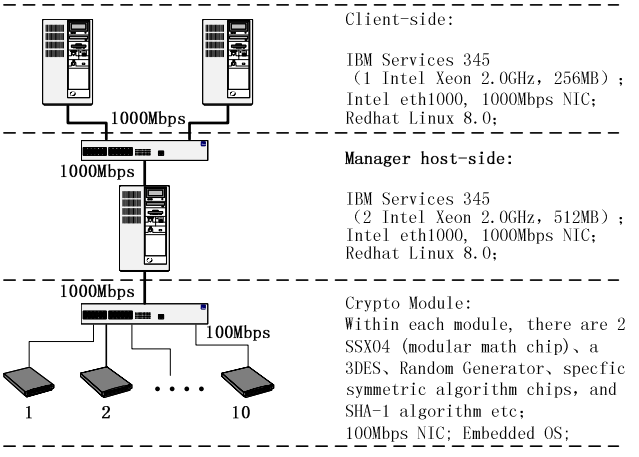


Fig. 3. Implementation and measurement environment of cryptographic service system

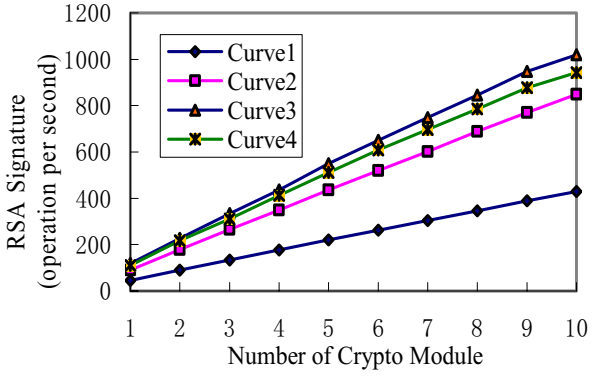


Fig. 4. Results of 1024-bit RSA digital signature performance

The Crypto Module mainly implements the functions of Crypto Device Communication Interface and Crypto Module levels in layered architecture which is illustrated in Fig.1. In each crypto module, there are multiple agent threads for every crypto chip or crypto function. The information of agents and their corresponding chips are collected by an on-module manager routine. When agent threads are ready to work, the on-module manager routine sends the information to host-side. The software of Manager host-side maintains the information in an *HandlID* Table, and creates corresponding threads to establish TCP/IP connections with the agent on-module. Meanwhile, the Manager host-side software maintains every socket handle into the *HandleID* Table. By establishing the *HandleID* Table, each algorithm agent which running on crypto module has a Manager host-side partner thread that initiates task requests, to which the module-side agent responds. Consequently, an on-module crypto-computing resource becomes visible to the Manager host-side crypto-service requesters.

RSA algorithm computing function is implemented by a modular math crypto-chip, which is SSX04. The SSX04 can generate 65 digital signatures per second (1024-bit RSA) within crypto module. We use different ways of abstraction resource (exclusive and share), and establish different amounts of agent (1/2/3 agents) for one SSX04-chip. The measurement results are shown in Fig.4.

Curve1. Exclusive way. In this situation, each module is only required by one signature task. Curve1 are the performance results of crypto server when there are 1, 2, … , and 10 modules participate in generating of signature, respectively. It is inter-module parallel processing.

Curve2. Exclusive way. On Client-side, the number of applications required signature service are 2, 4, … , and 20 , respectively. In this situation, each module is required by two signature tasks. Both of SSX04 in crypto module process signature tasks. It is a kind of inter-chip parallel processing.

Curve3. Share way (established 2 agents for each chip). On Client-side, the number of applications required signature services are 4, 8, … , and 40, respectively. Each module is required by four signature tasks. Each SSX04 in crypto module is represented by two agents in share way. It is a kind of inter-session parallel processing.

Curve4. Share way (established 3 agents for each chip). On Client-side, the number of applications required signature service are 6, 12, … , and 60, respectively. In this situation, each module is required by six signature tasks. Each SSX04 in crypto module is represented by three agents in share way. It is a kind of inter-session parallel processing.

The speed of generate signature in exclusive-way of abstracting resources is lower than that of in share-way of abstracting resources. Along with the increase of number of crypto modules in parallel computing, according to the results, every of the four situations in the measurement shows that the performance is improved linearly. Multiple crypto modules can be used in the architecture of parallel crypto server, and performance scales approximately linearly with the number of modules.

7 Conclusions

Cryptography is the crucial techniques of today's information security. Nevertheless, as the requirements of crypto algorithm become increasingly complex and the never-ending demands for increased speed, efficient processing and scalability architecture of crypto server at host-side become a constraint that prevents the achievement of acceptable secure services at large e-commerce and e-governments. In this paper, the authors proposed an innovative design in parallel cryptographic server, which based on the hardware of high performance and programmable secure crypto module. The architecture provides a well scalability framework by using a general device API, as well as obtains high performance by carrying cryptography computations in parallel between crypto chips in crypto modules.

References

1. Smith, S.W., Weingart, S.H.: Building a High Performance, Programmable Secure Coprocessors . Computer Networks (Special Issue on Computer Network Security), Vol. 31. (1999) 831-860.
2. Amold, T.W., Van Doorn, L.P.: The PCIXCC:A new cryptographic coprocessor for the IBM eServer. IBM Journal of Research and Development, Vol. 31.(2004) 475-487.
3. PKCS #11 v2.11. Cryptographic Token Interface Standard. RSA Laboratories.(2001).
4. Peter, G.: The Design of a Cryptographic Security Architecture. Proceedings of the 8th Usenix Security Symposium. USENIX, Washington, D.C (1999) 153-169.
5. John, L.: Generic Security Service Application Programming Interface. RFC 2078.(1997)
6. Johnson, D.B., Dolan, G.M.: Transaction security system extensions to the common cryptographic architecture. IBM Journal of System. Vol. 30.(1991) 230-243.
7. Wu, L.,Weaver,T.A.: CryptoManiac:A Fast Flexible Architecture for Secure Communication. Proceedings of the 28th International Symposium on Computer Architecture (ISCA 2001). IEEE CS Press. Sweden (2001) 110-119.
8. Kuo, H., Verbauwhede, I.M.: Architectural Optimization for a 1.82 Gb/s VLSI Implementation of the AES Rijndael Algorithm. In: Koc, Cetin K., Nacchae, D., Paar, Christof. (eds.): Cryptographic Hardware and Embedded Systems - CHES 2001. Lecture Notes in Computer Science, Vol. 2162. Springer-Verlag, Paris (2001) 51-64.
9. Lennon, R.E.: Cryptography architecture for information security. IBM Journal of System. Vol. 17.(1978) 138-150.

Survivability Computation of Networked Information Systems

Xuegang Lin[1], Rongsheng Xu[2], and Miaoliang Zhu[1]

[1] Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027, China
lxg@nsg.ihep.ac.cn, zhum@zju.edu.cn
[2] Institute of High Energy Physics, CAS, Beijing 100049, China
xurs@nsg.ihep.ac.cn

Abstract. Survivability should be considered beyond security for networked information systems, which emphasizes the ability of continuing providing services timely in malicious environment. As an open complex system, networked information system is presented by a service-oriented hierarchical structure, and fault scenarios are thought of as stimulations that environment impact on systems while changes of system services' performance as system reaction. Survivability test is done according to a test scheme which is composed of events, atomic missions, fault scenarios and services. Survivability is computed layer by layer based on the result of survivability test through three specifications (resistance, recognition and recovery). This computation method is more practicable than traditional state-based method for it converts direct defining system states and computing state transition into a layered computation process.

1 Introduction

As computers and networks are widely applied in society, more infrastructure is based on networked information systems. Though security is regarded as an important attribute for these systems, there are not perfect systems that can prevent any attack and achieve full security. Then, we should consider how these systems can survive when they are intruded. Survivability is system's ability of providing essential services timely despite attacks, failures and other accidents, and it is beyond the concept of security for system users concern more on the entire system's services that they have to use than individual component.

To establish more survival system, it's necessary to measure its survivability to estimate the performance of system's services under malicious environment. In the field of survivability, telecommunication network is studied mostly, and its survivability is mostly measured by network nodes and the links based on graph model[1]. For networked information system, Guo[2] presented a systematic method for survivability analysis based on relations between service and configuration, and gave theoretic quantitative description which was not applied in practice. Bao[3] used cut-set and Monte-Carl method to analyze network security management systems quantitatively. Jha[4] described each node in networked system using finite state machine, and utilized model checking, Bayesian

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 407–414, 2005.
© Springer-Verlag Berlin Heidelberg 2005

analysis, probabilistic system and other mathematic methods for qualitative and quantitative analysis. Krings[5] presented a five-step model to transform the survivability analysis into a parameterized graph. McDermott[6] used static and probability of system state transmission to compute survivability which is base on a fault activity model. SNA[7] method provided a framework to analyze information systems' survivability and have been applied in engineering. However, the above works mostly focus on qualitative analysis or theoretical computation.

Survivability computation is necessary besides qualitative analysis for it can give more accurate evaluation of system survivability and be used to define survivability grades. In our early work, we provided survivability analysis model by dividing it into three subproblem: environment, system and analysis model[8]. In this paper, based on analysis model[8], we present a general survivability computation method for networked information system, which is done by system structure definition, survivability test and computation. This method is more practicable for the survivability test can be done objectively and survivability can be computed layer by layer based on the survivability test.

2 System Model

For the complexity and scale of networked system, it's difficult and even impracticable to compute its survivability based on its physical structure. Survivability focuses whether services of the whole system can be survive under malicious environment but not individual components, so we needn't analyze its components all detail, then a model should be given to describe the system.

According to user requirement and system functions, essential service $\{ES_i\}$ can be defined, which is the service that system must provide under being attacked and is the ultimate object of system survivability analysis. We can think that essential service is delivered from the server to the user through a pipeline, which is service stream, then system components which are traversed by the pipeline can be identified. So, all these components involved in a service stream ES_i compose a component set $\{EC_{ij}\}$, that is, essential service ES_i can be achieved by services that components $\{EC_{ij}\}$ provide.

For essential service ES_i, it may suffer from many events such as attacks and others, and some events can be correlated and organized into a set that have the same intention, then this event set is called fault scenario. There are many fault scenarios influencing essential service ES_i with different weight, but we can select several typical scenarios according to the situation that system faces, and these fault scenarios compose a set $\{FS_{ij}\}$. Like the service pipeline above, fault stream FS_{ij} will also pass through some system components which compose a set $\{EC'_{ij}\}$. Generally, $\{EC'_{ij}\}$ is a subset of $\{EC_{ij}\}$. Then, a list of components can be organized by the corresponding essential service or fault scenario.

According to the relations among ES, FS and EC, the entire networked information system can be described by a hierarchical structure, where the center is system essential services and system components are linked according to its supporting services. This hierarchical structure presents the logical relation of

the system inner structure, displays the usual hierarchical idea of modern system's architecture, and differs from hierarchical model[9] where components are corrected by specification.

Furthermore, essential service may be completed by several other services, e.g., web service may be composed of apache service, some business components and database service, and the service that can not be divided is called atomic service. Then, essential service and essential component can be defined by formalization based on the above structure. Each essential service can be given as the following format: $ES_i = \{dss_i, \nu s_i, AS, EC, FS\}$, where dss_i is the description of ES_i, νs_i is the weight presenting the service's importance to user, AS is the atomic services set $\{AS_{ij}\}$ supporting ES_i, EC is the involved essential component set $\{EC_{ij}\}$, and FS is the fault scenario set $\{FS_{ij}\}$ influencing ES_i. Atomic service can be defined like essential service, $AS_{ij} = \{dsa_{ij}, \nu a_{ij}\}$, where νa_{ij} presents its importance. For each essential component, it can be defined as $EC_{ij} = \{dsc_{ij}, ECI, ES, FS\}$, where dsc_{ij} is the description, $ECI = \{ECI_{ij}\}$ is optional and identifies the component's information such as its operating system which will be used in the survivability test, ES is the services set that it supports, FS is the involved fault scenario set. So, the relation between set $\{ES_i\}$ and set $\{EC_{ij}\}$, set $\{FS_{ij}\}$ and set $\{EC'_{ij}\}$ is a many-to-many relationship, and relation between $\{ES_i\}$ and $\{FS_{ij}\}$ is one-to-many.

3 Survivability Test

Generally, system's performance can be analyzed by system test which is done through inputting a stimulation and observing its reaction. Since networked information system can be thought of as an open complex system in systems science, the environment's impact such as attacks and hardware faults can be regarded as stimulation, and the change of system services' state as system reaction, then, survivability can be measured through survivability test.

3.1 Fault Scenarios Definition

While system is running, it will be influenced by many events, and essential services that it should provide to users will be degraded or even down. According to the definition of fault scenario, the impact of environment can be presented by fault scenarios. We define that atomic mission as a mission that can be achieved by only one event having an actual target, e.g. root privilege. Then, fault scenario can be divided into several stages ultimately, each of which can be completed by an atomic mission.

Based on usage scenarios of essential services and those involved components, typical fault scenarios for essential services ES_i can be defined as: $FS_{ij} = \{dsf_{ij}, ES_i, P_{ij}, AM, EC'\}$, where dsf_{ij} is its description, ES_i is the essential service that it influences, P_{ij} is the weight describing its severity to ES_i, AM is the atomic mission set $\{AM_{ijk}\}$ that FS_{ij} can be divided into, and EC' is the essential component set $\{EC'_{ij}\}$ that FS_{ij} passes through.

For set AM, each atomic mission can be defined as $AM_{ijk} = \{dsa_{ijk}, I_{ijk}, E\}$, where dsa_{ijk} is its description, I_{ijk} is the weight presenting its importance to FS_{ij}, E is the event set $\{E_q\}$ that can implement AM_{ijk} and defined in the following test scheme. Using attack tree to describe the relation of elements in these sets, the relation between elements in $\{AM_{ijk}\}$ of fault scenario is AND for only all of atomic missions' success leads to the success of fault scenario, and elements in $\{E_q\}$ is OR for any of these event can complete the atomic mission.

3.2 Test Scheme Creation

In fault scenario's definition, it is divided into several key stages, i.e. atomic missions, but how to implement these atomic mission, i.e. the set $\{E_q\}$, is unknown. According to the vulnerabilities of services and information of components $\{ECI_{ij}\}$, atomic mission and its event set $\{E_q\}$ can be defined which will constitute a test scheme for the following survivability test.

However, there are different choices to achieve the same atomic mission. So we can classify and grade them, and different combination of these events present different environment, i.e. stimulations. To grade events that system may suffer, we can define its attributes $a_i (\in [0, 1])$ (such as availability of tools, difficulty of launching, necessary resource, vulnerabilities utilized and so on.) with weights $\nu_i (\sum_i \nu_i = 1)$, so its danger value can be defined as $w = \sum_i a_i * \nu_i$. Then, a series of critical value $L_q = \{l_1, \ldots, l_Q\}$ is defined, which grades events into Q grades. We can choose N events in each grade to achieve AM_{ijk}, and these events compose an event set $\{E_{qn}\}$ of grade q, then $N * Q$ events in all Q grades compose event set $\{E_q\}$ of atomic mission AM_{ijk}. For events in grade q, a distributing weight d_q is defined to describe the probability of being used to achieve the atomic mission, so we can adjust d_q, change L_q to shift grade and select different events in each grade to present different choice to implement atomic mission, i.e. test environment.

On the other hand, these events can be classified based on a certain taxonomy and be input in an event database. Then, events in set $\{E_{qn}\}$ can be selected by searching this database according essential component information and event intention, which test scheme creating more easily and objectively.

Finally, essential services, fault scenarios, atomic missions and events present a structure like linked list, and constitute the test scheme for survivability test.

3.3 Test

There are two kind of survivability test: actual test and simulative test. Generally, the former is done on the actual system with actual events, which can gain more accurate result, but can not be applied on uncompleted systems and may do some damage to systems. For example, SNA[7] method can be used in the actual test, and a "Red Team" can be organized to test the system. While, the latter is completed on a software platform through simulating events on the system model with some expert experiences or some mathematic theories, which can be applied in most systems but may be less believable. Kim[10] developed

a simulation platform Modsim III to test system survivability based on vulnerabilities. Christie[11] used Easel to model and simulate system survivability.

According the test scheme created above, actual test or simulative test can be done service by service which is actually completed hierarchically (event-atomic mission-fault scenario-essential service). In our research group, this test is achieved actual test and the simulation test platform is under developing. Then, test result can be gotten and used on the following survivability measure.

4 Survivability Computation

According to Ellison's definition of survivability[12], survivability can be described by 3"R": resistance, recognition, and recovery. Resistance is a basic attribute for survivability, and presents system's ability of preventing attacks. Recognition presents the ability of monitoring of system's change, based on which system can be aware of something happened, and recovery reflects system's ability of restoration after being damaged. For most survivable system, these three abilities present the idea of survivability design and control: preventing, monitoring and reacting. Then, system survivability can be computed in the following three aspects based on the result of the above survivability test.

4.1 Resistance

Resistance describes system essential services' ability of preventing attacks, software or hardware faults and other events, and presents the entire system's security, but not individual component's. According to fault scenario definition, a weight P_{ij} is given to describe the severity of fault scenario FS_{ij}, and weight $\{I_{ijk}\}$ is defined to describe the weight of its atomic mission $\{AM_{ijk}\}$. In the survivability test, we select N events for each grade to complete each atomic mission. If danger value of event E_{qn} in grade q is w_{qn}, the danger value of entire event set $\{E_{qn}\}$ of grade q can be defined as $W_q = (\sum_n w_{qn})/N$.

Then, the percent of events in grade q that success in achieving atomic mission's target, Pr_Fault_q, can be gotten by the survivability test. So resistance of atomic mission AM_{ijk} is: $Resis_AM_{ijk} = \sum_q (W_q * (1 - Pr_Fault_q) * d_q)$.

According to the logical relation of atomic missions in fault scenario FS_{ij}, resistance of FS_{ij} is: $Resis_FS_{ij} = \sum_k (I_{ijk} * Resis_AM_{ijk})$. Resistance of essential service ES_i can be gotten by considering all fault scenarios' resistance: $Resis_ES_i = \sum_j (P_{ij} * Resis_FS_{ij})$

Based on the importance of each essential service, system's resistance ability can be gotten by integrating all essential services: $Resis = \sum_i (Resis_ES_i * \nu s_i)$.

4.2 Recognition

Recognition in the field of survivability is different from that in Intrusion Detection Systems for it emphasizes on detecting the change of entire system or performance of system service, but not the individual event. Then, recognition

can be divided into three layers based on the relation between events, fault scenarios and essential services, i.e. recognition of malicious events, fault scenario and degradation of essential service performance.

Recognition ratio for the q grade events of atomic mission AM_{ijk} is defined as Pr_Rg_q, which can be simply pointed as the percent of the events recognized in total N events of set $\{E_{qn}\}$ in the survivability test. Furthermore, how long this recognition is done after event happens is another factor. Recognition time of event E_{qn} can be regularized into [0,1] and identified as t_{qn}, then the entire recognition time of grade q can be calculated as $T_q = (\sum_n T_{qn})/N$.

Then, recognition of atomic mission AM_{ijk} is: $Recog_AM_{ijk} = \sum_q (W_q * Pr_Rg_q * T_q * d_q)$, where W_q is danger value and t_q is distributing weight. According to the logical relations between atomic missions and its weight in fault scenario, recognition of FS_{ij} is: $Recog_FS_{ij} = \sum_k (I_{ijk} * Recog_AM_{ijk})$.

Summarize all fault scenarios, recognition of essential service ES_i is:

$Recog_ES_i = \sum_j (P_{ij} * Recog_FS_{ij})$. Finally, the entire system's ability of recognition can be gotten through all essential services: $Recog = \sum_i (Recog_ES_i * \nu s_i)$.

4.3 Recovery

Survivability focuses the recovery of system essential services and can stand that some components or services don't work. For survivable system, though its essential services may be degraded under fault scenarios, this phenomenon may be temporary and system may provide service to user in another way for system's reconfiguration and adaption.

For different system service has various requirement of recovery time, requirement time of atomic service AS_{ik} can be defined as $ReqT_{ik}$. Under FS_{ij} in the survivability test, actual recovery time can be gotten as $TrueT_{jk}$, and recovery percent of AS_{ik} is Pr_Rv_{kj} which can be defined by the response rate to all user's request for some kind of services such as Web service. So, recovery of atomic service is determined by the recovery time, recovery degree and the corresponding fault scenario: $Recov_AS_{ik} = \sum_j (P_{ij} * (Pr_Rv_{kj} * ReqT_{ik}/TrueT_{kj}))$.

Summary all atomic services' recovery, the recovery of ES_i is: $Recov_ES_i = \sum_k (\nu a_{ik} * Recov_AS_{ik})$.

Then, summarize all essential services, and system ability of recovery is $Recov = \sum_i (\nu s_i * Recov_ES_i)$. Finally, the survivability value can be gotten by a vector $SUR = [Resis, Recog, Recov]$.

Based on the above computation process, parameters in computation can be gotten by survivability test or defined in advance, and system survivability can be computed layer by layer which is corresponding with the test scheme structure. Though there are some models for survivability analysis, such as Markov chain[4], finite state machine[4], graph model[5] and state transmission based model[6], the above process of survivability computation shows that "3R" need not be computed based on direct state definition and transition analysis, but on the result of events tested or atomic services. This method avoids defining system states and computing state transition in Markov chain or graph model which is difficult to be realized by computer.

5 Demonstration

To demonstrate the above computation process, a campus course registration system is given here to be analyzed in brief. Essential services of the system are: ES_1, students browse course information that they can register, and $\nu s_1 = 0.5$; ES_2, professors browse their courses information that they should teach, $\nu s_2 = 0.3$; ES_3, administrators manage the system, $\nu s_3 = 0.2$. For the most purpose is to show the computation process, the system model and survivability test is omitted here and only resistance and recovery is computed.

To create test scheme, events in database are graded into 5 grades by critical values $\{L_q\} = \{0.2, 0, 4, 0, 6, 0.8\}$, and distributing weights are $\{d_q\} = \{0.3, 0.3, 0.2, 0.1, 0.1\}$. We select 10 events in each grades to complete each atomic mission, then all selected events of all fault scenarios constitute the test scheme that is used in survivability test.

In the survivability test, events generated can be identified whether are successful or recognized through IDS alerts, system logs, system state and other information. After testing based on the test scheme, 3R can be computed layer by layer. For each grade event of AM_{111}, test result shows that success ratio of events in each grade is $\{Pr_Fault_q\} = \{0.4, 0.2, 0.5, 0.3, 0.4\}$, and danger value of each grade can be computed as $\{W_q\} = \{0.16, 0.37, 0.51, 0.67, 0.83\}$, so $Resis_AM_{111} = \sum_{q=1}^{5}(W_q * (1 - Pr_Fault_q) * d_q) = 0.2813$. Other atomic missions can be computed in the same way, then, fault scenarios' resistance can be gotten $Resis_FS_{11} = \sum_{k=1}^{4}(I_{11k} * Resis_AM_{11k}) = 0.391255$, so $Resis_ES_1 = \sum_{j=1}^{2}(P_{1j} * Resis_FS_{1j}) = 0.525101$ and $Resis_ES_2$, $Resis_ES_3$ can be gotten similarly. Finally, system's resistance ability can be computed as $Resis = \sum_{i=1}^{3}(Resis_ES_i * \nu s_i) = 0.560487$.

For ES_1, it can divided into 2 atomic services: AS_{11}, web server's http service provided by IIS with $\nu a_{11} = 0.4$; AS_{12}, course information provided by SQL Server database with $\nu a_{12} = 0.6$, and both of them should be recovered in 5 hours($ReqT_{11} = ReqT_{12} = 5$). In the test, AS_{11} is degraded only under FS_{11}, and AS_{12} is degraded only under FS_{12} and FS_{13}. Test result shows that AS_{11} is recovered 70%($Pr_R\nu_{11} = 0.7$) in 2 hours($TrueT_{11} = 2$) under FS_{11}. Under FS_{12}, AS_{12} is recovered 100%($Pr_R\nu_{12} = 1.0$) in 10 hours($TureT_{12} = 10$). Under FS_{13}, AS_{12} recovers 100%($Pr_R\nu_{13} = 1.0$)in 4 hours($TureT_{12} = 4$).

So, recovery of $\{AS_{ij}\}$ can be computed as $Recov_AS_{11} = \sum_{j=1}^{3}(P_{1j} * (Pr_Rv_{1j} * ReqT_{11}/TureT_{1j})) = 1.125000$, $Recov_AS_{12} = 0.77500$.

Then $Recov_ES_1 = \sum_{k=1}^{2}(Recov_AS_{1k} * \nu a_{1k}) = 0.915000$, and $Recov_ES_2 = 0.758000$, $Recov_ES_3 = 0.806000$. Finally, system's ability of recovery is $Recov = \sum_{i=1}^{3}(Recov_ES_i * \nu s_i) = 0.846100$.

6 Conclusions

Beyond security, survivability is necessary for networked information systems to assure availability of essential service for users. In this method of survivability computation, parameters in computation can be gotten practicably and easily,

and the layered process of computation avoids dividing state space in Markov chain or other model based on system states. In the survivability test, the event database which records most general computer and network events makes creating the test scheme automatic and objectively which can be used in other security application. However, the measure method should be more refined for actual application, and based on this method, a software platform is being developed for managing the computation process and reducing the manual work.

References

1. Louca, S., Pitsillides, A. and Samaras, G.: On network survivability algorithms based on trellis graph transformations. Proceeding of the 4th IEEE Symposium on Computers and Communications. Red Sea, Egypt. (1999) 235–243
2. Guo, Y.B., Ma, J.F.: Quantifying survivability of services in distrusted system. Journal of Tongji University. **30** (2002) 1190–1193.
3. Bao, X.G., Hu, M.Z., Zhang, H.L. and Zhang, S.R.: Two survivability quantitative analysis methods for network security management systems. Journal of China Institute of Communications. **25** (2004) 34–41
4. Jha, S., Wing, J., Linger, R. and etc: Survivability analysis of network specifications. International Conference on Dependable Systems and Networks. New York, USA. (2000)
5. Krings, A.W., Azadmanesh, M.H.: A graph based model for survivability analysis. Technical Report, UI-CS-TR-02-024. (2004)
6. McDermott, J.: Attack-potential-based survivability modeling for high-consequence systems. Proceeding of Third IEEE International Workshop on Information Assurance. College Park, Maryland. (2005) 23–24
7. Mead, N.R., Ellison, R.J., Longstall, T. and etc: Survivable network analysis method. Technical Report, CMU/SEI-2000-TR-013. (2000)
8. Lin, X.G., Xu, R.S. and Zhu, M.L.: Survivability analysis for information systems. Proceeding of the 7th International Conference on Advanced Communication Technology. Phoenix Park, Korea. **1** (2005) 255–260
9. Gao, Z.X., Ong, C.H. and Tan, W.K.: Survivability assessment: modelling dependencies in information systems. Proceedings of the Information Survivability Workshop. Vancouver, BC. (2002)
10. Kim, H.J.: System specification based network modeling for survivability testing simulation, Proceeding of the 5th International Conference on Information Security and Cryptology. Seoul, Korea. LNCS **2587** (2002) 90–106
11. Christie, A.M.: Network survivability analysis using easel. Technical Report, CMU/SEI-2002-TR-039. (2002)
12. Ellison, R., Fisher, D., Linger, R. and etc: Survivable network systems: an emerging discipline. Technical Report, CMU/SEI-97-153. (1997)

Assessment of Windows System Security Using Vulnerability Relationship Graph

Yongzheng Zhang, Binxing Fang, Yue Chi, and Xiaochun Yun

Research Center of Computer Network and Information Security Technology,
Harbin Institute of Technology, Harbin 150001, Heilongjiang, China
zyz@pact518.hit.edu.cn

Abstract. To evaluate the security situation of Windows systems for different users on different security attributes, this paper proposes a quantitative assessment method based on vulnerability relationship graph (VRG) and an index-based assessment policy. Through introducing the correlative influences of vulnerabilities, VRG can be used to scientifically detect high risk vulnerabilities which can evoke multistage attacks although their threats on surface are very little. Analysis of 1085 vulnerabilities indicates that for trusted remote visitors, the security of Windows systems is lower while for distrusted remote visitors, they are relatively secure. But there is no obvious difference of the security risk on confidentiality, authenticity and availability of Windows systems. In several known versions, the security of Windows NT is almost lowest.

1 Introduction

When enjoying the convenience brought by Internet, people also have to suffer from various intrusions and attacks from Internet which make any system linked into Internet placed at risk[1]. Why attacks like worm, virus, malicious code and intrusion can badly compromise computer systems is due to the existence of security vulnerabilities in design, implementation and maintenance of computer and software systems. Therefore, vulnerability assessment is greatly significant for ensuring the security of computer systems.

By the running status of softwares, vulnerability assessment can be divided into dynamic assessment and static assessment. Dynamic assessment is to identify the known vulnerabilities in running softwares and evaluate the security risk of vulnerabilities. This method reflects the risk of a system in running time. Static assessment is to analyze the security risk of software systems according to the information of known vulnerabilities in the databases. It doesn't need the process of vulnerability detection. The sense of this method is to compare the security of different software systems and reflect the inherent risk of flaws occurring in the process of system design or implementation. The related work contains: Jiwnani and Zelkowitz[2] present a vulnerability taxonomy to analyze the genesis, location and impact of Windows NT and Linux vulnerabilities in order to develop more secure operating systems. Hedbom et al[3] compares the security of Windows NT and Unix in security features and vulnerabilities, and draws a conclusion that there are no significant

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 415 – 420, 2005.
© Springer-Verlag Berlin Heidelberg 2005

differences in the real level of security between these systems. Wang[4] proposes a sum-based quantitative assessment approach of Windows NT, Linux and Solaris vulnerabilities.

But this paper only focuses on the prevalent operating system Windows and conducts further research on the static assessment approach of Windows. To scientifically reflect the correlative impact of vulnerabilities, this paper introduces a conception and generating method of vulnerability relationship graph (VRG). Then we propose a VRG-based quantitative assessment approach and an index-based assessment policy. Our goal is to analyze and assess the security risk caused by vulnerabilities in the process of design and implementation of several prevalent versions of Windows systems. And we expect comparing and reviewing their security situation for different users on different security attributes.

2 Vulnerability Relationship Graph

To describe the characteristics of vulnerabilities, we have proposed a vulnerability taxonomy in the previous work[5], including six quantitative attributes as follows:

Premise privilege-set is the necessary privilege-set which an attacker needs for attempting to exploit a vulnerability. It is abbreviated to 'Ppre'.

Consequence privilege-set is the privilege-set which an attacker can obtain after successfully exploits a vulnerability to make a direct privilege escalation. It is abbreviated to 'Pcon'.

Attack-complexity denotes the difficult degree and probability of an attacker's exploiting a vulnerability. It is abbreviated to 'E'.

Confidentiality(C), auThenticity(T), Availability(A) is respectively used to express the influence of vulnerabilities on confidentiality, authenticity and availability of software systems in the process of a privilege escalation. Each one of them is quantified by the weight increment of premise privilege-set and consequence privilege-set. It is explained that 'authenticity' is more accurate and comprehensive than 'integrity'[6], so we replace 'authenticity' for 'integrity' as one of three dimensional security attributes.

The quantitative range of the above attributes is 0~1. Due to the length of the paper, we can't give unnecessary details of the classifying attributes and related conceptions which are discussed in the literature [5].

Referring to the theory of attack tree[7], we propose a notion of vulnerability relationship graph which is a directed graph describing the correlations between vulnerabilities. According to the characteristics of vulnerability correlation, a VRG comprises two basic components named AND-Structure and OR-Structure as shown in Fig.1. AND-Structure denotes that the precondition of exploiting the vulnerability v is to successfully attack $v_i(i=1\sim n \wedge n \geq 2)$ in a multistage attack. While OR-Structure indicates that if only any one of $v_i(i=1\sim n \wedge n \geq 1)$ is exploited successfully, attackers can attempt to exploit next successive vulnerability v. We let VRG G=(V,E), where V is a set of vulnerabilities, E is a set of directed edges. Moreover, we classify G to some subgraphs G[OS] by operating system types. An expression like "$v_i \rightarrow v_j \rightarrow v_k$" is stated to be a correlative vulnerability chain headed by v_i and tailed by v_k. And an isolated vulnerability can be regarded as a case of a correlative vulnerability chain.

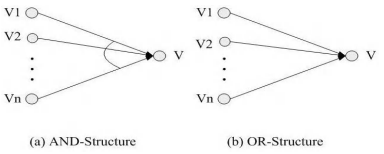


Fig. 1. Two basic components of VRG

Based on our vulnerability taxonomy stated before, we have proposed an automated mining method[8] which can generate different vulnerability relationship subgraphs with respect to different versions of Windows systems.

3 Quantitative Assessment Based on VRG

To estimate the security risk of operating systems, we give risk computation formulas of a vulnerability or a correlative chain:

$$SecThreat(X, V_h \rightarrow V_m \rightarrow V_t) = V_t.X.Pcon - V_h.X.Ppre \quad (X \in \{C, U, A\}) \tag{1}$$

$$SecRisk(X, V_h \rightarrow V_m \rightarrow V_t) = SecThreat(X, V_h \rightarrow V_m \rightarrow V_t) \times \prod_{i \in \{h,m,t\}} V_i.E \tag{2}$$

$$\begin{aligned} &VulSecRisk(X, V) = \\ &\quad Max(SecRisk(X, i) \mid i \in \{all\ correlative\ vulnerability\ chains\ headed\ by\ V\}) \end{aligned} \tag{3}$$

$$\begin{aligned} &ObjSecRisk(Obj, User) = (Y_1, Y_2, \cdots, Y_p), \\ &\quad Yi.X = Max^{(i)}(VulSecRisk(X, V) \mid V\ is\ related\ to\ Obj \wedge V.Ppre = User), \\ &\quad p \in [1, n], i \in \{1, \cdots, p\}, User \in \{CPphyaccess, CPaccess, CPuser, All\} \end{aligned} \tag{4}$$

Formula (1) and (2) are respectively security threat formula and security risk formula of the correlative vulnerability chain like "$V_h \rightarrow V_m \rightarrow V_t$" on attribute X. Where, $V_t.X.Pcon$ denotes the threat of consequence privilege-set of vulnerability V_t on attribute X. $V_h.X.Ppre$ indicates the threat of premise privilege-set of vulnerability V_h on attribute X. $V_i.E$ expresses the attack-complexity of vulnerability V_i. Formula (3) is used to compute the risk of a vulnerability on attribute X after introducing the correlative influences.

Referring to the assessment idea of "Dow Jones Index", we present an index-based assessment policy. The core idea of this policy is that a sequence of vulnerabilities with the highest risk is selected as evaluating evidence of a system. We use a matrix (Y1,Y2,…,Yn) to record the sequences, where, n is the number of vulnerabilities related to the object, Yi(i=1~n) denotes the i highest risk vector on confidentiality, authenticity and availability, noted as $Yi=(Yi.C,Yi.T,Yi.A)^T$, and $Yk \geq Yj$, namely, $Yk.C \geq Yj.C, Yk.T \geq Yj.T, Yk.A \geq Yj.A$ ($k \leq j$ *and* $j, k \in \{1,2, ..., n\}$). Formula (4) is used to calculate the risk of object 'Obj' faced by a user of 'User' on different attributes. Where, $V.Ppre$ is the premise privilege-set of vulnerability V. The relations

of 'User' and privilege-set can be determined by Table 1. $Max^{(i)}\{DataSet\}$ denotes the i largest element in the set of 'DataSet'. And '*p*' is a selective index which is used to control the length of the sequences. To compare the risks of two systems, we define a comparing mechanism of vulnerability sequences. Suppose the risk expressions of two vulnerability sequences on confidentiality are respectively $(Y_1.C,\cdots,Y_n.C,)$, $(Y_1^{'}.C,\cdots,Y_n^{'}.C)$. The rule is as follows:

(1) $(Y_1.C,\cdots,Y_n.C,) > (Y_1^{'}.C,\cdots,Y_n^{'}.C), if\ Y_1.C > Y_1^{'}.C$;

(2) $(Y_1.C,\cdots,Y_n.C,) < (Y_1^{'}.C,\cdots,Y_n^{'}.C), if\ Y_1.C < Y_1^{'}.C$;

(3) Consider the relations of $Y_2.C\ and\ Y_2^{'}.C$, *if* $Y_1.C = Y_1^{'}.C$; The rest may be deduced by analogy;

Thus, the comparing mechanisms about authenticity and availability are also deduced by the same theory.

4 Windows Vulnerability Assessment

We select 1085 Windows vulnerabilities with obvious privilege escalations in Bugtraq[9] to be the assessment evidences. The assessment processes are as follows:

- Apply our vulnerability taxonomy to classify and describe each vulnerability.
- Generate a VRG G. Specify Windows 98, Windows NT, Windows 2000 and Windows XP as assessment objects. Then respectively retrieve the corresponding vulnerability relationship subgraphs G[OSi] from G.
- Let *p*=8. Adopt the index policy to respectively compute $ObjSecRisk(Obj,User)$, where, $Obj \in \{Windows\,98, Windows\,NT,\ Windows\,2000, Windows\,XP\}$, $User \in \{DistrustedUsers, TrustedUsers, RegularUsers\}$.

To compare the index policy with other policies, we also give the statistical results about the number and risk sum of isolated vulnerabilities, as shown in Fig.2. The risk of an isolated vulnerability is determined by the product of its attack-complexity and the weight increment of its premise and consequence privilege-set. Because the theory of these two methods is similar, the shape of two subfigures is also alike. Therefore, they can reach the same conclusion that the security of XP, 98, 2000 and NT is in descending order.

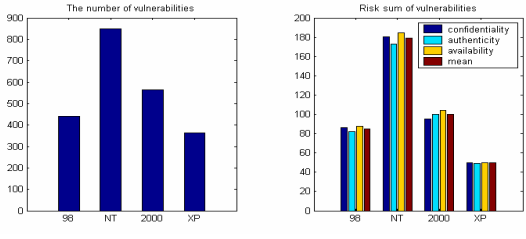


Fig. 2. The number and risk sum of vulnerabilities

Then, we give the assessment results as shown in Fig.3, where, three rows of subfigures respectively express the security risk on confidentiality, authenticity and availability of Windows systems, and there columns of subfigures respectively denote the security risk of Windows systems for distrusted remote visitors, trusted remote visitors and system regular users. A group of bars in a subfigure expresses a vulnerability sequence which reflects the risk of a system. The comparison of systems can be conducted by means of the comparing mechanism stated before.

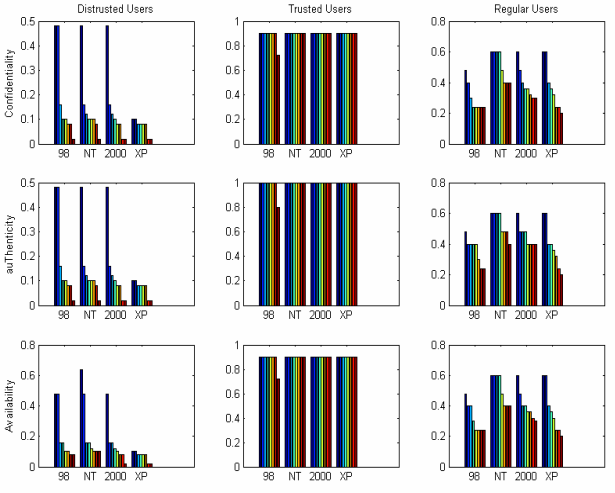


Fig. 3. Risk situation of Windows systems for different users on different attributes

Next, we will analyze the results from three aspects.

1) Overall analysis of Windows systems. From Fig.3 we can see that for trusted remote visitors, the security of Windows systems is lower because they contain large number of vulnerabilities whose risk is close to 1. But for distrusted remote visitors, they are relatively secure. This indicates that the existence of security devices like a firewall is active, to some extent. In addition, there is no obvious difference of the security risk on confidentiality, authenticity and availability of Windows systems.

2) Security comparison of Windows versions. Fig.3 shows that the security of Windows NT is almost lowest on all attributes. For trusted remote visitors and system regular users, the risk difference of Windows 98, 2000 and XP is very small. It is regarded that for distrusted remote visitors, the risk of Windows 98 is highest while Windows XP is distinctly lowest.

3) Analysis of assessment policy. As stated above, there is a marked divarication about the security of Windows XP and 2000 evaluated by the risk sum and index policies. To some extent, the traditional sum policy can reflect an overall risk situation of an operating system. But it probably leads to a problem that the risk sum of some vulnerabilities with low risk is greater than that of a vulnerability with high risk. In fact, although the risk sum of Windows XP is much less than that of Windows 2000, the number of high risk vulnerabilities in Windows XP for trusted remote visitors and system regular users is not less than the latter. This fact makes these two systems placed

at the same risk. Therefore, we believe that the number of vulnerabilities never reflects their risk. A catastrophic vulnerability can pose tremendous threats to an operating system. But the index policy can effectively avoid the above defect.

5 Conclusions

This paper proposes a VRG-based quantitative assessment method and an index-based assessment policy. Through introducing the correlative influences of vulnerabilities, VRG can be used to scientifically detect high risk vulnerabilities which can evoke multistage attacks though their threats on surface are very little. In addition, the index policy reflects the assessment idea that the risk of systems is not related to the risk sum and quantity of vulnerabilities, but to high risk vulnerabilities.

Analysis of 1085 vulnerabilities indicates that for trusted remote visitors, the security of Windows systems is lower while for distrusted remote visitors, they are relatively secure. But there is no obvious difference of the security risk on confidentiality, authenticity and availability of Windows systems. In addition, the security of Windows NT is almost lowest on all attributes. For trusted remote visitors and system regular users, the risk difference of Windows 98, 2000 and XP is very small. For distrusted remote visitors, the risk of Windows 98 is highest while Windows XP is distinctly lowest.

References

1. Michener, J.: System Insecurity in the Internet Age. IEEE Software. Vol.16, No.4, (1999) 62-69
2. Jiwnani, K., Zelkowitz, M.: Maintaining Software with A Security Perspective. ICSM'02. Montréal, (2002) 194-203
3. Hedbom, H., Lindskog, S., Axelsson, S., Jonsson, E.: A Comparison of the Security of Windows NT and UNIX. Third Nordic Workshop on Secure IT Systems, November (1998)
4. Wang, L.D.: Quantitative Security Risk Assessment Method for Computer System and Network (In Chinese). Ph.D. Thesis, Harbin Institute of Technology (2002)
5. Zhang, Y.Z., Yun, X.C.: A New Vulnerability Taxonomy Based on Privilege Escalation. Proceedings of the 6th ICEIS, Vol.3, (2004) 596-600
6. Fang, B.X.: Network and Information Security, and Its Novel Technology. (2001) http://pact518.hit.edu.cn/view/view/frame.htm
7. Schneier, B.: Attack Trees. Dr. Dobb's Journal, Vol.24, No.12, (1999) 21-29
8. Zhang, Y.Z., Yun, X.C., Fang, B.X. (eds.): A Mining Method for Computer Vulnerability Correlation. Int. J. Innovative Computing, Information and Control. Vol.1, No.1, (2005) 43-51
9. SecurityFocus: Bugtraq. (2005) Http://www.securityfocus.com/bid/

A New (t, n)-Threshold Multi-secret Sharing Scheme

HuiXian Li[1,2], ChunTian Cheng[2], and LiaoJun Pang[3]

[1] School of Electronic and Information Engineering, Dalian University of Technology, Dalian1 16024, China
hxli@student.dlut.edu.cn
[2] Institute of Hydroinformatics, Dalian University of Technology, Dalian1 16024, China
ctcheng@dlut.edu.cn
[3] National Key Labtorary of Integrated Service Networks, Xidian University, Xi'an 710071, China
pangliaojun_x@etang.com

Abstract. In a (t, n)-threshold multi-secret sharing scheme, at least t or more participants in n participants can reconstruct $p(p \geq 1)$ secrets simultaneously through pooling their secret shadows. Pang *et al*. proposed a multi-secret sharing scheme using an $(n+p-1)$th degree Lagrange interpolation polynomial. In their scheme, the degree of the polynomial is dynamic; with the increase in the number of the shared secrets p, the Lagrange interpolation operation becomes more and more complex, at the same time, computing time and storage requirement are large. Motivated by these concerns, we propose an alternative (t, n)-threshold multi-secret sharing scheme based on Shamir's secret sharing scheme, which uses a fixed nth degree Lagrange interpolation polynomial and has the same power as Pang *et al*.'s scheme. Furthermore, our scheme needs less computing time and less storage requirement than Pang *et al*.'s scheme.

1 Introduction

A secret sharing scheme is a technique to share a secret among a group of participants. It is mainly used to protect important information from being lost, destroyed, modified, or into wrong hands. The first secret sharing scheme is the (t, n)-threshold secret sharing scheme, which was introduced by Shamir [1] and Blakley [2] independently in 1979. In the (t, n)-threshold scheme, any subset of size at least t participants is a qualified subset, that is to say, at least t participants can reconstruct the shared secret, but $(t-1)$ or fewer participants can obtain nothing about the secret.

In the above schemes, each participant's secret shadow can be used in only one secret sharing session. Later, several schemes [3], [4] were proposed for multiple secrets sharing. In a multi-secret sharing scheme, every participant only needs to keep one secret shadow and many secrets can be shared independently without refreshing the secret shadow. In 2000, Chien *et al*. [5] proposed a new type of (t, n)-threshold multi-secret sharing scheme, in which $p(p \geq 1)$ secrets instead of only one secret can be shared in each sharing session. Their scheme can find important applications [6], but its computation is very complex because it needs to solve $(n+p-t)$ simultaneous equations. Later in 2004, Yang *et al*. [6] proposed an alternative implementation of

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 421 – 426, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Chien *et al.*'s scheme, which is easier in computation than their scheme, but needs more public values. Recently, Pang *et al.* [7] also proposed a new scheme, which is a combination of advantages of Chien *et al.*'s and Yang *et al.*'s schemes and avoids disadvantages of their schemes. However, Pang *et al.*'s scheme needs to reconstruct an $(n+p-1)$th degree polynomial in both the secret distribution and the secret reconstruction. The degree of the polynomial in Pang *et al.*'s scheme is dynamic, and it varies with the number of the shared secrets p. It is obvious that the lager p is, the more complex the Lagrange interpolation operation is. When p is very large, Pang *et al.*'s scheme will be very low in efficiency and it may be of no use on some occasions. The same problem exists in Chien *et al.*'s and Yang *et al.*'s schemes too. Motivated by these concerns, we propose a new (t, n)-threshold multi-secret sharing scheme based on Shamir's scheme [1], which can be used as an alternative implementation of Pang *et al.*'s scheme. In our scheme, to share p secrets, only an nth degree Lagrange interpolation polynomial is required, and thus the computation of our scheme is easier than that of Pang *et al.*'s scheme especially when p is very large.

The rest of this paper is organized as follows. In Section 2, we introduce a definition and a theorem relative to the proposed scheme. In Section 3, we present the proposed scheme in detail. Then we make an analysis and discussion about the scheme in Section 4. Finally, we give conclusions in Section 5.

2 Preliminary

The proposed scheme adopted a special function, the two-variable one-way function [3], [6]. We will first introduce the definition of this function, and then give a theorem that supports our scheme.

Definition 1. (Two-variable one-way function). *The two-variable one-way function $f(r, s)$ is a function that maps any r and s onto a bit string $f(r, s)$ of a fixed length. This function has the following properties*:

(a) *Given r and s, it is easy to compute $f(r, s)$;*
(b) *Given s and $f(r, s)$, it is hard to compute r ;*
(c) *Having no knowledge of s, it is hard to compute $f(r, s)$ for any r ;*
(d) *Given s, it is hard to find two different values r_1 and r_2 such that $f(r_1,s)=f(r_2,s)$;*
(e) *Given r and $f(r, s)$, it is hard to compute s;*
(f) *Given pairs of r_i and $f(r_i, s)$, it is hard to compute $f(r', s)$ for $r' = r_i$.*

Theorem 1. *Given $(m + 1)$ unknown variables $x_i \in GF(q)$ $(i = 0, 1, \ldots, m)$, and m equations $x_i' = x_0 \oplus x_i$ $(i = 1, 2, \ldots, m)$, where $GF(q)$ is a finite field. Here "$\oplus$" denotes exclusive-or bit by bit. Only the values of x_i' $(i = 1, 2, \ldots, m)$, are published. Given x_0, it is easy to find the remaining unknown symbols x_i $(i = 1, 2, \ldots, m)$; without any knowledge of x_0, it is computationally infeasible to determine the values of these unknown symbols.*

Proof of Theorem 1: We prove the theorem 1 in two steps:
(1) Given x_0, we can find x_i $(i = 1, 2, \ldots, m)$ easily by computing $x_i = x_0 \oplus x_i'$.

(2) Without any knowledge of x_0, Theorem 1 is equal to solve m simultaneous equations, $x_0 \oplus x_i = x_i'$ $(i = 1, 2, \ldots, m)$, with $(m+1)$ unknown symbols $x_i (i = 0, 1, \ldots, m)$. So, given these m equations, the values of these unknown symbols cannot be uniquely determined since the number of the available equations is less than that of the unknown symbols. The only thing for an adversary to do is to guess the value of any x_i $(i = 0, 1, \ldots, m)$. But the probability that the adversary succeed in doing it is only $1/q$ due to $x_i \in GF(q)$. If $GF(q)$ is a sufficiently large finite field, the successful probability tends to 0. So with no knowledge of x_0, it is computationally infeasible to determine the values of these unknown symbols.

3 The Proposed Scheme

Based on Theorem 1, we propose a new (t, n)-threshold multi-secret sharing scheme, which contains three phases:

3.1 System Parameters

Let $GF(q)$ denote a finite field, where q is a large prime number. All numbers are elements of $GF(q)$. Suppose that there is n participants $U_1, U_2, \ldots, U_n$ in the system. Let D $(D \notin \{U_1, U_2, \ldots, U_n\})$ denotes a dealer. In the whole paper, we assume that the dealer is a trusted third party. The dealer randomly selects n distinct integers, $s_1, s_2, \ldots, s_n$, from $GF(q)$ as participants' secret shadows. The public identifier for each participant is a well-known number, which can be used to represent each participant individually. The dealer randomly selects n distinct integers, $u_i \in [n-t+2, q]$, for $i = 1, 2, \ldots, n$, as participants' public identifiers. There are p secrets $k_1, k_2, \ldots, k_p$ to be shared among n participants. Let $f(r, s)$ be a two-variable one-way function defined above, which is used to compute pseudo shadows of participants.

3.2 Secret Distribution

The trusted dealer performs the following steps to implement the secret distribution:

a. Randomly choose an integer r and compute $f(r, s_i)$ for $i = 1, 2, \ldots, n$.

b. Use n pairs of $(0, k_1)$ and $(u_1, f(r, s_1))$, $(u_2, f(r, s_2))$, $\ldots$, $(u_n, f(r, s_n))$ to construct an nth degree polynomial $h(x) = a_0 + a_1 x + \ldots + a_n x^n$.

c. Compute $z_i = h(i) \bmod q$ for $i = 1, 2, \ldots, n-t+1$ and $k_i' = k_1 \oplus k_i \bmod q$ where $i = 2, 3, \ldots, p$.

d. Publish the values of $r, z_1, z_2, \cdots, z_{n-t+1}, k_2', k_3', \cdots, k_p'$.

3.3 Secret Reconstruction

In order to reconstruct the shared secrets, at least t participants pool their pseudo shadows $f(r, s_i)$ for $i = 1', 2', \ldots, t'$. From these t pseudo shadows, we have t pairs of $(u_i, f(r, s_i))$ for $i = 1', 2', \ldots, t'$. With the knowledge of the public values $z_1, z_2, \ldots, z_{n-t+1}$, we can get $(n-t+1)$ pairs of (i, z_i) for $i = 1, 2, \ldots, n-t+1$. Therefore, there are $(n+1)$ pairs obtained altogether, by which the nth degree polynomial $h(x)$ can be

uniquely determined. We use (X_i, Y_i) for $i=1, 2, \ldots, n+1$ to denote these $(n+1)$ pairs, respectively. So $h(0)$ can be reconstructed through the following Lagrange interpolation polynomial:

$$h(0) = \sum_{i=1}^{n+1} Y_i \prod_{j=1, j\neq i}^{n+1} \frac{-X_j}{X_i - X_j} \bmod q \ . \tag{1}$$

We have $k_1 = h(0)$, subsequently, the remained $(p-1)$ secrets can be easily found by $k_i = k_1 \oplus k_i' \bmod q$ for $i = 2, 3, \ldots, p$, respectively.

4 Analysis and Discussion

We examine the security of our scheme from the following different perspectives:

(1) From public values $k_2', k_3', \cdots, k_p'$, we can get $(p-1)$ simultaneous equations $k_1 \oplus k_i = k_i' \bmod q$ for $i = 2, 3, \ldots, p$ with p unknown symbols. According to Theorem 1, it is computationally infeasible for an attacker to get any knowledge of the shared secrets k_i (i=1, 2, …, p).

(2) Our scheme is based on Shamir's scheme, and any $(t-1)$ or fewer participants cannot reconstruct the polynomial $h(x)$. So they cannot uniquely determine the value of k_1. But t or more participants can easily determine the secret k_1 with their pseudo shadows computed from their real secret shadows, and then they can compute the remained $(p-1)$ secrets by computing $k_i = k_1 \oplus k_i' \bmod q$ ($i = 2, 3, \ldots, p$). Therefore, our scheme enforces the (t, n)-threshold multi-secret sharing rule.

(3) Each participant's secret shadow can be reused in our scheme, and it is not disclosed even after multiple secret reconstructions. Even though n pseudo shadows $f(r, s_i)(i = 1, 2, \ldots, n)$ have been exposed among many co-operating participants after a secret reconstruction, where r is randomly selected every time, the real secret shadow s_i is well protected by the properties of the two-variable one-way function. Therefore, in order to share the next p secrets, the secret dealer only needs to randomly choose a new integer r without redistributing every participant's secret shadow s_i (i=1, 2, …, n). So our scheme is a really multi-use scheme.

4.1 Performance Analysis

We shall evaluate the performance of our scheme in the following three aspects:

(1) In Pang *et al.*'s and our schemes, the Lagrange interpolation computation is the most time-consuming operation. When $p = 1$, our scheme is the same as Pang *et al.*'s scheme, that is, both schemes need to reconstruct an nth polynomial. But when $p > 1$, our scheme is more efficient than Pang *et al.*'s scheme. Their scheme needs to reconstruct an $(n+p-1)$th polynomial. However, our scheme still needs to reconstruct an nth polynomial in Eq.(1). Though $(p-1)$ "⊕" operations are added in our scheme, they are negligible compared with the Lagrange interpola-

tion computation. In Pang *et al.*'s scheme, the larger p is, the higher the degree of the polynomial is. As we know, a great deal of computing time is needed for reconstructing a polynomial with a high degree, which will reduce the efficiency of their scheme. Therefore, the performance of our scheme is better than that of Pang *et al.*'s scheme, especially when p is very large.

(2) The number of public values is also an important parameter that determines the performance of a scheme, because it affects the storage and communication complexity of the scheme [7]. In table 1, we list the number of public values needed to share p secrets in each scheme.

Table 1. The number of public values in each scheme

	Our scheme	Pang et al.'s scheme	Yang et al.'s scheme	Chien et al.'s scheme
$p > t$	$n+p-t+1$	$n+p-t+1$	$n+p-t+1$	$n+p-t+1$
$p \le t$	$n+p-t+1$	$n+p-t+1$	$n+1$	$n+p-t+1$

From table 1, we can see that when $p>t$, four schemes have the same number of public values, and when $p \le t$, $(n+p-t+1)$ public values, which is needed in our scheme, Pang *et al.*'s and Chien *et al.*'s scheme, is less than $(n+1)$, which is needed in Yang *et al.*'s scheme, especially when p is very close to 1 and t to n. If $p=1$ and $t=n$, only 2 public values are needed in Chien *et al.*'s, Pang *et al.*'s and our schemes, but $(n+1)$ public values are required in Yang *et al.*'s scheme.

(3) Our scheme use an nth polynomial, while Pang *et al.*'s scheme use an $(n+p-1)$th polynomial. If we use link list to store a polynomial, n storages are needed in our scheme while $(n+p-1)$ storages are needed in Pang *et al.*'s scheme. When $p=1$, the storages of our scheme is the same as that of Pang *et al.*'s scheme, i.e., both schemes need n storages. But when $p>1$, Pang *et al.*'s scheme needs $(n+p-1)$ storages, however, our scheme still needs n storages. It is obvious that $(n+p-1)>n$ where $p>1$. So the storage cost of our scheme is lower than that of Pang *et al.*'s scheme.

5 Conclusions

In the paper, we propose a new (t, n)-threshold multi-secret sharing scheme based on Shamir's scheme. In our scheme, $p(p \ge 1)$ secrets can be shared in a secret sharing session, and t or more participants can co-operate to reconstruct these secrets, but $(t-1)$ or fewer participants can derive nothing about these secrets. The secret dealer can dynamically determine the number of the distributed secrets. Each participant's secret shadow can be reused, so the dealer only needs to randomly select a new integer r without redistributing secret shadows in each secret sharing session. The computing time of our scheme is less than that of Pang *et al.*'s scheme, due to take an nth polynomial. Furthermore, our scheme needs less storage demand than Pang *et al.*'s scheme.

Acknowledgements

Part of this research was supported by the Chinese National Natural Science foundation No. 50479055 and the National Key 973 Project of China, under contract no. G1999035805.

References

1. Shamir, A.: How to Share a Secret. Communications of the ACM 22 (1979) 612–613
2. Blakley, G.: Safeguarding Cryptographic Key. In: Smith, M., Zanca J. T. (eds.): Proc. AFIPS 1979 Natl. Conf.. AFIPS Press, New York (1979) 313–317
3. He, J., Dawson, E.: Multistage Secret Sharing Based on One-way Function. Electronics Letters 30 (1994) 1591–1592
4. Harn, L.: Efficient Sharing (Broadcasting) of Multiple Secrets. IEE Proceedings—Computers and Digital Techniques 142 (1995) 237–240
5. Chien, H.-Y., Jan, J.-K., Tseng, Y. -M.: A Practical (t, n) Multi-secret Sharing Scheme. IEICE Transactions on Fundamentals E83-A (12) (2000) 2762–2765
6. Yang, C.-C., Chang, T.-Y., Hwang, M. S.: A (t, n) Multi-secret Sharing Scheme. Applied Mathematics and Computation 151 (2004) 483–490
7. Pang, L.-J., Wang, Y.-M.: A New (t, n) Multi-secret Sharing Scheme Based on Shamir's Secret Sharing. Applied Mathematics and Computation 167 (2005) 840–848
8. Tan, K.-J., Zhu, H.-W., Gu, S.-J.: Cheater Identification in (t, n) Threshold Scheme. Computer Communications 22 (1999) 762–765

An Efficient Message Broadcast Authentication Scheme for Sensor Networks*

Sang-ho Park and Taekyoung Kwon

Sejong University, Seoul 143-747, Korea
superh1@sju.ac.kr, tkwon@sejong.ac.kr

Abstract. A sensor network consists of a large number of small sensor nodes with wireless networking. An attacker can easily insert forged messages in wireless networking environment. Therefore, broadcast authentication mechanisms are necessary for a base station to broadcast the same messages to many sensor nodes. The famous broadcast authentication mechanisms, *TESLA*, *μTESLA*, and *Multilevel μTESLA*, are adaptable. We observe these mechanisms and point out problems of *Multilevel μTESLA* with regard to efficiency and scalability. Subsequently, we propose *Efficient Two-level μTESLA* to mitigate the problems.

1 Introduction

A sensor network consists of one or a few base stations and many sensor nodes with wireless networking. The sensor node is a small device that has low computational power, small memory size, small storage space, and exhaustible battery, while the base station is believed more robust than sensor nodes. Due to the wireless property of sensor devices, the sensor networks may easily be compromised by an attacker who sends forged messages or modified ones. Therefore, we must authenticate messages transmitted over the wireless sensor networks.

For broadcasting messages from the base station to sensor nodes, we need a broadcast authentication method such as a digital signature. However, the mathematical overhead of digital signature operations makes it impractical to be used in the sensor network. TESLA [1] is a protocol which uses MAC (Message Authentication Code) and broadcasting for message authentication, while μTESLA [2] and Multilevel μTESLA [3] are adapted versions for broadcast authentication in the wireless sensor network. In this paper, we observe these mechanisms and point out problems with regard to efficiency and scalability. Subsequently, we propose *Efficient Two-level μTESLA* to mitigate the problems.

* This research was supported in part by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment). This research was also supported in part by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 427–432, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Previous Mechanisms: Multilevel μTESLA

TESLA [1], μTESLA [2] and Multilevel μTESLA [3] are protocols that imitate asymmetric algorithm using symmetric one. They use MAC and broadcasting mechanisms. The base station generates MAC with a symmetric key and reveals the key after a certain period of time. It removes computational overheads of public-key operations from the digital signature based schemes. These schemes are efficient for broadcast authentication. However, they still have some overheads to distribute key chain commitment securely. TESLA uses a digital signature algorithm and broadcasting for solving this problem but it may cause computational overheads in the sensor network. So, μTESLA resolves it by unicasting but it lays communication load.

Multilevel μTESLA is a protocol that mitigates overheads of key chain commitment distribution. The basic idea is to predetermine and to broadcast the key chain commitments. It divides the life time of sensor network into n_0 time intervals in high-level (long time intervals). Each interval of high-level is divided into n_1 time interval in low-level (short time intervals). Long time intervals should cover the life time of sensor network without a large number of keys and memory requirements in the base station. Short time intervals help the sensor nodes have small buffer size. The high-level key chain is only used for distribution and authentication of low-level key chain commitments. The low-level key chain is used for broadcast authentication. Both levels use MAC for authentication.

It predetermines high-level key chain commitment K_0 and the first low-level key chain commitment $K_{1,0}$ to authenticate key chains. The other low-level key chain commitments are broadcasted by CDM (*Commitment Distribution Message*)s during the life time of high-level key chain. The following is the key chain scheme of Multilevel μTESLA with 2 levels.

- One-way pseudo random functions : F_0, F_1, F_{01}, F'
- High-level time intervals : $I_1, I_2, ..., I_{n_0-1}, I_{n_0}$
- High-level key chain : $K_1, K_2, ..., K_{n_0-1}, K_{n_0}$
- The last high-level key : $K_n \leftarrow PRNG$
- High-level key generation : $K_i = F_0(K_{i+1})$
- Low-level time intervals in ith high-level interval : $I_{i,1}, I_{i,2}, ..., I_{i,n_1-1}, I_{i,n_1}$
- Low-level key chain in ith high-level interval : $K_{i,1}, K_{i,2}, ..., K_{i,n_1-1}, K_{i,n_1}$
- The last low-level key in ith high-level interval : $K_{i,n_1} = F_{01}(K_{i+1})$
- Low-level key generation in ith high-level interval : $K_{i,j} = F_1(K_{i,j+1})$
- Low-level key chain commitment distribution : $K'_i = F'(K_i)$

  $CDM_i = i|K_{i+2,0}|MAC_{K'_{i+1}}(i|K_{i+2,0})|K_i$

CDMs provide information for authenticate key chains of both levels by containing high-level key and low-level key chain commitment. Sensor nodes must store 3 CDMs ($CDM_{i-2}, CDM_{i-1}, CDM_{i.}$) in their buffer for using $K_{i,0}$ in CDM_{i-2}. Multilevel μTESLA can add a few levels. It extends available time of key chains to cover long life time of sensor nodes. If higher-level is made of L keys, life time of sensor nodes is extended L times. It keeps short time intervals of the the lowest-level key chain and uses long life time of sensor network without a long key chain.

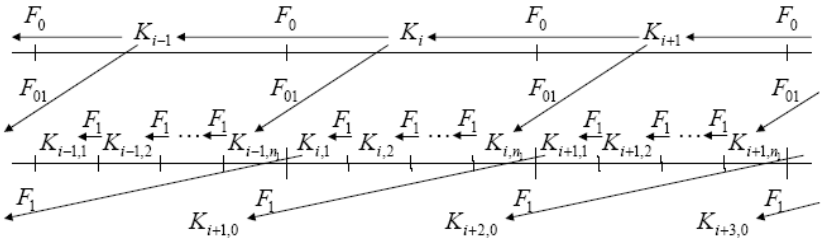


Fig. 1. Multilevel μTESLA (2 levels)

3 Efficient Two-Level μTESLA

3.1 Problems of Multilevel μTESLA

Multilevel μTESLA can use m levels to cover a long period of time. However, a number of levels may cause many overheads. Figure 2 shows key chains of 4-level μTESLA. We will show a full detail of overheads of figure 2 in Section 3.3.

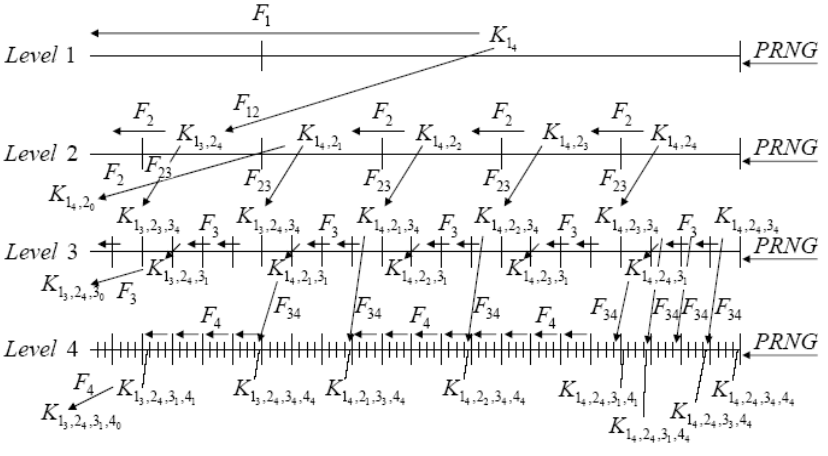


Fig. 2. Multilevel μTESLA (4 levels)

It can only cover a fixed period of time although sensor nodes are available for longer time. Sensor nodes can have enough battery to do more operations when we are incorrect in predicting the life time of sensor network. In this case, we waste some or many resources because of fixed life time of Multilevel μTESLA.

3.2 Efficient Two-Level μTESLA

We propose a new mechanism named *Efficient Two-level μTESLA* which uses new key chains to mitigate problems of Multilevel μTESLA. This protocol generates a new key chain $P_1, P_2, ..., P_{n_0-1}, P_{n_0}$ and broadcasts it by CDMs. It

provides flexible key chain length for variable life time of sensor network using new key chains. The following is CDMs of Multilevel μTESLA to broadcast the last low-level key chain commitment $K_{n_0,0}$.

$$CDM_{n_0-2} = n_0 - 2 \,|K_{n_0,0} \quad |MAC_{K'_{n_0-1}} \,|K_{n_0-2}$$
$$CDM_{n_0-1} = n_0 - 1 \,|NULL \,|NULL \quad\quad |K_{n_0-1}$$

K_{n_0-1} in CDM_{n_0-1} is only used to authenticate CDM_{n_0-2}. K_{n_0} is never used. We use K_{n_0-1} and K_{n_0} to broadcast new key chain commitments $P_0, P_{1,0}$. The following and figure 3 show our key distribution scheme.

$$P_{n_0} = PRNG$$
$$K_{n_0,n_1} = F_{01}(P_1)$$
$$CDM_{n_0-2} = n_0 - 2 \,|K_{n_0,0} \quad |MAC_{K'_{n_0-1}} \,|K_{n_0-2}$$
$$CDM_{n_0-1} = n_0 - 1 \,|P_{1,0}, P_0 \,|MAC_{K'_{n_0}} \quad |K_{n_0-1}$$
$$CDM_{n_0} \quad = 0 \quad\quad |P_{2,0} \quad |MAC_{P'_1} \quad |K_{n_0}$$
$$CDM_1 \quad = 1 \quad\quad |P_{3,0} \quad |MAC_{P'_2} \quad |P_1$$
$$CDM_i \quad = i \quad\quad |P_{i+2,0} \quad |MAC_{P'_{i+1}} \quad |P_i$$

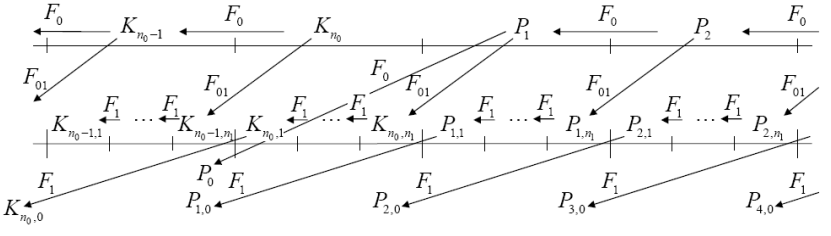


Fig. 3. Efficient Two-level μTESLA

Sensor nodes process the second field, $P_{1,0}, P_0$, in previous CDM if the first field in current CDM is 0. They should authenticate P_0 in previous CDM using P_1 in next CDM instead of authenticating K_{n_0} in current CDM. The remaining processes of our scheme are the same as Multilevel μTESLA.

3.3 Comparison of Efficient Two-Level μTESLA with Multilevel μTESLA

We assume that Multilevel μTESLA has m levels and each key chain consists of L keys. The lowest-level time interval is denoted I_{low}. The postulated system can then cover a $L^m \times I_{low}$ of time. $3(m-1)$ buffers for CDMs are necessary because it requires CDM between the higher-level and the lower-level. Each level n except the lowest-level sends CDMs to lower-level L^n times. Hence, a base station sends CDMs $\sum_{k=1}^{m-1} L^k$ times. F functions are needed for key chain, the last key of lower-level key and CDM. It must have the $2m-1$ number of distinct F functions for key generation and one F functions for CDM. $PRNG$ is operated M times for the last key of each level. Sensor nodes have to predetermine the first key chain commitment of each level. Each level n except the lowest-level runs F functions $3L^n$ times for key chain, the last key lower-level, and CDM. The

Table 1. Multilevel μTESLA vs. Efficient Two-level μTESLA

	Multilevel μTESLA	Efficient Two-level μTESLA
life time	$L^m(\times I_{low})$	$L^2 \times L^{m-2} = L^m(\times I_{low})$
CDM buffers	$3(m-1)$	3
CDM transmissions	$\sum_{k=1}^{m-1} L^k$	L^{m-1}
$PRNG$s	m	$L^{m-2}+1$
predetermination	m	2
F functions	$2m$	4
F operations	$\sum_{k=1}^{m-1} 3L^k + L^m$	$3L^{m-1} + L^m$

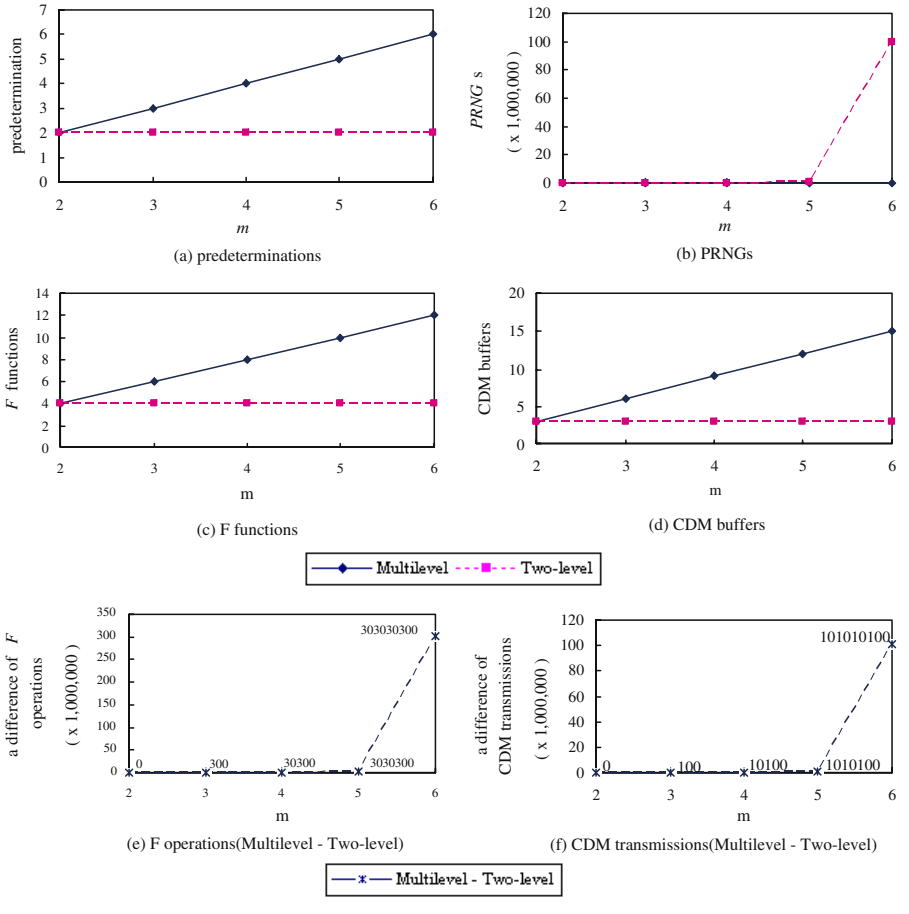


Fig. 4. Multilevel μTESLA and Efficient Two-level μTESLA with $L = 100$

lowest-level runs F functions L^m times for key chain generation. As a result, the base station and sensor nodes are operate F functions $\sum_{k=1}^{m-1} 3L^k + L^m$ times.

Efficient Two-level μTESLA can run broadcast authentication for $L^2 \times I_{low}$ intervals at once. Therefore, we repeat L^{m-2} times to cover the same life time as

Multilevel μTESLA with m-level. It has L^{m-1} high-level keys and L^m low-level keys. Sensor nodes require only 3 buffers for CDMs. A base station sends CDMs L^{m-1} times because the high-level consists of L^{m-1} keys. It can generate all keys and CDMs by F_0, F_{01}, F_1, F'. We must operate $PRNG$ L^{m-2} times to generate the last keys of each high-level key chain. The last key of the last low-level key chain is also generated by $PRNG$. The first key chain commitments of high-level and low-level have to be stored in sensor nodes. High-level runs F_0, F_{01} and F' functions L^{m-1} times each other and low-level operates F_1 functions L^m times. Finally, Efficient Two-level μTESLA executes F operations $3L^{m-1} + L^m$ times.

Therefore, our scheme is more efficient than Multilevel μTESLA except the number of $PRNG$ operations. Though there is more overhead than Multilevel μTESLA in terms of $PRNG$, we should remind that the base station only operates $PRNG$. There is no more overhead on sensor nodes. It is an insignificant overhead for sensor network because base stations generally are computationally robust machines and capable of a lot of $PRNG$ operations. Table 1 summarizes the comparison of Multilevel μTESLA and Efficient Two-level μTESLA with regard to efficiency. Figure 4 depicts the minute comparisons for $L = 100$.

4 Conclusion

We cannot use traditional broadcast authentication mechanisms to authenticate broadcasting messages in sensor network because of characteristics of sensor network such as low computational power, small memory size, wireless network, and etc. TESLA was proposed for adaptable broadcast authentication mechanism in traditional wired and wireless networks, but it is not suitable for sensor network. μTESLA mitigated problems of TESLA, and Multilevel μTESLA solved weak point of μTESLA. We pointed out overhead and lifetime limitation of Multilevel μTESLA, and proposed Efficient Two-level μTESLA for a solution. Efficient Two-level μTESLA has the same life time as μTESLA and reduces overhead of sensor network. We should mitigate overhead of $PRNG$ to generate new key chains on a base station in the future work.

References

1. A. Perrig, R. Canettid, D. Tygar and D. Song, "TESLA: Multicast source authentication transform," IRTF draft, draft-irtf-smug-tesla-00.txt.
2. A. Perrig, R. Szewczyk, V. Wen, D. Culler and D. Tygar, "SPINS: Security protocols for senor networks," Proceedings of the 7th annual international conference on Mobile computing and networking, pp.189-199, 2001.
3. D. Liu and P. Ning, "Multilevel μTESLA: Broadcast authentication for distributed sensor networks," ACM Transactions on Embedded Computing Systems (TECS), Vol. 3, Issue 4, pp.800-836, 2004.

Digital Image Authentication Based on Error-Correction Codes

Fan Zhang, Xinhong Zhang, and Zhiguo Chen

College of Computer & Information Engineering, Henan University, China
zhangfan@vip.sohu.com

Abstract. This paper presents a color image authentication algorithm based on convolutional coding. The message of each pixel is convolutional encoded with the encoder. After the process of parity check and block interleaving, the redundant bits are offset embedded in the image. The tamper can be detected and restored without accessing the original image. The experimental results show that the authentication algorithm based on convolutional coding has a good performance in the tamper detection, localization and the image reconstruction.

1 Introduction

Digital image manipulation software is now readily available on personal computers, and it is very easy both to tamper with any image and to make it available to others. Users have possibility to modify some features of the image or even the content of a scene, (e.g. to move, delete or add objects). Consequently, insuring digital image integrity becomes a major requirement for several applications and watermarking algorithms are promising techniques that allow to solve problem concerning image integrity checking.

There are several levels of image integrity that can be defined. The highest (but the most restrictive) level means that no alteration of the original (more precisely the watermarked original) image is allowed, e.g. the received data has to be identical to the original one. However, in real life situations, images can be transformed, compressed, etc, without modifying their semantic content. Therefore, and depending on different applications, several lower levels of image integrity may be considered, where the lowest one reject only images which were subject to malicious manipulation changing the content of the original image.

The content authentication determines whether an image has been tampered or not, and if necessary, locate malicious alterations made on the image. Authentication on a still image or a video is motivated by recipient's interest, and its principle is that a receiver must be able to reliably identify the source of this document. Several techniques and concepts based on data hiding or steganography have been designed as a means for the image authentication [1],[2]. The existing works on image authentication can be classified into two categories: authentication based on the digital signature and authentication based on the watermarking.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 433–438, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Digital signature schemes are built on the ideas of hash (or called message digest) and public key encryption that were originally developed for verifying the authenticity of text data in network communication. Friedman extended it to digital image [3]. A signature computed from the image data is stored separately for future verification. This image signature can be regarded as a special encrypted checksum. It is unlikely that two different natural images have same signature, and even if a single bit of image data changes, the signature may be totally different. Furthermore, public-key encryption makes it very difficult to forge signature, ensuring a high security level. There are many other digital signature based authentication schemes have been proposed [4],[5].

In the recent years, several authentication schemes based on the watermarking have been presented [6],[7], including fragile watermarks, semi-fragile watermarks, and self-embedding. The visual redundancy of typical images enables us to insert imperceptible additional information and make the images capable of authenticating themselves without accessing the original images.

When the tampers have been detected, several schemes can be used for the image reconstruction, such as the use of error-correcting codes in authentication proposed by Lee [8] and the self-recovery watermarking proposed by Fridrich [9].

This paper presents a color image authentication algorithm based on convolutional coding. The high bits of color digital image are coded by the convolutional codes for the tamper detection and localization. The authentication messages are hidden in the low bits of image in order to keep the invisibility of authentication. The tamper can be detected and restored without accessing the original image.

2 Convolutional Coding and Decoding

Convolutional codes are one type of code used for channel coding. Convolutional codes are usually described by two parameters: the code rate and the constraint length. The code rate k/n, is a measure of the amount of added redundancy. A rate $R = k/n$ convolutional encoder with memory order m can be realized as a k-input, n-output linear sequential circuit with input memory m, i.e., inputs remain in the encoder for an additional m time units after entering. Typically, n and k are small integers, $k < n$. The constraint length parameter K, denotes the length of the convolutional encoder. Closely related to K is the parameter m=K-1, which indicates how many encoder cycles an input bit is retained and used for encoding after it first appears at the input to the convolutional encoder. The m parameter can be thought as the memory length of the encoder. A (n, k, m) convolutional encoder consists of k shift register with m delay elements and with n modulo-2 adders.

The convolutional decoder regenerated the information by estimating the most likely path of state transition in the trellis. Maximum likelihood decoding means the decoder searches all the possible paths in the trellis and compares the metrics between each path and the input sequence. The path with the minimum metric is selected as the output. So maximum likelihood decoder is the optimum decoder. In general, a convolutional code (n, k, m) has $2(m\text{-}1)k$ pos-

sible states. At a sampling instant, there are $2k$ merging paths for each node and the one with the minimum distance is selected and called surviving path. At each instant, there are $2(m\text{-}1)k$ surviving paths are stored in the memory. When all the input sequences are processed, the decoder will select a path with the minimum metric and output it as the decoding result.

3 Image Authentication Based on Convolutional Coding

3.1 Embedding of Image Authentication

Because of the requirement of invisibility, allowable redundant bits embedded in the image are limited. This paper deals with 24-bit BMP images. Each pixel of 24-bit BMP image is represented by three-color channels (RGB), each of the channel is 8 bits. Only two highest bits of each color channels is convolutional encoded with the encoder, the code rate is 1/2. Then we embed the two redundant bits two lowest bits of a color channels of an appointed other pixel. The change of image quality caused by the change of two lowest bits is usually invisible to the human vision system. The reason we select two highest bits of each color channels is that they are the most significant bets (MSB) and have decisive influence to the image. When an image is tempered and need to be reconstructed, the outline of original image can be reconstructed by only recover two highest bits of each color channels.

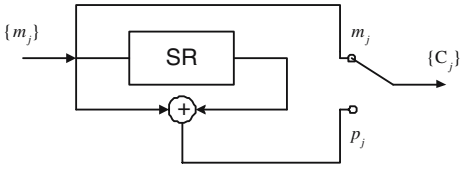


Fig. 1. Convolutional encoder

In this paper, a (2, 1, 2) convolutional code with the code rate 1/2 is used, and the redundant is 1 bit. This convolutional code can be used for the correction of one bit error from four bits.

The convolutional encoder we used is shown in figure 1. In figure 2, SR is shift register. The input message vector is $m = m_1, m_2, \ldots, m_j, \ldots$. The output codeword is $m_j \bigoplus p_j$, where $\bigoplus$ denote modulo-2 adders.

We extract two highest bits of RGB channel of each pixel, $\boldsymbol{m}$ = R[7], R[6], G[7], G[6], B[7], and input this vector to the convolutional encoder. Where the input vector is 5 bits, B[6] is not convolutional encoded by the encoder. The reason is that the convolutional encoder need a tail bit to flush the register and ensure that the tail end of the message is shifted from the register. And in order to embed two redundant bits to two lowest bits and keep the invisibility of authentication embedding, we shall limit the redundant of convolutional encoder as 6 bits. So the input vector is not 6 bits, B[6] is not convolutional coded by the encoder. The redundant bits of R[7], R[6], G[7], G[6], B[7] are respectively

embedded in R[0], R[1], G[0], G[1], B[0]. The redundant bit of tail bit is embedded in B[1].

Besides the convolutional encoding, we also use the parity check, block interleaving, and offset embedding to improve the sensitivity to tamper.

Parity check, also known as Vertical Redundancy Check (VRC), is a method of adding a parity bit to an array of binary digits to determine whether the number of 1 or 0 in the data is odd or even. In this paper, we use the odd parity check. Appending a bit that makes the total number of 1 in the message vector $\boldsymbol{m}$ is odd. If the number of 1 is already an odd number, the parity bit is set to 0. At the receiving end, the codeword is checked to see if he number of 1 is an odd number. If it is even, an error has occurred. We embed the odd parity check bit to the lower bit of blue channel, B[2].

Each convolutional encoded codeword is block interleaved. Every 12 pixels that select in sequence from the image are separated into a group. The last group is filled with zero pixels if the number of pixels is less than 12. The convolutional encoded codeword of each pixel is 12 bits. The block interleaving is simply done by arranging 12 codeword (12 pixels) into 12 row of an array and then transmitting them by column after column. If the convolutional code corrects random errors, the interleaved code corrects single bursts of length 12 or less.

We also use offset embedding to improve the reliability and robust of the image authentication system. Each redundant bit is offset embedded in 200×200 pixel. For example, the redundant bits of pixel (0, 0) are offset embedded in pixel (199, 199); the redundant bits of pixel (1, 1) are offset embedded in pixel (200, 200), and so on.

3.2 Detection of Image Authentication

The lower bits of RGB channel of each pixel are extracted and the redundant bits are recovered according to the offset and de-interleave process. Then we combine the corresponding higher bits with the redundant bits to make a codeword and convolutional decode it by decoder.

The convolutional decoder is shown as Figure 2. It includes two shift registers. The codeword m_j is the input of convolutional decoder, and the output is corrected codeword m'_j. S_0 is the correctional signal. If $S_0 = 1$, an error has occurred.

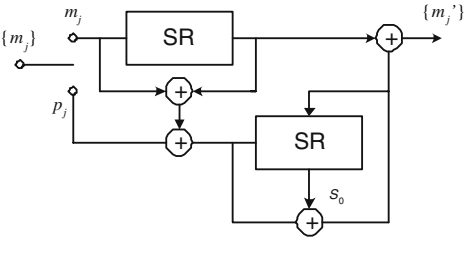


Fig. 2. Convolutional decoder

In the detection process of image authentication, using the odd parity check and comparing the redundant bits with the original bits, we can judge if this image has been tampered. If the image has been tampered, the redundant bits will replace the tampered bits, and the image is reconstructed.

4 Experimental Results

In the experiments, we deal with 256×256 24-bit BMP image for the image authentication. The experimental results of the convolutional decoded images are shown as Figure 3 (a). The authentication data that embedded in the image is invisible under normal viewing conditions.

Some common attacks are used to against this image authentication algorithm. The attack methods include painting, cropping, overlaying and so on. Some experimental results are shown as Figure 3 (b), (c), (d). According to the experimental results, the convolutional decoded images can exactly detect and localize the tamper in any pixels, and can correct tamper by itself.

Because of the requirement of invisibility, the code rate of this algorithm is only 1/2 and we only recover high bits in the color-channels of image, so the preferment of tamper recover is not very satisfying. In Figure 3 (d), when the

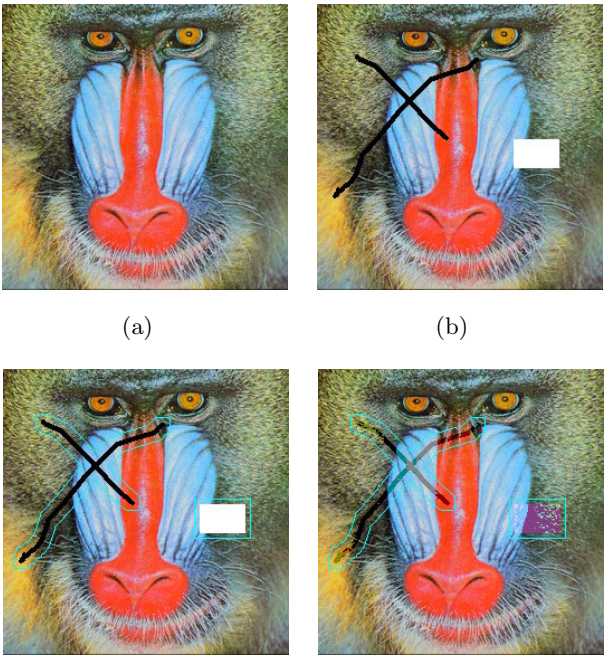


Fig. 3. (a) Authenticated image, (b) Tamper image, (c) Tamper detection and localization, (d) Reconstructed image

tamper is recovered, the color of the reconstructed tamper area is changed in despite of we can distinguish the outline of original image.

5 Conclusion

This paper presents a color image authentication algorithm based on convolutional coding. The message of each pixel is convolutional encoded with the encoder. After the process of parity check and block interleaving, the redundant bits are offset embedded in the image. The tamper can be detected and restored without accessing the original image.

References

1. Rey, C., Dugelay, J.: A Survey of Watermarking Algorithms for Image Authentication. EURASIP Journal on Applied Signal Processing **6** (2002) 613–621
2. Wu, J. H., Lin, F. Z.: Image Authentication Based on Digital Watermarking. Chinese Journal of Computers, **27** (9) (2004) 1153–1161
3. Friedman, G. L.: The trustworthy digital camera: Restoring credibility to the photographic image. IEEE Transactions on Consumer Electronics, **39** (4) (1993) 905–910
4. Wong, P., Memon, N.: Secret and public key image-watermarking schemes for image authentication and ownership verification. IEEE Transactions on Image Processing, **10** (10) (2001) 1593–1601
5. Lin, C. Y., Chang, S. F.: A robust image authentication method distinguishing JPEG compression from malicious manipulation. IEEE Transactions on Circuits and Systems for Video Technology, 2001, **11** (2) (2001) 153–168
6. Kundur, D., Hatzinakos, D.: Towards a telltale watermarking technique for tamper proofing. In: Proceedings of the IEEE International Conference on Image Processing, Chicago, USA, **2** (1998) 409–413
7. Abdulaziz, N., Pang, K. K.: Coding Techniques for Data Hiding in Images, Sixth International Symposium on Signal Processing and its Applications ISSPA 2001, Kuala-Lumpur, Malaysia, (2001) 170–173
8. Lee. J., Won. C. S.: A watermarking sequence using parities of error control coding for image authentication and correction. IEEE Transactions on Consumer Electronics, **46** (2) (2000) 313–317
9. Fridrich, J., Goljan. M.: Images with self-correcting capabilities. In: Proceedings of the IEEE International Conference on Image Processing, Kobe, Japan, **3** (1999) 792–796

Design and Implementation of Efficient Cipher Engine for IEEE 802.11i Compatible with IEEE 802.11n and IEEE 802.11e*

Duhyun Bae, Gwanyeon Kim, Jiho Kim, Sehyun Park**, and Ohyoung Song

School of Electrical and Electronic Engineering, Chung-Ang University,
221, HukSuk-Dong, DongJak-Gu, Seoul, Korea
{duhyunbae, cityhero, jihokim}@wm.cau.ac.kr,
{shpark, song}@cau.ac.kr

Abstract. For high data rate, new MAC mechanisms such as Block Ack in IEEE 802.11e and frame aggregation in IEEE 802.11n are being currently discussed and these mechanisms need short response time in each MPDU processing. In this paper, we propose a design of cipher engine for IEEE 802.11i to support these new mechanisms. We reduce the processing overhead of RC4 key scheduling by means of using the dual S-Box scheme. In CCMP design, parallel structure is proposed, and as a result, we reduce the processing time to 1/2 in comparison with the sequential structure and the response time independent of the size of the payload and only dependent on the clock frequency. In addition, we can expect to decrease power consumption in CMOS design process because the clock frequency reduces to 1/5 and the area doubles compared to the typical designs.

1 Introduction

IEEE 802.11 task group I presents the RSN (Robust Security Network) architecture that uses WEP, TKIP (Temporal Key Integrity Protocol) and CCMP (Counter with CBC-MAC Protocol) to protect WLAN traffic in MAC (Medium Access Control) layer [1][2]. In WEP and TKIP, RC4 stream cipher algorithm is used. This algorithm has an overhead of key scheduling which swaps S-Box memory 256 times before encryption [6]. Consequently, S-Box memory swapping delays encryption, reduces throughput and increases the response time is defined as an interval between the first data input time and the first processed data output time in this paper. On the other hand, AES (Advanced Encryption Standard) in CCM mode is used in CCMP. In AES-CCM, one AES core can be used to both MIC (Message Integrity Code) calculation and data encryption. So AES-CCM core can be implemented in hardware by two methods as follows: One is sequential structure that calculates MIC and then performs encryption. This method may lead to a significant savings in code and hardware size.

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC(Home Network Research Center)-ITRC support program supervised by the IITA(Institute of Information Technology Assessment).
** The Corresponding Author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 439 – 444, 2005.
© Springer-Verlag Berlin Heidelberg 2005

However, the first encrypted data will be ready after MIC calculation, so that the response time will increase. The other is the parallel structure design that computes the MIC and performs encryption simultaneously [3]. In the parallel structure, we can obtain the constant response time. But, the hardware size nearly doubles.

In this paper, we propose a design of the cipher engine for IEEE 802.11i to support new MAC mechanisms in other IEEE 802.11 standards by reducing the response time.

2 Design Considerations

In the IEEE 802.11 task group E, the Block Acknowledgement (Block Ack) mechanism is currently being discussed to reduce the acknowledgement overhead. The Block Ack mechanism allows a burst of frames to be transmitted before any acknowledgement. All the frames are separated by a short interframe space (SIFS) period [7]. In the IEEE 802.11n task group, single and multiple destination frame aggregation is being discussed. Some proposal specifies a standardized frame aggregation for both single and multiple destinations for high data rate [8].

In IEEE 802.11i, the encapsulation process is performed per MPDU. Because new MAC mechanisms in IEEE 802.11e and 802.11n have a short interval among MPDUs, the response time in a cipher core must be short. We may adopt faster clock frequency to reduce the response time. However it causes to increase power consumption. Therefore, to support new mechanisms in IEEE 802.11e and 802.11n and reduce power consumption, new design methods for 802.11i cipher core with the short response time are needed.

We may use SIFS as a factor to consider the maximum available response time. The total delay in MAC, PHY and the cipher core must range within the SIFS. Thus, the delay in the cipher core must be much smaller than the SIFS specified in other 802.11 standards. SIFS is 10μs in 802.11b and 16μs in 802.11a and 802.11n. In this paper, we assume that the maximum available response time in cipher core is about 3μs.

3 Algorithms and Response Time

The RC4 algorithm can be divided into 2 phases. The first phase is a key scheduling phase (KS-phase) and the second phase is a pseudorandom number generator (PRNG) phase. In WEP and TKIP, both the phases must be performed for every MPDU. In WEP and TKIP, the KS-phase initializes S-Box with WEP seed key and PRNG generates pseudorandom numbers and encrypts by XORing plaintexts and key streams [4][5].

In key scheduling process, as illustrated in Fig. 1, one loop in pseudo code takes 2 clock cycles with using dual-port ram. We read dual-port ram at the clock's falling edge and write at the clock's rising edge. Let i_n be the value of i in the n-th loop. In this phase, we don't need to wait for the end of swapping between S[i_n] and S[j_n] and can read S[i_{n+1}] because i value increases by one every loop and i_n can't be equal to i_{n+1}.

The PRNG phase is similar to the KS-phase as shown in Fig. 1. In a similar way to the KS-phase, we don't need to wait for the end of swapping and can read S[i_{n+1}]. In this design, the key scheduling takes 512 clock cycles and 8 bit encryption or decryption operation takes 2 clock cycles.

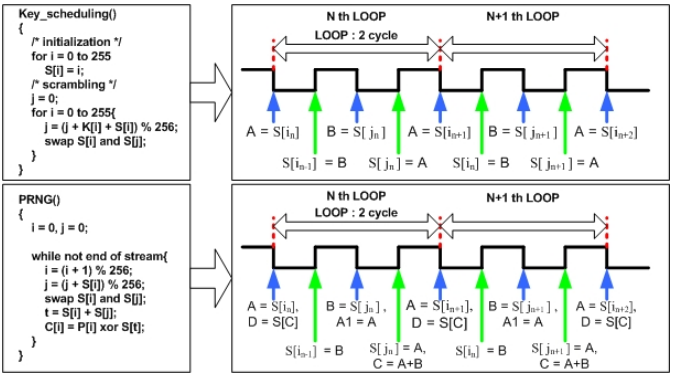


Fig. 1. Two phases of RC4 algorithms and S-Box memory access timing

As mentioned in Section 1, AES-CCM can be implemented in hardware by one of two methods. In the sequential structure design, the response time is proportional to the payload size. For example, if it takes 11 clock cycles for MIC calculation of 128 bits data, the total MIC calculation of 1024 bytes will take 748 clock cycles as described in Fig. 2.

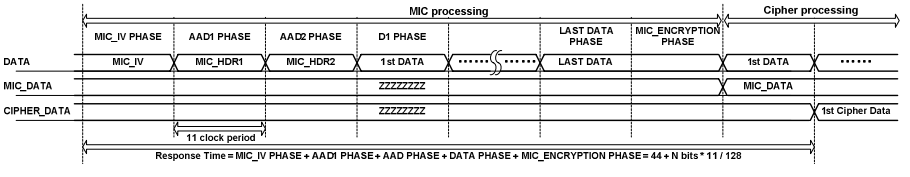


Fig. 2. The encapsulation timing diagram of a sequential AES-CCM core

In this paper, AES-CCM core is implemented in parallel structure to increase data throughput and reduce response time. The block diagram of AES-CCM core is illustrated in Fig. 5. The AES-CCM core consists of 5 blocks: AES-Cipher, AES-MIC, KEY_CTRL, INPUT_IF, and OUTPUT_IF. The AES-CIPHER block, which performs encryption, generates 128 bits cipher text every 11 clocks. The AES-MIC block computes MIC for NONCE, AAD (Additional Authentication Data) and MPDU data

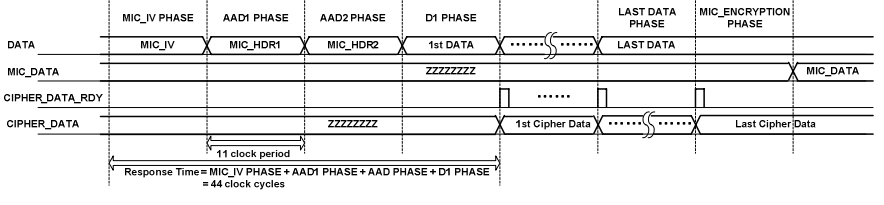


Fig. 3. The encapsulation timing diagram of a parallel AES-CCM core

field, and generates 64 bit encrypted MIC data. The KEY_CTRL block generates one round key from AES-CCM seed key per clock cycle. In Fig. 3, the timing diagram for encapsulation in our AES-CCM is shown. In our design, AES-CCM core needs only 44 clock cycles to obtain the first data output. *The important result is that the response time doesn't depend on the payload size.*

4 The Design of Protocol Core and Cipher Engine

In this paper, we introduce new method for the key scheduling, in which RC4 module has two S-Boxes and two key scheduling modules. When one pair of S-box and key scheduling module is used to perform encryption or decryption, the other pair is used to perform key scheduling about other RC4 key. To process the next MPDU, we can skip the key scheduling phase by just using pre-processed S-box and go to PRNG-phase. Thus, if the data field size of a MPDU currently being processed is more than 256 bytes, which it takes 512 clock cycles to encrypt or decrypt, the key scheduling for the next MPDU is performed at the same clock cycles in another S-Box and key scheduling module. The architecture of WEP and TKIP module is illustrated in Fig. 4.

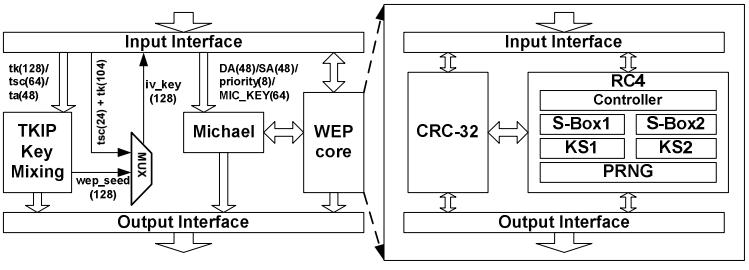


Fig. 4. The architecture of WEP and TKIP module

The CCMP module, as shown in Fig. 5, is designed by the parallel structure of AES-CCM core, explained in Section 3.

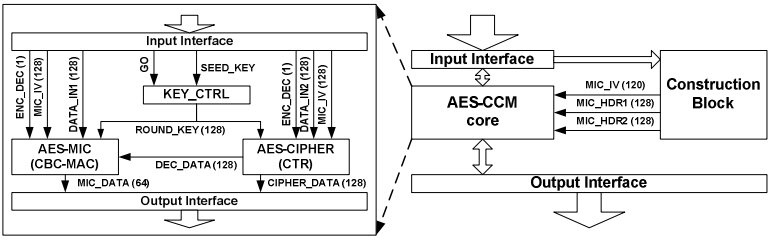


Fig. 5. The block diagram of AES-CCM core and CCMP module

The Construction Block creates the MIC initial vector and the AAD. The KEY_CTRL module generates the round keys from the AES-CCM seed key. The block diagram of the cipher engine for IEEE 802.11i is shown in Fig. 6.

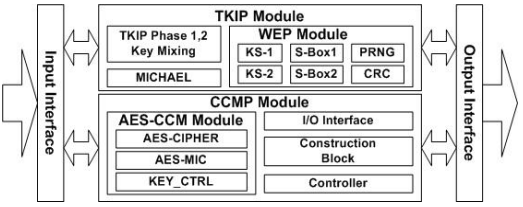


Fig. 6. The block diagram of the cipher engine designed for IEEE 802.11i

5 Performance Measurement and Verification

An implementation of the cipher engine for IEEE 802.11i was synthesized and implemented using Quartus II compiler and targeted to Altera Stratix FPGA device. The data throughput of our design satisfies MAC data processing speed of all standards. The data speed of MAC layer is 11 Mbps in IEEE 802.11b and 54 Mbps in IEEE 802.11a and 802.11g. Our design meets the requirements of all these standards [1].

Table 1 shows WEP, TKIP and CCMP performances. In WEP and TKIP, by using the dual S-box scheme, data throughput and response time have been improved. Specially, the response time has been 10 times improved. In CCMP, data throughput has been improved about 2 times than the sequential structure design and the response time has been reduced significantly by adopting the parallel structure. *In addition, the response times at WEP, TKIP and CCMP are not dependent on the payload size but only dependent on the clock frequency.*

Table 1. Performances for 1024 bytes payload at 50MHz clock frequency

Category	WEP	TKIP	Category	CCMP
Throughput W/O dual S-Box	160 Mbps	157 Mbps	Throughput in sequential structure	285 Mbps
Throughput with dual S-Box	195 Mbps	195 Mbps	Throughput in parallel structure	562 Mbps
Response Time W/O dual S-box	11.08 μs	11.92 μs	Response Time in sequential structure	14.96 μs
Response Time with dual S-box	0.8 μs	0.8 μs	Response Time in parallel structure	0.88 μs

Table 2 shows the design comparison between our design and a typical one with CCMP designed by a sequential structure and WEP by a single S-Box scheme. The number of logic gates used in our design is as about 2 times as in normal design. But, if the critical response time is 3 μs, the typical design should use 200 MHz clock frequency in TKIP and more than 250 MHz in CCMP to reduce the response time as the payload size becomes larger than 1024 bytes. But, our design can use the lower clock frequency considering the throughput of data, the response time, and the performance constraints specified in 802.11e and 802.11n. We can use 25 MHz clock, which results in about 100 Mbps data throughput and 1.6μs response time in TKIP.

The minimum of the clock frequency achievable in a typical design is as 10 times as our design's. Therefore, even if our design doubles the number of the logic gates, the total power consumption of our design is compatible with a typical design's.

Table 2. Design comparison

TKIP		CCMP	
Category	Logic Used	Category	Logic Used
Without pre-computing S-Box	4225	Sequential design	5565
With pre-computing S-Box	7718	Parallel design	9598

6 Conclusion

In this paper, we propose the cipher engine for IEEE 802.11i that supports WEP, TKIP and CCMP protocols and also supports new MAC mechanisms in other IEEE 802.11 standards. In WEP design, we reduce the response time by adopting the dual S-Box scheme. In AES-CCM module design, we obtain the constant response time by using parallel structured AES-CCM core. So our design can support new mechanisms that include Block Ack in 802.11e and frame aggregation in 802.11n and reduce power consumption because our design has the shorter response time and the throughput of data achievable at the lower clock frequency.

References

1. IEEE P802.11, The Working group for Wireless LANs, http://www.ieee802.org/11/
2. IEEE standard 802.11i, July 2004.
3. D.Whiting, Hifn, R.Housley, N.Ferguson, MacFergus: Counter with CBC-MAC(CCM) RFC 3610, September,(2003)
4. K.H. Tsoi, K.H. Lee and P.H.W Leong,: A Massively Parallel RC4 Key Search Engine. IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM'02) 1082-3409/02 (2002).
5. Jon Edney, William A. Arbaugh: Real 802.11 security: Wi-Fi Protected Access and 802.11i. Addison Wesley, (2003).
6. Wang Shunman, TaoRan, WangYue, ZhangJi: WLAN and It's Security Problems, IEEE (2003).
7. IEEE standard 802.11e/D13.0 January (2005)
8. Matthew S. Gast: 802.11 Wireless Networks – 2nd Edition. O'REILLY, (2005).

Secure Delegation-by-Warrant ID-Based Proxy Signcryption Scheme*

Shanshan Duan, Zhenfu Cao, and Yuan Zhou

Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai 200030, China
{dss, cao-zf, zhouyuan}@sjtu.edu.cn

Abstract. In this paper, we first construct a security model for delegation-by-warrant ID-based proxy signcryption schemes and formalize notions of security for them. To the best of our knowledge, no related work has been done. Then we present such a scheme based on the bilinear pairings, and show that it is provably secure in the random oracle model. Specifically, we prove its semantic security under the DBDH assumption and its unforgeability under the BDH assumption.

1 Introduction

BACKGROUND. The proxy signature primitive was introduced by Mambo, Usuda and Okamoto([1]) and it enjoyed a considerable amount of interest. However, most of the proxy signature schemes were informal and without complete proofs. The first work to formally define a model for them was by Boldyreva, Palacio, and Warinschi([2]). Another new type of cryptographic primitive called 'signcryption' was proposed by Zheng([3]) and a first example of formal security proof in a formal security model was published in 2002([4]). Motivated by the two cryptographic primitives, in 1999, Gamage et al.([5]) extended the proxy signature and introduced a proxy signcryption scheme by combining proxy signature and encryption algorithm. However, this scheme was not under a security model and thus no proof of security was provided. Identity based cryptosystem was introduced by Shamir in 1984([6]). A satisfying ID-based encryption scheme only appeared in 2001([7]), based on bilinear maps over elliptic curves. In this field, what's still missing is a proxy signcryption scheme with guaranteed security.

OUR CONTRIBUTIONS. We define a formal model for the security of delegation-by-warrant ID-based proxy signcryption schemes, which enables security analysis of such schemes. For confidentiality purpose, we define a notion of ciphertext indistinguishability under adaptive chosen-ciphertext attacks. We require that given a signcryption or a proxy signcryption, the adversary can not learn any information about the corresponding plaintext. For non-repudiation purpose, we define a notion of signature unforgeability under chosen-message attacks. We require that the adversary can not create a signcryption or a proxy signcryption for a new message. Then we present a scheme that satisfies our security model's requirement. We show that, in the random

* Partially supported by the National Natural Science Foundation of China for Distinguished Young Scholars under Grant No. 60225007, the National Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20020248024, and the Science and Technology Research Project of Shanghai under Grant No. 04JC14055 and No. 04DZ07067.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 445–450, 2005.
© Springer-Verlag Berlin Heidelberg 2005

oracle model([8]), its security can be proven under the DBDH assumption while its unforgeability can be proven under the BDH assumption.

2 Preliminaries

For the page limit, we only review the computational problems on which our proposed scheme are based.

Definition 1. *Given two groups G_1 and G_2 of the same prime order q, a bilinear map $\hat{e} : G_1 \times G_1 \rightarrow G_2$ and a generator P of G_1,*

- *The $Bilinear\ Diffie-Hellman\ Problem$(**BDH**) is, given (P, aP, bP, cP) for unknown $a, b, c \in F_q$, to compute $\hat{e}(P, P)^{abc}$,*
- *The $Decisional\ Bilinear\ Diffie-Hellman\ Problem$(**DBDH**) is, given a tuple (P, aP, bP, cP) and an element $h \in G_2$, to decide whether $h = \hat{e}(P, P)^{abc}$ or not.*

3 Definitions for Delegation-by-Warrant ID-Based Proxy Signcryption Schemes

The original signer signs a warrant that contains the information regarding a particular proxy signer. We call the signature a certificate. Then the original signer sends the warrant and the certificate to the corresponding proxy signer who can use them as well as its own private key to sign a message on behalf of the original signer. Here ID_{os} and $d_{ID_{os}}$ are denoted as an original signer's public and private key. ID_{ps} and $d_{ID_{ps}}$ are denoted as a proxy signer's public and private key. ID_r and d_{ID_r} are denoted as a receiver's public and private key.

A Delegation-by-warrant ID-based proxy signcryption scheme $IDPWSC$ is a tuple $(\mathcal{G}, \mathcal{K}, (\mathcal{D}, \mathcal{P}), \mathcal{PSC}, \mathcal{PVD}, \mathcal{PNR}, \mathcal{PV})$ defined as follows:

$\mathcal{G}$: The parameter generation algorithm, which takes a security parameter k as input, outputs $params$ and $master-key$. Intuitively, the system parameters will be published while the $master-key$ will be known only to the "Private Key Generator"(PKG).

$\mathcal{K}$: The key generation algorithm, which takes as input $params$, $master-key$ and an identity information $ID \in \{0,1\}^*$, outputs the corresponding private key $d_{ID} = sQ_{ID} = sH_1(ID)$.

$(\mathcal{D}, \mathcal{P})$: (interactive)Proxy-designation algorithms. Each algorithm takes ID_{os}, ID_{ps} and a warrant M_ω set in this delegation as input. $\mathcal{D}$ also takes $d_{ID_{os}}$ as input and $\mathcal{P}$ also takes $d_{ID_{ps}}$ as input. As result of the interaction, the output of $\mathcal{P}$ is a proxy signcryption key SC_p and $\mathcal{D}$ has no local output.

$\mathcal{PSC}$: The proxy signcryption scheme, which takes a proxy signcryption key SC_p, ID_r and a message m as input, outputs the proxy signcryption C_p.

$\mathcal{PVD}$: The proxy unsigncryption algorithm, which takes ID_{os}, ID_{ps}, d_{ID_r} and C_p as input, returns a clear text $m \in \mathcal{M}$ or the symbol $\perp$ if C_p was an invalid ciphertext.

$\mathcal{PNR}$: The proxy non-repudiation algorithm, which takes ID_{os}, ID_{ps}, the key pair $(ID_r, d_{ID_r})$and C_p as input, returns either a tuple (m, σ) or the $\perp$ symbol.

$\mathcal{PV}$: The proxy verification algorithm, which takes ID_{os}, ID_{ps}, the outputs (m, σ) of $\mathcal{PNR}$ and the certificate S_ω as input, outputs 1 if σ is a valid signature for m.

In our scheme, we assume that there is no secure channel between an original signer and a proxy signer. So, when they execute the proxy-designation protocol, an original signer sends to his proxy signer a warrant M_ω and a signature S_ω for it in a public channel. And the proxy signer sets his proxy signcryption key SC_p as $(M_\omega, S_\omega, ID_{os}, d_{IDps})$.

SECURITY NOTIONS. We consider an extreme case that the adversary is working against a single honest user, say ID_1, and can extract the private keys of all the other users. Since any attack can be carried out in the presence of more honest users, such an assumption is no loss of generality. Furthermore, according to the role of the attacked user, the adversary is allowed to issue not only standard signcryption and unsigncryption requests but also proxy signcryption and proxy unsigncrytion requests.

Definition 2. *We say that an $IDPWSC$ scheme has the indistinguishability against adaptive chosen ciphertext attacks property($IDPWSC-IND-CCA$) if no probability polynomial time adversary has a non-negligible advantage in the following game:*

Setup: The challenger generates system parameters by running algorithm $\mathcal{G}$. Then he runs $\mathcal{K}$ to generate the private key d_{ID_1} of the user ID_1. d_{ID_1} is kept secret while the system parameters are given to the adversary $\mathcal{A}$.

Phase 1: The adversary performs the following requests in any order and any number of times:

- Key Extraction requests: Given an identity $ID_i (i \neq 1)$, the challenger returns the private key d_{ID_i} corresponding to ID_i.
- ID_1 Designates ID_i requests: $\mathcal{A}$ requests with ID_1 running $\mathcal{D}(ID_1, ID_i, d_{ID_1})$ for $i \neq 1$, and plays the role of ID_i running $\mathcal{P}(ID_1, ID_i, d_{ID_i})$. After a successful run, the challenger $\mathcal{C}$ generates an warrant M_ω for ID_i, and a signature S_ω for M_ω, then responds (M_ω, S_ω).
- ID_i Designates ID_1 requests: $\mathcal{A}$ requests with ID_1 running $\mathcal{P}(ID_i, ID_1, d_{ID_1})$ for $i \neq 1$ and plays the role of ID_i running $\mathcal{D}(ID_i, ID_1, d_{ID_i})$. After a successful run, the output of $\mathcal{P}$ is the proxy signcryption key of ID_1 on behalf of ID_i and $\mathcal{A}$ can not get access to it.
- ID_1 Designates ID_1 requests: $\mathcal{A}$ requests the user ID_1 to run the designation protocol with itself and see the transcript of the interaction.
- Signcryption query(m, ID_r): $\mathcal{A}$ produces a message m, an arbitrary receiver's identity ID_r and requires the result of the algorithm $\mathcal{SC}(m, d_{ID_1}, ID_r)$.
- Unsigncryption query(c, ID_s): $\mathcal{A}$ produces a ciphertext c, the signer's identity ID_s and requires the result of the algorithm $\mathcal{VD}(c, d_{ID_1}, ID_s)$.
- Proxy signcryption query$(m, ID_{os}, ID_r, S_\omega, M_\omega)$: $\mathcal{A}$ produces a message m, a receiver's identity ID_r and an original sender's identity ID_{os}. If the ID_{os} designated ID_1 request has been issued, the challenger responds by running the algorithm $\mathcal{PSC}$ with the private key d_{ID_1} to signcrypt the message (m, S_ω). Otherwise, the query is said to be invalid and $\perp$ is returned.
- Proxy unsigncryption query$(C_p, S_\omega, ID_{os}, ID_{ps})$: $\mathcal{A}$ produces a ciphertext c_p, a proxy signer's identity ID_{ps}, the corresponding original sender's identity ID_{os} and his signature S_ω on M_ω. The challenger first verifies the signature S_ω, then executes the algorithm $\mathcal{PVD}$ with d_{ID_1}. If the operation is successful, the signed message is returned to $\mathcal{A}$. Otherwise, the $\perp$ symbol is returned as a result.

Challenge: Once the adversary decides that Phase 1 is over, he does as follows:

1. Produce two plaintext m_0, m_1 of equal size and an arbitrary signer's key pair (ID_s, d_{ID_s}). The challenger flips the coin b to compute $c = \mathcal{SC}(m_b, d_{ID_s}, ID_1)$. c is then sent to $\mathcal{A}$ as a challenge.
2. Produce two plaintext m_0, m_1 of equal size, an arbitrary proxy signer's key pair $(ID_{ps}, d_{ID_{ps}})$ and the corresponding original signer's signature S_ω on an appropriate warrant M_ω. The challenger flips the coin b_p to compute a proxy signcryption $c_p = \mathcal{PSC}\left(m_{b_p}, SC_p, ID_1\right)$. c_p is sent to $\mathcal{A}$ as a challenge.

Phase 2: $\mathcal{A}$ performs new queries as in Phase 1, but it may not ask the unsigncryption query of the challenge c or proxy unsigncryption query of the challenge c_p.

Guess: At the end of the game, $\mathcal{A}$ outputs b' or b'_p, and wins if $b = b'$ or $bp = b'_p$.
$\mathcal{A}'s$ advantage is defined to be

$$Adv^{IDPWSC-IND-CCA}_{IDPWSC,\mathcal{A}}(k) = \max\{\Pr[b = b'] - 1/2, \Pr[b_P = b'_P] - 1/2\}.$$

Definition 3. *We say that an $IDPWSC$ scheme is existentially unforgeable against chosen-message attacks($IDPWSC - UF - CMA$) if no probability polynomial time forger $\mathcal{F}$ has a non-negligible success probability in the following game:*

Setup: This phase is handled as in the proof of Definition 2.
Queries: All the oracle queries are handled as in the proof of Definition 2.
Output: Eventually, one of the following cases happens:

1. $\mathcal{F}$ outputs a ciphertext C and (ID_r, d_{ID_r}). If the result of $\mathcal{VD}(C, d_{ID_r}, ID_1)$ is a message m such that C is not the output of a signcryption query (m, ID_r), and $\mathcal{NR}(C, d_{ID_r}, ID_1)$ is (m, σ) such that $\mathcal{V}(m, \sigma, ID_1) = 1$ holds, then 1 is returned. The probability that such case happens is denoted by $Adv_{\mathcal{F},1}(k)$.
2. $\mathcal{F}$ outputs a tuple $(C_p, M_\omega, S_\omega, ID_{os})$ and (ID_r, d_{ID_r}). If the result of $\mathcal{PVD}$ is m such that C_p is not the output of a proxy signcryption query $(m, ID_{os}, ID_r, S_\omega, M_\omega)$ and the output of $\mathcal{PNR}$ is (m, σ_p) such that $\mathcal{PV}(ID_{os}, ID_1, m, S_\omega, \sigma_p) = 1$ holds, then 1 is returned. The probability that such case happens is denoted by $Adv_{\mathcal{F},2}(k)$. [forgery of a proxy signcryption by ID_1 on behalf of ID_{os}].
3. $\mathcal{F}$ outputs a tuple $(C_p, M_\omega, S_\omega, ID_1)$, ID_{ps} and (ID_r, d_{ID_r}). If the output of $\mathcal{PNR}$ is (m, σ_P) such that $\mathcal{PV}(ID_1, ID_{ps}, m, S_\omega, \sigma_P) = 1$ holds and the ID_1 designates ID_{ps} request has never been issued, then 1 is returned. The probability that such case happens is denoted by $Adv_{\mathcal{F},3}(k)$.[forgery of a proxy signcryption by ID_{ps} on behalf of ID_1].
4. Otherwise, $\mathcal{F}$ returns 0.

We define the success probability of forger $\mathcal{F}$ as

$$Adv^{IDPWSC-UF-CMA}_{IDPWSC,\mathcal{F}}(k) = \max\{Adv_{\mathcal{F},1}(k), Adv_{\mathcal{F},2}(k), Adv_{\mathcal{F},3}(k)\}.$$

4 A Delegation-by-Warrant ID-Based Proxy Signcryption Scheme

We present our scheme based on the bilinear pairings.

$-\mathcal{G}$: Assume k is a security parameter. G_1, G_2 are GDH groups of prime order $q > 2^k$(with G_1 additive and G_2 multiplicative). P is a generator of G_1 and $\hat{e}$:

$G_1 * G_2 \rightarrow G_2$ is a bilinear map. Picks a random master key $s \in F_q^*$ and computes $P_{pub} = sP$. Chooses hash functions $H_1 : \{0,1\}^* \rightarrow G_1$, $H_2 : G_2 \rightarrow F_q^*$, $H_3 : G_1 \times G_1 \times G_2 \rightarrow F_q^*$, $H_4 : \{0,1\}^n \times F_q^* \rightarrow \{0,1\}^m$, $H_5 : G_1 \times G_1 \times \{0,1\}^n \times F_q^* \rightarrow \{0,1\}^m$, $H_6 : \{0,1\}^* \times \{0,1\}^n \times G_1 \rightarrow G_1$(The plaintexts must have a fixed bitlength of n with $m + n < k \approx \log_2^q$.) The system's public parameters are $params = \{G_1, G_2, \hat{e}, P, P_{pub}, H_i(i = 1, ..., 6)\}$.

$-\mathcal{K}$: Given an identity ID, the PKG computes $Q_{ID} = H_1(ID)$ and the private key $d_{ID} = sQ_{ID}$.

$-(\mathcal{D}, \mathcal{P})$: In order to designate ID_{ps} as a proxy signer, the original signer ID_{os} first makes a warrant M_ω and publishes it. Then ID_{os} uses the following scheme to produce a signature on M_ω

1. Randomly pick $r_\omega \in F_q^*$, compute $U_\omega = r_\omega P \in G_1$, $H_\omega = H_6(ID_{os}, M_\omega, U_\omega)$
2. Compute $V_\omega = d_{ID_{os}} + r_\omega H_\omega$

The signature on M_ω is $\langle U_\omega, V_\omega \rangle$.

After receiving the proxy certificate $(U_\omega, V_\omega, M_\omega)$, the proxy signer verifies $\langle U_\omega, V_\omega \rangle$ on M_ω by first taking $Q_{ID_{os}} = H_1(ID_{os})$ and $H_\omega = H_6(ID_{os}, M_\omega, U_\omega)$, then checking if $\hat{e}(P, V_\omega) = \hat{e}(P_{pub}, Q_{ID_{os}})\hat{e}(U_\omega, H_\omega)$ holds. If it does, the proxy signer computes his proxy signcryption key as $(M_\omega, U_\omega, V_\omega, d_{IDps})$.

$-\mathcal{SC}$: To signcrypt a message m, the sender performs the following steps:

1. Compute $Q_{IDr} = H_1(ID_r)$
2. Randomly pick $x \in F_q^*$, and compute $k_1 = H_2(\hat{e}(P, P_{pub})^x)$, $k_2 = H_3(Q_{IDs}, Q_{IDr}, \hat{e}(P_{pub}, Q_{IDr})^x)$
3. Compute $r = (M||H_4(M, k_2))k_1k_2 \bmod q$, $s = xP_{pub} - rd_{IDs}$

The ciphertext is $\sigma = (r, s)$.

$-\mathcal{VD}$: When receiving $\sigma = (r, s)$, the receiver ID_r performs the following tasks:

1. Compute $Q_{IDs} = H_1(ID_s)$
2. Compute $k_1 = H_2(\hat{e}(P, s)\hat{e}(P_{pub}, Q_{IDs})^r)$, $f = \hat{e}(s, Q_{IDr})\hat{e}(Q_{IDs}, d_{IDr})^r$ and $k_2 = H_3(Q_{IDs}, Q_{IDr}, f)$
3. Compute $M' = r(k_1k_2)^{-1} \bmod q$
4. Take M_1 as the first n bits of M'. If $(M_1||H_4(M_1, k_2))$ are the first $m + n$ bits of M', accept the plaintext M_1

$-\mathcal{PSC}$: To signcrypt a message M on behalf of the original signer ID_{os}, the proxy signer ID_{ps} performs:

1. Compute $Q_{IDr} = H_1(ID_r)$
2. Randomly pick $x \in F_q^*$, and compute $k_1 = H_2(\hat{e}(P, P_{pub})^x)$ and $k_2 = H_3(Q_{IDps}, Q_{IDr}, \hat{e}(P_{pub}, Q_{IDr})^x)$
3. Compute $r = (M||H_5(U_\omega, V_\omega, M, k_2))k_1k_2 \bmod q$, $s = xP_{pub} - rd_{IDps}$

The ciphertext is $(M_\omega, U_\omega, V_\omega, r, s)$.

$-\mathcal{PVD}$: When receiving $C_p = (M_\omega, U_\omega, V_\omega, r, s)$, the receiver ID_r performs the following tasks:

1. Check if $\hat{e}(P, V_\omega) = \hat{e}(P_{pub}, Q_{ID_{os}})\hat{e}(U_\omega, H_\omega)$ holds. If the condition does not hold, reject the ciphertext.
2. Compute $Q_{IDps} = H_1(ID_{ps})$ and $k_1 = H_2(\hat{e}(P, s)\hat{e}(P_{pub}, Q_{IDps})^r)$
3. Compute $f = \hat{e}(s, Q_{IDr})\hat{e}(Q_{IDps}, d_{IDr})^r$, $k_2 = H_3(Q_{IDps}, Q_{IDr}, f)$, $M' = r(k_1 k_2)^{-1} \bmod q$
4. Take M_1 as the first n bits of M', accept the plaintext M_1 if and only if $(M_1||H_5(U_\omega, V_\omega, M_1, k_2))$ are the first $m + n$ bits of M'

$-\mathcal{PNR}$: (proxy non-repudiation algorithm) If the receiver wants to prove to a third party that the proxy sender has signed a message m, he forwards $(M, \sigma) = (M, M_\omega, U_\omega, V_\omega, f, r, s)$ by computing $f = \hat{e}(s, Q_{IDr})\hat{e}(Q_{IDps}, d_{IDr})^r$.

$-\mathcal{PV}$: (proxy verification algorithm) When the third party receives (M, σ), he performs similar steps in the $\mathcal{PVD}$ algorithm.

Theorem 1. *In the random oracle model, if an adversary $\mathcal{A}$ has a non-negligible advantage against the IDPWSC-IND-CCA security of the above scheme, then there exists a PPT distinguisher $\mathcal{C}$ that can solve the Decisional Bilinear Diffie-Hellman Problem.*

Proof. Omitted here for the page limit.

Theorem 2. *In the random oracle model, if a forger $\mathcal{F}$ has a non-negligible success probability against the IDPWSC-UF-CMA security of the above scheme, then there exists a PPT algorithm $\mathcal{C}$ that can solve the Bilinear Diffie-Hellman Problem.*

Proof. Omitted here for the page limit.

References

1. Mambo, M., Usuda, K., Okamoto,E.: Proxy signatures for delegating signing operation. In Proceedings of the 3rd ACM Conference on Computer and Communication Security(CCS), ACM (1996) 48-57
2. Boldyreva, A., Palacio, A., Warinschi, B.: Secure Proxy Signature Scheme for Delegation of Signing Rights, IACR ePrint Archive, available at http://eprint.iacr.org/2003/096/ (2003)
3. Zheng,Y.: Digital Signcryption or how to achieve cost(signature&encryption)≪ cost(signature)+cost(encryption). In: B.S. Kaliski Jr. (Ed.): Advance in Cryptology – CRYPTO'97, Lecture Notes of Computer Science, Vol. 1294. Springer-Verlag, Berlin Heidelberg New York (1997) 165-179
4. Baek, J., Steinfeld, R., Zheng, Y.: Formal Proofs for the Security of Signcryption. In: Naccache, D., Paillier, P. (ed.): PKC 2002. Lecture Notes in Computer Science, Vol. 2274. Springer-Verlag, Berlin Heidelberg New York (2002) 81-98
5. Gamage, C., Leiwo, J., Zheng, Y.: An efficient scheme for secure message transmission using proxy-signcryption. In: Edwards J. (ed.): Proceedings of the 22th Australasian Computer Science. Aukland:Springer-Verlag, Berlin Heidelberg New York (1999) 420-430.
6. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakley, G.R., Chaum, D. (ed.): Advance in Cryptology - CRYPTO 1984. Lecture Notes in Computer Science, Vol. 196. Springer-Verlag, Berlin Heidelberg New York (1984) 47-53
7. Boneh, D., Franklin, M.: Identity-based encryption from the Weil Paring. In: Kilian, J. (ed.): Advances in Cryptology - CRYPTO 2001. Lecture Notes in Computer Science, Vol. 2139. Springer-Verlag, Berlin Heidelberg New York (2001) 213-229
8. Bellare, M., Rogaway, P.: Random oracles are practical: A paradigm for designing efficient protocols. In Proc.Conf Computer and Communication Security (1993)

Building Security Requirements Using State Transition Diagram at Security Threat Location

Seong Chae Seo[1], Jin Ho You[1], Young Dae Kim[1], Jun Yong Choi[2], Sang Jun Lee[3], and Byung Ki Kim[1]

[1] Department of Computer Science, Chonnam National University, 300 Yongbong-dong, Buk-gu, Gwangju 500-757, Korea
{scseo, jhyou, utan, bgkim}@chonnam.ac.kr
[2] School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu 702-701, Korea
jychoi@ee.knu.ac.kr
[3] Department of Internet Information Communication, Shingyeong University, 1485 Namyang-dong, Hwaseong-si, Gyeonggi-do 445-852, Korea
aura88@empal.com

Abstract. The security requirements in the software life cycle has received some attention recently. However, it is not yet clear how to build security requirements. This paper describes and illustrates a process to build application specific security requirements from state transition diagrams at the security threat location. Using security failure data, we identify security threat locations which attackers could use to exploit software vulnerabilities. A state transition diagram is constructed to be used to protect, mitigate, and remove vulnerabilities relative to security threat locations. In the software development process, security requirements are obtained from state transition diagrams relative to the security threat location.

1 Introduction

The security engineering in the software development process has received some attention recently. The security incidents attacking software vulnerabilities have been increasingly apparent. The various kinds of threats that may be mounted by attackers do harm to assets (e.g., data, services, hardware, and personnel). Software which is not designed with security is vulnerable [2]. Attackers are becoming more malicious, and use more sophisticated attack technology than previously experienced. Therefore, software developers have recognized the need to identify security requirements at an early phase of the software development process [4, 6, 11, 12].

The security technology which protects, mitigates, and removes the vulnerabilities can be divided into two classes: software security and application security [6]. Application security is related to protecting software and the systems that the software runs in a post facto way, after the development is completed. In spite of application security, the number of security vulnerabilities are growing.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 451–456, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Security vulnerabilities are the result of bad software design and implementation [12]. Software security is about building secure software [6]. It is always better to design for security from scratch than to try to add security to an existing design. Some researchers [1, 5, 8, 9] have contributed to the body of work related to developing secure software from scratch. They identified security goals and requirements in the analysis phase and have provided methods to remove vulnerabilities while the software is being designed.

2 Related Work

2.1 Vulnerability Analysis

Vulnerabilities involve code bugs or design errors. They have originated from misunderstanding and the inconsistency of security requirements [13]. Information systems being built and managed today are prone to the same or similar vulnerabilities that have plagued them for years [1]. Bishop [3] describes that similar vulnerabilities are classified similarly and every vulnerability has a unique, sound characteristic(set of minimal size). He illustrates an approach that classifies, corrects, and removes security vulnerabilities from the software.

2.2 The Existing Method of Identification of Security Requirements

Identifying Security Requirements Using Extension of Existing Model. Some researchers [7, 10, 15] illustrate an approach that is an extension to the syntax and semantic of the Unified Modeling Language (UML) which models security properties of software. *Misuse cases* [7] or *abuse cases* [10] are specialized kinds of *use cases* that are used to analyze and specify security threats. UMLsec [15] is a UML profile for designing and modeling security aspects of software systems.

Identifying Security Requirements Using Ani-Moel. Some researchers illustrate an approach that models, specifies, and analyzes application specific security requirements using a goal-oriented framework for generating and resolving obstacles to goal satisfaction. Threat trees are built systematically through anti-goal refinement until leaf nodes are derived that are software vulnerabilities observable by the attacker. Security requirements are then obtained as countermeasures [1, 2, 13].

Identifying Security Requirements Using CC(Common Criteria). Common Criteria(CC) [4] is a scheme for defining security requirements and evaluating products to check if products meet the requirements.

Security Pattern. A security pattern [14] describes a particular recurring security problem that arises in specific contexts and presents a well-proven generic scheme for its solution.

3 Modeling State Transition Diagrams at the Security Threat Location

Developers involved in the software development process have to learn the principles for building secure software. It is better to design for security from scratch in the software development process [12]. The activities of identification of security requirements begin with the development of state transition diagrams, as shown on the left in Fig. 1.

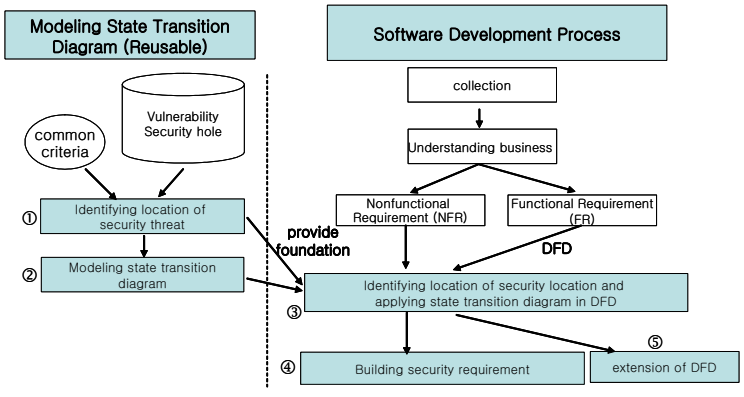


Fig. 1. Building Security Requirements at the Security Threat Location

3.1 Identifying the Location of the Security Threat

To build secure software, we need to document attack information (security failure data, etc) in a structured and reusable form. We first find the location of the security threat which could be exploited by attackers. We use security failure data to improve the security of software and particularly to identify the location of security threats.

We use abstracted DFD to identify the location of the security threat. Then we classify security failure data into equivalence classes using the equivalence relation and vulnerabilities analysis of Bishop [3]. We found the set of equivalence classes: user input process, network process and data store process. The location of the security threat is one of the equivalence classes.

3.2 Modeling the State Transition Diagram

Attackers exploit security vulnerabilities in software, which causes the change of assets in software. Therefore, we need to trace the change of state of assets. We use state transition diagrams of Unified Modeling Language (UML) to represent the change of state of assets at the security threat location. The state transition diagram should lead to the derivation of robust security requirements as anticipated countermeasures to such threats.

The procedure used for modeling state transition diagrams systematically is as follows:

1. Build sets of security vulnerabilities at specific locations of software security threats.
2. For each location, build a set of characteristics of a vulnerability, which is a condition that must hold for the vulnerability to exist (A more technical definition may be found in [3]).
3. Model a state in the state transition diagram for the commonality of the character set.
4. Document the event in the state transition diagram for the condition which triggers the character set.
5. Elaborate the state transition diagram. If the event and guard are correct, then a secure state is achieved; otherwise, a security violation state is achieved. The building process repeats steps 2 and 3 until it reaches a final state.

Each security threat location is modeled as a state transition diagram of user input processes, network connected processes (server side, client side), data stored processes, etc. This paper describes a state transition diagram that models the state of assets at user input processes but does not describe others.

4 Building Security Requirements

The process of software development deals with functional requirements and non-functional requirements. In this paper, we focus on the security requirements in the non-functional requirements. The activities of building security requirements are shown on the right of Fig. 1.

4.1 Identifying Location of Security Threat in DFD

The functional requirements of the software are decomposed into a DFD (Data Flow Diagram). We find the assets vulnerable to security accidents. The assets of software consist in data and service. Then, we mark the DFD with graphical notation on the security threat location(① to ⑨ in Fig. 2). The identified locations are: the class of the user input process(as shown as ①③⑨ in Fig. 2), the class of the network process(as shown as ② in Fig. 2), and the class of the data store process(as shown as ④⑤⑥⑦⑧ in Fig. 2).

4.2 Applying the State Transition Diagram

We now apply the state transition diagram as identified in Sect. 3.2 to the security threat locations, as shown in Fig. 2. We identify processes which handle valuable assets in our DFD. This is represented by ❶❷❸ in Fig. 2. These processes have more than one state transition diagram. If the process has more than two state transition diagrams, the order of applying the state transition diagrams corresponds to the order of data flow of processes within the DFD.

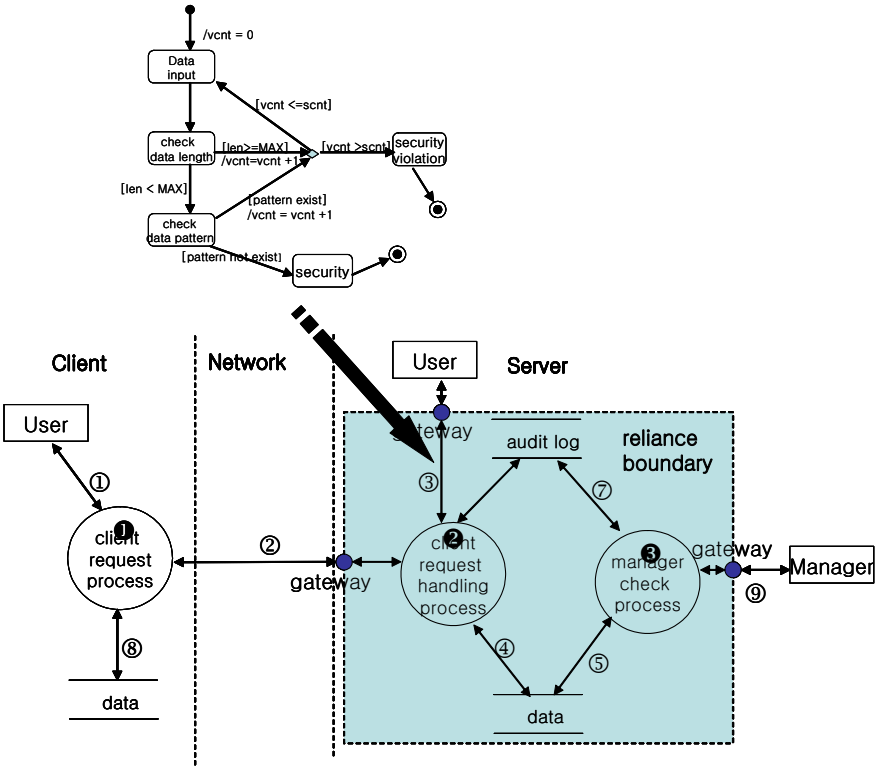


Fig. 2. Applying State Transition Diagram

Process ❶ in Fig. 2 has three state transition diagrams: the user input process, the network process (client side), and the data store process. The order of applying the state transition diagrams for process ❷ in Fig. 2 is network process (server side), user input process access control, and data store process. Figure 2 shows how to apply the state transition diagram of the user input process to the location ③ of process ❷ in Fig. 2.

4.3 Building the Security Requirements

In building the security requirements, we will make use of the change of state of valuable assets to be protected. Concrete values must be written into the state transition diagram of processes in the DFD. The rule for building the secuirty requirements is as following:

In the state transition diagram,

- a guard of state change creates a security requirement
- an event of state change creates a security requirement.

5 Conclusions

The security requirements in the software life cycle has received some attention recently to build secure software. It is recommended to design for security from scratch, rather than to try to add security to an existing design, and to identify the security requirements in an earlier phase of development. This paper describes and illustrates a process to build application specific security requirements from state transition diagrams at the security threat location. By using the reusable state transition diagrams, the security requirements can be identified easily at an early phase of software development. Additionally, the proposed process provides a cost-effective security analysis.

References

1. Moore, A. P., Ellison, R. J., Linger, R. C.: Attack modeling for information security and survivability. Technical Report CMU/SEI-2001-TN-001 (2001)
2. Lamsweerde, A.: Elaborating security requirements by construction of intentional anti-models. In: Proceedings of the 26th International Conference on Software Engineering(ICSE'04) (2004)
3. Bishop, M.: Vulnerabilities analysis. In: Web proceedings of the 2^{nd} International Workshop on Recent Advances in Intrusion Detection (RAID'99) (1999)
4. Common criteria for information technology security evaluation, Version 2.1. CCIMB-99-031 (1999)
5. Firesmith, D.: Specifying reusable security requirements. Journal of Object Technology **3** (2004)
6. McGraw, G.: Software security. IEEE Security & Privacy **2** (2004) 80–83
7. Alexander, I.: Misuse cases: Use cases with hostile intent. IEEE Software **20** (2003) 58–66
8. Krsul, I. V.: Computer vulnerability analysis. PhD thesis, Purdue University (1998)
9. McDermott, J.: Extracting security requirements by misuse cases. In: Proc. 27^{th} Technology of Objected-Oriented Languages and Systems (2000) 120-131
10. McDermott, J., Fox, C.: Using abuse case models for security requirements analysis. In: Proc. Annual Computer Security Applications Conference (ACSAC'99) (1999)
11. Whittacker, J. A., Howard, M.: Building more secure software with improved development processes. IEEE Security & Privacy **2** (2004) 63–65
12. Viega, J., McGraw, G.: Building secure software. Addison-Wesley (2004)
13. Howard, M., LeBlanc, D. C.: Writing secure code, Second edition. Microsoft (2003)
14. Schumacher, M., Roedig, U.: Security engineering with patterns. In: PLoP Proceedings (2001)
15. Jurjens, J.: UMLsec: Extending UML for secure systems development. In: UML 2002 (2002)

Study on Security iSCSI Based on SSH

Weiping Liu and Wandong Cai

Department of Computer Science & Engineering, Northwestern Polytechnical University, Xi'an 710072, China
sam2008cn@163.com

Abstract. The iSCSI protocol is becoming an important protocol to enable remote storage access through the ubiquitous TCP/IP networks. This paper analyzes the security and performance characteristics of the iSCSI protocol, points out the limitation of the security iSCSI scheme based on IPSec, and presents the security iSCSI scheme based on SSH. With application of SSH port forwarding, a secure tunnel can be built in TCP layer to ensure the security of iSCSI session. Experiments show that throughput of the security iSCSI based on SSH rises up 20% and CPU utilization greatly lowers 50% with the same encryption algorithm, compared with the security iSCSI based on IPSec. So the performance of the security iSCSI based on SSH is obviously superior to the one based on IPSec.

1 Introduction

For scalability, ease of management, cost-effectiveness, and higher performance, enterprise storage is increasing organized as distributed storage, usually in the form of Storage Area Networks (SANs). SANs are storage solutions built around block-addressed storage units located on a dedicated high-speed network, usually Fibre Channel (FC). But the connection distance of FC is limited to about 10KM, which can not meet the transmission distance requirements for some applications.

IP networks become more stable and ubiquitous. Thus there has been growing interest in carrying storage traffic across TCP/IP infrastructure to exploit the ubiquity and reduce cost. The Internet SCSI (iSCSI) protocol [1], developed by IETF, defines a means to enable block storage accesses over TCP/IP networks. By carrying SCSI commands and storage traffic over TCP/IP networks, iSCSI is used to facilitate data transfers and storage management over long distances. Combined with Gigabit and 10 Gigabit Ethernet transports, IP security and quality of service protocols, iSCSI opens new opportunities for highly scalable and secure shared storage networks.

While iSCSI makes use of the advantages of TCP/IP networks, it inherits their insecure characteristics as well. In RFC of iSCSI, IETF recommends IPSec to be used for iSCSI security. But researches show that IPSec could obviously reduce system throughput as well as greatly increase CPU utilization [2].

This paper first analyzes the security and performance characteristics of iSCSI, then describes a security iSCSI scheme based on SSH, and tests and proves the superiority of this scheme in performance through experiments.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 457 – 462, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 iSCSI Protocol and Security Analysis

The iSCSI protocol consists of an iSCSI initiator and an iSCSI target device. The initiator generates a SCSI command request for a target that can be reachable via an IP network. The steps show as follows: (1) An application residing on the initiator machine starts the transaction by requesting some data from or by sending some data to the iSCSI target device in the iSCSI target machine. (2) The operating system then generates SCSI command and date requests after receiving the raw data from the application. The commands and data are then encapsulated in iSCSI PDU and sent over an Ethernet connection. (3) The iSCSI PDU traveling over the IP network reaches the destination, where it is encapsulated separated SCSI command or data. (4) The SCSI command or data are then sent to the SCSI storage device. (5) The response to the above request will then be sent to the iSCSI initiator in a similar way. (6) The entire process happens through one or more TCP connections between the initiators and the target.

2.1 iSCSI Performance

TCP/IP is a protocol based on stream, which has large difficulty in implementing zero copy. Thus data have to read and write more than twice in memory, so system workloads are relatively heavy. In addition, the size of fragment of different protocols is different. The size of common fragment of Ethernet is 1.5K. While I/O package from storage applications usually ranges from 4K to 64K, which must be unpacked into smaller fragments to be able to transport on physical Ethernet network. When reaching the destination, they must be packed into I/O package with original size. The operation on package packing and unpacking would sharply increase system workloads. Therefore, the characteristics of TCP/IP prevent the further increase in performance of iSCSI.

2.2 iSCSI Security

In an iSCSI-based storage system, the storage data user (initiator) and storage device (target) may be separated physically. They communicate with each other through IP networks. Both the control and sensitive data packets are vulnerable to attack. The same attack in a normal IP network can be applied in an iSCSI environment. iSCSI security protocol must support confidentiality, data origin authentication and integrity on a per-packet basis [3].

iSCSI login phase provides mutual authentication of the iSCSI endpoints. An initiator must first establish an iSCSI session before it can conduct data access. A session is composed of one or more TCP connections. In the first TCP connection setup phase, a login phase begins. During this login phase, both participated parties need to exchange information for authenticating each other, negotiate the session's parameters, open security association protocol, and mark the connection as belonging to an iSCSI session. IETF recommends Secure Remote Password (SRP), Challenge Handshake Authentication Protocol (CHAP) to be used for iSCSI login management [3].

Although iSCSI supports authentication at stage of session building, iSCSI does not define authentication, integrity and confidentiality mechanisms per-packet basis. So the problem in need of solution is to ensure the security of iSCSI PUD in the phase of transmission. Many existing security schemes can be used with iSCSI, such as IPSec [4],[5] and Secure Sockets Layer (SSL) [6]. However, each scheme has its advantages and disadvantages.

IETF recommends IPSec to be used for iSCSI security. IPSec can protect any protocol running above IP. However, operations on packing and unpacking of storage data in TCP/IP have brought large system workloads. If IPSec fulfills encryption and authentication security mechanisms in IP layer, it means each IP package needs extra process, which would induce large system workloads and further lower the performance of iSCSI. But in this scheme IPSec is a separated part of iSCSI. The security handling in iSCSI is left to the underlying IPSec. iSCSI driver does not need to implement any complicatedly security-related protocol and the iSCSI application does not need to touch the underlying IPSec security service. This separation simplifies the iSCSI implementation [2].

SSL protocol has been universally accepted on the WWW for authenticated and encrypted communication between clients and servers. SSL provides secure communication between client and server in the transport level by allowing mutual authentication, digital signatures for integrity, and encryption for privacy. But SSL needs to be tightly incorporated into iSCSI driver. The iSCSI commands and data message need to be encrypted by SSL before delivering to TCP layer. Therefore, it need to implement SSL in iSCSI driver, it increase the difficulty of system implementation.

3 Security iSCSI Based on SSH

Secure Shell (SSH) [7] is an open protocol used to secure network communication between two entities. SSH connections provide highly secure authentication, encryption, and data integrity to combat security threats. SSH uses client/server architecture in its implementation. Fig.1 shows the SSH process: (1) The SSH client provides authentication to the SSH server. In the initial connection, the client receives a host key of the server; therefore, in all subsequent connections, the client will know it is connecting to the same SSH server. (2) The SSH server determines if the client is authorized to connect to the SSH service by verifying the username/password or public key that the client has presented for authentication. This process is completely encrypted. (3) If the SSH server authenticates the authorized client, the SSH session begins between the two entities. All communication is completely encrypted.

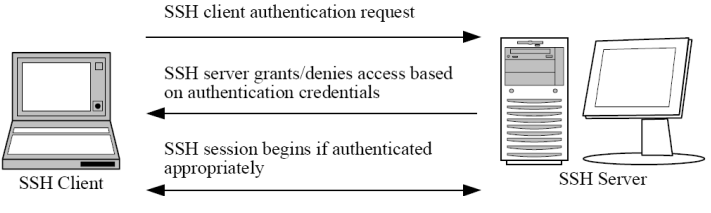


Fig. 1. SSH communication from an SSH client to an SSH server

SSH provides three main capabilities: secure command-shell, secure file transfer, and port forwarding. SSH port forwarding is a powerful tool that can provide security to TCP/IP applications. After port forwarding has been set up, SSH reroutes traffic from a program (usually a client) and sends it across the encrypted tunnel, then delivers it to a program on the other side (usually a server).

3.1 Security iSCSI Based on SSH

Normally, the iSCSI initiator establishes TCP connections with the port A of the iSCSI target, as Fig.2 (a) shows. Making use of SSH port forwarding, we turns the TCP connect between iSCSI initiator and target into a secure tunnel built between SSH client and SSH server, as Fig.2 (b) shows. First of all, on the iSCSI initiator computer, bind a SSH port B with iSCSI target port A. It mean that The SSH client sends a request for port forwarding to the SSH server (running on the same host with iSCSI target). The request contains information about what target port the application connection shall be forwarded to. Then the iSCSI initiator is reconfigured to connect to the port B of the SSH client, instead of connecting directly to the port A of the iSCSI Target. When the iSCSI initiator sends data to the SSH client, data is forwarded to the SSH server over the secure encrypted connection. The SSH server decrypts the data and forwards it to the iSCSI Target. Any data sent from the iSCSI Target is forwarded to the iSCSI initiator in the same manner.

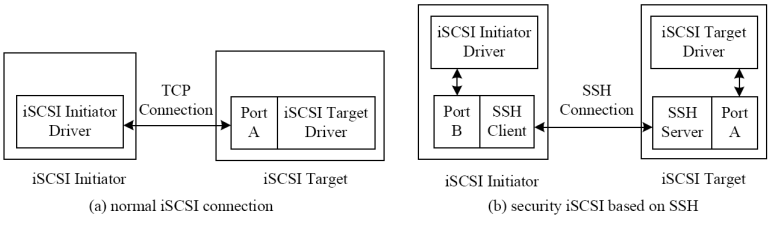


Fig. 2. Comparison between common iSCSI and iSCSI based on SSH

SSH fulfill security mechanism in TCP layer. In this way, the workloads of encrypt data package are much less than that of encrypt data package in IP layer. Therefore, it reduces system workloads resulting from security mechanism in the condition of ensuring equal security performance.

4 Test Environment and Performance Evaluation

This paper uses RedHat 8.0 with Linux kernel 2.4.18-7 as operating system platform, OpenSSH to enable SSH. UNH-iSCSI 1.5.04 as a base iSCSI implementation is used. Both initiator and target link up directly with 100Mbps Ethernet Switcher. Initiator is equipped with 2.4G Intel Pentium4 CPU and 512M memory and target with 1G Intel Pentium3 CPU and 1G memory. The network interface cards on these machines are RealTek RTL8139 PCI Ethernet adapters with 10/100 Mbps.

The experiment makes use of SSH port forwarding to build secure tunnel between the iSCSI initiator and the iSCSI target. Thereby, the whole iSCSI session is encrypted in TCP layer. Besides, the experiment also uses IPSec to encrypt data

package in IP layer in the same environment. In order to fairly compare, 3DES encryption algorithm is used in two security mechanisms.

4.1 Performance Evaluation

Considering iSCSI is a block transport technology, this paper uses IOmeter by Intel corporation to test the performance and CPU utilization of secure iSCSI, and sets each read operations and write operations occupying 50%, each sequence and random occupying 50% and the size of block ranging from 1k to 1024k.

Comoparing with iSCSI without security mechanism, the throughput of security iSCSI based on IPSec obviously falls approximately 40% and of security iSCSI based on SSH decreases obviously approximately 20%, which can be seen from Fig.3. Besides, CPU utilizations evidently increase in two security schemes, knowing from Fig.4. In iSCSI based on IPSec, the 100% rise in CPU utilization, while in iSCSI based on SSH, the rise reaches 50%. Thereby, we think that iSCSI based on TCP/IP is an obvious CPU intensive operation. Such practices as packing and unpacking of IP package have occupied lots of system resources, which will result in high rise in system CPU utilization. If certain security mechanism is introduced to TCP/IP, it will further escalate occupation in system resources and decrease of iSCSI performance. This is very obvious in security iSCSI based on IPSec. Therefore, security iSCSI based on IPSec implemented by hardware leaves the transaction special for packet to specialized processor and releases system resources for normal practice transaction. But this will cause steep rise in system cost. Comparing with the security iSCSI based on IPSec, I/O performance of the security iSCSI base on SSH can rise by 20%, while CPU utilization can fall by 50%. So, it has obvious advantages.

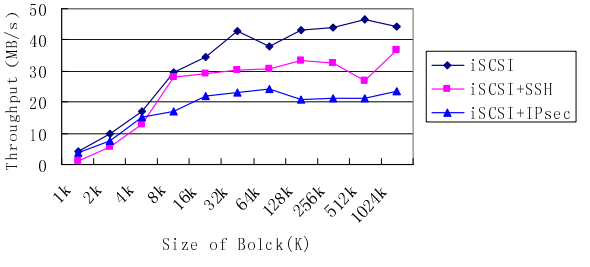


Fig. 3. Throughput comparison between iSCSI, iSCSI+SSH and iSCSI+IPSec

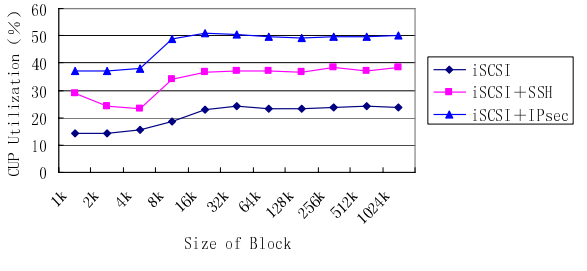


Fig. 4. CPU utilization comparison between iSCSI, iSCSI+SSH and iSCSI+IPSec

Considering iSCSI software implementation used in experiment, it has no optimization for performance. In reference [8], author presents raising iSCSI performance by adopting cache technology. So, we will carry out special optimize for performance particularly for iSCSI software implementation, study iSCSI security mechanism in Gigabit Ethernet network environment and expect to further improve security iSCSI performance.

5 Conclusions

iSCSI transports storage traffic over TCP/IP networks, it is very convenient to fulfill such applications as network storage, remote backup, remote mirror, data migration, duplicate and etc. While the characteristic also induces possibility that storage data is exposed to the public or attacked maliciously. So, effective security mechanism is an important component of iSCSI implementation. In RFC of iSCSI, IETF recommends IPSec to be used for iSCSI security. However, IPSec can sharply dilute iSCSI performance according to related studies. This paper presents the scheme of security iSCSI based on SSH. By comparing with the scheme based on IPSec, the advantages in performance of this scheme are obvious.

References

1. Satran, J., Smith, etc., Internet Draft, iSCSI Specification, http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-20.txt, Jan. (2003)
2. Shuang-Yi Tang, Ying-Ping Lu, and David H.C Du,: Performance study of Software-Based iSCSI Security, Proceedings of the First International IEEE Security in Storage Workshop, (2003)
3. Internet Draft, Securing Block Storage Protocols over IP, http://www.ietf.org/internet-drafts/draft-ietf-ips-security-14.txt, July (2002).
4. S. Kent, and R. Atkinson: IP Authentication Header, IETF RFC 2402, (1998)
5. S. Kent, and R. Atkinson: IP Encapsulating Security Payload, IETF RFC 2406, (1998)
6. Mraz, R., Secure blue: An Architecture for a Scalable, Reliable High Volume SSL Internet Server, Proceedings 17th Annual Conference on Computer Security Applications, (2001)
7. Ylonen, etc., SSH Protocol Architecture, http://www.ietf.org/internetdrafts/draft-ietf-secsh-architecture-07.txt., (2001)
8. Xubin He, Qing Yang, and Ming Zhang: A Caching Strategy to Improve iSCSI Performance, IEEE Annual Conference on Local Computer Networks, (2002)

A Scheduling Algorithm Based on a Trust Mechanism in Grid*

Kenli Li, Yan He**, Renfa Li, and Tao Yang

School of Computer and Communications, Hunan University, Changsha 410082, China
heheyan@126.com, lkl510@263.net

Abstract. Trust has been recognized as an important factor for scheduling in Grid. With a trust-aware model, task scheduling is crucial to achieving high performance. In this paper, a trust model adapted to the Grid environment is proposed and a scheduling algorithm based on the trust mechanism is developed. In the trust model, trust value is computed based on the execution experiences between users and resources. Based on the trust model Min-min algorithm are enhanced to ensure the resource and application security during the scheduling. Simulation results indicate that the algorithm can remarkably lessen the risks in the task scheduling, improve the load balance and decrease the completion time of tasks, therefore it is an efficient scheduling algorithm in trust aware Grid system.

1 Introduction

Grid[1] computing systems has attracted the attentions of research communities in recent years due to the unique ability of marshalling collections of heterogeneous computers and resources, enabling easy access to diverse resources and services[2]. Lots researches have been developed on the scheduling algorithm in Grid[3]. However, existing scheduling algorithms largely ignore the security induced risks involved in dispatching tasks to remote sites. In Grid computing systems, two major security scenarios are often encountered[4]. First, user programs may contain malicious codes that may endanger the Grid resources used. Second, shared Grid resources once infected by network attacks may damage the user applications running on the Grid.

As the resource nodes have the ultimate power of controlling the execution of the user program or task request, protection of user program remains a difficult problem[5]. The aim of the current security solutions is to provide the Grid system participants with certain degree of trust that their resource provider node or user program will be secure[6]. Trust is the *firm belief* in the competence of an entity to act as expected[7]. The *firm belief* is a dynamic value subject to the entity's behavior and applies within a specific context at a given time, and spans over a set of values ranging from *very trustworthy* to *very untrustworthy*. When making trust-based decisions, entities can rely on others for information pertaining to a specific entity.

There have been some previous research efforts related to our study. Abawajy[8] suggested a new scheduling approach to provide fault-tolerance to task execution in a

* This work is supported by the Key (Key grant) Project of Chinese Ministry of Education. NO. 105128.

** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 463 – 468, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Grid environment with high overhead for copying tasks into some resources. Azzedin[9] developed a security-aware model between resource providers and consumers and enhanced the scheduling algorithms based on the security model. A fuzzy inference approach to consolidate security measures for trusted Grid computing is introduced by Song[10]. The trust model was extended from security awareness to security assurance and the risky influences were tested for Min-min and Sufferage heuristics[4].

In this paper we present a trust model wherein each node is assigned a trust value that reflects the transaction experiences of all accessible nodes in Grid. Then TrustMin-Min algorithms are proposed to secure the tasks and resources, decrease the security overhead, and improve the completion time of the tasks.

This paper is organized as follows. The trust model is analyzed in Section 2. Section 3 presents the Trust Min-min algorithms and two trust partition schemes in the algorithm. The performance and the analysis of the proposed algorithms are examined in Section 4. Conclusion and future work is briefly discussed in Section 5.

2 A Trust Model

Sepandar[11] provided the eigenvector to compute global trust values in Peer to Peer system. Here we introduce into the Grid to guide tasks scheduling and propose a CalculateTrust algorithm fit for the Grid.

In Grid system, nodes may rate each other after each transaction. $Succ_{ij}$ is used to describe it: $Succ_{ij} = 1$ if *node i* executes a task from *node j* successfully, and $Succ_{ij} = -1$ otherwise. S_{ij} is defined as the sum of the ratings of the individual transactions that *node i* has executed tasks from *node j*: $S_{ij} = \sum Succ_{ij}$.

There are often some nodes that are known to be trustworthy beforehand. Define the pre-trust value $p_i = 1/n$ (n is the total number of accessible nodes) if node i is pre-trustworthy, and $p_i = 0$ otherwise. The aggregation of local trust value is defined as:

$$c_{ij} = \begin{cases} \dfrac{\max(S_{ij},0)}{\sum_j \max(S_{ij},0)}, & \sum_j \max(S_{ij},0) \neq 0 \\ p_j \quad , \quad otherwise \end{cases} \tag{1}$$

The normalized local trust values of node i must be aggregated by asking its acquaintances: $t_{ij} = \sum c_{ik} c_{kj}$, in matrix notation is: $\vec{t}_i = C^T \vec{c}_i$, where C = metric $[c_{ij}]$, $\vec{c}_i = [c_{i1},c_{i2},...c_{ij},\ldots c_{in}]^{\mathrm{T}}$ and $\vec{t}_i = [t_{i1}, t_{i2},...t_{ij},\ldots t_{in}]^{\mathrm{T}}$. In order to get a wider view, *node i* may ask his friends' friends and $\vec{t}_i = (C^T)^2 \vec{c}_i$ is got. Continue this until a complete view of the network under the assumptions is gained: $\vec{t}_i = (C^T)^n \vec{c}_i$, where C is irreducible and aperiodic.

In Grid systems, there may be potential malicious collectives which is a group of malicious nodes who know each other and give each other high local trust values in an attempt to subvert the system. We solve this by $\vec{t}^{(k+1)} = (1-\alpha) C^T \vec{t}^{(k)} + \alpha \vec{p}$, where α is a constant less than 1. This will break the collectives by increasing the

effectiveness of pre-trust value. In a distributed environment, C and $\vec{t}$ are stored in each node and node i can compute its own global trust value t_i by:

$$t_i^{(k+1)}=(1-\alpha)(c_{1i}t_1^{(k)}+c_{2i}t_2^{(k)}+\ldots+c_{ni}t_n^{(k)})+\alpha p_i \quad (2)$$

In Grid, t_i may be delayed or dropped in transmission and the computation of t_j is retarded, consequently increase the overhead of trust value computation. We solve this by limiting the waiting time. If the trust vector from node j dose not arrive timely, previous vector of node j is set. The modified algorithm is shown as follows.

```
CalculateTrust Algorithm:
Initial:
   p_i: the initial trust value of node i;
CalculateTrust
   for each node i do
      all nodes j, t_j^(0) = p_j;
      repeat
       wait  r_ji t_j^(k)  from node j for μ seconds, j≠i;
       if  r_ji t_j^(k)  from j does not arrive
             r_ji t_j^(k) = r_ji t_j^(k-1);
       compute t_i^(k+1) = (1 - α)(r_1i t_1^(k) + r_2i t_2^(k) + ... + r_ni t_n^(k)) + α p_i;
       send  r_ij t_i^(k+1)  to other nodes;
      until  | t_i^(k+1) - t_i^(k) | < ε
```

3 TrustMin-Min Algorithm

Min-min algorithm[12] is a fast, easy and efficient batch algorithm. It does not consider security and tasks are mapped to the fastest resource based on the concurrent status of each resource. In order to ensure the security of tasks and resources, our TrustMin-min algorithm begins with computing the trust value of each node by the CalculateTrust algorithms shown in session 2. Then tasks and resources with accordant trust value are selected and scheduled by the rule of Min-min algorithm. The skeleton of the TrustMin-min algorithm is shown as follows.

```
Initial:
    All nodes get the trust value by CalculateTrust;
TrustMin-min
(1) partition the tasks and resources into
                several sets by the trust value;
(2) for each trust value set
(3)   do
(4)     map the tasks to a resource by the Min-min;
(5)   until all tasks in the set are mapped
```

The partition of the tasks and resources based on the trust value is the key of the TrustMin-min algorithm. The trust value computed by the CalculateTrust is between 0 and 1, and [0,1] must be divided into several parts by a certain scheme. Here we introduce two natural schemes to solve this problem.

3.1 The Even-Segments Scheme

The Even-Segments scheme divides [0,1] into λ segments. The length of each segment is equal and the segment *i* is TL_i. All tasks with the trust value in TL_i are set in ST_i, and the resources compose the set SN_i. Tasks in ST_i must be mapped to the resources in SN_i.

The vale of λ influences the performance. With the increase of λ, more segments are set and tasks are mapped to the resources that they trust most. Simultaneously the probability of $SN_i = \Phi$ increases and results in increasing the waiting time. Less λ will argument the failure rate of task execution and increase the completion time. If $\lambda \neq 1$, $SN_i = \Phi$ may occur. Once there are some tasks but no resources in a trust value segment, the tasks must wait until the trust value of the tasks or the resources changed. Time wastes here badly and the system performance is depressed. To solve the problem, the Even-Segments scheme is modified as follows:

1. Initial: λ is set as a small integer Δ (Δ>0);
2. In each trust value segment, if $ST_i \neq \Phi$ and $SN_i \neq \Phi$, tasks are mapped to the resources by the Min-min algorithm;
3. If there are tasks that have not been mapped, $\lambda = 2\lambda$, go back to step 2 until all tasks have been scheduled.

The TrustMin-min algorithm based on the Even–Segments Scheme is described as follows.

```
Initial:
    All nodes get the trust value by CalculateTrust;
TrustMin-min with Even-Segments Scheme
  λ = ;
  While U ≠ Φ
    map trust value of all nodes to λ trust level;
    for each trust level i
        case (ST_i ≠ Φ) ⋂ (SN_i ≠ Φ)
            schedule the tasks in ST_i to SN_i by Min-min;
            U=U−ST_i;
        case (SN_i = Φ) ⋂ (ST_i ≠ Φ)
            λ =2 λ ;
```

3.2 The Central Extend Scheme

Different from the Even-Segments scheme, the Central Extend scheme randomly selects a task T_i and regards the trust value of this task TT_i as the center point. τ ($0<\tau<1$) is set as the extending of trust value matching range. All tasks and resources with trust value in $[\max(TT_i-\tau,0),\min(TT_i+\tau,1)]$ compose the set ST_i and SN_i. Tasks in ST_i are mapped to the resources in SN_i by the Min-min algorithm. The minimum of τ is set as $\min(|TD_j-TT_i|)$,j=1,2,…n,$j\neq i$. Repeat the operations upon until all tasks are mapped. Detail algorithm is shown in TrustMin-min with Central Extend Scheme.

```
Initial:
      All nodes get the trust value by CalculateTrust;
TrustMin-min with Central Extend Scheme
  While U ≠ Φ
    randomly select task T_i from U, T_i=(D_i,B_i,TT_i);
    compute the trust range:
                [max(0, TT_i-τ),min(TT_i+τ,1)];
    ST=tasks with trust value in the trust range;
    SN=resources with trust value in the trust range;
    schedule the tasks ST to resources SN by Min-min;
    U=U-ST;
    update the state of mapped resources;
```

The Central Extend scheme divides the range of [0,1] from the point of tasks. ST_i and SN_i must not be empty as $\tau \geqslant \min(|TD_j\text{-}TT_i|), j=1,2,\ldots n, j\neq i$, which means that one resource is at least selected to match the task. Appropriate value of τ will efficiently improves the success rate of tasks execution.

4 Simulation Results

We choose the GridSim[3] as the simulator to investigate the performance of the algorithms. 10 resource nodes and 50~300 tasks were simulated. 30% percent of pre-trusted nodes were set in the following experiments. The number of failed tasks N_{fail} of Min-min algorithm, TrustMin-min algorithm with Even-Segment scheme and TrustMin-min algorithm with Central Extend scheme were compared. As the TrustMin-min algorithms map the tasks to corresponding trust level nodes, N_{fail} was much lower than the Min-min algorithm, as shown in Figure 1. N_{fail} of the Central Extend scheme was lower than the Even-Segment scheme as at least one resource is selected to match the tasks' trust value.

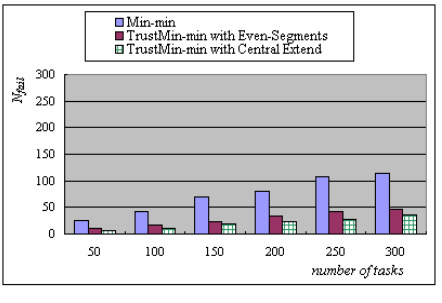


Fig. 1. Number of Failed Tasks

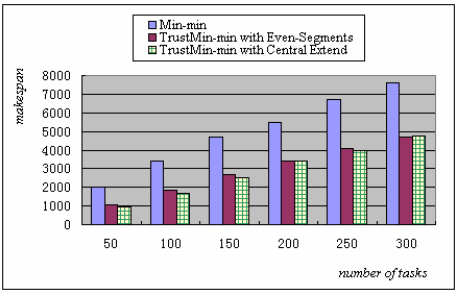


Fig. 2. Makespan of Algorithm

Figure 2 shows the makespan of the algorithms. Min-min algorithms without considering trust value will take more possibility to reschedule the tasks than the TrustMin-min algorithms. As a result, the makespan the Min-min were higher than the others. With lower N_{fail} of the TrustMin-min with Central Extend than the TrustMin-min with Even-Segments, makespan decreased as lesser tasks must be rescheduled.

5 Conclusion

Trust driven task scheduling is crucial to achieving high performance in an open Grid computing environment. In this paper, we analyze the trust model adapted to the Grid, and based on the trust model the most famous Min-min algorithm is modified to work under the risky Grid system. Simulation results show that the TrustMin-min algorithms can remarkably lessen the risks in the task scheduling, improve the load balance and decrease the completion time of tasks.

References

1. I. Foster and C. Kesselman. The Grid: Blueprint for a Future Computing Infrastructure. San Francisco: Morgan Kaufmann Publishers, USA (1999)
2. Ching Lin, Vijay Varadrajan, Yan Wang. Enhancing Grid Security with Trust Management. In: Proceedings of the 2004 IEEE International Conference on Services Computing.(2004)
3. Rajkumar Buyya, Manzur Murshed. A Deadline and Budget Constrained Cost-Time Optimisation Algorithm for Scheduling Task. Farming Applications on Global Grids CoRR cs.DC/0203020: (2002).
4. S. Song, Y. Kwok, K. Hwang. Security-Driven Heuristics and a Fast Genetic Algorithm for Trusted Grid Job Scheduling. In: Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium, IEEE Computer Society (2005).
5. Butt, Ali Raza; Adabala, Sumalatha; Kapadia, Nirav H. Figueiredo, Renato J. Fortes, Jose A.B. Grid-computing portals and security issues. Journal of Parallel and Distributed Computing, 63(10), October, (2003) 1006-1014.
6. N. K. R. F. Butt, S. Adabala and J. Fortes. Fine-grain access control for securing shared resources in computational grids. In: Proc. of Int Parallel and Distributed Processing Symposium, (IPDPS02), April (2002).
7. F. Azzedin and M. Maheswaran. Integrating Trust into Grid Resource Management Systems. In: Proceedings of the International Conference on Parallel Processing, (2002).
8. J. H. Abawajy. Fault-Tolerant Scheduling Policy for Grid Computing Systems. In: Proceedings of IEEE 18th International Parallel & Distributed Processing Symposium, IEEE Computer Society (2004).
9. F. Azzedin and M. Maheswaran. Towards Trust-Aware Resource Management in Grid Computing Systems. In: Proc. of Int'l Symp. on Cluster Computing and the Grid, Germany, (2002).
10. S. Song, K. Hwang, and M. Macwan. Fuzzy Trust Integration for Security Enforcement in Grid Computing. In: Proceeding of IFIP Int'l conference on Network and Parallel Computing. Lecture Notes in Computer Science. Berlin: Springer-Verlag, (2004)9-21.
11. S.D. Kamvar, M. T. Schlosser, and H. Garcia-Molna. The EigenTrust Algorithm for Reputation Management in P2P Networks. In: Proceedings of the 12th Intemational World Wide Web Conference (WWW '03). New York: ACM Press, (2003) 1273 - 1279.
12. Sivakumar Viswanathan, Bharadwaj Veeravalli. Design and Analysis of a Dynamic Scheduling Strategy with Resource Estimation for Large-Scale Grid Systems. In: Proceedings of the Fifth IEEE/ACM International Workshop on Grid Computing (GRID'04) (2004).

Enhanced Security and Privacy Mechanism of RFID Service for Pervasive Mobile Device

Byungil Lee and Howon Kim

Electronics and Telecommunications Research Institute, Korea
bglee@etri.re.kr
http://www.etri.re.kr
161 Gajeong-dong, Yuseong-gu, Daejeon, Korea

Abstract. The RFID (Radio Frequency Identification) technology has drawn much attention for not only its efficiency in inventory and production management, such as WMS, ERP, but also its significant potential for improving our daily lives. Also mobile phone combined RFID reader will popular and widely using device in our ubiquitous society. Mobile RFID technology holds great promise, but it also raises significant security and privacy concerns for the ubiquitous world. This paper suggests a new architecture for mobile RFID service, which support secure RFID service and uses the privacy aware access control for collection or gathering of tagged information in RFID system.

1 Introduction

Recently, low-cost radio frequency identification (RFID) has attracted growing interes from both industry and academic institutes. The RFID industry is very active, with numerous companies developing RFID tags of varying capabilities. With the growing importance of RFID for the ubiquitous infrastructure, numerous works on RFID technology including development of related applications have been conducted in a variety of fields. RFID can be potentially utilized in a broad range of everyday applications because of its capacity to readily relay information. Using RFID, the information of tagged products is given from information server that utilizes data identified by a RF reader via the emission of an RF signal. While people often think of RFID tags as simply an updated replacement of the familiar bar code, they differ in several important ways, including usage of memory, identity individual items, writing of new data, interfacing with sensors, and digital data source. In order to adapt limited computing resource, such as RFID to secure sensor networks, new security approaches to cope with new threat have also been studied[1].

Also mobile phones have become extremely popular and widely used in networked society. The mobile phone with RFID reader and RFID inter-worked mobile network, so called mobile RFID system is an innovative and convenient business area. The mobile RFID systems are different from other means of identification because mobile RFID communication is non-contact and non-line-of-sight, whereas other means of it is more difficult for the owner of the tag to physically impede communication with the 915 MHz tag. Yet, as the information stored on the tag becomes more and more

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 469 – 475, 2005.
© Springer-Verlag Berlin Heidelberg 2005

valuable, it is necessary to think through some of the security and privacy related issues in RFID[2][3].

However, unless this mobile RFID system is properly designed and constructed, they can cause massive collateral damage to consumer privacy and company security. A store's inventory labeled with unprotected tags may be monitored by unauthorized readers on behalf of competitors. In addition, personalized tags would radiate identifying personal information to any tag reader anywhere, thus making their every movement traceable by mobile device. Despite RFID's many advantages, there are many problems in deploying RFID technology extensively. In addition, mobile RFID network includes mobile operator's network. The mobile RFID network is not any more closed network. Thus, we have to design different security model for secure mobile RFID service. Based on the pervasive deployment of RFID tags, we propose a novel secure and privacy-aware mobile RFID architecture.

The remainder of this paper is organized as follow. Section 2 gives a brief architecture for mobile RFID service. Section 3 delineates a model of privacy protection and a technical working mechanism. Section 4 states security model and scenarios for mobile RFID service, Finally, conclusions and a discussion of potential areas for future research are presented in Section 5.

2 Mobile RFID Service Architectures

2.1 Current RFID Service and Mobile RFID Service

Many companies and organizations, including EPC Global Network, have developed RFID systems and international standardization is ongoing. The architecture of an RFID system is shown in Figure 1.

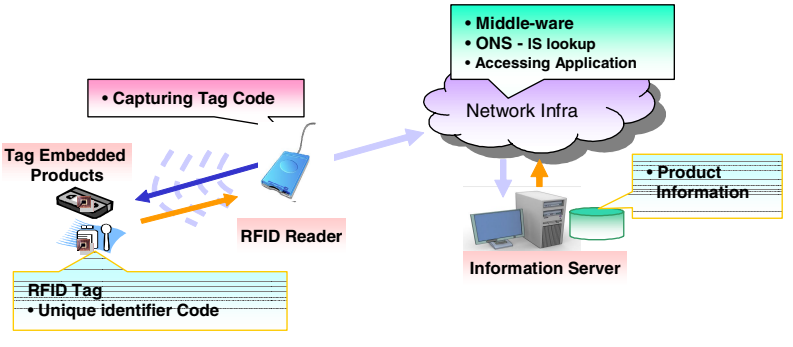


Fig. 1. RFID System Architecture

As tags are more widely utilized, the amount of information around us will increase dramatically, thus posing a critical security and privacy problem using mobile RFID device. From this point of view, we define requirements of security and privacy protection in a mobile RFID system.

- Provide confidentiality of RFID security data(key value etc), data integrity, authentication and authorization of accessing application, non-repudiation of RFID transaction[5][6][7].
- Avoid collecting unnecessary private information in the mobile RFID system
- Employ a controllable access control mechanism to the collected data in the mobile RFID system.
- Avoid collecting unnecessary private tracing in the mobile RFID system.
- Mobile Operators should connect to RFID service provider securely.
- Secure communication should be established for all mobile communication links with the RFID system.

For this reason, we propose a mobile RFID model that provides secure mobile service and protects personal privacy in the mobile RFID system using a personal privacy policy in order to administer information more flexibly and securely as well as mitigate the aforementioned problems.

In first, we introduce network architecture for mobile RFID system such as Figure 2. This architecture includes mobile operator's network such as 2G or 3G mobile network. The mobile RFID device provides the functions of RFID Device Handler for HAL(Handset Adaptation Layer) to the RFID reader and common RFID middleware platform for secure application, so called RFID secure WIPI(Wireless Internet Platform for Interoperability)[4] API and security and privacy application.

As highly coupled limited resource problem of a low-cost RFID tag, security and privacy technologies (reader, middleware and server side) can be added or enhanced to overcome the security weaknesses in an RFID system.

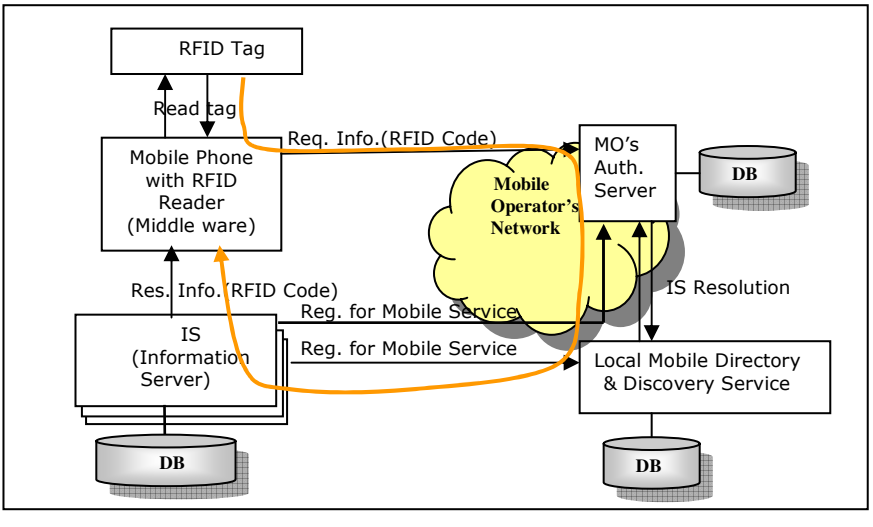


Fig. 2. Mobile RFID System Architecture

3 Proposed Privacy and Security Scheme for Mobile RFID Service

The tag memory is insecure and susceptible to physical attacks. This includes a myriad of attacks such as shaped charges, laser etching, ion-probes, tempest attacks and many others[7]. Fortunately, these attacks require physical tag access and are not easily carried out in public or on a wide scale without detection[8-9]. The user memory bank of RFID tag can store the encrypted data of privacy policy owner configured, as shown in Figure 3. If an owner decides his privacy policy such as in Figure 3, access control policy(P_L) against pre-defined access group is 4,4,3,2,1 and the policy data is stored in tag. When anyone read the tag, the policy data must be transport to server securely and the information of tag can be provided by the access policy. Access control group, A_G of the information of the tag attached product can describe the enclosed description document of the product.

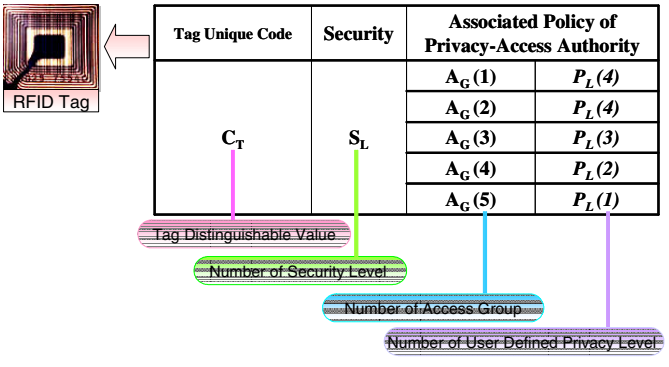


Fig. 3. Data Structure of Tag for Privacy and Security

The security level is a key factor to ensure the privacy policy and related data along with the tag, reader, middleware, and server. Table 1 represents the additionally required phases according to the request. A variety of access control methods can be added as the importance of the information increases. The enforcement of the access control should be proportionally complicated and stronger according to the security level. The information about a product can transfer as a XML format with security(WS Security) and also can be subject to access controlled by a standard technology such as XACML.

Table 1. Security Level for Secure Communication in RFID system

Security Level	1	2	3	4
Public(no secrity)	o	x	x	x
Password based Security(no anonymity of ID)	x	o	x	x
Symmetric based Security(anonymity of ID), Protection of Location based Traceability	x	x	o	x
Public Key based Encryption, Protection of Location based Traceability	x	x	x	o

4 Proposed Scenarios for Secure and Privacy Aware Mobile RFID Service

This section describes the trust model and the scenarios for secure and privacy aware mobile RFID service. Establishing an efficient and a convincing trust model is required to ensure secure transactions and service. Figure 4 depicts a security association of our proposed mobile RFID system.

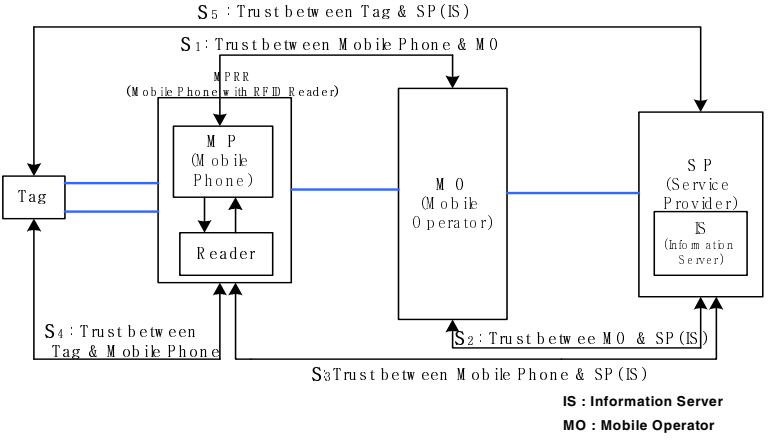


Fig. 4. A trust model for mobile RFID system

We assume that S_2(MO and IS) are high-computing and resource rich entities. During the protocol execution they can easily and very efficiently perform expensive PKI-based tasks like public-key encryption, decryption, and digital certificate and signature verifications. The RFID service provider have a security and privacy policy by user's P(L). The procedures of mobile RFID service are shown in Figure 5.

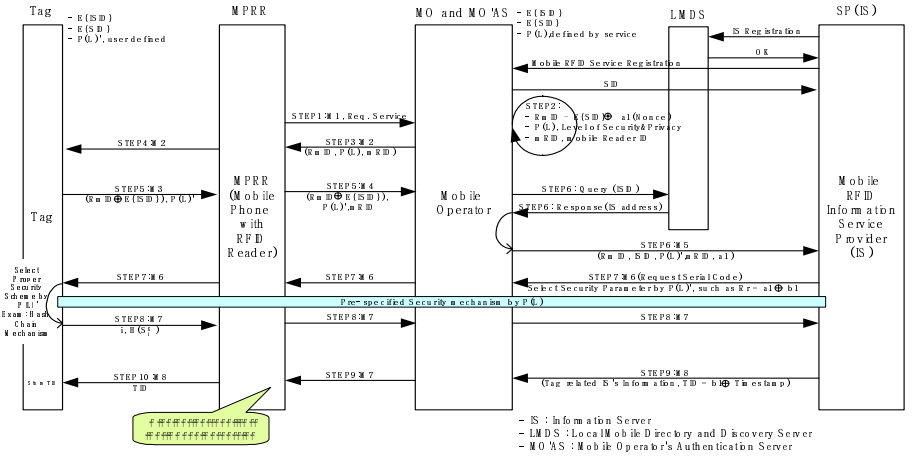


Fig. 5. Scenarios for secure mobile RFID service

We also assume that S_1 (Mobile phone and MO) can perform symmetric key based task, encryption, decryption, verification as pre-shared mobile network security.

We proposed a secure mobile RFID model to be deployed for implementation of RFID service through the mobile network. The authentication server of a mobile operator takes charge of security for mobile RFID service and query to directory service center for IS connection and the procedures for mobile RFID service are as follows.

1) Prior Operation
- IS registers to the Local Mobile Directory and Discovery Server with an IS-ID.
- IS registers to the MO'AS with a mobile service ID(=SID) and the SID is stored in Tag's memory.

2) STEP 0
- User enters PIN to mobile phone for service authentication.
- User selects mobile RFID service, and then connects to RFID reader and display the result.
- User selects specified mobile RFID service from lists of mobile RFID service.

3) STEP 1
- MPRR(Mobile phone with RFID reader) request service to MO by sending M1 message to MO.
- M1 contains timestamp, types of reader, serial of the reader, mobile phone number, and selected service.
- The mobile phone uses the master shared secret key(MK) and performs symmetric-key encryption.

4) STEP 2
- MO's AS finished the authentication process successfully and get mobile RFID service ID by user's selected service and create a new unique reader ID(=mRID) and Nonce, a1.
- MO's AS also select P(L), privacy and security level by specified mobile RFID service.
- MO's AS compute exclusive OR with a1 and encrypted mobile service ID.(result = Rmid)

5) STEP 3, 4
- MO's AS envelops the value(Rmid, P(L) and RID) and send to Tag through MPRR.

6) STEP 5, 6
- Tag extracts the a1 from encrypted SID, stored in memory and compute exclusive OR with Encrypted ISID, a1 and send the computed value and P(L)' to MO.
  P(L)' : If it differs from received P(L), the proper security mechanism is describes in the previous chapter we proposed.
- MO'AS decrypts ISID and query to LMDS with ISID and receive the real address.
- MO'AS send M5 message to IS.

7) STEP 7
- MO select security mechanism by P(L)' and send to M6 message to Tag for real serial code.

8) STEP 8
- Tag checks the M5, and compute encryption or hash function with serial code and new b1.
- Tag send M6 message and IS extract the serial code by decryption or exclusive OR of received data.

9) STEP 9, 10
- IS calculates TID for context based authentication, by following method. (TID = b1 $\oplus$ Timestamps)
- IS send tag information to MPRR and user can check the requested information on mobile device.
- Tag and IS store a TID value for authentication against cloning.

The advantages of this procedure are as follows: first, MO and IS can communicate with MPRR and Tag securely in the mobile RFID service. Second, the owner of the tagged product able to set the privacy level and security level on tag data fields and configures the level of the information that will be accessible to the public and the access control. Third, the query of IS with a Tag's code occurs in MO, instead of MPRR, consume mobile resource, it is cost effective solution. Our approach is very practical and easily deployable solution in current mobile communications infrastructure. Because our approaches provide a separated security structure between MO and IS.

5 Conclusion

In this paper, we propose a mobile RFID model to be deployed for implementation of secure and privacy conscious RFID system through the mobile network.

The proposed algorithms were presented in terms of privacy policy whereas we introduced a privacy protection scheme using owner configured privacy policy in Tag. In the proposed mechanism, the owner of the tagged product sets the privacy level and security level on tag data fields and configures the level of the information that will be accessible to the public and the access control. And also we propose a scenario of the mobile RFID service, MO and IS can communicate with MPRR and Tag securely. By doing so, the customized security and privacy protection can be achieved. In this regard, the suggested mechanism is an effective solution for security and privacy protection in a mobile RFID system.

References

1. Seunghun Jin, et al.: Cluster-based Trust Evaluation Scheme in Ad Hoc Network. ETRI Journal, Vol.27, No.4 (2005)
2. S. E. Sarma, S. A. Weis, and D.W. Engels. RFID systems, security and privacy implications Technical Report MIT-AUTOID-WH-014, AutoID Center, MIT (2002)
3. Weis, S. et al. "Security and Privacy Aspects of Low-Cost Radio Frequency identification Systems", First Intern. Conference on Security in Pervasive Computing (SPC) (2003)
4. WooYoung Han, et. Al.: A Gateway and Framework for Telematics Systems Independent on Mobile Networks. ETRI Journal, Vol.27, No.2 (2005)
5. M. Ohkubo, K. Suzuki and S. Kinoshita, Cryptographic Approach to "Privacy-Friendly Tags, RFID Privacy Workshop2003 (2003)
6. Jan E. Hennig, Peter B. Ladkin, Bern sieker, Privacy Enhancing Technology Concepts for RFID Technology Scrutinised, RVS-RR-04-02, 28 October (2004)
7. Simson Garfinkel, Beth Rosenberg, RFID Applications, Security, and Privacy, Addison-Wesly (2005)
8. Stephen A. et al., Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems, Security in Pervasive Computing 2003, LNCS 2802, (2004) pp. 201–212
9. Sanjay E. Sarma, et al., RFID Systems and Security and Privacy Implications, CHES 2002, LNCS 2523, pp. 454–469, (2003).

Worm Propagation Modeling and Analysis on Network

Yunkai Zhang[1,2], Fangwei Wang[2], Changguang Wang[3], and Jianfeng Ma[1]

[1] Computer Science and Technology, Xidian University, Xi'an 710071, China
zhyk@hebtu.edu.cn, ejfma@hotmail.com
[2] Network Center, Hebei Normal University, Shijiazhuang, Hebei 050016, China
{zhyk, fw_wang}@hebtu.edu.cn
[3] College of Physics Science and Information Engineering, Hebei Normal University, Shijiazhuang, Hebei 050016, China
wangcg@hebtu.edu.cn

Abstract. In recent years, network worms that had a dramatic increase in the frequency and virulence of such outbreaks have become one of the major threats to the security of the Internet. This paper provides a worm propagation model based on the SEIR deterministic model. The model adopts the birth rate and death rate so that it can provide a more realistic portrait of the worm propagation. In the process of defending worm, dynamic quarantine strategy, dynamic infecting rate and removing rate are adopted. The analysis shows that the worm propagation speed can be efficiently reduced to give people more precious time to defend it. So the negative influence of the worm can be reduced. The simulation results verify the effectiveness of the model.

1 Introduction

Network worms are programs that self-propagate through a network exploiting security or policy vulnerabilities in widely-used services [1]. In recent years, much more new vulnerabilities are exploited by network worms than before, such as Code Red worm and Nimda 2001, SQL Slammer worm in 2002, Blaster in 2003, Sasser and Witty in 2004, Sober in 2005. Worms have become increasingly easy to distribute and significantly more threatening to the security of the Internet. The time required for the infection of global targets has shrunk from days to minutes. For example, flash worm can infect almost all vulnerable servers on the Internet in less than 30 seconds and can infect all ones of a enterprise within milliseconds [2].

Many scholars abroad have conducted thorough research to worms, and proposed some models [3,4]. However, many models' worm scanning rate is assumed to be constant over time. In addition, when considering worm's removal, we should not only calculate infected hosts, but also the susceptible ones. Lastly, an important factor is the recovered rate that is assumed to be constant in present references. Many current papers based on SIS and SIR assume the number of hosts to be constant. However, worms may cause enough deaths to influence the population size. In view of the problems of present models, we propose a new model in this paper. It has the characteristics as follows: (1) Using birth rate and death rate; (2) Using the dynamic infection rate; (3) Using the dynamic removal rate; (4) Using dynamic quarantine measures to isolate some hosts if their certain service ports are detected to abnormal traffic.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 476 – 481, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The rest of this paper is organized as follows. In Section 2 we provide a new worm model: a modified SEIR model. We present simulations based on the new worm model in Section 3. In the end, Section 4 concludes this paper.

2 New Worm Propagation Model

Simple computer worms are similar to biological viruses in their self-replicating and propagation behaviors. Thus the mathematical methods can be adapted to the study of computer worm propagation. Deterministic models, which are compartmental ones, attempt to describe and explain what happens on the average at the population scale. The SEIR model includes four compartments represented by the Susceptible, Exposed, Infectious, and Recovered. There exist many mathematical models at present; this article first briefly introduces the classical deterministic SEIR epidemic model [5]. Although the model has some flaws, for example, it does not consider the user's countermeasures, a constant infection rate and so on, it is the foundation of other models such as references [6, 7].

Traditional SEIR model considers the recovering process of infectious hosts. It assumes that during an epidemic of a contagious disease, some infectious hosts either recover or die; once a host recovers from the disease, it will be immune to the disease forever—it is in Recovered state. Thus every host stays in one of four states: Susceptible, Exposed (Infected), Infectious, and Recovered. Any host has either the state transition Susceptible→Exposed→ Infectious→Recovered or stays Susceptible state forever. The SEIR model is not suitable for modeling network worm propagation, because it has some flaws: Firstly, the model assumes the infection rate and recovering rate to be constant, which are not accurate for a rampant network worm. Secondly, countermeasures will recover both susceptible and infectious hosts from circulation. However, SEIR model only accounts for the recovered of the birth rate; high birth rate will generate more infectious hosts. Thirdly, the death rate is not considered in SEIR model, because worms may cause enough deaths to affect the initial number of hosts.

2.1 Worm Propagation Model Based on SEIR

The SEIR model assumes that each host stays in one of four states. It considers the removal process of infectious host. Considering the actual situation that the removal hosts should comprise two parts: the removal hosts from infectious ones and the removal hosts from susceptible ones. Birth rate and death rate are important factors in worm propagation. Altering the differential equations to deal with births and deaths changes the population dynamics because new hosts can get diseased continuously.

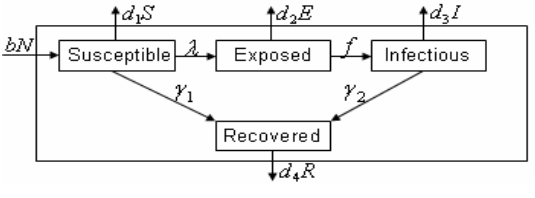


Fig. 1. The graph of state transition in modified SEIR

Therefore, we show that recovering new susceptible hosts by births generate the oscillation of most worms. The frequency of oscillation over time will depend on the birth rate; elevated birth rates will entice tight oscillation, and low birth rate loose oscillation. The relation of these four parts is shown in Fig.1.

Assuming initial number of hosts N and a birth rate b per unit time, the number of new susceptible hosts per unit time is given by bN . Moreover, $d_1S(t)$, $d_2S(t)$, $d_3S(t)$ and $d_4S(t)$ reflect the number of susceptible hosts, exposed hosts, infectious hosts, and recovered hosts, respectively. Thus, in this model the rate of change per unit time in the number of these parts is given by the following differential equations:

$$\begin{cases} dS(t)/dt = bN - (\lambda + d_1 + \gamma_1(t))S(t), \quad dE(t)/dt = \lambda S(t) - (f + d_2)E(t) \\ dI(t)/dt = fE(t) - (\gamma_2(t) + d_3)I(t), \quad dR_1(t)/dt = \gamma_1(t)S(t) - d_4R_1(t) \\ dR_2(t)/dt = \gamma_2(t)I(t) - d_4R_2(t) \end{cases} . \tag{1}$$

Where $\lambda = \beta(t)I(t)$, $f = 1/T_l$, $\gamma_2 = 1/T_p$, $\gamma_1 = 1/T_p{}^2$.

$R(t)$ is composed of two parts: the removal hosts from infectious ones and the removal hosts from susceptible ones. Denote d_1, d_2, d_3 and d_4 as the death rates from susceptible, exposed, infectious, recovered hosts respectively.

2.2 Worm Modeling Based on Quarantined Strategy

In this model, once a host is infected, it is immediately quarantined by the automatic defense system. Some certain hosts are immediately quarantined if there is abnormal traffic in their ports. In order not to affect the user's normal use, the propagation time cannot be excessively long. The new relation of these four parts is shown in Fig. 2.

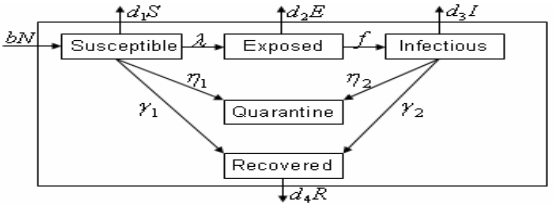


Fig. 2. The graph of state transition in modified SEIR based on quarantine strategy

Denote T as the quarantine time; $R_1(t)$ 、 $G_1(t)$ the number of susceptible hosts that are removed and quarantined at time t ; $R_2(t)$ 、 $G_2(t)$ the number of infectious hosts that are removed and quarantined at time t ; Denote η_1 as the quarantine rate of susceptible hosts. The larger the quarantine rate η_1 is the higher the abnormal detection system is. Denote η_2 as the quarantine rate of infectious hosts. The larger the quarantine rate η_2 is, the better the performance of the abnormal detection system. γ_1 is the recovered rate of the susceptible hosts, γ_2 is the recovered rate of the infectious

hosts. Along with the enhancement of people's understanding the worms, the treatments are more pertinent. Therefore, these two removal rates are also variables which are denoted $\gamma_1(t)$ and $\gamma_2(t)$ respectively. For this dynamic quarantine system, we suppose that we remove hosts uniformly from $S(t), I(t)$. In other words, all hosts are equal probability to be recovered. At time t, all hosts in $G_1(t)$ are susceptible hosts that are quarantined before $t-T$ have been released from $G_1(t)$. At any time τ, there are $S(\tau)-G_1(\tau)$ susceptible hosts that are not quarantined. Due to considering equal probability to recover hosts, the number of removal host from quarantined $G_1(\tau)$ should be $\gamma_1(t)G_1(\tau)$. Therefore, we can obtain

$$G_1(t)=\int_{t-T}^{t}[S(\tau)-G_1(\tau)]\eta_1 d\tau-\int_{t-T}^{t}\gamma_1(t)G_1(\tau)d\tau. \tag{2}$$

The quarantine time T being too small, thus during the time interval T, $S(t)$ and $G_1(t)$ do not change much, we can approximately treat them as constants during the time interval T as

$$S(\tau)\cong S(t), G_1(\tau)\cong G_1(t) \quad \forall\tau\in[t-T,t]. \tag{3}$$

From (3) and (4), we obtain $G_1(t)=p_1S(t)$, where $p_1=\eta_1T/(1+(\eta_1+\gamma_1(t))T)$.

Using the same procedure as (4) by replacing $S(t)$ and $G_1(t)$ to $I(t)$ and $G_2(t)$ respectively, we derive $G_2(t)=p_2I(t)$, where $p_2=\eta_2T/(1+(\eta_2+\gamma_2(t))T)$.

A summary of the mathematical model for the worm propagation is given by the following system of four ordinary differential equations:

$$\begin{cases} dS(t)/dt=bN-(\lambda^{'}+d_1+\gamma_1(t))S(t), \quad dE(t)/dt=\lambda^{'}S(t)-(f+d_2)E(t) \\ dI(t)/dt=fE(t)-(\gamma_2(t)+d_3)I(t), \quad dR_1(t)/dt=\gamma_1(t)S(t)-d_4R_1(t) \\ dR_2(t)/dt=\gamma_2(t)I(t)-d_4R_2(t) \end{cases} \tag{4}$$

Where $\lambda^{'}=(1-p_1)(1-p_2)\beta(t)I(t)$. We refer to it as the SEIRQSA model.

In the above model (4), all hosts have an equal probability to be removed. However, because the security staffs are limited, they are not able to inspect all hosts. A feasible and available approach under the situation is only inspect and cure the hosts that have been quarantined in infected hosts and have raised alarm in susceptible ones.

Now the worm propagation model based on quarantine strategy is:

$$\begin{cases} dS(t)/dt=bN-(\lambda^{'}+d_1+\gamma_1(t)^{'})S(t), \quad dE(t)/dt=\lambda^{'}S(t)-(f+d_2)E(t) \\ dI(t)/dt=fE(t)-(\gamma_2(t)^{'}+d_3)I(t), \quad dR_1(t)/dt=\gamma_1(t)^{'}S(t)-d_4G_1(t) \\ dR_2(t)/dt=\gamma_2(t)^{'}I(t)-d_4G_2(t) \end{cases}. \tag{5}$$

Where $\gamma_1^{'}(t)=p_1\gamma_1(t)$, $\gamma_2^{'}(t)=p_1\gamma_2(t)$. We refer to it as the SEIRQSB model.

3 Simulation Experiments and Numerical Analysis

There are several undetermined dynamic parameters in the above model, such as $\beta(t), G_1(t), G_2(t), \gamma_1(t), \gamma_2(t)$, thus we can only obtain the numerical solutions of the differential equations by using Matlab Simulink. The experiment environment is as follows: We suppose that the initial vulnerable is $S(0)=360000$, $\phi=358$ /minute, $I(0)=100$ .The quarantine rate of susceptible and infectious hosts are $\eta_1=0.00002315$ /minute and $\eta_2=0.2$ /minute respectively. The quarantine time $T=10$ minutes. We denote worm's infection rate as $\beta(t)=\beta_0[1-J(t)/N]^{\omega}$, where $J(t)=I(t)+R_2(t)$ is the number of infected hosts by time t. We set $\omega=3$ and denote $\beta_0=\phi/\Omega=8.34\times10^{-8}$ as the initial infection rate.

The worm's removal rate is $\gamma_1(t)=\gamma_1^0[1+R_1(t)/N]^{\sigma}$, where $\gamma_1^0=0.0001$ is the initial removal rate of susceptible hosts. The more the removal susceptible hosts are, the larger the $\gamma_1(t)$ is. Homoplastically, the worm's removal rate is $\gamma_2(t)=\gamma_2^0[1+R_2(t)/N]^{\alpha}$, where $\gamma_2^0=0.01$ is the initial removal rate of infectious hosts. $\sigma=3, \alpha=3$. Death rate $d_1=0.00005$, $d_2=0.00003$, $d_3=0.00005$, $d_4=0.00002$, birth rate $b=0.00003$. Average latent period $T_l=5$ minutes. Average duration of patching $T_p=12$ minutes. The rate $f=1/T_l=1/5=0.2$. Initial recovered rate from infectious hosts $\gamma_2^0=1/T_p=1/12$.

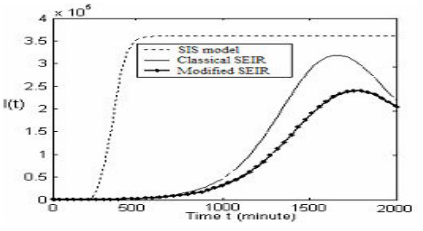


Fig. 3. The comparison of worm model

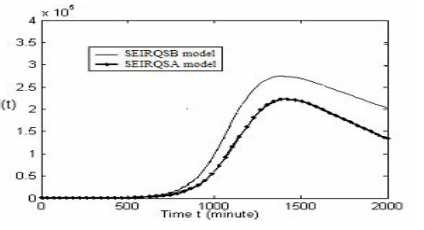


Fig. 4. The comparison of SEIRQSA and SEIRQSB

It can be clearly seen from Fig. 3 that the total hosts to the peak of this paper's modified SEIR model are less than that of classical SEIR, moreover, the time to the peak is postponed. Furthermore, the dropping tendency is quite quicker.

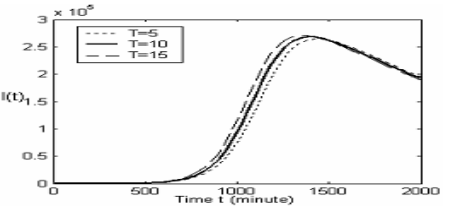


Fig. 5. The comparison of various quarantine time in SEIRQSA model

From Fig. 4 we can see that the total hosts to the peak of SEIRQSA model reduce by 15,975 than that of SEIRSQSB model, the responding time reduces by 57 minutes.

Next, we will illuminate the effect of quarantine time T to these models. We run three simulations with $T = 5$ minutes, $T = 10$ minutes, $T = 15$ minutes respectively.

From Fig. 5, we can see that the larger the quarantine time is, the faster the hosts are infected, the faster the worm propagates and the quicker the dropping speed is.

4 Conclusion

This article proposes a new worm propagation model based on the dynamic quarantine strategy and its characteristics are as follows: (1) Using the dynamic quarantine strategy. (2) Using the dynamic infection rate and the dynamic removal rate. (3)We consider the birth rate and death rate so that the model closely reaches the real circumstances. The simulation results show that this model can effectively reduce worm's propagation speed.

Acknowledgements

This work was supported by a grant from the National Natural Science Foundation of China (No.90204012), the Natural Science Foundation of Hebei Education Department of Hebei China under Grant No. 2004445.

References

1. N.Weaver, V.Paxson, S.Staniford, et al: A Taxonomy of Computer Worms, ACM CCS Workshop on Rapid Malcode, ACM Press, Washington, DC, USA, (2003)11-18
2. S. Staniford, V. Paxson, and N. Weaver: How to Own the Internet in Your Spare Time. Proc. of 11th USENIX Security Symposium, ACM Press, San Francisco, (2002)149-167
3. Cliff C.Zou, Weibo G., Donald F. T.: Code Red Worm Propagation Modeling and Analysis. Proc. of the 9th ACM Symp, ACM Press, Washington, USA, (2002)138-147
4. Cliff C.Zou, Weibo G., Donald F.T.: Code Red Worm Propagation Modeling and Analysis under Dynamic Quarantine Defense. ACM Conference on Computer and Communications Security, ACM Press, Washington, USA, (2003)51-60
5. Trottier, H. & Philippe, P.: Deterministic Modeling Of Infectious Diseases: Theory And Methods. The Internet Journal of Infectious Diseases, Vol.1, No.2, 2001, http:// www.ispub.com/ostia/
6. Trottier, H. & Philippe, P.: Deterministic Modeling of Infectious Diseases: Applications to Measles and Other Similar Infections. The Internet Journal of Infectious Diseases, Vol.2, No.1, 2002, http:// www.ispub.com/ostia/
7. Earn DJD, Rohani P, Bolker BM, et al.: A Simple Model for Complex Dynamical Transitions in Epidemics. Science, Vol.287, No.5453, (2000)667-670

An Extensible AAA Infrastructure for IPv6*

Hong Zhang, Haixin Duan, Wu Liu, and Jianping Wu

Network Research Center, Tsinghua University, Beijing 100084, China
zhang-hong04@mails.tsinghua.edu.cn
{dhx, jianping}@cernet.edu.cn
liu-w@mail.tsinghua.edu.cn

Abstract. AAA (Authentication, Authorization, and Accounting) is an effective component in IP network to control and manage network entities. It has been widely used in IPv4 network and will continuously play an important role in IPv6 network. This paper proposes a new extensible AAA infrastructure which is performed within the CNGI (China Next Generation Internet) project and has the following merits: (1) provide a uniform AAA mechanism; (2) support user roaming in global IPv6 network; (3) introduce for the first time the concepts of both PDN (Personal Domain Name) and DDN (Device Domain Name), to assign and manage the lengthy and complex IPv6 addresses. We discuss and implement the concrete procedures of this infrastructure, and then point out it is a suitable solution for IPv6 network to obtain enhanced level of security.

1 Introduction

The IPv6 (Internet Protocol, version 6) is conceived with two main goals: increase the address space and improve security, relative to IPv4. The community achieved the first goal by increasing the IP address length from 32 bits to 128 bits [1]. However, it is annoying to remember the IPv6 address due to its length and its complicated expression. In the CNGI (China Next Generation Internet) Project, to assign and manage the lengthy and complex IPv6 addresses, we introduce the concepts of both PDN (Personal Domain Name) and DDN (Device Domain Name), which can be used as globally unique identifiers for users and devices in the IPv6 Internet.

IETF intends to solve the second goal by mandating the inclusion of IPSec. However, as presented in [2], IPv6 is facing some threats the same with IPv4, we still need to deploy some additional security mechanisms to protect IPv6 network application from certain kinds of attacks (e.g. unauthorized access). The AAA infrastructure provides a security mechanism for the Internet Service Providers (ISPs) to manage network entities and resources. The Authentication process is used to determine both the identity and validity of the requestor; The Authorization process is performed to determine whether the access to services or resources should be granted to a requesting entity or not; The Accounting process is fulfilled to collect resource usage information which will be used for accounting information on service user by the ISPs [3].

* This work is supported by grants from CNGI, 973, 863 and the National Natural Science Foundation of China (Grant No. #90104002 & #2003CB314805 & #2003AA142080 & #60203044).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 482 – 487, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The main purpose of our infrastructure is to address the following issues:

- To provide a uniform AAA mechanism in the global native IPv6 network. The user can access the global network at any attachment point by providing his globally unique PDN and corresponding attestation.
- To support user roaming in the global IPv6 network. When user roams from one domain to another, he will be assigned with a new IPv6 address which is frequently changed and annoying to remember. In our infrastructure, the change of IPv6 address is transparent to the user.
- To make the AAA mechanism be easy and convenient for extensible and scalable in the IPv6 network by supporting a variety of authentication protocols and by simplifying the cross-domain authentication procedure.

The rest of this paper is organized as follows. In section 2, we will briefly introduce some schemes related to our work. In section 3, we will describe our infrastructure in detail, including the functionality of components in the infrastructure and the AAA procedures. Finally, the paper is concluded in Section 4, in which we will point out the advantages of this infrastructure and the future work.

2 Related Works

Rafael Marín López, etc. have proposed a mechanism in [3] to support the roaming service among multi-domain networks and heterogeneous access technologies. It discussed some issues related with the deployment of some authentication protocol in IPv6-based network, and illustrated the authentication procedures about Radius scenario, Radius-Diameter scenario and Diameter scenario, respectively.

Andrea Floris and Luca Veltri have proposed a scheme in [4] to support IPv6-based terminals to roam among heterogeneous networks. They integrated the IPv6 autoconfiguration mechanism with the functionality provided by AAA framework to manage user roaming and access control. It discussed the scenarios about access control with stateless IPv6 autoconfiguration and stateful IPv6 autoconfiguration.

The schemes described above can realize the authentication roaming among different domains. However, this has also induced some problems. For example, the administrators have to write the related information (including IP address and the shared secret, etc.) into the configuration file in advance, which is quite inefficient and error-prone for authentication in large distributed IPv6 network. In our solution, we settle this problem by introducing the concept of PDN and DDN in combination with DNSSEC, in which we can easily get the IPv6 address of an Authentication Relayer in another domain by looking up the DNS server with its DDN.

3 Our Solution

In this section, we will introduce our infrastructure in detail. Firstly, we present the architecture and explain the functionality of each component, and then we will describe the concrete procedure for each scenario.

3.1 Architecture of the Extensible AAA Infrastructure

Fig. 1 depicts the architecture of our infrastructure which is composed of two domains. As our research is performed within the CNGI, we deploy our implementation in CERNET2 (China Education and Research NET 2).

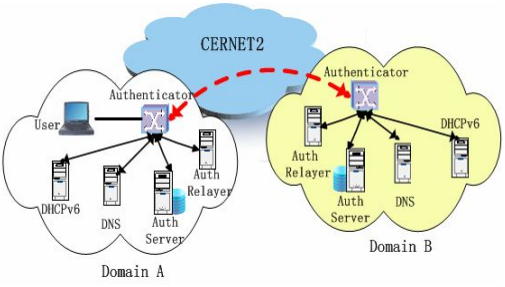


Fig. 1. Architecture of extensible AAA infrastructure

We introduce Authentication Relayer, DNS server and DHCPv6 server into the AAA infrastructure in addition to the basic equipments specified in RFC2865[5][6]. The functionality of the components is explained as follows:

Authenticator. It is used to send AAA request to the authentication server on behalf of the users. We implement the Authenticator in the 802.1x switch.

Authentication Relayer. Its main function is to determine whether a requesting user belongs to the local domain or another according to his PDN, and then decides to perform intra-domain authentication procedure or inter-domain authentication procedure. In addition, It has to translate the one authentication protocol data into another authentication protocol data (e.g. Radius to Diameter) and vice versa.

Authentication Server. It is used to authenticate the users' requests and grant access of resource to them according to the corresponding entry in database. It is also responsible for accounting.

DHCPv6 Server. This server is used for IPv6 address auto-configuration. In our implementation, the DHCPv6 server has to support another function, that is, it has to inform the Authentication Relayer of the parameters assigned to an authenticated user.

DNS Server. It is used to support PDN and DDN which are very important for the infrastructure to be scalable. PDN is a Domain Name assigned to a user. We define it in the form of *UserName@DomainName*. It means that the specified user belongs to the specified domain. Similarly, DDN is a Domain Name assigned to a network Device such as routers and servers, etc. It is formed as DeviceName.DomainName.

3.2 Intra-domain Authentication Scenario

In this scenario, we intend to give a detailed description of intra-domain authentication. Fig.2 shows the message flow chart for the intra-domain scenario.

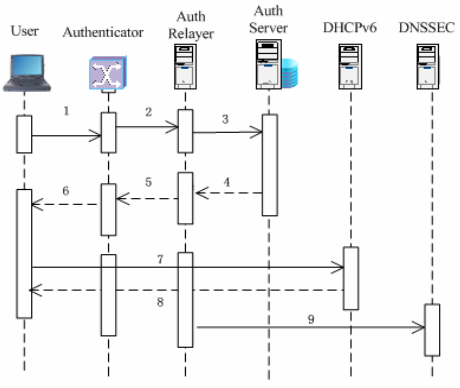


Fig. 2. Intra-domain authentication procedure

The authentication procedure is described step-by-step in the following:

Step (1). The user initializes the AAA procedure by sending an EAPoL-Start message to the Authenticator, providing his PDN and the corresponding attestation.

Step (2). The Authenticator sends an AAA Request message on behalf of the user to the Authentication Relayer through a secure channel (e.g. EAP-TLS).

Step (3). The Authentication Relayer analyze the PDN, determining whether the user belongs to the local domain or not. In this scenario, we assume that this is a local user. Then, the authentication Relayer sends the AAA request to the Authentication server through a secure channel (e.g. EAP-TLS).

Step (4). The Authentication server extract the use's PDN and attestation from the request to authenticate the user. Then it will send an AAA response message to the Authentication Relayer, which contains the privileges granted to the requesting user.

Step (5). The Authentication Relayer decides whether to grant access to the user based on the AAA response message and local policy, and then sends an AAA reply to the Authenticator, specifying the user's granted rights.

Step (6). The Authenticator will decide whether to allow user to access network according to the information in the AAA relay, and then it will sends an EAP reply message to the user.

Step (7). The user has to send multicast DHCPv6 SOLICIT message in order to discover the location of the local DHCPv6 server and the related parameters.

Step (8). The DHCPv6 server will send a DHCPv6 ADVERTISE message to the host, which contains the offered network configuration parameters. The Authenticator will bind the IPv6 address of the host with the corresponding port.

Step (9). The Authentication Relayer will send a DNS Binding Update message containing the user's PDN and his current IPv6 address to the DNS server, which notifies that the user has been assigned with the new IPv6 address.

3.3 Inter-domain Authentication Scenario

In this scenario, we intend to give a detailed description about inter-domain authentication. The scenario is shown is Fig. 3. We assume that the user belongs to domain B, and he is now roaming in domain A. For simplicity, we adopt Radius as the default authentication protocol in both domains.

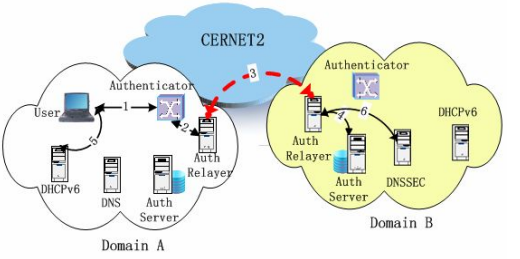


Fig. 3. Inter-domain authentication scenario

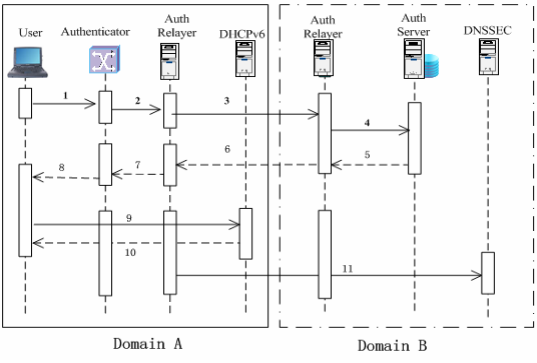


Fig. 4. Inter-domain authentication procedure

Fig. 4 demonstrates the related message flow and the authentication procedure is briefly described in the following.

Step (1) and (2) are similar to the ones in the previous scenario.

Step (3). The Authentication Relayer in domain A analyze the user's PDN, finding that the user not belong to the local domain, then it will deliver the AAA request to the Authentication Relayer of domain B through a secure channel (e.g. IPSec).

Step (4). The Authentication Relayer in domain B will check the validity of the AAA request and forward it to the Authentication server in its domain.

Step (5). The Authentication server in domain B will extract the use's PDN and attestation from the request to authenticate the user. After that, an AAA response mesage will be generated and sent to the Authentication Relayer in domain B.

Step (6). The Authentication Relayer in domain B will forward the AAA response message to the Authentication Relayer in domain A through IPSec.

Step (7), (8), (9) and (10) are similar to step (5), (6), (7) and (8) discussed in the intra-domain scenario.

Step (11). The Authentication Relayer of domain A will send a DNS Binding Update message containing the user's PDN and the current IPv6 address to the Authentication Relayer of domain B, which will forward this message to DNSSEC server in domain B to notify that the user has been assigned with this new IPv6 address.

4 Conclusion and Future Work

IPv6 is a good solution to some problems faced by IPv4, but it can not get everything done once and forever. It has to deploy some additional security mechanisms besides the only inclusion of IPSec. AAA is an import mechanism to protect the network resource and service. It has proved its effectiveness in IPv4-based network, and it should also be very important for IPv6 network. We proposed a new AAA infrastructure which is generic and scalable enough for future development in IPv6 network. It could achieve the goals described in section 1.

Our future work concentrates on two aspects. First, we will improve the current scenarios by adding new functions, for example, realizing the mutual transition between two different authentication protocols. Second, we will address the security issues in our solution and evaluate the performance of the signaling load in order to make our solution work more efficiently.

References

1. Rolf Oppliger: Security at the Internet Layer. Computer, vol. 31, no. 9 September (1998) 43-47
2. Sean Convery, Darrin Miller: IPv6 and IPv4 Threat Comparison and Best Practice Evaluation. Cisco Systems. March (2004) 43
3. Rafael Marín López, Gregorio Martínez Pérez, Antonio F. Gómez-Skarmeta: Implementing RADIUS and Diameter AAA Systems in IPv6-Based Scenarios. AINA (2005) 851-855
4. Andrea Floris, Luca Veltri: Access Control in IPv6-based Roaming Scenarios. Communications, 2003. ICC '03. IEEE International Conference on Volume 2, 11-15 May (2003) 913 – 917
5. L.Blunk, J.Vollbrecht: PPP Extension Authentication Protocol. IETF RFC2284, March (1998)
6. C. Rigney, S. Willens, A. Rubens, W. Simpson: Remote Authentication Dial In User Service (RADIUS). IETF RFC 2865, June(2000)
7. LAN/MAN Standards Committee of the IEEE Computer Society: Port Based Access Control. IEEE Std 802.1x-2001, October (2001).
8. Mattbew S. Gast: 802.11 Wireless Networks: The Definitive Guide. O'Reilly & Associates, Inc. (2002)
9. B.Aboba, D.Simon: PPP EAP TLS Authentication Protocol. IETF RFC2716. October (1999).
10. Joshua Hill: An Analysis of the RADIUS Authentication Protocol. InfoGard Laboratories.
11. Anand R. Prasad, Henri Moelard and Jan Kruys: Security Architecture for Wireless LANs: Corporate & Public Environment. VTC 2000-Spring Tokyo. 2000 IEEE 51st. Volume 1, 15-18 May (2000) 283 – 287
12. Rojas, O.R., Othman, J.B., Sfar, S.: A new approach to manage roaming in IPv6. Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference (2005) 56
13. H. Eertink, A. Peddemors, R. Arends, and K. Wierenga, "Combining RADIUS with Secure DNS for Dynamic Trust Establishment between Domains", Extended abstract accepted to TERENA Networking Conference (TNC'05), June (2005)
14. Wikipedia ,http://en.wikipedia.org/wiki/RADIUS
15. IEEE 802 LAN/MAN Standards Committee,http://www.ieee802.org

The Security Proof of a 4-Way Handshake Protocol in IEEE 802.11i*

Fan Zhang[1], Jianfeng Ma[1,2], and SangJae Moon[3]

[1] Key Laboratory of Computer Networks and Information Security (Ministry of Education), Xidian University, Xi'an 710071, China
fzhang@mail.xidian.edu.cn
[2] School of Computing and Automatization, Tianjin Polytechnic University, Tianjin 300160, China
jfma@mail.xidian.edu.cn
[3] Mobile Network Security Technology Research Center, Kyungpook National University, Sankyuk-dong, Buk-ku, Daegu 702-701, Korea
sjmoon@ee.knu.ac.kr

Abstract. The IEEE 802.11i is the security standard to solve the security problems of WLAN, in which, the protocol 4-way handshake plays a very important role in the authentication and key agreement process. In this paper, we analyzed the security of protocol 4-way handshake with the Canetti-Krawczyk (CK) model, a general framework for constructing and analyzing authentication protocols in realistic models of communication networks. The results show that 4-way handshake protocol can not only satisfy the definition of Session Key security defined in the CK model, but also the universal composition security, a stronger definition of security. So it can be securely used as the basic model of the authentication and key agreement of WLAN.

1 Introduction

The protocol 4-way handshake (in short, 4WHS) presented within 802.11i plays a very important role in the authentication and key agreement process [1-7]. Till now, some works have been done on its security analysis [8]. The main contribution of this paper is to analyze its security with the Canetti-Krawczyk (CK) model [9], a general framework for constructing and analyzing authentication protocols in realistic models of communication networks. The CK model can provide a sound formalization for the authentication problem and suggest simple and attractive design principles for general authentication and key exchange protocols, thus is popular used today. The results show that 4-way handshake protocol not only satisfies the definition of session key security presented in the CK model, but also the universal composition security, a higher level of security than SK security [10]. So it can be securely used as the basic model of the authentication and key agreement of WLAN.

* Research supported by the National Natural Science Foundation of China (Grant No. 90204012), the National "863" High-tech Project of China (Grant No. 2002AA143021), the Excellent Young Teachers Program of Chinese Ministry of Education, the Key Project of Chinese Ministry of Education, and the University IT Research Center Project of Korea.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 488 – 493, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Due to space limitations we will not go into the details of the proofs (any interested reader should contact the first author), but illustrate the results. The whole proofs will appear in the full version of the paper.

2 Canetti-Krawczyk Model

The CK model includes three main components: authenticated-link adversarial model (AM), unauthenticated-link adversarial model (UM) and authenticator [8].

–AM defines an idealized adversary that is not allowed to generate, inject, modify, replay and deliver messages of its choice except if the message is purported to come from a corrupted party. Thus, an AM–adversary can only activate parties using incoming messages that were generated by other parties in π.

–UM is a more realistic model in which the adversary does not have the above restriction. Thus, a UM–adversary can fabricate messages and deliver any messages of its choice.

–An authenticator is a protocol translator C that takes as input a protocol π and outputs another protocol $\pi_0 = C(\pi)$. Authenticators can be constructed by applying a Message Transmission authenticator to each of the messages of the input protocol.

Definition 1[9]**:** A KE protocol π is called SK-secure without perfect forward secrecy in the AM if the following properties are satisfied for any AM-adversary $\mathcal{A}$.

1. If two uncorrupted parties complete matching sessions then they both output the same key;
2. The probability that $\mathcal{A}$ can distinguish the session key from a random value is no more than 0.5 plus a negligible fraction in the security parameter.

The definition of SK-secure protocols in the UM is done analogously.

Definition 2[9]**:** Let π and π' be n-party message-driven protocols. We say that π' emulates π in the unauthenticated-links model if for any UM-adversary μ there exists an Am-adversary A such that $AUTH_{\pi,A}$ and $UNAUTH_{\pi',U}$ are computationally indistinguishable.

Definition 3[9]**:** A compiler **C** is an algorithm that takes for input descriptions of protocols and outputs descriptions of protocols. An authenticator is a compiler **C** where for any protocol π, the protocol $C(\pi)$ emulates π in the unauthenticated-links model.

Theorem 1[9]**:** Let λ be an MT-authenticator. Then C_λ is an authenticator. □

Theorem 2[9]**:** Let π be a SK-secure key-exchange protocol in the AM and let λ be an MT-authenticator. Then $C_\lambda(\pi)$ is a SK–secure key-exchange protocol in the UM. □

Definition 4 [9]**:** We say that a KE protocol satisfies SK-security without PFS if it enjoys SK-security relative to any KE-adversary in the UM that is not allowed to expire keys.

Theorem 3: A KE protocol satisfying SK-security can provide the security attributes like loss of information, key-compromise impersonation, known-key security and unknown-key share, etc[9]. □

For key-exchange protocols, Universal composable (UC) security is stronger than SK-security[11-14]. But a SK-secure key exchange protocol can be transformed into a UC-secure one through the addition of an ACK property to it. A key-exchange protocol π has the ACK property if by the time that the first party outputs the session key, the internal states of both parties are "simulatable" given only the session key and the public information. Following is the formal definition of ACK property.

Definition 5 [10]**:** Let $\mathcal{F}$ be an ideal functionality and let π be an SK-secure key exchange protocol in the $\mathcal{F}$-hybrid model. An algorithm I is said to be an internal state simulator for π if for any environment machine Z and any adversary $\mathcal{A}$ we have $\mathrm{HYB}^{\mathcal{F}}_{\pi,\mathcal{A},Z} \approx \mathrm{HYB}^{\mathcal{F}}_{\pi,\mathcal{A},Z,I}$.

Protocol π is said to have the ACK property if there exists a good internal state simulator for π.Here, $\mathrm{HYB}^{\mathcal{F}}_{\pi,\mathcal{A},Z}$ denote the output of Z after interacting with adversary $\mathcal{A}$ and parties running π in the $\mathcal{F}$-hybrid model. Let $\mathrm{HYB}^{\mathcal{F}}_{\pi,\mathcal{A},Z,I}$ denote the output of Z after engaging in the above modified interaction with π, $\mathcal{A}$ and I in the F-hybrid model.

Theorem 4 [10]**:** Let π be a key exchange protocol that has the ACK property, and is SK-secure. Then π is UC-secure. (This holds both in the AM and in the UM, both with and without PFS.) □

3 The Security Analysis of 4-Way Handshake Protocol

According to the semantics of CK, the transactions of 4-way handshake protocol are written as follows with redundant information eliminated.

Let *k* be a security parameter. $first_{n_1}(\ )$ and $\sec ond_{n_2}(\ )$ are pairwise key hierarchy functions mentioned in the IEEE 802.11i, that is, $first_{n_1}(k_0)$ gets the second 128 bits of k_0 as the Encryption Key (EK), and $\sec ond_{n_2}(k_0)$ gets the first 128 bits of k_0 as the MIC key (MK).

1. Both players pre-share a key k_{ij}.
2. The initiator, p_i, on input (p_i, p_j, s), chooses $r_i \xleftarrow{R} \{0,1\}^k$ and sends (p_i, s, r_i) to p_j;
3. Upon receipt of (p_i, s, r_i), the responder p_j chooses $r_j \xleftarrow{R} \{0,1\}^k$, where $r_j \neq r_i$, computes $k_{0=} prf_{k_{ij}}(r_i, r_j, p_i, p_j)$, $k_1 = first_{n_1}(k_0)$, $k_2 = \sec ond_{n_2}(k_0)$, $t_j = MAC_{k_2}(s, r_j)$ and sends (p_j, s, r_j, t_j) to p_i.

4. Upon receipt of (p_j, s, r_j, t_j), player p_i computers $k_{0=} prf_{k_{ij}}(r_i, r_j, p_i, p_j)$, $k_1 = first_{n_1}(k_0)$, $k_2 = \sec ond_{n_2}(k_0)$,then it verifies the authentication tag t_j and if successful it computes $t_i = MAC_{k_2}(s, r_i)$ and sends (p_i, s, r_i, t_i) to p_j and outputs session key k_1.
5. Upon receipt of (p_i, s, r_i, t_i), player p_j verifies the authentication tag t_i and if successful it outputs session key k_1.

We will extract the parts which responsible for the confidentiality and for the authentication respectively from the original protocol 4-way handshake and further analyze their security.

3.1 Protocol $4WHS_{AM}$

This protocol is described as follows:

1. Both players pre-share a key k_{ij}.
2. The initiator, p_i, on input (p_i, p_j, s), chooses $r_i \xleftarrow{R} \{0,1\}^k$ and sends (p_i, s, r_i) to p_j;
3. Upon receipt of (p_i, s, r_i), the responder p_j chooses $r_j \xleftarrow{R} \{0,1\}^k$, where $r_j \neq r_i$, and sends (p_j, s, r_j) to p_i. Then p_j outputs session key $prf_{k_{ij}}(r_i, r_j)$.
4. Upon receipt of (p_j, s, r_j), player p_j outputs session key $prf_{k_{ij}}(r_i, r_j)$.

Theorem 5: If the pseudorandom function f is secure against chosen message attacks, the protocol $4WHS_{AM}$ is SK-secure without PFS in the AM. □

3.2 Protocol λ_{prf}

The protocol λ_{prf}, based on pseudorandom function and MAC function, is constructed as follows: At first, initiator I was activated, who distributes a sharing key k_{ij} for each pair of participants. When the participant p_i is activated by the request of sending a message m to p_j, λ_{prf} will activate a sub-protocol $\hat{\lambda}_{prf}$, which processes as follows (since $\hat{\lambda}_{prf}$ only involve two participants, we use p_A and p_B to replace p_i and p_j):

1. p_A sends a message *m* to the recipient p_B.Upon receipt of message *m* from p_A, party p_B chooses a random value $N_B \xleftarrow{R} \{0,1\}^k$, and sends challenge $\{m, N_B\}$ to p_A.
2. Upon receipt of the challenge from p_B, party p_A calculates $k = prf_{k_{AB}}(N_B, P_A, P_B)$ and sends response $\{m, MAC_k(m)\}$ to p_B.
3. Upon receipt of the response from p_A, party p_B calculates $k = prf_{k_{AB}}(N_B, P_A, P_B)$ and verifies $MAC_k(m)$. If the MAC is successfully verified, p_B accepts *m*.

Theorem 6: Assume that the pseudorandom function and MAC in use are secure against chosen message attacks. Then protocol λ_{prf} emulates protocol MT in unauthenticated networks. □

3.3 The Security Analysis of 4-Way Handshake Protocol in the UM

We have proved that the protocol $4WHS_{AM}$ is SK-Secure without PFS in the AM, and the protocol λ_{prf} is a MT-Authenticator, thus we get the result of security analysis of 4WHS in the UM.

Theorem 7: If the pseudorandom function and MAC function in use are secure against chosen message attacks, protocol 4-way handshake is SK-Secure in the UM. □

Inference 1: If p_i and p_j in the protocol $4WHS_{UM}$ don't get the same session key, the probability of both sides completing a matching session can be neglected. □

3.4 The 4-Way Handshake Protocol Is UC-Secure

We have proved that 4WHS is SK-secure. According to theorem 4, now we prove that it has the ACK property, thus also satisfies the definition of UC-secure.

Theorem 8: The protocol 4WHS has the ACK property. □

Theorem 9: If the pseudorandom function and MAC function in use are secure against chosen message attacks, protocol 4-way handshake is UC-Secure. □

4 Conclusion

In this paper, we analyzed the security of protocol 4-way handshake with the Canetti-Krawczyk（CK）model. We proved that 4-way handshake protocol can satisfy the definition of SK-security in the UM and also the UC security with a higher level of security. So it can be securely used as the basic model of the authentication and key agreement of WLAN.

References

1. 802.11-1999., I.S., Information technology -Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements - Part 11: Wireless LAN Medium Access Control and Physical Layer Specifications. (1999).
2. 802.11b-1999, I.S., Higher-Speed Physical Layer Extension in the 2.4 GHz Band, Supplement to IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. (1999).
3. IEEE P802.11i D3.0, Specification for Enhanced Security. (2002).
4. Aboba, B., and Simon D., *PPP EAP TLS Authentication Protocol.*(1999) (RFC 2716)
5. 802.1X-2001, I.S., IEEE Standard for Local and metropolitan area networks - Port-Based Network Access Control. (2001).
6. Rigney, C., et al., Remote Authentication Dial in User Service (RADIUS).RFC 2865. (2000).
7. Borisov, N., Goldberg, I., and Wagner, D., Intercepting mobile communications: the insecurity of 802.11. In The 7th Annual International Conference on Mobile Computing and Networking. (2001). Italy.
8. Changhua, H., Mitchell, C J. Analysis of the 802.11i 4-Way Handshake. In proceedings of WiSe'04, October 1, (2004), Philadelphia, Pennsylvania, USA.
9. Canetti, R., and Krawczyk, H., Analysis of Key-Exchange Protocols and Their Use for Building Secure Channels. In proceedings of Eurocrypt 01. (2001).
10. Canetti, R. and Krawczyk, H., Universally Composable Notions of Key Exchange and Secure Channels. In proceedings of Eurocrypt 02. (2002).
11. Lindell, Y., Composition of secure multi-party protocols - A comprehensive study. Lecture Notes in Computer Science, (2003). Springer-Verlag 2815.
12. Lindell, Y., General Composition and Universal Composability in Secure Multi-Party Computation. In 44th FOCS, (2003).
13. Canetti, R. Universally composable security: a new paradigm for cryptographyic protocol. In 42nd FOCS. (2001).
14. Canetti, R., et al. Universally Composable Two-party and Muti-party Secure Computation. In 34th STOC. (2002).

A Noble Key Pre-distribution Scheme with LU Matrix for Secure Wireless Sensor Networks*

Chang Won Park, Sung Jin Choi, and Hee Yong Youn**

School of Information and Communications Engineering,
Sungkyunkwan University, Suwon, Korea
parkcw@keti.re.kr, {choisj, youn}@ece.skku.ac.kr

Abstract. In wireless sensor network security is as important as performance and energy efficiency for many applications. In a recently proposed key pre-distribution scheme suitable for power and resource constrained sensor nodes, a common key is guaranteed to be found between two nodes wanting to communicate and mutual authentication is supported. However, it has a shortcoming that the time overhead is high for performing LU decomposition required in the key pre-distribution step. This paper proposes a new scheme which significantly reduce the overhead by avoiding the LU decomposition. The proposed scheme requires O(k) time complexity to find a common key while the earlier schemes of different approaches require O(k^2) when there exist k keys to compare. The proposed scheme thus displays a significant improvement in the performance and energy efficiency of the sensor nodes.

1 Introduction

Wireless sensor networks (WSNs) applied to the monitoring of physical environments have recently emerged as an important infrastructure. They result from the fusion of wireless communications and embedded computing technologies [1]. Sensor networks consist of hundred or thousands of sensor nodes, the low power devices equipped with one or more sensors. Besides the sensors, a sensor node typically contains signal processing circuit, microcontroller, and wireless transmitter/receiver. By feeding the information on the physical world into the existing information infrastructure, the networks are expected to lead to a future where computing is closely coupled with the physical world and even used to affect it via the actuators. The potential applications include monitoring remote or inhospitable locations, target tracking in battlefields, disaster relief networks, and environmental monitoring, etc. While the research on the WSNs has been focusing on energy efficiency, network protocols, and distributed databases, much less attention has been given to the security issue. However, security is as important as performance and energy efficiency in many applications. Besides the battlefield applications, security is critical for premise protection and surveillance, and the critical communities such as airports, hospitals, etc. Sensor networks have

* This research was supported in part by the Ubiquitous Autonomic Computing and Network Project, 21st Century Frontier R&D Program in Korea and the Brain Korea 21 Project in 2005.

** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 494 – 499, 2005.
© Springer-Verlag Berlin Heidelberg 2005

distinctive features, and the most important ones are constrained energy and computational resources. The existing security mechanisms must be adapted or new ones need to be developed to take such conditions into account [2, 3].

To address this issue a scheme has been recently proposed which is based on random key pre-distribution. It guarantees that any pair of nodes can find a common key between them by using a pool of keys formed in a symmetric matrix and LU decomposition of it [9]. However, it has a shortcoming that the time overhead is high due to LU decomposition. This paper thus proposes a new scheme which allows the same property as [9] without LU decomposition and thereby reducing the time overhead significantly. The time complexity of the proposed scheme is $O(k)$ for checking if there exists any common key between the nodes wanting to communicate, while the earlier schemes of different approaches [4−6] require $O(k^2)$ where k is the number of keys stored in each sensor node. The proposed scheme thus displays a significant improvement in terms of both the performance and energy aspect.

The rest of the paper is organized as follows. Section 2 discusses the existing key distribution approaches for sensor network, and Section 3 presents the proposed scheme. Finally, concluding remark is given in Section 4.

2 Related Works

The traditional key exchange and distribution protocols based on infrastructures in the internet using trusted third parties are impractical for large scale WSNs because of unknown network topology prior to deployment, communication range limitation, intermittent sensor-node operation, and network dynamics, etc. To date, the only practical option for the distribution of keys to the sensor nodes of large-scale WSNs would have to rely on key pre-distribution. Here, keys are pre-installed in the sensor nodes and the nodes having a common key are assumed to have secure connectivity between them.

There exist a number of key pre-distribution schemes for the WSNs. One solution is to let all the nodes carry a master secret key. Any pair of nodes can use this global master secret key to achieve key agreement and obtain a new pairwise key. This scheme does not exhibit desirable network resiliency: if one node is compromised, the security of the entire sensor network will be compromised. Some existing studies suggest storing the master key in tamper-resistant hardware to reduce the risk [4], but this increases the cost and energy consumption of each sensor node. Furthermore, tamper-resistant hardware might not always be safe. Du et al. [5] proposed another key pre-distribution scheme which substantially improves the resiliency of the network compared to other schemes. This scheme displays a threshold property; when the number of compromised nodes is smaller than the threshold, the probability that any node other than the compromised nodes is affected is close to zero. This desirable property lowers initial payoff of small scale network breaches to an adversary, and makes it necessary for the adversary to attack a significant portion of the network.

Recently, Eschenauer and Gligor [6] proposed a random key pre-distribution scheme. Here a pool of random keys are selected from a key space. Each sensor node receives a subset of the random keys from the pool before deployment. Any two nodes able to find one common key within their respective subsets can initiate

communication assuming their connectivity is secure. Based on this scheme, Choi and Youn [9] proposed a key pre-distribution scheme, which guarantees that any pair of nodes can find a common key between themselves. This was achieved by using the keys assigned from the lower and upper triangular matrix decomposed from a symmetric matrix of a pool of keys. However, it has a shortcoming that the time overhead is high for performing LU decomposition. We next present the proposed scheme greatly reducing the overhead by avoiding the LU decomposition while preserving the same property.

3 The Proposed Scheme

First, we briefly describe how the proposed key pre-distribution scheme works. The proposed scheme uses a random graph like the Eschenauer's method [6]. It, however, guarantees that any pair of nodes can find a common key between them with simple calculation. The basic idea is as follows.

Assume that node_x has *i*th row of matrix A, A_{r_i}, and *i*th column of matrix B, B_{c_i}, respectively. Note here that the position of row and column are same (*i*). Similarly, node_y has *j*th row and column of matrix A and B, A_{r_j} and B_{c_j}. If the two nodes exchange their columns and calculate vector product of the original row and the newly received column from the other node, they will get (*i*, *j*) and (*j*, *i*) element of the matrix produced by the product of matrix A and B. If the product matrix is a symmetric matrix, then the two elements are same and the two nodes are guaranteed to find a same key.

In order to implement this property, in [9], a symmetric matrix of randomly generated keys was first constructed, and then decomposed into lower triangular(L) matrix and upper triangular(U) matrix. Then one pair of row and column are randomly selected and assigned to each node. As mentioned earlier, however, matrix decomposition is a computation intensive process. Especially, when the number of keys is large as required in a typical WSN, the time overhead will be very significant. To solve this problem, we propose to construct an L matrix first using a pool of random keys, and then obtain a U matrix such that product of L and U matrix results in a symmetric matrix. Note here that almost half of the elements of L and U matrix are 0, and the values of the elements of the symmetric matrix are not fixed but only one constraint of symmetricity. Therefore, the U matrix can be quickly determined using the given L matrix.

We now explain the proposed key pre-distribution scheme. It consists of three off-line steps, namely generation of a large pool of keys; generation of an L and U matrix whose product results in a symmetric matrix; key assignment to each sensor node.

The first step is the generation of a large pool of keys (e.g., $2^{17} \sim 2^{20}$ keys). We propose a key pre-distribution scheme using the random key approach. Here each sensor node receives a subset of random keys before deployment. For communication, the two nodes need to find one common key to assure the security. Therefore, the base station first needs to generate a large pool of keys in this step.

The second step is the formation of an L matrix using the pool of keys generated in the first step. Then, a U matrix is obtained such that product of them generates a symmetric matrix. We present the computation step using a case of 3×3 matrix for readability.

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{12} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \cdot \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{12} & k_{22} & k_{23} \\ k_{12} & k_{23} & k_{33} \end{bmatrix} \text{(K: a symmetric matrix)} \tag{1}$$

$$\begin{aligned} & l_{11} \cdot u_{11} = k_{11},\ l_{11} \cdot u_{12} = k_{12},\ l_{11} \cdot u_{13} = k_{13} \\ & l_{21} \cdot u_{11} = k_{12},\ l_{21} \cdot u_{12} + l_{22} \cdot u_{22} = k_{22},\ l_{21} \cdot u_{13} + l_{22} \cdot u_{23} = k_{23} \\ & l_{31} \cdot u_{11} = k_{13},\ l_{31} \cdot u_{12} + l_{32} \cdot u_{22} = k_{23},\ l_{31} \cdot u_{13} + l_{32} \cdot u_{23} + l_{33} \cdot u_{33} = k_{33} \end{aligned} \tag{2}$$

In Equation (2), the *k* values are arbitrary and thus the equations required to be solved becomes as follows.

$$\begin{aligned} & l_{11} \cdot u_{12} = l_{21} \cdot u_{11} \ \Rightarrow\ u_{12} = \frac{l_{21}}{l_{11}} \cdot u_{11} \\ & l_{11} \cdot u_{13} = l_{31} \cdot u_{11} \ \Rightarrow\ u_{13} = \frac{l_{31}}{l_{11}} \cdot u_{11} \\ & l_{21} \cdot u_{13} + l_{22} \cdot u_{23} = l_{31} \cdot u_{12} + l_{32} \cdot u_{22} \ \Rightarrow\ u_{23} = \frac{l_{31} \cdot u_{12} + l_{32} \cdot u_{22} - l_{21} \cdot u_{13}}{l_{22}} \end{aligned} \tag{3}$$

Note here that there exist three equations with five unknown variables (u_{11}, u_{12}, u_{13}, u_{22}, and u_{23}). Therefore, there exist many solutions for this linear system, and thus we randomly set the values of two variables (here u_{11} and u_{22}, the diagonal elements). Then the remaining elements can be decided using Equation (3). Note that the last diagonal element of U matrix, u_{33}, can be arbitrarily decided since k_{33} in Equation (2) is also arbitrary.

The final step is key pre-distribution. In this step one random row of L and one column of U are assigned to every node. One and only one requirement here is that the same row and column position are selected such that L_{r_i}(*i*th row of L) and U_{c_i}(*i*th column of U) are selected for a node.

(Finding a common key): Assume that node_x and node_y contains (L_{r_i} and U_{c_i}) and (L_{r_j} and U_{c_j}), respectively. When node_x and node_y need to check if they have a common secret key between them, they first exchange their columns, and then compute a vector product. In node_x, $L_{r_i} \times U_{c_j} = K_{ij}$ is obtained, while $L_{r_j} \times U_{c_i} = K_{ji}$ is obtained in node_y. Recall that K is a symmetric matrix, and thus $K_{ij} = K_{ji}$. node_x and node_y can then confirm they have a common key. Note that the proposed scheme allows any pair of nodes to find a common key between them. By exchanging the columns and computing the key value, the sensor nodes can also mutually authenticate each other. The following is an example of the proposed scheme.

Example:

Step 1: Generation of a large pool of keys using a random graph. Assume here that a pool of keys is generated in the range of S (−3~3).

Step 2: After selecting (−1, 1, 2, 3) from the pool of keys, randomly construct an L matrix.

$$\mathrm{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ -1 & 1 & 2 \end{bmatrix}$$

Assume that u_{11}=1, u_{22}=2, and u_{33}=3.

$$u_{12} = \frac{l_{21}}{l_{11}} \cdot u_{11} = \frac{2}{1} \cdot 1 = 2,\ u_{13} = \frac{l_{31}}{l_{11}} \cdot u_{11} = \frac{-1}{1} \cdot 1 = -1$$

$$u_{23} = \frac{l_{31} \cdot u_{12} + l_{32} \cdot u_{22} - l_{21} \cdot u_{13}}{l_{22}} = \frac{-1 \cdot 2 + 1 \cdot 2 - 2 \cdot -1}{3} = \frac{2}{3}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 2 \\ -1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 2/3 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 10 & 0 \\ -1 & 0 & 23/3 \end{bmatrix}$$

Note here that the product of L and U matrix is a symmetric matrix as expected. Each element of U matrix can be directly computed using the elements of L matrix or randomly decided the diagonal elements, and this will be incomparably simple and fast than LU decomposition.

Step 3: Assume that L_{r_2} (2, 3, 0) and U_{c_2} (2, 2, 0)are stored at node_x. Similarly, L_{r_3} (−1, 1, 2) and U_{c_3} (−1, 2/3, 3) are stored at node_y. The process of key authentication between node_x and node_y is summarized in Table I. Observe that they find a common key of a value of 0.

Table I. The operations of key authentication

Operation	node_x	node_y
Key pre-distribution	L_{r_2} (2, 3, 0) U_{c_2} (2, 2, 0)	L_{r_3} (−1, 1, 2) U_{c_3} (−1, 2/3, 3)
Column-exchange	L_{r_2} (2, 3, 0) U_{c_3} (−1, 2/3, 3)	L_{r_3} (−1, 1, 2) U_{c_2} (2, 2, 0)
Key-computation	$L_{r_2} \times U_{c_3} = 0$ (=K_{23})	$L_{r_3} \times U_{c_2} = 0$ (=K_{32})

4 Conclusion and Future Works

Most earlier schemes proposed for security of distributed sensor network have used asymmetric cryptography such as Diffie-Hellman key agreement or RSA. However, these schemes are often not suitable for WSN due to limited computation and energy resources of sensor node. In an existing key pre-distribution scheme suitable for power and resource constrained sensor nodes, shared key is guaranteed to be found between two nodes wanting to communicate and mutual authentication is supported. However, it has a shortcoming that the time overhead is quite high to perform LU decomposition required in the key-distribution step. This paper has presented a new scheme which significantly reduces the time overhead by avoiding the LU decomposition in the key-

distribution step. The proposed scheme requires O(k) to check if there exists a common key between two nodes while the earlier schemes of different approaches require O(k^2) when there exist k key values to compare. The proposed scheme thus displays a significant improvement in the performance and energy efficiency of the sensor nodes. A new approach considering not only the connectivity but also the energy consumption issue in a more formal way will be developed in the future.

References

1. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci.: A survey on sensor networks: IEEE Communications Magazine, Vol. 40, no. 8, (2002) 102-114
2. F. Stajano.: Security for Ubiquitous Computing: Jhon Wiley and Sons, ISBN 0-470-84493-0, (2002)
3. H. Cam, S. Ozdemir, P.Nair, and D. Muthuavinashiappan.: ESPDA: Energy-efficient and secure pattern based data aggregation for wireless sensor networks: IEEE Sensor Networks (2003)
4. R. Anderson and M. Kuhn.: Tamper resistance – a cautionary note: Proceeding of the Second Usenix Workshop On Electronic Commerce, (1996) 1-11
5. D. Liu and P. Ning.: Establishing pairwise keys in distributed sensor networks: Proceedings of the 10th ACM Conference on Computer and Communications Security, (2003) 52-61
6. L. Eschenauer and V.D. Gligor.: A key-management scheme for distributed sensor networks: Proceeding of the 9th ACM Conference on Computer and Communication security, (2002) 41-47
7. H. Chan, A. Perrig, and D. Song.: Random key pre-distribution schemes for sensor networks: IEEE Symposium on Security and Privacy, (2003) 197-213
8. Erdos and Renyi.: On random graphs I: Publ. Math. Debrecen, Vol 6, (1959) 290-297
9. S.J. Choi and H.Y. Youn.: An Efficient Key Pre-distribution Scheme for Secure Distributed Sensor Network: The 2005 IFIP International Conference on Embedded And Ubiquitous Computing (EUC'2005), LNCS (2005)

A Virtual Bridge Certificate Authority Model*

Haibo Tian, Xi Sun, and Yumin Wang

Key Laboratory of Computer Networks and Information Security (Ministry of Education),
Xidian University, Xi'an 710071, China
hb_tian@hotmail.com

Abstract. Considering the PKI (public key infrastructure) interoperability problem, we bring out a VBCA (virtual bridge certificate authority) model and detail the construction, maintenance and usage of the model. Two basic tools are used: one is the well-exploited threshold signature technique and the other is a data structure that is called DsCert (double signature certificate). Benefit from these tools, one can use the VBCA to bridge two trust points, and then end entities relying on these points can establish trust relationship. A VBCA model is featured by local CA (certificate authority) autonomy, democratic decision, and efficient path processing. This model overcomes the BCA (bridge certificate authority) compromise problem and removes the cross certificates among trust domains.

1 Introduction

Steven Lloyd sets up the PKI interoperability framework [1] and summaries seven approaches to solve the inter-domain interoperability problem. The approaches include the cross certificate model, strict hierarchy model, and BCA (bridge certificate authority) model etc [2]. A BCA model has been proved to be a successful idea for the PKIs (public key infrastructures) interoperability [3]. John Marchesini and Sean Smith [4] have proposed a virtual hierarchy model to provide a resilient and efficient method for multiple PKIs collaboration. More recently, another hierarchy construction method without cross certificates is proposed by Satoshi Koga and Kouichi sakurai [5].

In a common deployment of the BCA model, a BCA is created by several cross-certified independent CAs which are called main CAs here. Each main CA can independently cross-certify with other CAs on behalf of the whole bridge CA. Hence a compromised main CA can create a trust relationship with a bogus CA and thus fool the legitimate CAs into trusting the bogus CA. The main CA compromise problem can be weakened at the cost of off-line operations and working in a multi-person controlled environment as argued in [3]. However besides these costs, the characteristic of one main CA on behalf of all main CAs to cross-certify with other non-main CAs is itself not a good property. We believe that a non-main CA who wants to cross-certify with a BCA should be certified by a majority of main CAs in the membrane of the BCA. This is called democratic decision property in this paper. With this property, a compromised CA is difficult to introduce a bogus CA into a bridged trust domain.

* The work is supported by the National Nature Science Foundation of China under the reference number 60473027.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 500 – 507, 2005.
© Springer-Verlag Berlin Heidelberg 2005

This paper brings out a VBCA (virtual bridge certificate authority) model. Fig. 1 presents an example of the model, where three types CA players are illustrated.

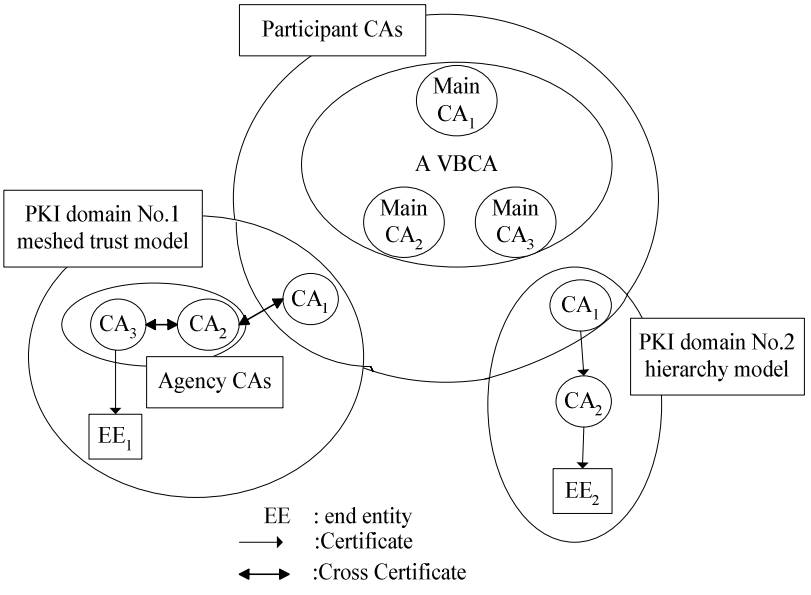


Fig. 1. A VBCA (Virtual Bridge CA) consists of main CAs CA_1-CA_3. The CA_1s in the No.1 and No.2 PKI domains and the three main CAs are all participant CAs. The CA_2 and CA_3 in the No.1 domain are agency CAs. The EE_1 and EE_2 denote end entities in different trust domains. The cross certificates are used within the No.1 and No.2 domain.

The VBCA is the core of this model. A VBCA is established and maintained by a group of main CAs in a distributed way for end entities of participant CAs and agency CAs in different trust domains to establish trust relationship. Distributed main CAs of a VBCA are liked together by threshold signature schemes. The security requirements of these schemes are proactive and adaptive secure for the assumption about main CAs belonging to different competitive organizations or enterprises. The democratic decision is mainly due to these schemes. These schemes are used here in a black-box way whose interface is defined in the section 2. Distributed participant CAs related to a VBCA has a DsCert which contains the signature of the VBCA. The DsCert is a special designed data structure to reduce the cross certificate number and accelerate the certificate path processing. According to different scenarios, a DsCert can be treated as a self-sighed certificate or cross certificates. The self-signed certificate function of a DsCert contributes to the autonomy property of a VBCA model. The data structure and usage of a DsCert is specified in section 3.1 and 3.2.

The remainder sections are dedicated for the VBCA model. Section 2 gives out relevant high level interfaces of the threshold signature technique. Section 3 presents the building, usage, and maintenance of the VBCA model. Some points are discussed in section 4.

2 Threshold Signature

This section defines general interfaces of sub-protocols of the proactive and adaptive threshold signature schemes. The defined interfaces describe the DKG (distributed key generation) protocols, DSG (distributed signature generation) protocols, SR (share refreshing) protocols. The detail scheme of an interface can be obtained in the reference papers [6], [7], [8], and [9].

The DKG protocol allows a set of n main CAs to jointly generate a pair of public and private keys according to the distribution defined by the underlying cryptosystem A high level interface for this protocol is as follows.

$$\begin{aligned}&((P_1, SK_1, PK, [Pub_\text{inf}]),\ (P_2, SK_2, PK, [Pub_\text{inf}]), ...,\ (P_n, SK_n, PK, [Pub_\text{inf}]))\\ &\leftarrow DKG(P_1, P_2, ..., P_n, [public_parameters])\end{aligned}. \tag{1}$$

The inputs to a DKG protocol include the identities of n main CAs and the optional public parameters such as the group information. These parameters are specified by each concrete scheme. The outputs from a DKG protocol include a VBCA public key PK, n shares $SK_1,\ldots,SK_n$ and some relevant public information (such as commitments for verification) for each main CA. Any t (the threshold value) main CAs can reconstruct a secret key, the VBCA private key. The secret sub-share for each main CA is secret itself.

The DSG protocol allows t (the threshold value) main CAs to jointly generate a signature of a message m using their separately holding sub-shares. A high level interface for this protocol is as follows:

$$(Sig) \leftarrow DSG((P_1, SK_1), (P_2, SK_2), ..., (P_t, SK_t), m, [public_parameters]). \tag{2}$$

The computation of the signature involves t main CAs which use their sub-shares to compute partial signatures for the message m. An option parameter field is included in the input, which is specified by an implementation. The output is the signature of the message m, which is the combination of partial signatures.

The SR protocol is mainly for the maintenance purpose. It allows a set of t (the threshold value) main CAs to jointly generate new sub-shares for n' main CAs with a threshold t' and an invariant VBCA private key, where n' and t' may or may not be the same as n and t. A high level interface is given out:

$$\begin{aligned}&((P_1', SK_1', [Pub_\text{inf}]),\ (P_2', SK_2', [Pub_\text{inf}]), ...,\ (P_{n'}', SK_{n'}', [Pub_\text{inf}]))\\ &\leftarrow SR((P_1, [SK_1]), (P_2, [SK_2]), ..., (P_t, [SK_t]), [public_parameters])\end{aligned}. \tag{3}$$

The inputs to a SR protocol include t main CA identities and the public parameters which are specified by each concrete scheme. The outputs from a SR protocol are locally computed new sub-shares $SK_1', \ldots, SK_{n'}'$ of a value zero and some other public information for the scheme robustness. A new sub-share of a main CA will be the combination of the new sub-share and its old sub-share. Whereas the intent is to transfer a VBCA's private key to a different access structure (n', t'), the secret sub-shares of t main CAs are also included in the inputs. In this case, t' new sub-shares can reconstruct the VBCA's private key. The new sub-share of a main CA will replace its existing old sub-share.

3 Details of the VBCA Model

3.1 Building the VBCA Model

1. Suppose n main CAs P_1, P_2,…, P_n. They agree on a set of concrete threshold signature protocols and execute the agreed DKG protocol to generate key materials for a VBCA.

$$\begin{aligned}&((P_1, SK_1, PK, [Pub_\inf]),\ (P_2, SK_2, PK, [Pub_\inf]),...,\\&(P_n, SK_n, PK, [Pub_\inf])) \leftarrow DKG(P_1, P_2, ..., P_n, [public_parameters])\end{aligned}. \tag{4}$$

2. Now each participant CA constructs its message m_i:

$$\begin{aligned}m_i = &Verion \| serialNumber \| msignature \| Issuer \| validity \| msubject\\&\| msubjectPublicKeyInfo \| IssuerUniqueID \| msubjectUniqueID\\&\| ssignature \| ssubject \| ssubjectPublicKeyInfo \| ssubjectUniqueID\\&\| AssuranceLevel\end{aligned}. \tag{5}$$

The symbol "‖" denotes concatenation. The first nine fields are the same as those in a common X.509 version 3 certificate [10]. The last five fields are unique to a DsCert. The symbol *ssignature* indicates the signature algorithm of a VBCA. The name of the VBCA is specified in the *ssbubject* field. The *ssubjectUniqueID* is the unique name of the VBCA. The public key (*PK*) generated in step 1 is specified in the *ssubjectPublicKeyInfo* field. The Symbol *AssuranceLevel* denotes the assurance level of a participant CA evaluated by a majority of main CAs. The last five fields can be defined as Version 3 extensions. Note that participant CAs include main CAs.

3. Each participant CA initializes the agreed DSG protocol to sign its constructed message m_i.

$$\begin{aligned}(sSignature) \leftarrow DSG((P_1, SK_1), (P_2, SK_2), ..., (P_t, SK_t),\ m_i,\\[public_parameters])\end{aligned}. \tag{6}$$

4. The participant CA constructs a DsCert locally. The participant CA P_i signs the message (m_i‖*sSignature*) with its private key using a signature algorithm indicated by the *msignature* field. The resulting certificate may looks like:

$$\text{DsCert} = m_i \| sSignature \| mSignature\ . \tag{7}$$

5. The participant CA replaces the self-signed certificate with the new DsCert for end entities in its own domain. New certificate policies and new certificate policy statements can be defined in each local domain to regulate the usage of a VBCA's public key.

3.2 Usage of the VBCA Model

We claim that end entities in a PKI trust domain can access to a DsCert efficiently. For a participant CA, if it is a root CA in its own PKI domain, the step 5 in the building procedure assures that all end entities in that trust domain can access a DsCert

efficiently. If the participant CA is in a meshed trust model, the step 5 can assures that end entities of the CA can access a DsCert. Other end entities relying on the agency CAs in the meshed model can depend on their trust points to construct a certificate path to the participant CA on demand or beforehand. Since the agency CAs and the participant CA are in a same PKI domain, it is an efficient thing to use a directory service or even to configure manually to build the certificate path.

We consider a simple case to show how to build the trust relationship between two end entities A and B. Suppose A has a certificate $Cert_A$, and its root CA is an agency CA in a PKI domain. When entity A decides to send a certificate list to B, the client software of A can construct a whole certificate path from locally stored information or from a directory service in the PKI domain. A whole certificate list is as follows:

$$DsCert \| Cert_1 \| \cdots \| Cert_L \| Cert_A . \tag{8}$$

For the entity A, the symbol $Cert_1 \| \cdots \| Cert_L$ denotes L intermediary cross certificates. The certificate $Cert_L$ is issued to A's root CA. The certificate $Cert_1$ is issued by the participant CA in the PKI domain to an intermediary agency CA.

When such a certificate list is sent to the entity B, B takes the following steps:

1. The entity B verifies the *mSignature* in the DsCert using the public key of B's root CA. If this verification results in *True*, B knows that A and B are in a same PKI domain. Then without the help of a VBCA, B can use its root CA's public key and polices in its domain to verify the received certificate list. Except for the first verification, any other verification which outputs *False* in this procedure results in a wrong message and a halt.
2. If the first verification results in *False*, B will verify the *sSignature* in the DsCert using the VBCA's public key in the DsCert available to B in B's local domain. Suppose that the entity B is in a meshed trust domain. Then the entity B can query the DsCert on demand in its trust domain or access it from local storage. If the verification results in *True*, the entity B knows that A and B have the same VBCA. Since the VBCA's public key is signed by a participant CA in B's trust domain and the participant CA is cross-certified by B's root CA through a cross certificate list in B's domain, and since the public key of A's root CA is signed by the VBCA's public key in the received DsCert, B can now trust A's root CA under the certificate policies of B's trust domain. Then B can use the public key of A's root CA and the policies in B's domain to verify the remainder certificates in the certificate list. Any verification which outputs *False* in this procedure results in a wrong message and a halt.
3. If the returned wrong message by the above verification procedure indicates that B and A are not in a same domain and have no same VBCA, B will send a wrong indication to A. The entity A will inform this indication to A's root CA to accelerate its root CA to join the bridged trust domain.

3.3 Maintenance of the VBCA Model

The routine work of the VBCA model maintenance includes the periodical refreshment of the sub-shares, the changing of the threshold value, the enrollment of a new participant CA, the update of the certificate revocation lists etc.

The sub-shares periodical refreshment procedure is initiated by t main CAs jointly. The procedure involves n main CAs which may verify the public values and update their own sub-shares. At the agreed update time, the main CAs execute the following SR protocol:

$$\begin{aligned}&((P_1,\Delta K_1,[Pub_\inf]),\ (P_2,\Delta K_2,[Pub_\inf]),...,\\&(P_n,\Delta K_n,[Pub_\inf]))\leftarrow SR(P_1,P_2,...,P_t,[public_parameters])\end{aligned}. \tag{9}$$

After the execution, each main CA updates its own sub-share:

$$SK_i = SK_i + \Delta K_i, i = \{1,2,\cdots,n\}. \tag{10}$$

The three tasks of changing threshold value, new main CAs enrollment, and revocation of main CAs can be fulfilled in a similar way. On demand, all involved main CAs execute the following SR protocol:

$$\begin{aligned}&((P_1',SK_1',[Pub_\inf]),\ (P_2',SK_2',[Pub_\inf]),...,\ (P_{n'}',SK_{n'}',[Pub_\inf]))\\&\leftarrow SR((P_1,SK_1),(P_2,SK_2),...,(P_t,SK_t),[public_parameters])\end{aligned}. \tag{11}$$

Identities $P_1,\cdots,P_t$ are t main CAs before the execution. These CAs form a routine committee or an on demand ad-hoc committee which should inform the involved CAs and initiate the protocol. The intersection of the CA identity set $\{P_1',P_2',\cdots,P_{n'}'\}$ and set $\{P_1,\cdots,P_n\}$ may be empty or not. Hence it is flexible enough to deal with the three tasks. If a main CA has an old sub-share, the CA will replace the old sub-share with a new one.

For a new participant CA, if the CA is not a main CA, it has to contact at least t main CAs to sign its DsCert. After the evaluation of the assurance level, the CA will construct message m_i, query for the signature of the constructed message, sign the *mSignature* itself, and update the self-signed certificate in its own domain as in steps 2-5 of the VBCA model building procedure (section 4.1).

The revocation procedure involves the CA certificate revocation and the end entity certificate revocation. The participant CA certificate revocation involves at least t main CAs of a VBCA. These main CAs decide whether a participant CA should be revoked, and collaborate to sign new authority revocation lists. The new authority revocation lists can be distributed by an email system to the participant CAs to propagate in its domain. The agency CA revocation is done by each PKI domain itself. The revocation list of the agency CAs is signed by a participant CA in the PKI domain, and can be distributed by an email system to other participant CAs. The received agency CA revocation lists are verified and their contents are re-signed and propagated by a participant CA in its domain.

The end entity certificate revocation is a more frequent event than the above events. Hence a mechanism without a collaborative computation of main CAs is more suitable. Considering a meshed trust model, we suggest a mechanism where a participant CA in a PKI domain serves as a gateway. At first, all new CRLs should be periodically centralized to a participant CA of the PKI domain. After verification of the CRLs, the participant CA will extract the contents of the CRLs and re-sign them with its own private key. Secondly, the re-signed CRL is distributed by an email system to other participant CAs relevant to a VBCA. Thirdly, each participant CA verifies the

received CRLs, extracts the contents of these CRLs, catenates the identity of the VBCA with the extracted contents, and re-signs that concatenation contents. At last, the new CRL is propagated by a participant CA in its domain.

4 Discussion

1. Certification path processing

The path length in a VBCA model is similar to that in a BCA model. The difference is that in a BCA model, the constructed path is completed by one end entity in a PKI domain via a global directory service of the BCA model, whereas in a VBCA model the path is constructed by end entities in two PKI domains separately so that the complexity of certificate path construction is limited into one PKI domain. If the participant CA of a VBCA model is in a hierarchy model and a DsCert is stored locally by its end entity, then half path construction can be omitted and the verification path length is just the length of the received certification list. In addition, a cache mechanism can fast the certificate path processing.

2. Issuance of new certificates

A cross certificate model requires $n(n\text{-}1)$ cross certificates for n CAs. In the BCA model, suppose that n full cross-certified main CAs form a BCA and m CAs are cross-certified with the BCA, and then the number of new cross-certificates is $n(n\text{-}1)+2m$. Comparatively, a VBCA model needs each participant CA to have a DsCert to replace its old self-signed certificate. Hence, the total number of certificates is the same as the number when the PKI domains have not been bridged.

3. Security policy

A Certificate Policy (CP) is defined as "named set of rules that indicates the applicability of a certificate to a particular community and/or class of application with common security requirements." Additionally, the rules describing how the CA are constrained and operated are defined in a document that is called Certification Practices Statement (CPS). A VBCA has a unique key pair, so the common CP and CPS rules for the VBCA are necessary. These rules can be negotiated by all main CA operators. A minimal common rule set should be defined and published. A duplicate of these rules should be stored in a policy authority of each PKI domain.

References

1. Lloyd, S. et al.: PKI Interoperability Framework. PKI Forum, (2001) Available at http://www.pkiforum.org
2. Lloyd, S. et al.: CA-CA Interoperability. PKI Forum, (2001) Available at http://www.pkiforum.org
3. Electronic Messaging Association Challenge 2000: Report of Federal Bridge Certification Authority Initiative and Demonstration. Draft 101500. (2000) Available at http://csrc.nist.gov
4. Marchesini, J., and Smith, S.: Virtual Hierarchies: An Architecture for Building and Maintaining Efficient and Resilient Trust Chains. Karlstad University Studies, In: Proceedings of the 7th Nordic Workshop on Secure IT Systems—NORDSEC (2002) Available at http://www.ists.dartmouth.edu

5. Koga, S., and Sakurai, K.: A merging method of certification authorities without using cross-certifications. IEEE computer society, Advanced Information Networking and Applications, 2004. AINA 2004. 18th International Conference on, Volume 2, 29-31 (2004) 174 – 177
6. Canetti, R., Gennaro, R., Jarecki, S., Krawczyk, H., and Rabin, T.: Adaptive security for threshold cryptosystems. Springer-Verlag, In proceedings of CRYPTO '99, LNCS 1666, (1999) 98-115
7. Abe, M., and Fehr, S.: Adaptively secure Feldman VSS and applications to universally composable threshold cryptography. Springer-Verlag, In M. Franklin, editor, Advances in Cryptology -CRYPTO '04. LNCS 3152, (2004) 317-335
8. Herzberg, A., Jarecki, S., Krawczyk, H., and Yung, M.: Proactive secret sharing or how to cope with perpetual leakage. Springer-Verlag, In Advances in Cryptology: CRYPTO '95, (1995) 339--352
9. Desmedt, Y., and Jajodia, S.: Redistributing Secret Shares to New Access Structures and its Applications. Technical Report ISSE TR-97-01. George Mason University, (1997)
10. Network Working Group: Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. Standards Track, RFC 3280, (2002)

Weak Signals in Information Security Management

Jorma Kajava[1], Reijo Savola[2], and Rauno Varonen[1]

[1] University of Oulu, Oulu, Finland
{Jorma.Kajava, Rauno.Varonen}@oulu.fi
[2] VTT Technical Research Centre of Finland, Oulu, Finland
Reijo.Savola@vtt.fi

Abstract. Usually, information security management practices do not explicitly take account of weak signals, factors that lie below the detection surface, which may, however, constitute a huge security threat. This study analyses what kinds of weak signals are present in information security, followed by a discussion on their detection. Responses to weak signals are also considered as well as certain privacy concerns related to the issue. These issues are of great urgency not only for government officials responsible of public security and dealing with the current wave of terrorism, but also to corporate information security and top managers running the day to day business of their companies.

1 Introduction

An important aspect in solving practical problems within fields such as electrical engineering and automation is known as the systemic approach [13]. This approach is based on analyzing the problem situation and creating an analytical model of it. A fairly straightforward way of achieving that goal involves using *strong signals* to implement a mathematical simulation environment and a testable prototype. Use of the systemic approach enabled the development of the modern industrial production environment.

In the systemic approach, *weak signals* were generally ignored as they were thought to have no more than a marginal effect on the final result. A new interest in weak signals (analogous to early warnings in military intelligence) arose in communications studies during the 1970s, when Igor Ansoff [1], [2] contended that strategic surprises from steady progress seemed to appear with increased frequency in corporate environments. He also noted that surprises were invariably preceded by small, almost undetectable, yet telltale signs: weak signals. With proper scanning and analysis of the operative environment, such signals could be interpreted and responded to in a timely fashion. For example in telecommunications information theoretic approaches have been used to detect weak signals from pseudo random noise.

Michelle Codet defines a weak signal in the following way: "*A weak signal is a factor for change hardly perceptible at present but which will constitute a strong trend in the future*." Another definition for weak signals is by Pierre Masse: "*A weak signal is a sign which is slight in present dimensions but huge in terms of its virtual consequences*" [6]. For the purposes of information security management, the word "trend" in the former definition could be replaced by "threat".

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 508 – 517, 2005.
© Springer-Verlag Berlin Heidelberg 2005

During the past decades, the pace of development in information technology has been tremendous. Constantly improving technology has allowed the detection and analysis of weak signals to take giant leaps. Nowhere is this trend clearer than in current space research, where the construction of models and simulations has been taken to a completely new level. Images that were previously deemed as fuzzy and impossible to interpret, have been analyzed by new equipment and found to reveal valuable information concerning new phenomena, such as magnetic storms. Another case in point is the computer simulation of Saturn's rings. Thus, through improved resolution and accuracy of detection and analysis, the progress of information technology has served to bring us closer to optimum solutions.

To elucidate phenomena within information security, different abstractions have been developed during the last 20 years. These abstractions include standards, documentation on best practices and conceptual models:

- **Information security management standards:** British Standard BS 7799 [5], published also as ISO/IEC 17799 [10], Code of Practice [4], ISO/IEC JTC1/SC27 [12], information security guidelines by the Royal Canadian Mounted Police [15];
- **Evaluation criteria:** Common Criteria (ISO 15408) [9], TCSEC (Trusted Computer System Evaluation Criteria) [17] and ITSEC (Information Technology Security Evaluation Criteria) [7];
- **Capability maturity models:** Systems Security Engineering Capability Maturity Model SSE-CMM (ISO/IEC Standard 21827) [11], INFOSEC Assessment Capability Maturity Model IA-CMM [8], Information Security Program Maturity Grid [16], and Murine-Carpenter Software Security Metrics [14];
- **Best practices:** e.g. Finnish VAHTI (Information Security Management Guidelines by the Finnish Government Information Security Management Board) [18];
- **Conceptual models:** e.g. "CIA" model (Confidentiality, Integrity, Availability, (Non-repudiation)) and various academic taxonomies.

It is important to note that these models are only meant to help us understand the nature of information security, but none of them describe it fully. More importantly, they do not include explicit guidelines for the management of weak signals. If weak signals could be integrated into the existing models and descriptions of information security, our understanding of information security would be that much more mature. According to current knowledge, every event in information security is a unique one-off event. Undoubtedly, the integration of weak signals into security models would allow us to derive more generic rules.

The rest of this paper is organised in the following way. Section 2 analyzes some types of weak signals relevant to information security, Section 3 concentrates on their detection and Section 4 discusses weak signals from both the public and information security perspectives. Finally, Section 5 gives conclusions and directions for further research.

2 Weak Signals in Information Security – A Threat Perspective

In the following, we investigate the types of weak signals present in information security and its management, classifying them to categories based on their relation to security threats. Threats are at the core of information security management; all of its practices are designed to protect information systems from identified threats. In the following discussion, "threat" refers to any identified threat discovered in threat analysis conducted by information security management. Please note that all the categories mentioned here are mutually dependent and overlapping, for such is the nature of information security.

2.1 Cross Relationships

In information security, weak signals are often regarded as manifestations of *cross relationships*. In this respect, information security threats can be seen as forming a complex net of threat factors, some of which are easy to point out, while others are extremely difficult to spot. Different technological, socio-economical and societal disciplines present in the information security environment contribute to the overall set of threats.

The embodiment of a discipline might be weak, and even a rigorous threat analysis within it may leave a number of threats unidentified. If just one of these threats materializes, the effects can be enormous. Developing mechanisms to detect weak signals of this type is a challenging undertaking that typically requires expertise across a range of disciplines.

Example 1. We denote the original set of identified threats in a system S (e.g., an organization or a product) by T, consisting of the threat factors T_0, T_1, T_2,…T_n. Later, the effect of discipline D_x is introduced into the system. This effect manifests itself as a weak signal type of threat τ_x, which can or cannot be identified. In the former case, T is updated to $T := T \cup \tau_x$. In the latter case, the effect of D_x remains a hidden threat represented by the undetected weak signal τ_x.

2.2 Granularity of Threat Analysis

If all important threats cannot be identified in threat analysis, the unidentified threats can be regarded as weak signals. Problems in defining the granularity level of threat analysis might arise from the scale and number of threats – some threats are alarming by their very nature, whereas others may be found only in a more detailed investigation. However, both kinds of information security threats may have similar effects. Moreover, threats may also be discontinuous, making it a real challenge to define explicit identification rules.

Example 2. Let R denote the collection of rules used to identify threats to a system S. If a threat presented by a weak signal τ_x is identified by a rule in R, it is a known threat. However, if no rule in R is able to identify τ_x, it remains undetected and represents a hidden threat to the system. However, if the identification of a threat T_y includes a cross relation with τ_x using a rule in R, τ_x may become known when the rules in R are applied again.

2.3 Changes in Time and Trends

The collection of information security threats to a system is not static. Security algorithms and other solutions are cracked and new vulnerabilities are found every now and then. Even complete platforms or communication protocol structures can be compromised. As a consequence, a system's threat landscape is constantly changing, possibly reflecting different kinds of trends.

Some weak signals can represent on-going or anticipated changes in the threat landscape. The actual change in time can happen in small steps or in one leap. In the former case, the trend could be exposed, if weak signals presenting the steps could be detected.

Example 3. Let τ_0 be a weak signal representing a security-threatening trend at time instant t_0, while τ_1 is a weak signal representing the same trend at time instant t_1, etc. There can be similar patterns in all τ_t, $t = 0, 1, 2, \ldots$. If a monitoring algorithm with the capacity to detect these similarities can be designed, the trend can be identified, allowing appropriate countermeasures to be taken.

2.4 Decision Making Based on Subjective and Deficient Information

Information security management utilizes a great number of abstract specialist statements regarding threats. These statements can vary a lot in content and specificity, and include references to obvious threat factors (strong signals), but possibly even buried weak signals. Nevertheless, the information used in threat analysis may be very subjective and even deficient.

We can model the process of information propagation using a filter model by Ansoff [2], see Fig. 1. A weak signal has to pass three different filters to have an impact on threat analysis. First, a surveillance filter defines the area we are observing. However, the challenge here is that discontinuities normally come from outside our "own" discipline. Then, a mentality filter is used for evaluating novel issues that have passed the surveillance filter. Finally, a power filter is applied when the cognitive filter is triggered by the threat analysis rules.

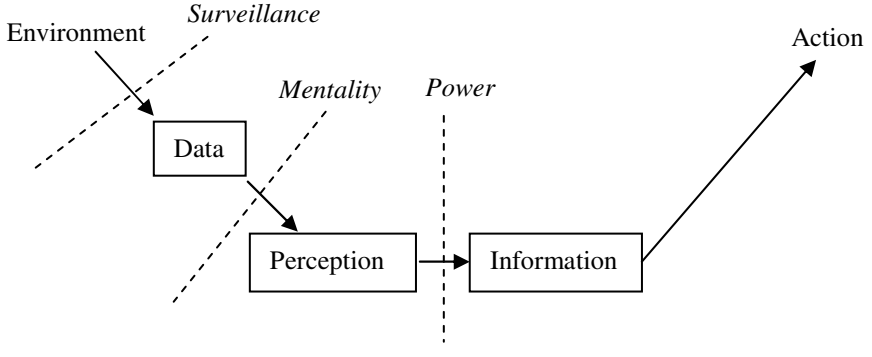


Fig. 1. A filter model for information propagation in decision making

Example 4. In a company's IT management department, the surveillance filter may be a security specialist's perception of a common information security standard, let us say [5]. The mentality filter of the company's IT security manager is used when the security specialist presents a novel security issue that passed the surveillance filter. Now, the manager might ask questions such as "Has this issue any meaning to us?" or "Is it important?" If the security manager thinks that the new issue is relevant, he might present it to the company's top management and possibly request financing for countermeasures.

2.5 Human Behaviour

In our activities, we all transmit a variety of information to our surroundings, only some of which we are consciously aware. In this way, we also emit weak signals. Our interactors typically pick up at least part of this information and use it to facilitate communication with us. The downside of this normal mode of human interaction is that we may inadvertantly reveal information that can be exploited for criminal or otherwise harmful purposes. From the viewpoint of the information transmitted, the significance of weak signals is always dependent on the person transmitting the signals, the situation, the interests at stake and vision.

3 Detection of Weak Signals

Responding to weak signals is a difficult issue. By definition, such signals are hard to detect and even harder to interpret. The core question in signal detection involves deciding, whether the picked up signals represent a message, a change in the prevailing situation, or whether they represent merely arbitrary changes or scattered information. A corollary to that question includes detecting the signals so early that sufficient time remains for an adequate response.

Information security issues can be analyzed, detected and monitored in various ways. Some known practices are risk management, vulnerability analysis and threat analysis. Online activities include intrusion detection and prevention (IDS/IPS), antivirus detection and firewall protection. All of these activities are based on measurements and a collection of security metrics.

3.1 Security Metrics in General

The management of information security becomes easier, if suitable metrics can be developed to offer evidence of a system's security level or performance (e.g., a product or an organization). Measurement results can then be used as a foundation for decision-making to ensure the desired security level or performance. The *measurement object* of security metrics can be a technical system, an organization or a business process – or all of them at the same time. Systematic security metrics provide a means for the comprehensive management of information security and can be applied in many different ways:

- **Security metrics in product R&D.** During the research and development phase, security metrics can be used to define security objectives (risk, threat and vulnerability analysis), compare security objects and validate the security

management process. This allows the security properties of a product to be predicted before actual product development. Information security can also be seen as a quality factor – in this case, quality metrics emphasizing security characteristics.

- **Security metrics as a basis for functionality.** The actual functionality of a product or system can be based on the use of different kinds of security metrics, for example, in Intrusion Prevention/Detection Systems (IPS/IDSs) or anti-virus software.
- **Security metrics for organisations.** Organizational security solutions can be measured, for example, by auditing. These audits are also based on suitable security metrics.
- **Process measurements.** Also the effectiveness and level of business and security management and engineering processes can be measured using security metrics.

The wide majority of available approaches to security metrics have been developed for evaluating the maturity of information security management *processes*. Most widely used of these maturity models is the Systems Security Engineering Capability Maturity Model SSE-CMM (ISO/IEC Standard 21827) [11].

3.2 Security Metrics for Weak Signals

The definition of security metrics for weak signals typically requires pre-recorded event material. Using this material, signals that would otherwise fall below the detection threshold of common security measurement systems can be uncovered. One method involves trying to find patterns in the event material with respect to time or some other dimension – *amplifying* the weak signals. Predictive algorithms such as Bayesian algorithms can be used.

Data gathering presents a challenge for the detection and measurement of weak signals as it may include privacy concerns (see Section 3.3).

The actual definition of security metrics can be carried out using the following compositional and iterative approach:

1. **Define security objectives:** security objectives can be defined on the basis of knowledge concerning the security environment, assumptions and threats. Among other things, the objectives also determine the required security level;
2. **Select component metrics (for both strong and weak signals)** based partly on the security objectives and partly on an investigation of the event material, possibly using an adaptive Bayesian algorithm or a pattern recognition algorithm;
3. **Find cross-relationships** (dependencies) between the component metrics and, if necessary, re-define them as independently as possible;
4. **Compose integrated security level information:** the final composition depends mainly on the measurement method and can be used for both *quantitative* and *qualitative* security metrics.

Composing an integrated metrics framework is far from a trivial task. The chief problem stems from our inability to understand what security actually consists of. Voas [19] states that we do not know *a priori* whether the security of a system composed of two components, *A* and *B*, can be determined merely from knowledge of the

security of *A* and *B*. Rather, the security of the composite is based on more than just the security of the individual components – it hinges on the cross-relationships.

3.3 Privacy Concerns

In summer of 2003, an information technology company presented its new IT system with enhanced security features. Central to this system is that it collects data pertaining to potential risks to company information systems and analyzes them automatically. One subsystem is dedicated to combining data from a variety of sources, such as employees' use of e-mail, net services and their physical movements. In effect, the system maintains a systematic behaviour profile of each individual on the payroll. Another of its conspicuous features is that surveillance also covers the company's WLAN. Thus, the system sounds the alarm when an employee either receives e-mail from competitors or sends them an e-mail message, or logs onto the company's network from outside. Also, the system compiles data on how much time the employee spends away from his or her post.

Obviously, this is a system that surveys its operational environment and collects data on weak signals. How these data will be analyzed and interpreted, what conclusions will be drawn and what responses initiated, is apparently decided by each customer who decides to purchase and install the system. As the use of surveillance systems is motivated on the basis of increasing security, there will be a conflict between security and privacy. Not to put too fine a point on it, the abuse and illegal use of interception of communications is becoming easier and more prevalent than many of us would like to have it.

Information security is a balancing act between good and bad. The essential question to be answered at all levels (such as corporate and governmental) involves determining the ultimate aim of security work. Other pertinent questions include whether information security measures violate basic civil rights such as those concerning privacy. The chief concern here is can a law abiding citizen trust the surveillance systems that record and analyze his or her activities?

What is important at the moment, is that these issues are discussed openly, especially the relationship between security and privacy, which represents a balancing act that is becoming increasingly difficult due to enhanced use of systems utilizing weak signals.

4 Weak Signals in Public and Information Security

On one view, authorities, be they government officials or corporate managers, consider their subjects at best as irresponsible accomplices to potential terrorists or at worst as wilful rogues waiting for an opportunity. However, from another perspective, authorities and citizens can be seen to work together, one supporting the other, to avert outside threats and presenting a unified front against terrorism and crime. Protecting human lives is the basic duty of our society and a critical prerequisite for its survival. Key roles in this undertaking are played by the detection, analysis and interpretation of weak signals.

A rigorous discussion has been conducted on why authorities failed to respond to a set of weak signals preceding September 11 attacks. Among the wealth of information amassed by the various governmental offices in the United States, were positive indications that airplanes would be hijacked and used to cause great damage. Although a few percipient individuals reacted to that data, the sheer number of potential threats and acts of terrorism had desensitized American security officials with the result that these weak signals were largely ignored and no preparations were made to prevent a critical, and — as it turned out — highly fatal threat.

This has changed completely. Lately, the news has been rife with reports on situations, where authorities have reacted very strongly to weak signals. The aviation industry has been particularly affected by this change of attitude. A number of flights have been cancelled during the past few years due to intercepted messages interpreted as indicating a potential threat to passenger safety. As it turned out, most of these situations were harmless, but the fact that the authorities responded so intensely is indicative of a raised awareness level concerning weak signals. And quire rightly, every critical signal should be responded to, no matter how feeble it is.

The London bomb blasts on July 7, 2005, have reverberated across the globe. Some reports of the events have included a small mention of the London Metropolitan Police having prevented 8 similar bombing attacks from taking place during the past two years. We may only guess what kind of weak signals led the police to the trail of the would-be terrorists in each case. At any rate, news such as this one are indicative of the increasingly prominent role of the use of weak signals in our society. Needless to say, weak signals are not only meaningful for our personal safety as shown by the examples above, but also to the security of the information we possess and use.

For the last 25 years, information security management has made steady progress. Ideas inherent in the model presented by the Royal Canadian Mounted Police [15] have been applied in various forms to the management of corporate security and information security. The model has served to instil in people a sense of the different layers of information security and their effects and interconnections. Thus, abstract and oftentimes abstruse notions have turned into more tangible descriptions that are easier to grasp. In reality, however, every information security related event is unique, a singular occurrence [3].

Information security can be characterized as a continuous balancing act between security and privacy. However, this already precarious situation is being further complicated by a new factor which has emerged in combination with the increased use of weak signals.

Introducing weak signals into the management of information security involves realizing that each dimension of information security has a weighted effect on all the other dimensions. As a result, a new era is dawning also in information security management. To meet the challenge, information security and information security management must be constructed from the bottom up, on the basis of the systems and their component processes. Functional information security cannot be purchased as a separate add-on module, it must be incorporated into all information systems during the initial planning process.

5 Conclusions and Future Work

This paper has focused on discussing weak signals in the context of information security, particularly from the threat perspective. A presentation of types of weak signals included a discussion on the significance of their cross-relationships, problems inherent in the granularity of threat analysis, changes in time and trends, the subjective and often deficient nature of information available to threat analysis as well as human behaviour. The detection of weak signals was found to rely heavily on security metrics providing proof of information systems' level of security or performance. Collecting and analyzing pre-recorded event material is of utmost importance for lowering the detection threshold and enabling the perception of faint patterns that may emerge from the data.

Responding to weak signals is just as important as detecting them, albeit often no less difficult. One problem is the sheer number of weak signals present in any system. Other difficulties include assessing and validating the value of the detected signals, which in many cases is done in a hierarchical structure, i.e., the person who detects a potential threat does not have the authority to act upon the matter and initiate a quick response.

Both the detection and analysis of weak signals require further work, but there are also related questions, pertinent to privacy and other civil rights concerns that must be addressed. This is particularly important today, after the recent wave of bomb blasts which have heightened the detection sensitivity of authorities across the globe.

References

1. Ansoff, I.: Managing Strategic Surprise by Response to Weak Signals. In: California Management Review, Vol. XVII, No. 2 (1975)
2. Ansoff, I.: Implementing Strategic Management, Prentice Hall (1985)
3. Anttila, J., Kajava, J., Varonen, R.: Balanced Integration of Information Security into Business Management. In: Proceedings of the Euromicro 2004 Conference. IEEE Computer Society, Los Alamitos, California, USA (2004)
4. British Standards Institution: A Code of Practice for Information Security Management, Department of Trade and Industry, DISC PD003 (1993)
5. British Standards Institution: BS 7799-2. Information Security Management Systems – Specification with Guidance for Use, Part 2, London (2002)
6. Gustafsson et al.: Uuden sukupolven teknologiaohjelmia etsimässä (In Finnish, Executive Summary in English available), TEKES National Technology Agency of Finland (2003)
7. Information Technology Security Evaluation Criteria (ITSEC) Version 1.2, Commission for the European Communities (1991)
8. INFOSEC Assurance Capability Maturity Model (IA-CMM), Version 3.0 (2003)
9. ISO/IEC 15408. Information Technology – Security Techniques – Evaluation Criteria for IT Security (1999)
10. ISO/IEC 17799. Information Technology – Code of Practice for Information Security Management (2001)
11. ISO/IEC 21827. Information Technology – Systems Security Engineering – Capability Maturity Model SSE-CMM (2002)
12. ISO/IEC JTC1/SC27. Guidelines for the Management of IT Security (1995)

13. London, K.: The People Side of System – The Human Aspects of Computer Systems, McGraw-Hill, London (1976)
14. Murine, G. E. and Carpenter, C. L. Measuring Computer System Security Using Software Security Metrics, in Finch, J. H. and Dougall, E. G. (eds), In Computer Security: A Global Challenge, Elsevier Science Publisher, Barking (1984)
15. Royal Canadian Mounted Police: Security in the EDP Environment. Security Information Publication, Second Edition. Gendarmerie Royale du Canada, Canada (1981)
16. Stacey, T. R. Information Security Program Maturity Grid, In Information Systems Security, Vol. 5, No. 2. (1996)
17. Trusted Computer System Evaluation Criteria (TCSEC) "Orange Book", U.S. Department of Defense Standard, DoD 5200.28-std (1985)
18. VAHTI – Valtionhallinnon tietoturvallisuuden kehitysohjelma 2004-2006 (The Finnish Government Information Security Development Program 2004-2006), Finnish Ministry of Finance, In Finnish, English summary available (2004)
19. Voas, J. Why is it so Hard to Predict Software System Trustworthiness from Sofware Component Trustworthiness? In Proceedings of the 20th IEEE Symposium on Reliable Distributed Systems, p. 179 (2001)

PDTM: A Policy-Driven Trust Management Framework in Distributed Systems*

Wu Liu, Haixin Duan, Jianping Wu, and Xing Li

Network Research Center of Tsinghua University, Beijing 100084, China
liuwu@ccert.edu.cn

Abstract. This paper presents a policy-driven trust management framework (PDTM) which is composed of five interfaces to feature the fully decentralized and policy-driven framework. The transmission interface allows trust instances to be exchanged between principals. The trust induction interface encapsulates the evaluation of policies and answers queries made against these policies. The trust management interface allows the trust instances including collection, storage and retrieval to be downloaded from small mobile devices, where resources are limited. The policy inquiry interface is designed to facilitate communication between strangers, so that unknown policies can be discovered through a query-based process. The trust agent interface, on the other hand, is designed to automate communication between strangers.

1 Introduction

In general, traditional distributed authorization approaches are based on either identity or capability. The identity-based approaches focus on authentication, and it stimulated much research on cryptographic protocols [1] that allow communicating parties to identify each other and often also establish a shared secret for securing communication sessions. While capability-based systems [2][3] take a different approach which rely on capabilities contained in credentials to grant or deny access.

Decentralizing policy management based on authority delegation is one of the key features of current trust management. The ultimate principal may present the set of credentials at the resource provider. Delegation of authority is not unique to trust management. It also forms the basis of key-oriented access control, whose representative systems include DSI [5] and PKI [6]. The key idea of such access control systems is to treat public keys as principal identifiers and it naturally relies on the use of public key cryptography to provide principal authentication. While such systems did not focus on the design of unified security framework, because of their well-defined compliance computation [4], they may be considered as a form of trust management system. Although the current distributed authorization is a major improvement over the traditional methods, today's modern distributed applications generate new requirements that need to be addressed.

* This work is supported by grants from the National Natural Science Foundation of China (Grant No. #60203044 & #2003AA142080 & #90104002) and China Postdoctoral Science Foundation #2005037070.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 518 – 525, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper we address issues previously unresolved by the current state-of-the-art with a new trust management system, called PDTM (Policy-Driven Trust Management framework) which features a fully decentralized and policy-driven framework.

This paper is organized as follows. Section 2 describes its architecture which consists of a collection of nodes implementing interfaces. Section 3 describes in detail all the interfaces that facilitate trust management. And finally, we conclude in Section 4.

2 Architecture of PDTM

Seen in figure 1, the PDTM consists of a collection of nodes [8] which is a processing entity for messages, and may generate messages for other nodes. Each node may provide services as methods. These methods are mapped directly into PDTM actions, where the method name maps to the action name, and arguments of a method invocation map to parameters of an action instance. A node may also implement a number of interfaces to support trust management services, in addition to its own methods. These interfaces include interfaces which are collectively referred to as the PDTM interfaces:

(1) Transmission interface
(2) Trust induction interface
(3) Trust management interface
(4) Policy inquiry interface
(5) Trust agent interface

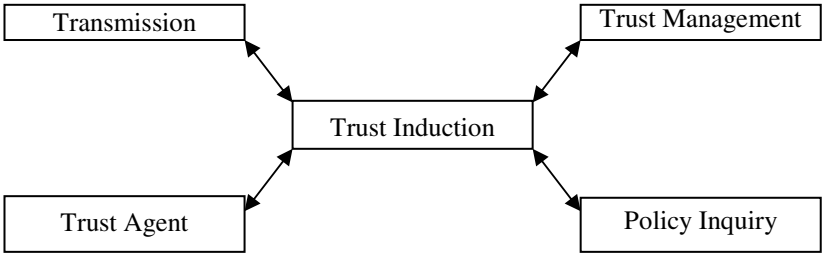


Fig. 1. Architecture of PDTM

These five interfaces are collectively known as the PDTM interfaces which facilitate trust management.

- The transmission interface allows trust instances to be exchanged between principals.
- The trust induction interface is the core of the architecture, which encapsulates the evaluation of policies and answers queries made against these policies.
- The trust management interface allows the management of trust instances, including collection, storage and retrieval. It is designed specifically to download these tasks from small mobile devices, where resources are limited.

- The policy inquiry interface is designed to facilitate communication between strangers, so that unknown policies can be discovered through a query-based process.
- The trust agent interface, on the other hand, is designed to automate communication between strangers.

3 Interfaces of PDTM

This section describes all the interfaces in detail.

3.1 The Trust Induction Interface

The trust induction interface encapsulates the evaluation of both trust policies and action policies.It has a single method, *induc*:

InfResult induc (CredentialSet trust_instances,QueryTemplate template, Environment environment, int flag)

The argument trust_instances is a set of trust instances, the template is either a trust or an action template, the environment is a set of name-value bindings, and the flag specifies additional settings. The return value contains either a set of new trust instances or a Boolean value, and optionally an induction trace. The evaluation of this method depends solely on the given arguments, which are either provided by the requester or collected from the context. There is no state maintained between invocations. The provided mechanism for inducing trust decisions is therefore stateless.

The representation of a trust template can be written as:

name (param; …) : truster →subject

Where *name* gives the name of a trust statement or a wildcard, *param* may either be a value or a variable, *truster* and *subject* may either be a principal identifier, a variable or a wildcard. When name is a wildcard, *param* will become irrelevant. An action template can be written similarly as: *name (param; …) : requester*

Where *name* must refer to the name of an action and no wildcard allowed, *param* may either be a value or a variable, and the requester may be a principal identifier, a variable or a wildcard. Depending on the information in a template, induc answers six types of query:

(1) **Trust establishment.** Given name, truster and subject, it determines if the named trust statement could be instantiated. If successful, return parameter values of the result trust instance.
(2) **Horizontal coverage.** Given name and truster, it determines the complete set of principals who may obtain trust statement name from the specified truster, and the parameter values for the result trust instances.
(3) **Vertical coverage.** Given either truster or subject, or both, it computes the complete set of trust instances between them. This determines the maximum trust that a truster can assert at the time the induction is executed.
(4) **Complete coverage.** Given a blank trust template, it computes the complete set of trust instances the policies may give.

(5) **Action decision.** Given name and param, it determines whether it can be satisfied. This is the typical type of query for determining access control decisions, and is also called authorization.
(6) **Action coverage.** Given name, it determines the complete set of action instances the given set of trust instances satisfy.

A trust instance passed to *induc* may either be an actual instance or a reference to locate it. A mechanism for automated credential collection is described in the next section. The method argument flag may indicate if the caller wishes to obtain the induction trace. However, the processing node may refuse or limit how much of the trace should be provided. It may pose a security risk if the complete trace is fed back since it enables probing of internal policies. Nevertheless, the induction trace is valuable for auditing purposes, especially if an induction node is only used internally.

A node supporting the trust induction interface is associated with a set of trust policies. These could either be directly provided to the node at deployment, or be retrieved from a policy inquiry node, seen in Section 3.4.

3.2 The Transmission Interface

The transmission interface defines the mechanisms for trust transmission, supporting point-wise transfer of trust instances. It defines two styles of interfaces: push and pull. For the push interface, the transmission source initiates the transfer, while for the pull interface, the transmission target requests certain trust instances. A node may offer either or both styles. The push-style interface is suitable for a principal to actively distribute trust knowledge as it is gained, whereas the pull-style interface is suitable for a principal to passively share trust knowledge.

The pull interface defines a *getTrustInstance* method that takes a source identifier, a target identifier and a trust template1. A trust template can be thought of as a trust instance with unfilled parameters and/or truster and subject. When invoked, the node first determines whether it represents the source principal. If so, it returns the trust instances matching the trust template. The target identifier is not directly used, but gives supplementary information that may be useful for audit or security purposes.

For the push interface, a source sends trust instances asynchronously. The target principal first registers for transmission, expressing its interest in certain trust instances. A registration request consists of a source identifier, a target identifier, a trust template, reply addresses of the target principal and a registration policy. Once a registration is received, the node determines if the requested transmission is allowed and raises an exception if not. Otherwise, it adds the registration into a registration table, which contains entries of registration, indexed by the source principal and the name of trust instances. When the node learns about a new trust instance owned by principals it represents, it checks through the table and initiates the transmission process according to the policy of each registration if matches are found. A node learns about new trust instances from a number of sources: directly from those principals it represents transmission from other principals, or as results of trust induction, seen in Section 3.1. The registration policy is a collection of name-value pairs, expressed in an XML [7] format. It adjusts how a registration is handled. A node should respect the registration policies it understands, and may discard those it does not. This processing semantics allows application-specific, extended policies to be supported.

Note that the processing of certain registration policies may require the source principal to maintain states about the interested target principals. However, this is different from the mechanism for processing trust instances, which is stateless. This will be further clarified in the next section.

3.3 The Trust Management Interface

Typically, a principal would manage its own trust instances. Under some circumstances, it is desirable to delegate these tasks to another trusted node that offers the trust management interface.

Conceptually, a trust management node maintains a collection of credential bags, where each bag contains credentials owned by a principal. The trust management interface offers two categories of methods: privileged and public. Privileged methods can only be invoked by the owners, while public methods are available to anyone. A request to a privileged method needs to be signed by the requester. The node, upon receiving the request, needs to determine its integrity and freshness, and whether the requester is permitted for the requested method. There are four privileged methods:

- *addCredential*
- *removeCredential*
- *getMatchedCredentials*
- *getAll-Credentials*

These methods are defined as follows:

CredentialRef addCredential (Credential trust_instance)
void removeCredential (CredentialRef ref)
CredentialSet getMatchedCredentials (QueryTemplate template)
CredentialSet getAllCredentials ()

All these methods identify the credential bag to operate on using the requester's identity, and as invocation of these methods must be signed by the requester, the requester identity is always known. It is hence unnecessary to explicitly add an argument to these methods to identify the requester.

The addCredential method adds a credential to the bag belonging to the requester and returns a reference key.

The removeCredential method is used to remove the credential referenced by the key given as an argument from the requester's bag. Both getMatchedCredentials and getAllCredentials return multiple credentials of the requester.

The getAllCredentials method returns all credentials belonging to the requester.

The getMatchedCredentials method takes a trust template as argument, and returns all credentials matched by the template. This allows selective retrieval of credentials.

Public methods include getCredential and getCredentials, defined below:

Credential getCredential (CredentialRef ref)
CredeentialSet getCredentials (CredentialRefSet ref_set)

The getCredential method takes a single reference key and returns the credential, and the getCredentials method works similarly but for a set of credentials. The security of these methods lies in the quality of the reference keys. While the format for keys is implementation-specific, the keys should not be predictable to prevent creden-

tial retrieval based on key guessing. The recommended approach for producing the keys is to use cryptographic hash algorithms on the credentials which will provide appealing uniqueness and unpredictability properties.

The trust management interface is designed to facilitate automated credential collection. Recall from the previous section that trust instances passed to an induction interface may either be concrete instances or references. A reference consists of a pair of URL and key, where the URL addresses a trust management node and the key is the local reference at the node to locate the trust instance. A compliant trust induction node automatically fetches trust instances using getCredential and/or getCredentials prior to performing the induction.

3.4 The Policy Inquiry Interface

The policy inquiry interface specifies methods for querying and retrieving policies. It is designed to facilitate communication between strangers and enable centralized management of policies. There are two approaches to achieve this:

(1) **A node may publish its ontology and policies.** This could either be distributed at some well-known location or retrieved directly from the node, provided it supports a retrieval interface. The policy document should be described as a PDTM Policy Interchange document, which has an XML-structured format.
(2) Alternatively, **a node may support programming interfaces for interrogating and discovering its policies by supporting the interface described in this section.** This approach provides an opportunity for automating the process of communication establishment between strangers. This will be explained later in the section.

The former is suitable for which authority hierarchy exists, while the latter is much more dynamic. It allows strangers to gain understanding and form trust relationships at runtime. This therefore requires more runtime and deployment support. It also enables centralized policy management. Centralized policy management is particularly suitable in two situations:

(1) For an organization, policies often tend to be large and complex. Centralized management helps reduce administrative burdens and errors because policies can be specified and analyzed at a single location, in a consistent manner.
(2) For mobile computing, where resources are scarce and constrained, managing policies on a separate, perhaps non-mobile, node helps reduce storage and bandwidth usage.

The policy specification framework supported by the inquiry interface is based on the PDTM Policy Language. Recall that policies include trust policies or action policies. We use the term metadata to refer to the specification of trust statements and actions. The method *getTrustSpec* and *getActionSpec* both take a name, and retrieve the specification of the named trust statement and action respectively. The specification is given as a fragment of a PDTM Policy Interchange document. For example, suppose A1 is declared as A1 (string x, float y) in the PDTM Policy Language. The equivalent declaration in PDTM Policy Interchange document would be:

```
<Statement name="A1">
<Parameter name="x" type="xsd:string" />
<Parameter name="y" type="xsd:float" />
</Statement>
```

This fragment creates a trust statement type which may be referenced by the policy specification returned by method *getTrustPolicies* and *getActionPolicies*. The method getTrustPolicies and getActionPolicies respectively take a trust template and an action template as argument, and return a set of policies matching the template. The rule for determining the relevance of a policy with regard to a template is based on static matching. For trust policies, the template is matched against the trust template in the assert clause; for action policies, the action template in the grants clause is matched.

3.5 The Trust Agent Interface

As previously mentioned, when strangers make contact, there are several issues to be resolved, e.g. unknown policies and credential ontology, limited mutual trust, etc. Even when policies are known, it is still in the interest of a requester to disclose the least trust instances for a request, especially if they contain sensitive information. The trust agent interface is designed to encapsulate a principal, providing an active interface on behalf of the principal. It automates the process of policy inquiry and negotiation, and computes the disclosure set of credentials for requests. It supports the use of meta-policies to control and constrain the automation. Meta-policies are just like other policies and may also be expressed in the PDTM Policy Language. However, unlike other types of policy which are about trust relationships or actions, meta-policies are about policies.

A trust agent may provide assistance on several aspects. It may act as a front-end for a principal, providing a high-level interface for services. In this configuration, the principal delegates the task of selecting trust instances to the trust agent, and it issues a high-level request for service without attaching trust instances to the trust agent. The trust agent then examines the action policies, selects trust instances and finally issues the actual request with those trust instances to the service node.

One aspect of meta-policies is to allow principals to dictate the rules for choosing credentials for requests. The profile is designed to express four types of conditions:

- Designated principal disclosure
- Context-specific disclosure
- Trust-directed disclosure
- Mutual exclusion

A principal may extend this profile or use other proprietary policies to express its meta-policies if needed.

Trust agents can also automate negotiated requests. A negotiated request is an approach to mechanize the policy negotiation process. The idea of negotiated request is that a pair of trust agents carries out a negotiation conversation, gradually exchanging trust instances. When sufficient trust is gained on both sides, the request will be performed.

Note that meta-policies in a negotiated request session play two different roles. On the requester side, meta-policies are used to specify what credentials may be disclosed

and their conditions. On the responder side, in addition, meta-policies specify conditions for disclosing policies.

4 Conclusion

This paper presents a policy-driven trust management framework (PDTM) which is composed of five interfaces. The transmission interface allows trust instances to be exchanged between principals. The trust induction interface is the core of the architecture, which encapsulates the evaluation of policies and answers queries made against these policies. The trust management interface allows the management of trust instances, including collection, storage and retrieval. It is designed specifically to download these tasks from small mobile devices, where resources are limited. The policy inquiry interface is designed to facilitate communication between strangers, so that unknown policies can be discovered through a query-based process. The trust agent interface, on the other hand, is designed to automate communication between strangers.

References

1. B. Lampson, M. Abadi, M. Burrows, and E. Wobber: Authentication in distributed systems: Theory and practice.ACM Transactions on Computer Systems, vol. 10, Nov. (2002)265-310
2. J. A. Bull, L. Gong, and K. R. Sollins: Towards security in an open systems federation.in European Symposium on Research in Computer Security (ESORICS) (2002)3-20
3. R. Hayton, OASIS: An Open Architecture for Secure Interworking Services. Univeristy of Cambridge Computer Laboratory(2002)
4. M. Blaze, J. Feigenbaum, J. Ioannidis, and A. D. Keromytis: The role of trust management in distributed systems security. in Proceedings of Fourth International Workshop on Mobile Object Systems: Secure Internet Mobile Computations, No. 1603 in Lecture Notes in Computer Science, Springer-Verlag, July(1999)185-210,
5. R. L. Rivest and B. Lampson: SDSI-A simple distributed security infrastructure.http: //theory.lcs.mit.edu/~rivest/sdsi10.ps, Aug. (2004)
6. C. M. Ellison, B. Frantz, B. Lampson, R. Rivest, B. Thomas, and T. Ylonen: SPKI certificate theory. RFC 2693, Internet Engineering Task Force, Sept. 1999. See http://www.ietf.org/rfc/ rfc2693.txt
7. World Wide Web Consortium, Extensible Markup Language (XML) 1.0, 2nd Ed., Oct. 2000. http://www.w3.org/TR/2000/REC-xml-20001006
8. M.Gudgin, M. Hadley, J.-J. Moreau, and H. F. Nielsen, SOAP Version 1.2 Part 1: Messaging Framework (W3C Working Draft 17 December 2001). World Wide Web

Methodology of Quantitative Risk Assessment for Information System Security

Mengquan Lin[1], Qiangmin Wang[2], and Jianhua Li[1]

[1] Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200030, China
linmq@cdgdc.edu.cn, lijianhua888@sjtu.edu.cn
[2] School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200030, China
qmwang@sjtu.edu.cn

Abstract. This paper proposes a security assessment method of information system based on mixed methods of constructing weights of criteria, which indicate how to evaluate the overall security of information system in a synthetic and quantitative way from the aspect of confidentiality, integrity, availability and controllability of the information system security.

1 Introduction

With the development of information technology, Internet has been an important tool of information spread. We are afflicted by the security problems of information while benefiting much from it. So there is a requirement to evaluate the security of information system to make sure its robust resistance to the potential risks.

Risk management has been introduced to solve these problems. It concerns the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level. Although all elements of the risk management cycle are important, risk assessments provide the foundation for other elements of the cycle. In particular, risk assessments provide a basis for establishing appropriate policies and selecting cost-effective techniques to implement these policies. Most of the risk assessment methodologies have been qualitative - that is, their assessment of probability and consequence of risks is based on a "low/medium/high" characterization. Qualitative information security risk management standards are included in the US Federal standards [1]. Recent guidelines that recommend qualitative risk analysis techniques are included in [2-3].

Quantitative risk analysis is used extensively in disciplines other than information security, including finance, healthcare, and safety[4]. Information risk assessment is still a relatively immature and evolving field within information security and information systems. This may account for why almost all of the major methodologies are qualitative, which would allow for cost/benefit analyst. This paper will focus on quantitative risk analysis of information system and a risk assessment model will be built based on the analysis.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 526 – 531, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 The Risk Analysis of Information System

The objective of information security is to preserve the agency's information assets and the business processes in the context of the following four attributes, i.e. confidentiality, integrity, availability and controllability.

When one of the four attributes is faced with threats, the information system security will be at risk. So in our risk assessment model, the overall objective is information system security. The first level criteria are the four attributes. It is shown in figure 1. Then we construct hierarchy structure of the risk of the four criteria

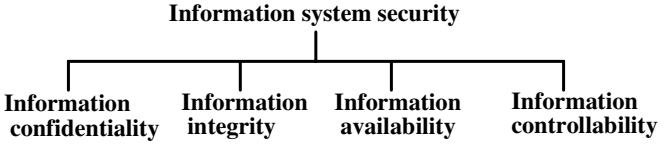


Fig. 1. The model of information system risk assessment

respectively. Here we give an example of risk of information confidentiality. The hierarchy structure of the risk in information confidentiality is shown in figure 2. When the hierarchy structure of the risk is constructed, the ISO/IEC 17799 standard is used[5]. After constructing the risk model, the most important thing of risk assessment is to assign accurately the weights for criteria. In next section, we will describe the methodology of constructing weights for the criteria.

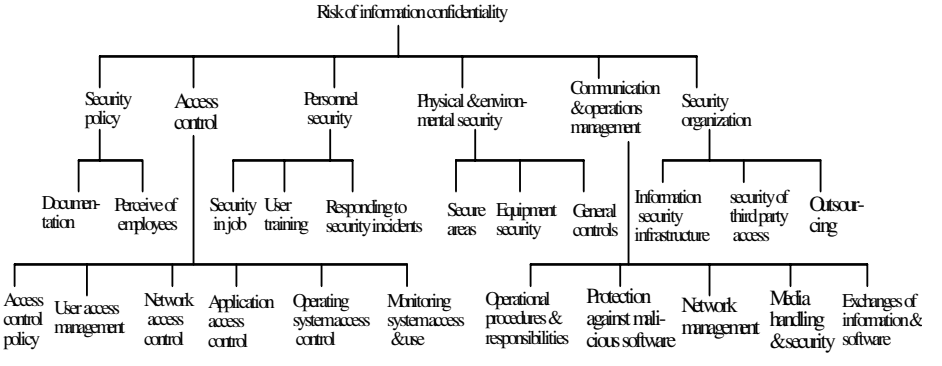


Fig. 2. The hierarchy structure of the risk in information confidentiality

3 The Methodology of Constructing Weights for the Criteria

It is important to estimate accurately the weights of the criteria. The adverse impacts of threats are calculated by using those weights. When determining those weights, one can rely on his knowledge of specialty, which is not a general method. It is impossible to assign those weights exactly.

In this paper, a mixed method of constructing weights of criteria based on Delphi Method and Analytic Hierarchy Process (AHP)[6-7] is proposed. The method is abbreviated as MMCW. In MMCW, based on their attributes, different methods of constructing weights are used in different level of criteria. The first level criteria are confidentiality, integrity, availability and controllability. The lower level of criteria affects the result of risk assessment through affecting the upper level criteria level by level. The higher the level is, the more important the weights of criteria are for risk assessment. The Delphi method allows us a better use of experts and is more practical and accurate. Therefore, the Delphi Method and AHP are used to construct the criteria weights for the first level criteria and the lower level criteria respectively. Another advantage of MMCW is that we can adjust the weights of the first four level criteria for different information systems. Different information systems may emphasize particularly on different criteria. For example, the information systems of government departments may emphasize particularly on the confidentiality, while finance group may focus on the confidentiality and the integrity.

3.1 The Delphi Method

The Delphi method may help leaders get better forecasts and advice from their advisors. It recognizes unencumbered human judgment not only as legitimate but possibly superior as a necessary input to decision-making in multiple contexts. By using anonymity in the reaction process from the participants it bypasses or avoids the typical socio-psychological barriers and traps, created by the well-known phenomenon of dominating or domineering personalities. Since there are several "rounds" of a Delphi, the final results are often a consensus forecast or judgment. The Delphi technique is a very useful method to obtain refined, hence better, results from expert opinions. It might be of practical use for deciding the first level weights of the criteria in risk assessment. Suppose that there are 4 criteria and 10 experts taking part in the work. The table distributed to experts in the first round is shown in Table 1.

Table 1. The consultation table for the weights of system information security

	Very important (3)	Important (2)	General (1)	Unimportant (0)
Confidentiality				
Integrity				
Availability				
Controllability				

The number in ( ) represents relative importance.

After the table in the first round received, it needs to be dealt with statistical analysis. The two steps of the statistical analysis are described as follows:

Step 1. Calculate average weight $\overline{w_i}$ of the weight w_i of the *i*th criterion in the first round consultation using Formula 1. w_{ij} represents the weight of the *i*th criterion given by the *j*th expert.

$$\overline{w_i} = \frac{1}{m}\sum_{j=1}^{m} w_{ij} \quad (i=1,2,\ldots, n) \tag{1}$$

Step 2. Calculate the deviation Δ_{ij} using Formula 2, which is the weight w_{ij} minus $\overline{w_i}$

$$\Delta_{ij} = w_{ij} - \overline{w_i} \quad (i=1,2,3,4), (i=1,2,\ldots, 10) \tag{2}$$

Then the above results are sent back to experts. The second round consultation begins and at the same time requests that those experts, who give the bigger Δ_{ij} in the first round, give a new judgment. After several rounds of a Delphi, the final results are often a consensus judgment. Then the final results of experts consultation are normalized, the weights of first level criteria are gained.

Although it can gain more accurate results by constructing weights of criteria based on Delphi Method through several rounds of expert consultation, Delphi Method takes long time to get results. It is not suitable to those projects that must be done in a short time. So AHP is proposed to deal with the weights of the lower level of criteria.

3.2 Analytic Hierarchy Process

We use AHP to estimate the weights of the lower level of criteria. AHP is used to determine the weights, which can avoid some complicated calculation. The AHP divides problem spaces into hierarchies and is visualized by the use of tree maps. It is designed for less tangible, qualitative considerations such as values and preferences. A pair-wise rating of criteria (attributes) is performed, then normalized. A scale of 1 (equal importance) to 9 (absolutely more important) is used. The weights are averaged for each criterion. See reference [8] for further information. The weights of the lower level of criteria are given through using matrix $\boldsymbol{A}$: $\boldsymbol{A} = (a_1, a_2, ., a_p)$.

4 Determination of the Synthetic Judgment

In above section, we have analyzed the risk factors of an information system, and constructed the hierarchy structure of the risk of an information system. Then the methodology of constructing weights for the criteria is described. In order to quantify the risk of an information system, the last step is to calculate the value of risk assessment.

4.1 Determining the Subjection Degree of the Criteria

When the risk of an information system is assessed, several experts need to evaluate the risk of the criteria based on a given comment level group. The given comment level group may be {very low(5), low(4), medium(3), high(2), very high(1)}. The results of experts' evaluation are shown with subjection degree matrix $\boldsymbol{R}$: $\boldsymbol{R}=(\mathbf{r}_{ij})_{p\times l}$ In the matrix $\mathbf{R}$, $\mathbf{r}_{ij}$ represents that for the ith criterion the percentage of experts who

give the *j*th level evaluation, i.e., $\mathbf{r_{ij}=d_{ij}/D}$. $\mathbf{d_{ij}}$ represents for the *i*th criterion the number of experts who give the *j*th level evaluation, D is the total number of experts. *p* is the number of the criteria, and *l* is the number of comment levels.

4.2 Determining the Synthetic Judgment of the Risk Assessment of Information System

After gaining the subjection degree matrix **R,** we need calculate the synthetic value of risk assessment of information system. All factors should be considered in the synthetic risk assessment. So the weighted-mean model is selected. Formula 3 is used to calculate the synthetic risk of the assessed information system. The results of the calculation determine the risk level of the assessed information system.

$$B = A \circ R = (a_1, a_2, \cdots a_p) \circ \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{bmatrix} = (b_1, b_2, \cdots b_m) \tag{3}$$

$$b_j = \sum_{i=1}^{p} a_j r_{ij} \tag{4}$$

5 Results

In this section, an example is given to show how the AHP is used to estimate the weights of the lower level of criteria . According to the method of reference [8], when estimating the weights of criteria, the criteria are subject to a series of paired comparisons (ranging from 1/9 to 9) and a pairwise comparison matrix is built. The pairwise comparison matrix is shown in table2. In table2, ISI and STPA are the abbreviations for information security infrastructure and security of third party access respectively. The first step is to calculate the normalized eigenvectors, the results are shown in table 2. The normalized eigenvectors are weights of the criteria we want to estimate. The second step is to calculate the maximum latent root of the matrix by using Formula 5. The third step is to calculate the Consistency Index and the Consistency Ratio by using Formula 6 and Formula 7 respectively. The consistency ratio shows how consistent the evaluation of the criteria is during the comparison. A low value (ideally below 0.1) is proof of stable consistency. In this way all of the weights of the lower levels of criteria can be calculated.

Table 2. The pairwise comparison matrix for the criteria of criterion security organization

	ISI	STPA	Outsourcing	Weights
ISI	1	2	5	0.321
STPA	1/2	1	3	0.109
Outsourcing	1/5	1/3	1	0.570

$$\lambda_{\max} = \frac{1}{m}\sum_{i=1}^{m}\frac{(TA)_i}{w_i} = 3.006 \tag{5}$$

$$C.I. = \frac{\lambda_{\max} - m}{m-1} = \frac{3.006 - 3}{3-1} = 0.003 \tag{6}$$

$$C.R. = \frac{C.I.}{R.I.} = \frac{0.003}{0.58} = 0.005 \prec 0.1 \tag{7}$$

where C.I. = Consistency Index; C.R. = Consistency Ratio; R.I. = Randomize Index; $\lambda_{\max}$ =the maximum latent root of the matrix; m = dimension of the matrix.

6 Conclusion

In this paper, the methodology of quantitative risk assessment for information system security is proposed. In the proposed method, the four security attributes of an information system security are assessed separately. So the method is very flexible. The determination of weights is a subjective process. Therefore, how to assign the weights correctly is the key to the successful use of the method. Delphi Method and AHP are used to determine the weights for different level of criteria.

References

1. US Department of Commerce/National Institute of Standards and Technology: Guideline for the Analysis of Local Area Network Security (1994)
2. US General Accounting Office: Information Security Risk Assessment: Practices of Leading Organizations (1999). URL: http://www.gao.gov/special.pubs/ai00033.pdf
3. NIST Special Publication SP800-30: Risk Management Guide for Information Technology Systems (2002) URL:http://csrc.nist.gov/publications/nistpubs/800-30/sp800-30.pdf
4. Jacobson Robert: Quantifying IT Risks, ITAudit (2002). URL: http://www.theiia.org/itaudit/index.cfm?fuseaction=forum&fid=479
5. BS7799-1:1999 Information security management-Part1: Code of practice for information security management; Part2: Specification for information security management
6. Harold A. Linstone and Murray Turoff: The Delphi Method: Techniques and Applications. Massachusetts: Addison-Wesley (1975)
7. Xu Shubo: Practical decision-making method: the theory of analytic hierarchy process. Tianjin University press (1988)

A Secure and Efficient (t, n) Threshold Verifiable Multi-secret Sharing Scheme*

Mei-juan Huang[1], Jian-zhong Zhang[1], and Shu-cui Xie[2]

[1] College of Mathematics and Information, Shaanxi Normal University, Xi'an 710061, China
huangmeijuan@stu.snnu.edu.cn
[2] Department of Applied Mathematics and Physics, Xi'an Institute of Post and Telecommunication, Xi'an 710061, China

Abstract. Ting-Yi Chang et al.(2005) have proposed an efficient (t, n) threshold verifiable multi-secret sharing (VMSS) scheme, which is more secure than the one adopted in Lin and Wu (1999) and it can provide more efficient performance than the other VMSS schemes in terms of computational complexity. However, this paper will show that Chang et al.'s scheme is in fact insecure by presenting a conspiracy attack on it. Furthermore, a more secure scheme is proposed.

1 Introduction

In 1979, the first (t, n) threshold secret sharing schemes were proposed by Shamir and Blakley, respectively. However, there are several common drawbacks in both the secret sharing schemes. For example, only one secret can be shared during one secret sharing process; the cheating by the dealer and the cheating by any participant cannot be detected [4].

Later, taking all the drawbacks into consideration, several (t, n) threshold verifiable multi-secret sharing (VMSS) schemes were proposed. In a VMSS scheme, there are multiple secrets to be shared during one secret sharing process, and the cheating by the dealer and that by any participant can be detected. In 1995, Harn proposed an efficient threshold verifiable multi-secret sharing scheme [1], which is based on Lagrange interpolating polynomial and DSA-type digital signatures. In 1997, Chen et al. proposed an alternative (t, n)threshold VMSS scheme [2]. In 1999, Lin and Wu have further proposed a (t, n) threshold VMSS scheme [3] based on the intractability of factorization and the problem of discrete logarithm module a composite. However, there are different drawbacks (refer to [4] for more details) in the schemes [1, 2, 3]. Recently, Chang et al. (2005) have proposed an improvement (t, n) threshold VMSS scheme [4] based on Lin-Wu's scheme [3], Chang et al.'s scheme not only successfully overcomes the drawbacks of Lin-Wu's scheme, but also provides more

* Supported by the National Natural Science Foundation of China under Grant No.10271069; the Natural Science Research Program of Shaanxi Province of China under Grant No.2004A14; Postgraduates initiative funds of Shaanxi Normal University.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 532 – 537, 2005.
© Springer-Verlag Berlin Heidelberg 2005

efficient performance than other VMSS schemes in terms of computational complexity. However, we will point out that Chang's VMSS scheme cannot withstand conspiracy attack.

In this paper, we will analyze Chang et al.'s scheme and show that arbitrary $t+1$ participants can conspire to obtain the system's secret R with a high probability, then; these malicious participants can reconstruct the shared secret independently. Furthermore, we will try to improve the security of Chang et al.'s scheme while maintaining the original advantages.

2 Review of Chang et al.'s Scheme and Its Weakness

2.1 Review of Chang et al.'s Scheme

We will briefly describe the initialization, shadow generation and secret reconstruction stages of Chang et al.'s scheme (refer to [4] for more details).

2.1.1 Initialization Stage

The dealer (denoted as U_D) creates a public notice board (NB). The parameters are defined by U_D as follows: $N = p \cdot q$, p and q are two large prime numbers, where $p = 2p' + 1$, $q = 2q' + 1$, with themselves prime; $R = p' \cdot q'$; g is a generator with order R in Z_N, $e \cdot d \equiv 1 \bmod \varphi(N)$. U_D puts $\{N, g, e\}$ on the NB and keeps $\{R, d\}$ secret.

2.1.2 Shadow Generation Stage

Let $G = \{U_1, U_2, \cdots, U_n\}$ be a group of n participants and $S = \{S_1, S_2, \cdots, S_m\}$ be a set of m secrets. Every U_i has her/his identity $ID_i (i = 1,2, \cdots, n)$. U_D performs the following steps:

Step 1. Randomly generate a secret polynomial function

$f(x) = a_0 + a_1 x + \cdots + a_{t-1} x^{t-1} \bmod R$, where $a_k \in Z_R$, for $k = 0,1, \cdots, t-1$.

Step 2. Compute a secret shadow x_i for every $U_i \in G$ as $x_i = f(ID_i) \cdot p_i^{-1} \bmod R$, where $p_i = \prod_{U_k \in G, U_i \neq U_j} (ID_i - ID_k) \bmod R$, and distribute x_i to every $U_i \in G$ over a secret channel.

Step 3. Randomly choose m distinct integers $r_1, r_2, \cdots r_m \in Z_R$, for each secret $S_1, S_2, \cdots S_m \in S$; compute $C_j = g^{r_j \cdot d} \bmod N$ and $h_j = (g^{a_0 \cdot r_j \cdot d} \bmod N) \oplus S_j$ for $j = 1,2, \cdots, m$. Then, put $\{r_j, C_j, h_j\}$ on the NB.

2.1.3 Secret Reconstruction Stage

Let $W(|W| = t \le n)$ be any subset of t participants in G. Assume that t participants $U_i \in W$ cooperate to reconstruct a secret $S_j \in S$. Every $U_i \in W$ obtains the $\{r_j, C_j, h_j\}$ from the NB and uses his/her secret shadow x_i to compute a subshadow A_{ij} as $A_{ij} = (C_j)^{x_i} \bmod N$. Then, U_i broadcasts A_{ij} to the other participants in W. Every participant in W can reconstruct S_j as

$$S_j = h_j \oplus (\prod_{U_i \in W} (A_{ij})^{\Delta_i} \bmod N),$$

where

$$\Delta_i = (\prod_{U_k \in W, U_k \neq U_i} (-ID_k)) \cdot (\prod_{U_k \in G, U_k \notin W} (ID_i - ID_k)).$$

Then, all the secrets $S_1, S_2, \ldots, S_m \in S$ can be reconstructed by performing this stage, repetitively.

2.2 Our Attack on Chang et al.'s Scheme

In this section, we will show that any $t+1$ participants can conspire to compute the system's secret R or $\varphi(N)$ with a high probability. This attack refers to the observation of Chuan-Ming Li et al.

Theorem [5]**:** Let $w_1 \equiv y \bmod R$, $w_2 \equiv y \bmod R$, $w_3 \equiv y \bmod R$. If all the values of w_1, w_2 and w_3 are known, then, for some integers z and z', the one can have $w_3 - w_2 = z \cdot R$, $w_2 - w_1 = z' \cdot R$. Therefore, R will be revealed with a high probability by finding the greatest common divisor of $(w_3 - w_2)$ and $(w_2 - w_1)$.

According to the combination theorem, there are $t+1$ possible ways to choose any t participants from $t+1$ participants. Let each choice forms a subset Q_j of t participants. If t participants conspire, then they can compute

$$w_j = \sum_{i \in Q_j} x_i \Delta_i, \quad 1 \le j \le t+1$$

By the theorem, we have $w_1 \equiv a_0 \bmod R$, $w_2 \equiv a_0 \bmod R$, $w_3 \equiv a_0 \bmod R$, …, $w_{t+1} \equiv a_0 \bmod R$. Thus the system secret R will be revealed with a high probability by finding the $\gcd((w_2 - w_1), (w_3 - w_2), \cdots, (w_{t+1} - w_t))$. Once the system secret R is revealed by these $t+1$ participants, they are able to compute a_0 by any one of equations above, then, these malicious participants obtain $\{C_j, h_j, r_j\}$ from the NB and they can reconstruct the secret S_j independently as

$$S_j = h_j \oplus ((g^{a_0} \cdot C_j) \bmod N).$$

3 Improved (t, n) Threshold VMSS Scheme

Initialization stage of the improved VMSS scheme is the same as in 2.2.1.

3.1 Shadow Generation and Verification Stage

Let $G=\{U_1,U_2,\ldots,U_n\}$ be a group of n participants and $S=\{S_1,S_2,\ldots,S_m\}$ be a set of m secrets. Every U_i has her/his identity $ID_i(i=1,2,\ldots,n)$. U_D performs the following steps:

Step 1. Randomly generate a secret polynomial function

$$f(x)=a_0+a_1x+\cdots+a_{t-1}x^{t-1} \bmod R,$$

where $a_k \in Z_R$, and compute a check vector $V=[V_0,V_1,\ldots,V_{t-1}]$ for each coefficient a_k as $V_k=g^{a_k} \bmod N$, for $k=0,1,\cdots,(t-1)$ and put V on the NB.

Step 2. Randomly choose n distinct integers $Z_1,Z_2,\ldots,Z_n \in Z_R$, compute a secret shadow x_i for every $U_i \in G$ as

$$x_i=(f(ID_i)+Z_i)\cdot p_i^{-1} \bmod R$$

where $p_i=\prod_{U_k\in G,U_i\neq U_j}(ID_i-ID_k) \bmod R$, and compute the associated $y_i=g^{x_i} \bmod N$ as this U_i's public key to be put on the NB.

Step 3. Distribute $\{g^{p_i} \bmod N, g^{-Z_i} \bmod N, x_i\}$ to every $U_i \in G$ over a secret channel. When $U_i \in G$ receives the secret shadow x_i, he/she can check the following equation to verify the validity of x_i:

$$(g^{p_i})^{x_i}\cdot g^{-Z_i}=\prod_{k=o}^{t-1}(V_k)^{(ID_i)^k} (\bmod N). \tag{1}$$

If equation (1) does not hold, the secret shadow x_i distributed by U_D is not valid.

3.2 Credit Ticket Generation Stage

In this stage, let $W_k(|W_k|=t\le n)$ is one subset of t participants in G, for $k=1,2,\ldots,C_n^t$. U_D performs the following steps to compute m credit tickets $C_1,C_2,\ldots,C_m$ for each secret $S_1,S_2,\ldots,S_m \in S$.

Step 1. Randomly choose m distinct integers $r_1, r_2, \ldots r_m \in Z_R$, for each secret $S_1, S_2, \ldots, S_m \in S$.

Step 2. Compute a credible ticket C_j and a value h_j as

$$C_j = g^{r_j \cdot d} \bmod N, \quad h_j = (g^{a_0 \cdot r_j \cdot d} \bmod N) \oplus S_j,$$

for $j = 1,2,\ldots,m$. Then, put $\{r_j, C_j, h_j\}$ on the NB.

Step 3. Compute a value $B_j = \{B_{j1}, B_{j2}, \ldots, B_{jC_n^t}\}$, $B_{jk} = C_j^{-\sum_{U_i \in W_k} Z_i p_i^{-1} \Delta_i} \pmod N$, for $j = 1,2,\ldots,m$; $k = 1,2,\ldots,C_n^t$, where

$$\Delta_i = \Big(\prod_{U_k \in W_k, U_p \neq U_i} (-ID_p)\Big) \cdot \Big(\prod_{U_p \in G, U_p \notin W_k} (ID_i - ID_p)\Big).$$

Then, put $\{B_j\}$ on the NB.

In addition, if U_D wants to add a new secret S_{new} for sharing, he/she only needs to generate a new parameters $\{r_{new}, C_{new}, h_{new}, B_{new}\}$ for S_{new} and put it on the NB without interfering with the results generated in the previous stages.

3.3 Subshadow Verification and Secret Reconstruction Stage

Without loss of generality, assume that t participants $U_i \in W_k$ cooperate to reconstruct a secret $S_j \in S$, every $U_i \in W_k$ obtains the $\{r_j, C_j, h_j, B_j\}$ from the NB, and uses his/her secret shadow x_i to compute a subshadow A_{ij} as

$$A_{ij} = (C_j)^{x_i} \bmod N.$$

Then, U_i broadcasts A_{ij} to the other participants in W. Any other cooperator in W obtains U_i's public key y_i from the NB to verify the validity of A_{ij} as

$$(A_{ij})^e = (y_i)^{r_j} \bmod N. \tag{2}$$

If equation (2) does not hold, then they can stop this stage and announce that cheating by U_i has been identified. If all A_{ij}'s released by the t participants in W are valid, every participant in W can reconstruct S_j as

$$S_j = h_j \oplus \Big(\prod_{U_i \in W_k} (A_{ij})^{\Delta_i} \cdot B_{jk} \pmod N\Big)$$

where $\Delta_i = \Big(\prod_{U_k \in W_k, U_p \neq U_i} (-ID_p)\Big) \cdot \Big(\prod_{U_p \in G, U_p \notin W_k} (ID_i - ID_p)\Big)$. Consequently, all the secrets $S_1, S_2, \ldots S_m \in S$ can be reconstructed by performing this stage repetitively.

3.4 Security Analysis

In this section, we will prove our scheme can withstand conspiracy attack that is proposed in 2.2. The other security of our scheme is the same as that of Chang et al's scheme (refer to [4] for more details).

Conspiracy attack: Assume that $t+1$ participants $P=\{U_1,U_2,\ldots,U_t,U_{t+1}\}$ conspire to reconstruct the system secret R,a_0. Obviously, there are $t+1$ ways to choose any t participants from P. Each choice forms a subset P_j, $(j=1,2,\ldots,t+1)$, suppose that t participants $U_{ij}\in P_j$ release their shadows $x_{ij},(i=1,2,\ldots,t)$, and then they can compute

$$\sum_{U_{ij}} x_{ij}\Delta_{ij}=(a_0+\sum_{U_i\in P_j} z_{ij}\Delta_{ij}) \bmod R, \text{ for } i=1,2,\ldots,t;\ j=1,2,\ldots,(t+1)$$

Because $\sum_{u_i\in P_j} z_{ij}\Delta_{ij}$ $(j=1,2,\ldots,t+1)$ are different from each other, R will not be revealed according to the theorem in 2.2. Consequently, a_0 also fails to be worked out. Meanwhile, if the participants want to obtain z_i from the known parameters $\{g^{-z_i} \bmod N, g, N\}$, it is as difficult as breaking the discrete logarithm module a composite (DLMC) problem [6].

4 Conclusion

In this paper, we have analyzed Chang et al.'s (*t*, *n*) threshold VMSS scheme, and then showed that the scheme can't withstand conspiracy attack. Meanwhile, we have proposed an improved (*t*, *n*) threshold VMSS scheme, which can successfully make up deficiency of Chang et al.'s scheme, and the original advantages are maintained.

References

1. L. Harn: Efficient sharing (broadcasting) of multiple secret. IEE Proc. Comput. Digit. Tech. 142(3) (1995) 237-240.
2. L.Chen, D. Gollmann, C.J. Mitchell, P. Wild: Secret sharing with reusable polynomials. In: Proceedings of ACISP's97, 1997, 183-193.
3. T.Y. Lin, T. C. Wu: (t, n) threshold verifiable multi-secret sharing scheme based on factorization intractability and discrete logarithm modulo a composite problems. IEE Proc. Comput. Digit. Tech. 146(5) (1999) 264-268.
4. Ting-Yi Chang, Min-Shiang Hwang, Wei-Pang Yang: An improvement on the Lin-Wu (t, n) threshold verifiable multi-secret sharing scheme. Applied Mathematics and Computation 163 (2005) 169-178.
5. Chuan-Ming Li, Tzonelih Hwang, Narn-Yih Lee: Remark on the Threshold RSA Signature Scheme. In: Stinson D R ed. Advances in Cryptology—Crypto'93 Proceedings. Berlin: Springer-Verlag, (1993) 413-419.
6. L. Adleman, K. McCurley: Open Problems in Number Theoretic Complexity. 2', Lecture Notes Comput. Sci. 877 (1994) 291-322.

Improvement on an Optimized Protocol for Mobile Network Authentication and Security

ChinChen Chang[1,2] and JungSan Lee[2]

[1] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan, China
ccc@cs.ccu.edu.tw
[2] Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan, China
ljs@cs.ccu.edu.tw

Abstract. In 1998, Yi et al. proposed an authenticated key transport protocol for providing secure communications between the base station and the mobile user based on DSA signature scheme. Unfortunately, Laih and Chiou soon showed that Yi et al.'s scheme suffered from the forgery attack which an intruder can forge a valid certificate of a legal user. In this article, we present a simpler attack on Yi et al.'s scheme than that of Laih and Chiou. Furthermore, we also propose an improvement to repair Yi et al.'s scheme. The security of our proposed improvement is also based on DSA signature scheme.

1 Introduction

Due to the rapid development of wireless LAN and mobile networks, the mobile user is allowed to access the services without bounded areas after he/she is authenticated [4, 5, 6]. Yi et al. proposed an optimized authentication protocol for mobile networks based on DSA signature scheme in 1998 [1]. Yi et al.'s scheme consists of two phases: Registration Phase and Service Phase. The certificate of the transmitted message is generated in the Registration Phase while the authentication and key distribution are completed in the Service Phase. However, Martin and Mitchell showed that Yi et al.'s scheme has some security flaws in Service Phase in 1999. Soon, Yi replied to the claim made presented by Martin and Mitchell and showed that the indicated security flaws were not true [2, 3].

Unfortunately, Laih and Chiou pointed out that Yi et al.'s scheme suffered from the forgery attack which an intruder Eve can counterfeit a certificate of the legal user in Registration Phase as long as Eve can get some public information [7]. In this article, we propose a simpler attack on Yi et al.'s scheme than that of Laih and Chiou. Without sophisticated calculation, our proposed attack allows Eve be able to forge a certificate of the legal user easily in the Registration Phase. Therefore, Eve can pretend to be the legal user to communicate with others. What is more, we propose an improvement on Registration Phase of Yi et al.'s scheme, based on DSA signature scheme [10].

The rest of this paper is organized as follows. A review of Yi et al.'s optimized authentication protocol for mobile networks is described in Section 2, followed by a

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 538 – 541, 2005.
© Springer-Verlag Berlin Heidelberg 2005

simple forgery attack on Yi et al.'s scheme in Section 3. Our proposed improvement is shown in Section 4. Next, the security of our scheme is analyzed in Section 5. Finally, we make some conclusions in Section 6.

2 Related Works

In this section, we are going to introduce Yi et al.'s scheme. There are three participants, certificate authority (CA), base station (BS) and mobile user (MS), in their scheme. Their proposed scheme consists of two phases: the Registration Phase and the Service Phase. Before describing the details of these two phases, we first define the notations used in their scheme as follows [1, 8, 9, 10].

p, q: two large primes, and q | (p-1)
g: a prime element in GF(p)
x_c: the private key of CA
y_c: the public key of CA, $y_c = g^{x_c} \bmod p$
x_b: the private key of the base station
y_b: the public key of the base station, $y_b = g^{x_b} \bmod p$
x_m: the private key of the mobile user
y_m: the public key of the mobile user, $y_m = g^{x_m} \bmod p$
$h(\cdot)$: a public one-way hash function
C_i: a unique public information for the involved participant i

Registration Phase:

Step 1: Each participant computes the hash value of C_i and sends the computation result $h(C_i)$ to CA.

Step 2: CA generates a certificate for each involved participant i as follows,

$$s_i = g^{r_i} \bmod p,$$
$$t_i = -s_i - h(C_i)*r_i*x_c^{-1} \bmod q,$$

where r_i is a random number generated from GF(p).

Step 3: Then, other participants can verify the validity of the signature (s_i, t_i) by means of the following equation,

$$y_c^{s_i+t_i} \cdot s_i^{h(C_i)} = 1 \bmod p \quad (1)$$

If it holds, any two participants can carry on with mutual authentication and key distribution in the Service Phase.

Service Phase:

Step 1: While the mobile user MS wants to ask for communication services, he/she has to sends his/her certificate $Cert_M = (C_M, s_M, t_M)$ to the base station BS, where C_M is the public information of MS and (s_M, t_M) is the signature of C_M.

Step 2: After receiving the message sent from MS, BS checks the validity of $Cert_M$ by means of Equation 1. If it does not hold, the request is rejected; otherwise, BS randomly generates a session key k and computes as

$$y_m^{-x_b} \cdot k \bmod p .$$

BS then sends $Cert_B = (C_B, s_B, t_B)$ to MS as well as $y_m^{-x_b} \cdot k \bmod p$, where C_B is the public information of BS and (s_B, t_B) is the signature of C_B.

Step 3: Upon receiving the messages sent by BS, MS first checks the validity of $Cert_B$ by means of Equation 1. If it does not hold, the connection is terminated; otherwise, MS retrieves the session key

$$k = y_b^{x_m} \cdot (y_m^{-x_b} \cdot k) \bmod p .$$

Finally, both of them can use k to ensure the subsequent communications.

3 A Simple Attack on Yi et al.'s Scheme

In the following, we are going to present a simple attack on the Registration Phase of Yi et al.'s scheme, which an attacker Eve can easily make up a valid certificate of the legal participant to fool others. The details of our proposed attack are described as follows.

First, Eve generates a random number e from GF(q) and computes the certificate of the legal participant as follows,

$$s_i' = y_c^e \bmod p,$$
$$t_i' = -s_i' - h(C_i) * e \bmod q.$$

By Equation 1, we have

$$y_c^{s_i'+t_i'} \cdot (s_i')^{h(C_i)} = [y_c^{y_c^e + (-y_c^e - e \cdot h(C_i))}] \cdot (y_c^e)^{h(C_i)}$$
$$= (y_c^{-e \cdot h(C_i)}) \cdot (y_c^e)^{h(C_i)} = y_c^0 = 1 \bmod p.$$

Then, (C_i, s_i', t_i') can convince everyone of its validity; that is, Eve can use it to fool others successfully.

4 The Proposed Improvement

In this section, we propose an improvement on the Registration Phase of Yi et al.'s scheme to repair the security flaw shown in Section 3, based on DSA signature scheme. The improved Registration Phase is described as follows.

Step 1: Each participant computes the hash value of C_i and sends the computation result $h(C_i)$ to CA.

Step 2: CA generates a certificate for each involved participant as follows,

$$s_i = g^{r_i} \bmod p,$$
$$t_i = r_i^{-1}[x_c s_i + h(C_i)] \bmod q,$$

where r_i is a random number generated from GF(q).

Step 3: Then, other participant can verify the validity of the signature (s_i, t_i) by means of the following equation,

$$y_c^{t_i^{-1} \cdot s_i} \cdot g^{h(C_i) \cdot t_i^{-1}} = s_i \bmod p. \quad (2)$$

If it holds, any two participants can carry on achieving mutual authentication and key distribution in Service Phase.

5 Security Analysis of the Proposed Improvement

At first, we assume that DSA signature scheme is secure. That is, no one can successfully forge a certificate of the legal user to pass Equation 2. Hence, without the knowledge of the random number r_i and the private key x_c of CA, it is infeasible for the attacker Eve to make up a valid certificate of the legal user to fool others. As a result, our proposed improvement can certainly repair the security flaw of Yi et al.'s scheme such that it could be applied in the real world.

6 Conclusions

In this article, we have presented a simple attack on Yi et al.'s optimized authentication protocol for mobile networks. By our attack, an attacker Eve can easily forge a certificate of the legal participant and pretend to be the legal entity to fool others. What is more, we also propose an improvement on Registration Phase of Yi et al.'s scheme to repair the security flaws. The security of our proposed improvement is based on DSA signature scheme. Therefore, our improvement can withstand the proposed attack.

References

1. X. Yi, E. Okamoto and K.Y. Lam: An optimized protocol for mobile network authentication and security. ACM Mobile Computing and Communications Review, vol. 2, no. 3 (1998) 37-39.
2. K. Martin and C. L. Mitchell: Comments on an optimized protocol for mobile authentication and security. ACM Mobile Computing and Communications Review, vol. 3, no. 2. (1999) 37-
3. X. Yi: Author's reply to Comments on an optimized protocol for mobile authentication and security. ACM Mobile Computing and Communications Review, vol. 3, no. 2 (1999) 38-
4. C. Boyd and A. Mathuria: Key establishment protocols for secure mobile communications: A critical survey. Computer Communications, vol. 23 (2000) 575-587
5. D. S. Wang: On the design and analysis of authenticated key exchange schemes for low power wireless computing platforms. Ph.D. Thesis, July (2002).
6. D. S. Wang: An optimized authentication protocol for mobile network reconsidered. ACM Mobile Computing and Communications Review, vol. 6, no. 4 (2002) 74-76.
7. C.S. Laih and S.Y. Chiou: Cryptanalysis of an optimized protocol for mobile network authentication and security. Information Processing Letters, vol. 85 (2003) 339-341.
8. T. ElGamal: A public-key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, vol. 33, no. 2 (1985) 469-472.
9. W. Diffie and M. E. Hellman: New directions in cryptography. IEEE Transactions on Information Theory, vol. 22 (1976) 644-654.
10. National Institute of Standards and Technology, Digital Signature Standard (DSS): Federal Information Processing Standards Publication. FIPS PUB 186-2, Reaffirmed, January (2000).

Neural Network Based Flow Forecast and Diagnosis

Qianmu Li[1,2], Manwu Xu[1], Hong Zhang[2], and Fengyu Liu[2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
{liqianmu, xu.manwu}@126.com
[2] Computer Science Department, Nanjing University of Science and Technology, Nanjing 210094, China
{liqianmu, njust, liu.fengyu}@126.com

Abstract. Much of the earlier work presented in the area of on-line flow diagnosis focuses on knowledge based and qualitatively reasoning principles and attempts to present possible root causes and consequences in terms of various measured data. However, forecasting flow is an un-measurable operating variable in diagnosis processes that define the state of the network. Forecasting flow essentially characterize the efficiency and really need to be known in order to diagnose possible malfunction and provide a basis for deciding on appropriate action to be taken by network manager. This paper proposes a novel flow-predictable system (FPS) based on fuzzy neural network. The features of dynamic trends of the network flow are extracted using a decision matrix transform and a qualitative interpretation, and then are used as inputs in the neural network, so that it can be used to fit the smooth curves perfectly. It is adopts to deal with the mapping relation and categorizing the network faults. The experiment system implemented by this method shows the proposed system is an open and efficient flow-fault forecast engine.

1 Introduction

Traditional Flow Forecast/ Diagnosis methods need to establish mathematics model, but the complexity and accuracy of the model are hard to satisfy the real-time requirement of high-speed network. On the other hand, the mathematics model, which is simplified, couldn't bring on effective actual result. Moreover, traditional methods demand the understanding on recondite traffic representations (e.g. bit ratio of peak value, average bit ratio, abrupt length...etc.), however these representations couldn't describe the relativity of those paroxysmal flows. At the same time, it's difficult to implement effective control, and so that it causes the serious descent of the network performance. Leading the fuzzy neural network into network flow flow-fault forecast, this paper proposed a novel method: FPS. FPS is good at solving indetermination problem and self-study ability.

2 The Parameter of Network Flow

Network flow state can be classified: adding attribution, multiply attribution and min-max attribution. For some section of network $P(E_1, E_2, \cdots, E_{s-1})$, in which E be the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 542 – 547, 2005.
© Springer-Verlag Berlin Heidelberg 2005

states of network entities, if $e_{i,i+1} \in E_i$ (i=1,2,… S-1), define the j property of $e_{i,i+1}$ as $f_{i,i+1}^{j}$, and define the j property of the whole net-segment P as f_p^j, we get hold of the definitions as follows:

（1） adding attribution: If $f_p^j = \sum_{i=1}^{s-1} f_{i,i+1}^{j}$, name the j property of P a adding attribution, such as receive-datagram number, send-datagram number, jump number, delay time, delay tingle, cost, etc.

（2） multiply attribution. If $f_p^j = \prod_{i=1}^{s-1} f_{i,i+1}^{j}$, name the j property of P a multiply attribution, such as error rate, losing rate, node utilization rate, etc.

（3） min-max attribution. If $f_p^j = \min_{i=1,2,...,s-1} \{f_{i,i+1}^{j}\}$, name the j property of P a min attribution, such as fluff rate, fee, etc. If $f_p^j = \max_{i=1,2,...,s-1} \{f_{i,i+1}^{j}\}$, we name it a max attribution, such as port utilization rate, flux, bandwidth, etc.

The essence of network flow flow-fault forecast is to capture a set of situation information on network entities, and deduce the fault reason or fault equipments. So that flow-fault forecast is a kind of mapping from the situation information to the type of fault.

3 Neural Network Structure

Lots of literature have proceeded with the research of fuzzy logical neuron model [7]-[9], especially in 2002, Gavalas put forward the concept of weak T/ S norm, which looses the constraint of associative law in T norm in order to make some controllable operation combination with simple form as and/or operation. We improve his model, and realize not only weak T/ S norm cluster, but also both equivalence operation and sequential average operation, which meet the requirement of flow fault prediction.

Let $A_1' \otimes A_2' \otimes \cdots \otimes A_n'(u_1, u_2, u_n) = \overset{n}{\underset{i=1}{\wedge}} g_i(A_i'(u_i))$, in which g_i content $g_i(1) = 1$ in [0, 1]. The elasticity of every rule: $g_1, g_2, \cdots, g_n$ is set on through its importance. Let $(u_{1k}, u_{2k}, \cdots, u_{nk}) \in U_1 \times U_2 \times \cdots \times U_n$, where $u_{ik} \in U_i$, $k = 1,2,\cdots,L$, $|U_1 \times U_2 \times \cdots \times U_n|$ =L. We design the neural network shown as figure 1. Suppose $R_{(k)(j)} = R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v_j))$, we get: $v \in V$,

$$B'(v) = \bigvee_{(u_{1k},\cdots,u_{nk})} \{p_1(v)[\overset{n}{\underset{i=1}{\wedge}} g_i(A_i'(u_{ik})) \wedge R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]$$

$$+ p_2(v)[\overset{n}{\underset{i=1}{\wedge}} g_i(A_i'(u_{ik})) \vee R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]\}.$$

Let $p_1(v)=1 \quad p_2(v)=0 \quad g_i(A_i(u_{ik}))=A_i(u_{ik})$, then

$$B'(v)=\bigvee_{(u_{1k},\cdots,u_{nk})}[\overset{n}{\underset{i=1}{\wedge}}A_i'(u_{ik})\wedge R(\overset{n}{\underset{i=1}{\wedge}}A_i(u_{ik}),B(v))]\cdot$$

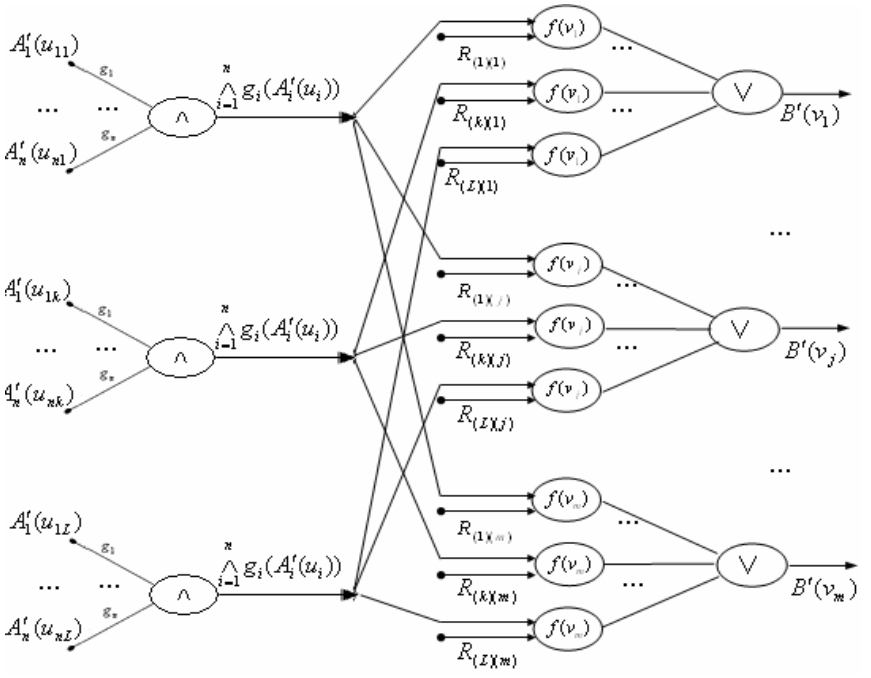


Fig. 1. Neural network structure for network flow fault prediction

Let $A_1 \otimes A_2 \otimes \cdots \otimes A_n \rightarrow B$ be a known rule, $R(A,B)$ be discretional implication relation. When training the neural network (fig.1), always be $p_1, p_2 \in F(V)$, if $A_i' = A_i$, then $\forall v \in V$, $B' = B$ iff

$$B(v) \le \bigvee_{(u_{1k},\cdots,u_{nk})}[\overset{n}{\underset{i=1}{\wedge}} g_i(\ (A_i(\ u_{ik})) + R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})),B(v))].$$

Proof. *If* $A_i' = A_i$, $\exists\ p_1, p_2 \in F(V)$, *which makes* $B' = B$. *That is,* $B(v) =$

$$\overset{n}{\underset{i=1}{\vee}}\{p_1(v)[A(u_i)\wedge R(A(u_i),B(v))]. + p_2(v)(A(u_i)\vee R(A(u_i),B(v))]\}$$

Because $p_1(v), p_2(v) \in [0,1]$, so $B(v) \le \overset{n}{\underset{i=1}{\vee}}[A(u_i) + R(A(u_i),B(v))]$.

Conversely: $f(x,y) = \bigvee_{(u_{1k},\cdots,u_{nk})} \{x[\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})) \wedge R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]$

$+ y[\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})) \vee R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]\}$.

Through mathematics analysis, the domain of $f(x,y)$ be $[f(0,0), f(1,1)]$. $f(0,0) = 0$, and $f(1,1) = \bigvee_{(u_{1k},\cdots,u_{nk})} [\overset{n}{\underset{i=1}{\wedge}} g_i((A_i(u_{ik})) + R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]$.

If $0 \le B(v) \le \bigvee_{(u_{1k},\cdots,u_{nk})} [\overset{n}{\underset{i=1}{\wedge}} g_i((A_i(u_{ik})) + R(\overset{n}{\underset{i=1}{\wedge}} g_i(A_i(u_{ik})), B(v))]$, obeying continuous function interpose-value theorem $\exists \ (p_1', p_2')$ leads to $f(p_1', p_2') = B(v)$, then let $p_1(v) = p_1'$, $p_2(v) = p_2'$. For $\forall \ v$, it always $\exists$ $p_1(v)$ and $p_2(v)$, which hold reasoning results $B' = B$, when $A_i' = A_i$.

4 Network Flow Fault Prediction

Using FPS to single-step forecast, we choose the structure with two inputs and one output. Carve inputs variable into seven fuzzy subsets: {NL, NM, NS, ZO, PS, PM, PL}, which express negative large, negative middle, negative small, zero, positive small, positive middle and positive large respectively. The value of x is disposed through normalization. So we get hold of 49 fuzzy rules, marking R:

R1: if x1 is NL, x2 is NL, then y1= $C_0^1 + C_1^1 x_1 + C_2^1 x_2$

…

R49: if x1 is PL, x2 is PL, then y49= $C_0^{49} + C_1^{49} x_1 + C_2^{49} x_2$

x1 and x2 mean the node loading of the former two time segments (Ti-2, Ti-1) (Ti-1, Ti) at time of Ti. Let the network desired output be B^d, then output error be $J = \frac{1}{2}(B^d - B)^2$. According to Gabarit approximation [6], data flow could be represented by continuum condition discrete time AR Markov model. If we let $\lambda(n)$ express the bit rate of No.n packets, then one rank AR Markov equation is shown as follows by the using of recursion relation: $\lambda(n-1) = l \cdot \lambda(n) + m \cdot \varpi(n)$, l and m are influence genes. Following experience, we can evaluate $l = 0.8781$. $m = 0.1108$. $\varpi(n)$ is the independence gauss white noise sequence, and its mean is 0.572, its variance is 1. Every node loading could be expressed by the formula: $L = \sqrt{\sum_{i=1}^{n} (k_i a_i^2)}$, L represents the loading value of local node; $a_1, ..., a_n$, are loading targets; $k_1, k_2, ..., k_n$, are weights. The simulation environment in this paper is NS2.

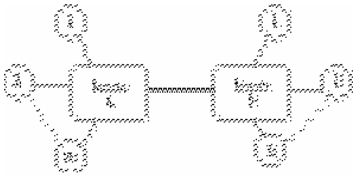


Fig. 2. Simulation network topology: Design a share bottleneck connection A and B in the topology, and the bandwidth of link is 100Mbit/s. Other link bandwidths are all 5Mbit/s. The link delay between node A and B is a variable, whose value is between 10ms and 300ms. Let packets length be 1024bytes. The buffer length in router A and B are all 40M. m=n=40. The send-velocity minimum of every node is 300kbits/s, and the maximum is 1500kbits/s.

We choose flow traffic prediction as an example. Set sampling cycle be 40ms and get 400s sampled data to train. Get eight sets at random once more. In each sets, 320s data is to test. Based upon these hypothesis and parameter, the result of training is shown as figure 3.

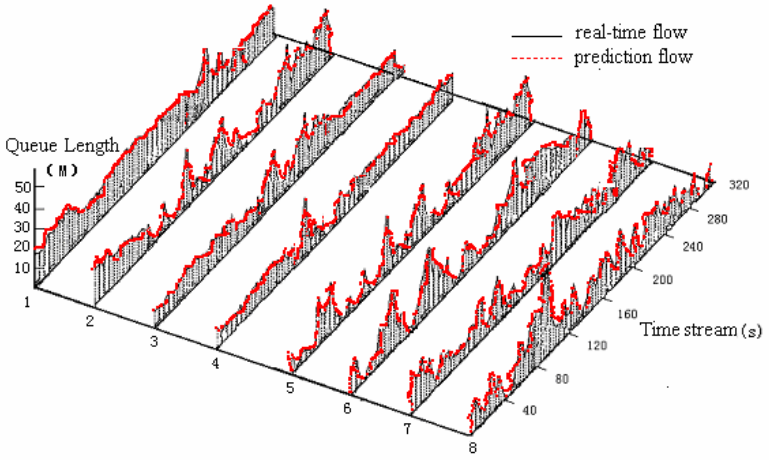


Fig. 3. Prediction result of node data flow

The results of comparing BPN with the neural network of FPS are shown in table 1. The first row represents TSE (total squared error), and the second and the third row represent iteration degree, when BPN or the neural network of FPS reach corresponding TSE. We also can hold MSE (mean square error, MSE＝TSE／400).

Table 1. Iteration degrees of BPN and the neural network of FPS

TSE	BPN	the neural network of FPS
2.43	238	61
2.23	633	78
2.03	13232	90
1.83		115
1.63		142

From table 1, we can find that convergence rate of the neural network of FPS is more quickly than BPN in evidence, and the neural network of FPS could gain wee TSE and MSE in tolerable iteration degree.

5 Conclusion

FPS is combined by neural network and fuzzy logic. This strategy is propitious to catch the flow features with mutative time exactly. Because the weight of network is adjustable, this strategy can satisfy the consistency requirement of fuzzy consequence for any implication relation $R(A,B)$. The experiment shows that it is effective to the adjustment of network rate and the of loss rate reduction. The strategy could show extensive adaptability with the adjustment of weight.

References

1. Jacobson V.: Flow Fault Avoidance and Control. IEEE/ACM Transaction Networking, 3 (1998) 314–329
2. Caserri C, Meo M.: A New Approach to Model the Stationary Behavior of TCP Connections. In: Tel Aviv. (eds.): IEEE Computer Society, Proc IEEE INFOCOM2000, Israel, CA(2000) 245–251
3. Floyd S, Fall K.: Promoting the Use of End-to-End Network Flow-fault forecast in the Internet. IEEE/ACM Transaction Networking, 4(2002) 458–472
4. Harris B, Hunt R.: TCP/IP Security Threats and Attack Methods. Computer Communications, 10(2002) 885–897
5. Skoundrianos E N, Tzafestas S G.: Flow-fault forecast via Local Neural Networks. Mathematics and Computers in Simulation , 60 (2002) 169–180
6. Bullell P, Inman D.: An Expert System for the Analysis of Faults in an Electricity Supply Network: Problems and Achievements. Computer in Industry, 37 (1998) 113–123
7. Tagliaferri R, Eleuteri A, Meneganti M, Barone F.: Fuzzy Min-Max Neural Network: from Classification to Regression. Soft Computing, 5 (2001) 69–76
8. Gavalas D, Greenwood D, Ghanbari M.: Advanced Network Monitoring Applications Based on Mobile/Intelligent Agent Technology. Computer Communications, 23 (2002) 720–730
9. Li Q M, Qi Y, Zhang H, Liu F Y.: New Network Flow-fault forecast Method Based on RS-Neural Network. Computer Research and Development, 10(2004) 1696–1702

Protecting Personal Data with Various Granularities: A Logic-Based Access Control Approach

Bat-Odon Purevjii[1], Masayoshi Aritsugi[1], Sayaka Imai[1], Yoshinari Kanamori[1], and Cherri M. Pancake[2]

[1] Department of Computer Science, Faculty of Engineering, Gunma University, 1-5-1 Tenjin-cho, Kiryu, Gunma 376-8515, Japan
{batodon, aritsugi, sayaka, kanamori}@dbms.cs.gunma-u.ac.jp
[2] School of Electrical Engineering & Computer Science, Oregon State University, Corvallis, OR 97331, USA
pancake@eecs.oregonstate.edu

Abstract. In this paper, we present a rule-based approach to fine-grained data-dependent access control for database systems. Authorization rules in this framework are described in a logical language that allows us to specify policies systematically and easily. The language expresses authorization rules based on the values, types, and semantics of data elements common to the relational data model. We demonstrate the applicability of our approach by describing several data-dependent policies using an example drawn from a medical information system.

1 Introduction

Some fine-grained access control mechanisms have been proposed for databases [3, 5, 10]. The database or security administrators are typically forced to manipulate fine-grained authorization objects separately from authorization rules defined on them. Thus, administering cell- or column-level protection on a database table is complicated in those systems. Because of this complexity, database security administration primarily occurs at the level of entire tables, rather than columns or individual data elements [6]. There is an inherent demand for tools that can reduce the complexity of access control administration so that it can be feasible at the level of individual columns, rows, and cells within database tables. We believe logical, yet simple, authorization description languages can provide a practical foundation for such tools.

Many high-level logical access control models and languages have been proposed so far. Their advantages include clean foundations, flexibility, expressiveness, and declarative nature. However, most of them ignore the more practical issues, which typically must be specific to the underlying data model available in particular systems [6]. Thus, there is a significant gap between the database enforcement mechanisms and the high-level policy specifications [8]. To be practical, the policy specification must be brought conceptually "nearer" the resources it is designed to control [6].

In this paper, we present a rule-based fine-grained data-dependent access control approach that enables to administer policies in a systematic and easy way. The authorization rules are described in a logical language based on Datalog [7]. The lan-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 548 – 553, 2005.
© Springer-Verlag Berlin Heidelberg 2005

guage expresses policies that vary according to the values, types and semantics of the data elements common to the relational data model. Because we define policies directly in terms of the underlying data elements, our language addresses a key concern of database implementers. We show the applicability of our approach by describing several security policies using an example drawn from a medical information system.

The remainder of the paper is organized as follows. Section 2 describes related work. We introduce our access control framework with its authorization description language in Section 3. In Section 4, we describe the corresponding policy definitions, using as an example the security requirements typical of medical information systems. Section 5 presents a brief conclusion and consideration of future work.

2 Related Work

With the mechanism in [3], cell- or row-level protection is enforced through a virtual view superimposed on some base tables. Because of its maintenance overhead, however, fine-grained access control using this view-based mechanisms has been avoided in practice. The Oracle Virtual Private Database (VPD) [10] enforces row-level protection through program codes written in the proprietary procedural language PL/SQL. The mechanism in [5] enforces row-level protection by employing their inference rules. Both of these two mechanisms are required to employ views to enforce cell-level protection. The concept "parameterized-view" introduced in [5] naturally simplifies the view maintenance process. However, fine-grained authorization objects and rules on them are manipulated separately in all these mechanisms. Instead, our framework provides a simple language that enables us to specify and manipulate authorization objects and rules in a unified way.

Significant work has also been done in security specification languages. Their semantics is complicated and is difficult to be implemented in real systems. Some candidate models [4, 2] provide richer high-level security policies; however, authorization object granularity in [2] is defined at the table level only. We could not specify fine-grained authorization objects by employing the language in [4] that is based on function-free Datalog. Instead, we employ Datalog with function symbols to express fine-grained authorization objects within authorization rules.

The reader is referred to [9, 1] as examples which provide thorough descriptions of the types of security requirements found in Medical Information Systems (MIS).

3 Access Control Framework

3.1 Preliminaries

Datalog [7] is a version of Prolog and is employed databases as a declarative language. Datalog programs are built from *atomic formulas* (*atoms*), which are predicate symbols with a list of arguments, e.g., $p(A_1,...,A_n)$, where p is the predicate symbol and $A_1,...,A_n$ are arguments. A predicate t corresponds to a database table T and the arguments correspond to column names of the table. The predicate t is true for its arguments if those arguments form a row/tuple of that corresponding table. That means a row in a table corresponds to a *fact* in a predicate. An argument in an extended form

of Datalog can be a variable, constant or a function. Like Prolog predicates, functions and constants begin with lowercase letters and variables begin with uppercase letters, with numbers used as constants. Logical statements, called *rules*, are written in the form $q \leftarrow p_1 \& ... \& p_n$, where q is the *head* and $p_1 \& ... \& p_n$ is the *body* of the rule. Thus, a body consists of one or more atoms, called *subgoals*, connected by (logical) AND. We translate the head q as true only if all subgoals $p_1 \& ... \& p_n$ in the body are true. Atomic formulas can also be constructed with the arithmetic comparison predicates (=, ≤, and so on); these are called *built-in* predicates. Atomic formulas with built-in predicates are written in the usual infix notation, e.g., $X<Y$ instead of $<(X, Y)$.

3.2 Authorization Description Language

As building blocks of our access control framework we assume a set of database users, a set of fine-grained objects derived from database tables, and a set of SQL privilege modes, or just privileges. In this framework *authorization rules* describe who can execute which privilege on what objects.

We define our authorization description language (ADL) as follows. Because of the page limitation, we provide a brief introduction of the language. The language relies on a small set of constructs. Based on the constructs, we give the definition of authorization rule in this subsection as well. ADL consists of the following constructs:

A set of constant symbols $a, b, c, d, e, \ldots,$.
A set of variable symbols $X, Y, Z, V, U, \ldots,$.
A set of function symbols $f, g, \ldots,$. The function symbols help us to describe fine-grained authorization objects within the *permitted* predicate.

A set of predicate symbols. The symbols correspond to tables of a database. The arity of a predicate must be fixed and is equal to the number of columns of the corresponding table. These predicate symbols appear only in the body of logical rules. In addition, the following three predicates and built-in predicates are employed in our framework. We give the definitions of these predicates separately as follows.

A predicate *permitted*. It is a 3-ary predicate that takes the form *permitted*(s, o, m) where s and m are expressions, that can be a constant or variable symbol, and o is a function symbol. The domain, or type, of s is a set of user IDs and the domain of m is privileges of SQL. The latter domain has fixed number of members, namely, *select*, *insert*, *delete*, and *update*. Let us assume a fixed database schema with k number of tables and h to be the total number of columns of the all tables. Then o is formed as $o(g_1(X_1,X_2,\ldots,X_j),\ldots,g_k(Y_1,Y_2,\ldots,Y_l))$, where the function symbols $g_1,\ldots,g_k$ correspond to the tables and the variable symbols $X_1,X_2,\ldots,X_j$ and $Y_1,Y_2,\ldots,Y_l$ correspond columns of the tables. Then we may assume that function o takes h domain elements and returns an "object" of type "record of h elements". When specifying authorization rules we omit function symbols $g_1,\ldots,g_k$ and unnecessary components from them.

A predicate *oType*. It is formed as *oType*(A_r, d) where A_r denotes an attribute of a table and d denotes a type. The predicate returns true if the type of A_r is d. Note that we will use the terms "column" and "attribute" interchangeably throughout the paper.

A predicate *oSemanticExclude*. It is formed *oSemanticExclude*(A_r, *txt*) where A_r denotes an attribute of a table and *txt* denotes an arbitrary text. The predicate returns true if A_r does not contain the given text.

A set of built-in predicates. =, ≠, <, >, ≤, and ≥.

Now we define our authorization rule based on the language constructs. An authorization rule takes the form:

$$permitted(s, o(g_1(X_1,X_2,\ldots,X_j),\ldots,g_k(Y_1,Y_2,\ldots,Y_l)), m) \leftarrow N_1,\ldots, N_k.$$

where s denotes user IDs, m denotes privileges, o is a function symbol that returns authorization objects, and $N_1,\ldots, N_k$ are subgoals. A subgoal N_i can be a predicate denoting a table, an *oType* predicate, an *oSemanticExclude* predicate or a *built-in* predicate. The authorization rule enforces constraints on type, semantic, value of attributes of tables. The body can be empty when we assign privileges without any constraints. Examples of authorization rules are provided in the next section.

4 Application to Medical Information Systems

In this section, we provide several examples of policies in a hospital database and the corresponding authorization rule specifications in ADL. We assume the schema of database HOSPITAL that is represented by the ER diagram as shown in Figure 1. The diagram consists of five entities and seven relationships. Note that when a relationship is converted to a table it takes all primary key attributes of the corresponding entities as its attributes. The following authorization rules are arranged from coarse-grained to fine-grained protection levels. The meaning of an authorization rule is given in an example policy. The readers may understand the meanings of other rules similarly.

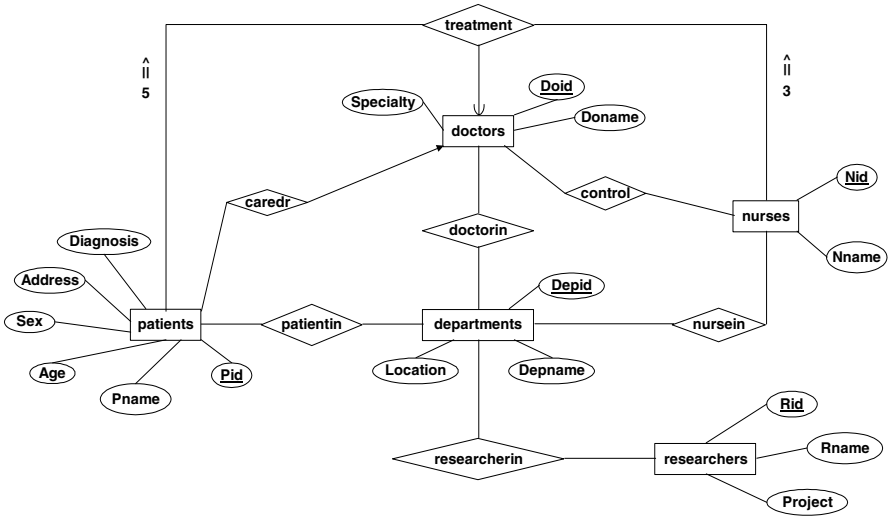


Fig. 1. A hospital database schema described in an ER diagram

Policy 1. *Doctors* who belong to the same department as a patient should be permitted to execute *select* privileges on the patient's record iff the *diagnosis* attribute does not contain the text "cancer":

permitted(Doid, patients(Pid, Pname, Age, Sex, Address, Diagnosis), select) ←
patients(Pid, Pname,...,Diagnosis) &
doctors(Doid, Doname, Specialty) &
patientin(Pid, Depid) &
doctorin(Doid, Depid) &
oSemanticExclude(Diagnosis, cancer).

Under the rule doctors are able to select only some rows from Patients table. We call this kind of rule *row-level protection*.

Policy 2a. A patient's record should partially be *selected* by *nurses*:

permitted(Nid, patients(Pname, Age, Address), select) ←.

Policy 2b. *Researchers* are permitted to *select* attributes of *patients* table that have integer type:

permitted(Rid, patients(A_i), select) ← *oType(A_i, integer).*

Under these two rules, 2a and 2b *nurses/researchers* are able to select only some columns from the table. We call this kind of rule *column-level protection*.

Policy 3a. *Researchers* who belong to the same department with some patients should be permitted to make research on data of patients who are not younger than 15:

permitted(Rid, patients(Age, Diagnosis), select) ←
patients(Pid, Pname,..., Diagnosis) &
researchers(Rid, Rname, Project) &
patientin(Pid, Depid) &
researcherin(Rid, Depid) &
Age ≥ *15.*

Meaning of the rule: For simplicity, we briefly define the meaning in natural language. Let us assume the particular rows (a_1,a_2,a_3,a_4,a_5,a_6), (b_1,b_2,b_3), (c_1,c_2), and (d_1,d_2) are stored in tables *patients*, *researchers*, *patieintin*, and *researcherin*, respectively. Then the rule says "If $a_1 = c_1$, $b_1 = d_1$ and $a_3 \geq 15$ are all true then the statement 'researcher b_1 is permitted to select a_3 and a_6 of the row (a_1,a_2,a_3,a_4,a_5,a_6)' is true".

Policy 3b. A *Doctor* who is taking care of a patient should be permitted to *update* the *Diagnosis* attribute of the patient's record:

permitted(Doid, patients(Diagnosis), update) ←
caredr(Did, Pid) &
doctors(Doid, Doname, Specialty) &
patients(Pid, Pname,…, Diagnosis).

Under the rules 3a and 3b, *researchers/doctors* are able to *select/update* only two columns/a column of some rows from the table. We call this type of rule *row-column-level protection*. When a doctor is taking care of only one patient at a time, *cell-level*

protection is described through the rule. At all the protection levels, the security administrator is able to describe and manipulate various kinds of authorization rules easily as shown.

5 Conclusion and Future Work

In this paper, we proposed a rule-based access control approach that enables us to manipulate fine-grained authorization objects and rules in a unified and easy way. The authorization rules are described in a simple language that we proposed. Since the language expresses policies in terms of the underlying data elements common to relational databases, it has no high-level semantics that is hard to implement in real systems. It was easy to manage authorization rules at all protection levels: from coarse-grained to fine-grained. This was shown by describing several data-dependent security policies needed in a medical information system.

In order to verify the complete correctness of the proposed language, we still need to employ higher-order logic to fully specify its constructs. Although it would be hard to implement such a specification in a real system, we believe that all necessary functionality could be implemented if we impose a few simple restrictions. Thus, the next phase in our continuing project is to implement the access control framework.

Acknowledgement

This research was partially supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B) 15300029, and Special Project for Earthquake Disaster Mitigation in Urban Areas (MEXT).

References

1. R.J. Anderson. A Security Policy Model for Clinical Information Systems. In *Proc. IEEE Symp. on Security and Privacy* (1996) 30-45
2. E. Bertino, S. Jajodia, and P. Samarati. Supporting Multiple Access Control Policies in Data-base Systems. In *Proc. IEEE Symp. on Security and Privacy* (1996) 94-108
3. P.P. Griffiths and B.W. Wade. An Authorization Mechanism for a Relational Database System. *ACM Transactions on Database Systems*, 1(3): 242-255, September 1976
4. S. Jajodia, P. Samarati, V. S. Subrahmanian, and E. Bertino. A Unified Framework for Enforcing Multiple Access Control Policies. In *Proc. ACM SIGMOD* (1997) 474-485
5. S. Rizvi, A.O. Mendelzon, S. Sudarshan, and P. Roy. Extending Query Rewriting Techniques for Fine-Grained Access Control. In *Proc. ACM SIGMOD*, Paris, France (2004) 551-562
6. A. Rosenthal and M. Winslett. Security of Shared Data in Large Systems: State of the Art and Research Directions. In *Proc. of ACM SIGMOD* (2004) 962 – 964
7. J. Ullman. Principles of Database and Knowledge-Base Systems. Vol 1&2. Computer Science Press (1988)
8. T. Yu. Fine-grained Access Control. *Fall Privacy Place Workshop Proceedings*, http://theprivacyplace.org/Workshops/fall_2004.html
9. L. Zhang, G. Ahn, and B. Chu. A Role-Based Delegation Framework for Healthcare Information Systems. In *Proc. ACM SACMAT* (2002) 125-134
10. The Virtual Private Database in Oracle9ir2: An Oracle Technical White Paper (2002). http://www.oracle.com/technology/deploy/security/oracle9ir2/pdf/ VPD9ir2twp.pdf

Enhancement of an Authenticated Multiple-Key Agreement Protocol Without Using Conventional One-Way Function

Huifeng Huang and Chinchen Chang

Department of Information Management,
National Taichung Institute of Technology, Taichung 404, Taiwan, China
phoenix@ntit.edu.tw
Department of Information Engineering and Computer Science,
Feng Chia University, Taichung 40724, Taiwan, China
ccc@cs.ccu.edu.tw

Abstract. The authenticated multiple-key agreement protocol provides two entities to authenticate each other and establish multiple common keys in a two-pass interaction, a protocol without using a conventional hash function simplifies its security assumption on only public hard problem. In 2004, Chien and Jan proposed an authenticated multiple-key agreement protocol to overcome the shortcomings that break the previous variants. This paper shows that Chien and Jan's scheme has a weakness that is vulnerable to forgery. To remedy this weakness, we improve Chien and Jan's scheme such that the newly improved scheme has authenticated property and does not significantly affect the efficiency of the original scheme. Compared to the previous schemes, our improved method also achieves better key utilization.

1 Introduction

Harn and Lin [6] noticed that the conventional one-way function is widely employed in many digital signature schemes [2][3][4]. These schemes become insecure if the conventional one-way function is not used [5][6][7][10]. They also observed that the security of these one-way functions, like MD5 [7], is based on the complexity of analysis of iterated functions but is not on a public hard problem [2][3][10] (the discrete logarithm problem is a public hard problem, and can be seen as a one-way function). So, it may become insecure to some special attacks later [7]. Therefore, Harn and Lin [6] first proposed the authenticated key agreement protocol without using the conventional one-way functions. In addition, their scheme greatly enhances the efficiency of key agreement protocol by allowing two entities to establish multiple keys instead of one common key in a two-pass interaction [6]. In their first pass, each entity generates and exchanges n public values in an authenticated manner. After exchanging the authenticated messages, two entities verify the received messages and then generate (n^2-1) common keys during the second pass. However, Yen and Joye [8] showed that an attacker could easily forge the signature of the exchanged public keys in Harn and Lin's scheme. Yen and Joye also proposed their modified version.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 554 – 559, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Interestingly, Wu et al. [9] found the same weakness in Yen and Joye's modified scheme. Later, to improve Harn and Lin's scheme, some authenticated multiple-key agreement protocols without using one-way function were presented [8][9][11]. Unfortunately, these improved schemes are also insecure [12]. In 2004, Chien and Jan proposed an improved scheme which can achieve better key utilization, compared to the previous schemes [12]. However, there's no real authenticated property in Chien and Jan's scheme. In their scheme, they use two entities A and B, yet entity B could easily pretend to be A and deliver forged information to B and would never be denied by A. Similarly, entity A could also easily pretend to be B to send other forgery to A and wouldn't be denied by B. That is, entity A/B can easily forge the other entity's message without being detected. Therefore, their scheme is unreliable and does not achieve authenticated property.

In this paper, we show that Chien and Jan's method is not a real authenticated multiple keys agreement protocol. We will then present a simple way to improve their scheme so as to verify the identity of the user and does not significantly affect the efficiency of the original scheme. This paper is organized as follows. Section 2 will briefly review Chien and Jan's scheme. In Section 3, we will present an approach to forge an authentic message without knowing the entity's secret key. In Section 4, we improve Chien and Jan's scheme to achieve the authenticated property. And some conclusions will be made in the last section.

2 Review of Chien and Jan's Scheme

Using the concept of Harn-Lin's method [6], Chien and Jan [12] proposed multiple-key agreement protocol which provided two entities to authenticate each other and create multiple common keys in two-pass interaction. There are three phases in their scheme: system initialization phase, key and signature generation phase, and verification phase. In the key and signature generation phase, each entity generates and exchanges n public values in an authenticated manner. After exchanging the authenticated messages, two entities verify the received messages and then generate n^2 keys in the verification phase, like the Diffie-Hellman approach [1]. By taking a simple example of n=2, we describe the idea of Chien-Jan's scheme as follows [12]:

System Initialization Phase

The system authority (SA) chooses a large prime number p and a primitive element g over $GF(P)$. Then, SA publishes p and g. Suppose that A and B are the two entities to authenticate each other and share multiple keys. The long-term secret key for A is x_a, and $cert(y_a)$ is the certificate of A's long-term public key $y_a = g^{x_a} \bmod p$. The long-term secret key, long-term public key, and the certificate of the public key for B are x_b, y_b, and $cert(y_b)$, where $y_b = g^{x_b} \bmod p$.

Key and Signature Generation Phase

Without loss of generality, entity A randomly selects two secret numbers k_{a1} and k_{a2}, then computes their corresponding public values $r_{a1} = g^{k_{a1}}$ and $r_{a2} = g^{k_{a2}} \bmod p$. Next, A signs his signature on these two values by generating the equation as

$$s_a \oplus K_{AB} = Time_a x_a - (r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1, \quad (1)$$

where $K_{AB} = g^{x_a x_b} = (y_a)^{x_b} = (y_b)^{x_a} \mod p$ is the long-term secret key between A and B. $Time_a$ is A's current timestamp. Then, A sends $(r_{a1}, r_{a2}, s_a, Time_a, cert(y_a))$ to B. Proceeding in a similar approach, B computes and sends $(r_{b1}, r_{b2}, s_b, Time_b, cert(y_b))$ to A.

Verification phase

After receiving information $(r_{a1}, r_{a2}, s_a, Time_a, cert(y_a))$ from entity A, entity B verifies if the following equation holds:

$$y_a^{Time_a} = (r_{a1} r_{a2})^{r_{a1} \oplus r_{a2}} g^{s_a \oplus K_{AB}} \mod p. \quad (2)$$

Entity A also verifies the information $(r_{b1}, r_{b2}, s_b, Time_b, cert(y_b))$ received from B. If both A and B succeed in their verifications, then they can derive four common keys $K_1 = (r_{a1})^{k_{b1}} = (r_{b1})^{k_{a1}} = g^{k_{a1}k_{b1}}$, $K_2 = (r_{a1})^{k_{b2}} = (r_{b2})^{k_{a1}} = g^{k_{a1}k_{b2}}$, $K_3 = (r_{a2})^{k_{b1}} = (r_{b1})^{k_{a2}} = g^{k_{a2}k_{b1}}$, and $K_4 = (r_{a2})^{k_{b2}} = (r_{b2})^{k_{a2}} = g^{k_{a2}k_{b2}} \mod p$.

3 The Weakness of Chien and Jan's Scheme

In this section, we present a strategy to show that Chien and Jan's scheme is not a real authenticated multiple-key agreement protocol. In Section 4, we will suggest a way to remedy the weakness of Chien and Jan's scheme. First, we describe the way to generate valid information by a masquerader as follows:

Step 1. After receiving the information $(r_{a1}, r_{a2}, s_a, Time_a, cert(y_a))$ from entity A, it has $s_a \oplus K_{AB} = Time_a x_a - (r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1$. Entity B can forge A's message by finding r'_{a1} and r'_{a2} such that $r'_{a1} r'_{a2} = r_{a1} r_{a2} \mod p$ and $\gcd((r'_{a1} \oplus r'_{a2}), p-1) = 1$.

Step 2. B finds an integer $w \in Z_{p-1}$ such that $(r'_{a1} \oplus r'_{a2}) = w(r_{a1} \oplus r_{a2}) \mod p\text{-}1$. Then, let $Time'_a = wTime_a \mod p\text{-}1$.

Step 3. B derives s'_a by generating the equation

$$s'_a \oplus K_{AB} = Time'_a x_a - (r'_{a1} \oplus r'_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1.$$

Finally, B can forge valid information $(r'_{a1}, r'_{a2}, s'_a, Time'_a, cert(y_a))$ of A.

In Step 1, we know that r_{a1}, r_{a2}, s_a, and $Time_a$ are known information. Applying the extended Euclidean Algorithm, for two values $(r'_{a1} \oplus r'_{a2})$ and $(r_{a1} \oplus r_{a2})$, B can easily find a unique integer $w \in Z_{p-1}$ such that $(r'_{a1} \oplus r'_{a2}) = w(r_{a1} \oplus r_{a2}) \mod p - 1$ in Step 2. From Equation (1), it provides

$$w(s_a \oplus K_{AB}) = wTime_a x_a - w(r_{a1} \oplus r_{a2})(k_{a1} + k_{a2})$$
$$= Time'_a x_a - (r'_{a1} \oplus r'_{a2})(k_{a1} + k_{a2}) \bmod p\text{-}1, \text{ where } Time'_a = wTime_a.$$

Because B knows $K_{AB} = g^{x_a x_b} = (y_a)^{x_b} = (y_b)^{x_a} \bmod p$, he can compute the value $w(s_a \oplus K_{AB}) \bmod p\text{-}1$. Hence, in Step 3, entity B can find s'_a such that the equation $s'_a \oplus K_{AB} = w(s_a \oplus K_{AB}) = Time'_a x_a - (r'_{a1} \oplus r'_{a2})(k_{a1} + k_{a2}) \bmod p\text{-}1$ holds.

Therefore, B can construct valid information $(r'_{a1}, r'_{a2}, s'_a, Time'_a, cert(y_a))$ of A even if he does not know A's secret key x_a. Both A and B can verify the following equation: $y_a^{Time'_a} = (r_{a1} r_{a2})^{r'_{a1} \oplus r'_{a2}} g^{s'_a \oplus K_{AB}} = (r'_{a1} r'_{a2})^{r'_{a1} \oplus r'_{a2}} g^{s'_a \oplus K_{AB}} \bmod p$.

In this situation, A cannot prove that the information is a forgery. Thus, B can masquerade as the information $(r'_{a1}, r'_{a2}, s'_a, Time'_a, cert(y_a))$ that actually came from A. Therefore, the malicious B can fool A into accepting the forged information as valid, without knowing the secret key x_a of entity A. Entity B can find as many such sets as he wishes and choose the suitable ones, waiting for the valid timestamp. In a similar approach, A can fool B into accepting the forged information as valid, without knowing the secret key x_b of entity B. Hence, Chien and Jan's scheme is easily confused.

4 Our Improved Scheme

Here, we propose an improved method to remedy the weakness of Chien and Jan's method in order to achieve authenticated property. The key point of our improved scheme is that we do not use timestamp as in Chien and Jan's scheme for verification of information. Our method also consists of three phases: system initialization phase, key generation and signature phase, and verification phase. The system parameters and the system initialization phase are the same as those in Chien and Jan's scheme.

Key and Signature Generation Phase

Entity A randomly selects two secret numbers k_{a1} and k_{a2}, then computes their corresponding public values $r_{a1} = g^{k_{a1}} \bmod p$ and $r_{a2} = g^{k_{a2}} \bmod p$. Next, A signs his signature on these two values by generating the equation as

$$s_a \oplus K_{AB} = (r_{a1} - r_{a2}) x_a - (r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \bmod p\text{-}1, \quad (3)$$

where $K_{AB} = g^{x_a x_b} = (y_a)^{x_b} = (y_b)^{x_a} \bmod p$ is the long-term secret key between A and B. Then, A sends $(r_{a1}, r_{a2}, s_a, cert(y_a))$ to B. Proceeding in a similar approach, B computes and sends $(r_{b1}, r_{b2}, s_b, cert(y_b))$ to A.

Verification Phase

After receiving information $(r_{a1}, r_{a2}, s_a, cert(y_a))$ from entity A, entity B verifies if the following equation holds:

$$y_a^{(r_{a1} - r_{a2})} = (r_{a1} r_{a2})^{r_{a1} \oplus r_{a2}} g^{s_a \oplus K_{AB}} \bmod p. \quad (4)$$

Entity A also verifies the information $(r_{b1}, r_{b2}, s_b, cert(y_b))$ received from B. If both A and B succeed in their verifications, they then can derive four common keys

$K_1 = (r_{a1})^{k_{b1}} = (r_{b1})^{k_{a1}} = g^{k_{a1}k_{b1}} \mod p$, $K_2 = (r_{a1})^{k_{b2}} = (r_{b2})^{k_{a1}} = g^{k_{a1}k_{b2}} \mod p$,
$K_3 = (r_{a2})^{k_{b1}} = (r_{b1})^{k_{a2}} = g^{k_{a2}k_{b1}} \mod p$, and $K_4 = (r_{a2})^{k_{b2}} = (r_{b2})^{k_{a2}} = g^{k_{a2}k_{b2}} \mod p$.

Next, the analysis of the security of our improved scheme is as follows. Based on Chien and Jan's scheme [12], the security of our improved scheme is the same as theirs except that our improved scheme does not have the same weakness. In Equation (3) of our improvement, it provides

$s_a \oplus K_{AB} = (r_{a1} - r_{a2})x_a - (r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1$. Using the same concept in Section 3, assume one constructs r'_{a1} and r'_{a2} such that $r'_{a1}r'_{a2} = r_{a1}r_{a2} \mod p$, and then derives w so that $(r'_{a1} \oplus r'_{a2}) = w(r_{a1} \oplus r_{a2}) \mod p\text{-}1$. According to Equation (3), it has

$$w(s_a \oplus K_{AB}) = w(r_{a1} - r_{a2})x_a - w(r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1$$
$$= w(r_{a1} - r_{a2})x_a - (r'_{a1} \oplus r'_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1. \quad (5)$$

Since $K_{AB} = g^{x_a x_b} = (y_a)^{x_b} = (y_b)^{x_a} \mod p$ is the long-term secret key between A and B. In Equation (5), if $w(r_{a1} - r_{a2}) = (r'_{a1} - r'_{a2}) \mod p\text{-}1$ holds, then entity B can easily derive s'_a such that

$$s'_a \oplus K_{AB} = w(s_a \oplus K_{AB}) = (r'_{a1} - r'_{a2})x_a - (r'_{a1} \oplus r'_{a2})(k_{a1} + k_{a2}) \mod \text{p-}1. \quad (6)$$

In this situation, B can forge A to generate the valid information $(r'_{a1}, r'_{a2}, s_a, cert(y_a))$ without knowing the secret key x_a of A. However, it is very difficult for B to obtain w, r'_{a1}, and r'_{a2} such that $r'_{a1}r'_{a2} = r_{a1}r_{a2} \mod p$, $(r'_{a1} \oplus r'_{a2}) = w(r_{a1} \oplus r_{a2}) \mod p\text{-}1$, and $w(r_{a1} - r_{a2}) = (r'_{a1} - r'_{a2}) \mod p\text{-}1$ holds. The probability is equivalent to performing an exhaustive search on w, r'_{a1}, and r'_{a2}. Relying on probability, B cannot forge the valid information of A. Similarly, A cannot forge the valid information of B. Thus, in our improved scheme, entity A/B cannot use the same method in Section 3 to cheat the other entity of passing the verification. Hence, our improved method can achieve authenticated property.

Next, the key utilization of our improved scheme is the same as Chien and Jan's scheme [12]. The above mentioned schemes [6][8][9][11] only use three of the four common keys to preserve the perfect forward secrecy. According to our signature generation equation, we have

$s_a \oplus K_{AB} = (r_{a1} - r_{a2})x_a - (r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) \mod p\text{-}1$ and
$s_b \oplus K_{AB} = (r_{b1} - r_{b2})x_b - (r_{b1} \oplus r_{b2})(k_{b1} + k_{b2}) \mod p\text{-}1$.

Then, we can compute

$$(r_{a1} - r_{a2})(r_{b1} - r_{b2})x_a x_b = (s_a \oplus K_{AB})(s_b \oplus K_{AB}) + (s_a \oplus K_{AB})(r_{b1} \oplus r_{b2})(k_{b1} + k_{b2}) +$$
$$(s_b \oplus K_{AB})(r_{a1} \oplus r_{a2})(k_{a1} + k_{a2}) + (r_{a1} \oplus r_{a2})(r_{b1} \oplus r_{b2})(k_{a1} + k_{a2})(k_{b1} + k_{b2}) \mod p\text{-}1,$$

and derive

$$K_{AB}^{(r_{a1}-r_{a2})(r_{b1}-r_{b2})} =$$
$$g^{(s_a \oplus K_{AB})(s_b \oplus K_{AB})}(r_{b1}r_{b2})^{(s_a \oplus K_{AB})(r_{b1} \oplus r_{b2})}(r_{a1}r_{a2})^{(s_b \oplus K_{AB})(r_{a1} \oplus r_{a2})}(K_1K_2K_3K_4)^{(r_{a1} \oplus r_{a2})(r_{b1} \oplus r_{b2})} \bmod p.$$

From the equation, we see the adversary cannot derive the long-term secret key K_{AB} even if he gets the session keys K_1, K_2, K_3, and K_4. Therefore, our scheme can use all the four session keys and achieve better key utilization.

5 Conclusions

In this paper, we have shown that Chien and Jan's scheme does not achieve authenticated property. We also propose a secure multiple-key agreement protocol without using any one-way functions, which simplifies its security assumption. In our method, the malicious entity has no chance to forge other entity's signature to send out the forgery. Thus, our method can achieve the authenticated property. Compared to the previous schemes, our scheme has better key utilization.

References

1. W. Diffie and M. E. Hellman: New directions in cryptography. IEEE Transactions on Information Theory. Vol. 22, No. 6. (1976) 644-654.
2. T. ElGamal: A public-key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, Vol. 33, No. 2 (1985) 469-472.
3. K. Nyberg and R. A. Rueppel: Message recovery for signature schemes based on the discrete logarithm. Advance in Cryptology-EUROCRYPT'94, Springer-Verlag, (1994) 175-190.
4. A. Arazi: Integrating a key cryptosystem into the digital signature standard. *Electronic Letters*, Vol. 29, No. 11 (1993) 966-967.
5. K. Nyberg and R. A. Rueppel: Weakness in some recent key agreement protocols. *Electronic Letters*, Vol. 30, No. 1 (1994) 26-27.
6. L. Harn and H. Y. Lin: An authenticated key agreement protocol without using one-way function. Proceedings of the 8th National Conference of Information Security, Kaohsiung, Taiwan, May (1998) 155-160.
7. H. Dobbertin: The status of MD5 after a recent attack. CyptoBytes Vol. 2, No, 2 (1996) 1-6.
8. S. M. Yen and M. Joye: Improved authenticated multiple-key agreement protocol. Electronic Letters, Vol. 34, No. 18 (1998) 1738-1739.
9. T. S. Wu, W. H. He, and C. L. Hsu: Security of authenticated multiple-key agreement protocols. Electronic Letters, Vol. 35, No. 5(1999) 391-392
10. A. Menezes, P. Van Oorschot, and S. Vanstone: Handbook of applied cryptography, CRC Press, (1997).
11. H. T. Yen, H. M. Sun, and T. Hwang: Improved authenticated multiple-key agreement protocol. Proceedings of the 11th National Conference of Information Security, Tainan, Taiwan, May(2001) 229-231.
12. H. Y. Chien and J. K. Jan: Improved authenticated multiple-key agreement protocol without using conventional one-way function. Applied Mathematics and Computation, Vol. 147, (2004) 491-497.

Topology-Based Macroscopical Response and Control Technology for Network Security Event

Hui He, Mingzeng Hu, Weizhe Zhang, Hongli Zhang, and Zhi Yang

Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
{hehui, mzh, zwz, zhl, yangzhi}@pact518.hit.edu.cn

Abstract. The large-scale network security events are becoming a major threat to internet. How to quickly detect and effectively control the network security events' spreading has become the research focus among network security experts. By combining active topology measurement with distributed anomaly detection, a large-scale network security events' discovery and cooperative system is proposed, which focuses on macroscopical alert analysis, control point selection, creating control suggestion etc. After the process of visualization, it exhibits preferable application effect. The experimental result proved that it offers administrators the direct decisive advice to prevent network security event from overspreading.

1 Introduction

The large-scale network security events are becoming a major threat to Internet. Since 2001, worms such as Code Red, Nimda etc., have spread all over the world in a few minutes, which led to striking cost to our society. Hence, network security events' discovery and response system is essential in fighting against the Internet disasters, a just-in-time warning and quarantine can be invaluable in saving money and circumscribing damages.

This led to the emergence of distributed IDSs such as DIDS[1], GrIDS[2], EMERALD[3], IDCA[4] and AAFID[5] etc. DIDS uses the paradigm of distributed data collection and data processing. GrIDS detects network abnormal by creating the active attacking hierarchy graph. AAFID is an agent-based DIDS which uses a three-level hierarchy. Low-level agents perform data collection and low-level analysis. IDCA is a benefit-driven framework of distribute IDS based on autonomy agents.

Above all, there are three factors that result in recently detection and response system incapable of preventing the network security events from spreading. First, the limitation of detection point placement. They directly affect the effect of controlling process afterwards. Next, the limitation of control points selection. After the detection of network events, it's crucial to cut off the infecting path. At last, the control policy employed in the control process. After finding out the proper control routers, it is vital how to prevent the events from overspreading. We need to design the control policy practically and effectually but not specially.

Different from the recent researches we propose to use Internet topology for efficient and macrotropical response and control of network abnormal events. It's a

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 560 – 566, 2005.
© Springer-Verlag Berlin Heidelberg 2005

novel approach which eliminates the problems mentioned above and enables the administrators to gain the whole viewpoint of network security events. Instead of analyzing the text alarm information, the distribution of the events on the topology is investigated.

The paper is organized as follows: in section 2 the architecture of NSEDCS (network security event discovery and collaboration system) is introduced and in section 3 the new control technologies is described, which contain router clustering algorithm, control router selection algorithm and control policy design. Section 4 will present some experimental results. Finally, section 5 concluded the paper.

2 The Architecture of NSEDCS

The design of NSEDCS (network security event discovery and collaboration system) can gain effectiveness when pouring the active abnormal events into the large-scale network. The architecture of system consists of four parts including detection subsystem, topology auto-discovering subsystem, analyzing subsystem and visualizing subsystem, which is shown in Figure 1.

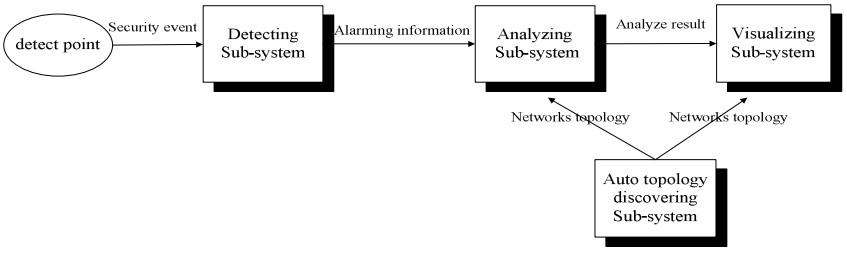


Fig. 1. The architecture of NSEDCS (network security event discovery and collaboration system)

The detection subsystem consists of distributed detect points placed on multiple network interface, which are responsible for the local detection of abnormal events and provide alarm information. Topology auto-discovering subsystem is one of the research efforts to construct a full map of the internet [6]. The analyzing subsystem acts as a set of synthetically processors in finding the key control routers from measured real route-level topology in order to restrict the overspread of the abnormal events on Internet. The details of these topics are discussed in Section 3.The visualizing subsystem shows the network topology, macro distribution graph of security events and control routers. It will give the network administrator good macroscopical viewpoint and control advice to cut off the spreading path of the abnormal events effectively.

In our paper, we mainly focus on the key control technology and algorithms based on topology applied to the analyzing subsystem. Now we have obtained the data by the multiple vantage points topology discovering subsystem [7] and the worms' data from distribute detecting subsystem. Thus, in the next section we will discuss the control technologies employed.

3 Control Technologies

In order to find out the concentration area of the abnormal events, we propose the control technology, which consists of three phases: router clustering, key control router selection algorithm and control policy.

3.1 Router Clustering

In this paper, we propose two definitions related to the problem of the router clustering, which is equivalent to the problem of topology graph partition.

Definition 1. (Router partition): For a router set V, the subset $\{V_1, V_2,\ldots V_n\}$ is called router partition when $V=V_1 \cup V_2 \cup \cdots \cup V_n$ and $\forall\, i, j \in [1,n]$, $V_i \cap V_j = \Phi$.

Definition 2. (Optimization partition rule): For a partition P, the number of cluster partitioned is denoted as N, and $e_{ij} = \{(m,n) | \forall\, m \in S_i, \forall\, n \in S_j\ ,\ (m,n) \in E\}$ is defined as the number of edges between set S_i and S_j, which $S=\{S_i | S_i$ is a set of router cluster partitioned$\}$. $|S_i|$ denoted the number of routers included in the set S_i.

A router partition is called optimum if it fulfills five demands: Connectivity is considered as a primary criterion for clustering; There are not too much routers in S_i, that is, $|S_i|$ is limited to a certain value; There are not too much clusters after partitioning; the number of edge between two clusters is as little as possible; it is not exist that one cluster with too much routers and the other with too little.

Combining all of the above demands, we propose approximate optimization algorithm for router partition. The algorithm is described as following:

```
Input: RouterSet (router-level topology)
Output: A set of cluster
Begin
      ClusterId = FIRSTCLUSTER
      for each route in RouterSet
        if route is UNCLASSIFIED then
            identify route with ClusterID
            put route in Queue
            while(Queue not Empty)
               pop route from Queue
               for each route1 in RouterSet
                  if route1 is UNCLASSIFIED then
                     if Distance(route, route1) < Threshold then
                        identify route1 with ClusterID
                        put route1 in Queue
                     End if
                  End if
               End for
            End while
            ClusterID ++
        End if
      End for
      End
```

Fig. 2. Router clustering algorithm based on DBSCAN

3.2 Control Router Selection Algorithm

The objective of control point selection is to provide network administrators an effective method for preventing abnormal events from spreading. Control routers

related to abnormal routers are picked up based on network topology. Based on different policy employed, we propose two algorithms to identify control routers in a specific topology.

3.2.1 Reduced Control Router Selecting

We have five definition related to the problem of reduced control router selecting.

Definition 3. (Abnormal hosts set): Let $SickH(h)$ denotes host h has been infected and can infect others. Then abnormal hosts set is defined as $AHS= \{h| SickH(h)\}$.

Definition 4. (Router gateway): Router r is the gateway of host h if and only if r is the first router traveled into Internet by h, denoted $r = RouteGate(h)$.

Definition 5. (Abnormal router sets): Router r is abnormal router if and only if there exist host h, exists $SickH(h)$ and $r=RouteGate(h)$, denoted $SickR(r)$, then we have abnormal router sets is denoted as $ARS=\{r| SickR(r)\}$.

Definition 6. (Neighbor routers): Let Adjoining $(r_1, r2)$ indicates neighbor routers if and only if there exist router r_1 and router r_2, where r_1 is connected with r_2 directly.

Definition 7. (Reduced control router sets): Let $CRS=\{r| \exists\ r_1, SickR(r_1), \text{Adjoining } (r_1, r_2)\}$ is the control router sets. Let $RCRS=\{ r| r \in CRS, r \notin ARS\}$ is the reduced control router sets.

The reduced control router selecting algorithm aims to select routers which are connected to the abnormal routers directly, but not included in the abnormal router sets. Although it will have greater number of routers, it is higher veracity to the path for events overspreading.

3.2.2 Key Control Router Selecting

Taking the reduced control router as event control point is the most accurate way to limit its spreading. But for the number of the reduced control router set, it is not the most effective method, especially in emergency. Thus, we proposed another type of control router sets called key control router set, which only contain important routers.

Concerning this criterion, there are two ideas, coarse selecting and fine selecting. Coarse selecting is to search the routers in the vital point in common sense, such as in the center of a cluster or at the cluster boundary. Based on the diverse topology structure, we put forward different selecting strategies. They are Near Neighbors Rule, High Degree Tendency Rule and Children Collapse Rule.

- Near Neighbors Rule: In Figure 3(a), Node A is denoted as an abnormal router and node B is the key control router. Because node B is connected with A directly just like node A's near Neighbor in the graph. So we called it as Near Neighbors Rule. Obviously, we can control node B to prevent node A from infecting node C.
- High Degree Tendency Rule: It is depicted in Figure 3(b) and Figure 3(c). As above, node A is denoted as a victim router and node C is the key control router in the graph. Figure 3(b) shows the node A, node B and node C is connected as a ring. Although node B and node C are both connected with node A, node C has the higher node-degree than node B, in the case, we consider node C as our selection. So we named the rule as High Degree Tendency Rule. Figure 3(c) is the same as Figure 3(b).
- Children Collapse Rule: Figure 3(d) illustrate the principle of Children Collapse Rule. In the graph, node A and node B are both abnormal routers, node C is the key

control router. We can find that node B and node A have the same root. If we want to prevent node A from spreading, we have to control node D, and in order to restrict node B, we have to control node C as well. Because node B and node A have the same root within a certain hops, so we consider their root as the key control point. In the case, it not only decrease the number of control point but also cut off the infecting path efficiently.

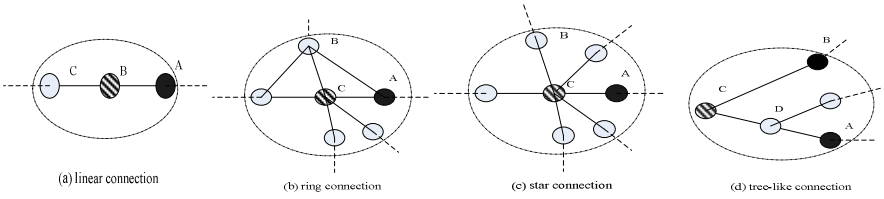


Fig. 3. Searching rules for Key Control Router based on diversity structure of topology

We performed to synthesize the above rules to build up our key control router sets. Finding out the router sets which have a great effect on the performance of network transmission based on diversity structure in topology is more convenient. Especially in emergency, when the network security events infect others too rapidly to response in time, we can employ control policy in those routers primarily, and then turn to the abnormal routers one by one.

3.3 Control Policy Based on Subnet Merging

Router can restrict some certain IP access to the network by AccessList [8]. Control policy creating technology processes as shown in Figure 4.

```
Input: HostSet   // Hosts set for merging
Output: Sets of subnet and hosts
Begin
   insert each host in HostSet into NetHash using hash function f
   attach each net in NetHash on C[INITMASKLEN]
   for i from INITMASKLEN to 1
      for each net1,net2 in C[i]
            if CanMerge(net1, net2) then
              net = Merge(net1, net2)
              delete net1,net2
              attach net on C[i-1]
            End if
      End for
   End for
   for i from 0 to INITMASKLEN
      print each net in C[i]
   End for
   End
```

Fig. 4. Control router merging algorithm

This method can speed up the response time for network administrator in emergency. The control policy based on subnet merging provided in the paper is not the only way to harness the large-scale network abnormal events, some control policy are studied such as in [9], aiming to prevent DDoS attack, they limited the DDoS traffic through the routers. These technologies are mostly to solve some certain attack. But the policy proposed in our paper is simple and more general.

4 Experimental Results

We conducted experiment based on the real network topology measured by auto topology discovering subsystem[7], which contains 19,847 distinct routers and 24,864 links between them from 31 multiple sources distributed over the entire CHINANET.

The clustering result shown in Figure 5, we can see that the whole topology network is divided into several clusters, which followed the demands of definition 4. For visualizing the topology in vector mode, we can see the routers of a certain cluster in detail by zooming out the graph. And we can see the effect of reduced control router algorithm in Figure 6 and effect of key control router selecting algorithm in Figure 7. As shown in Figure 6, the red node represents the control routers which enclose the black node, abnormal routers. In this way, the further spreading will be effectively controlled. But in the emergency, we have to response quickly to the abnormal event, such as worms or DDoS. We need to cut off the key routers in the path lest it infect other network. In Figure 7, the green node is the right routers mentioned above.

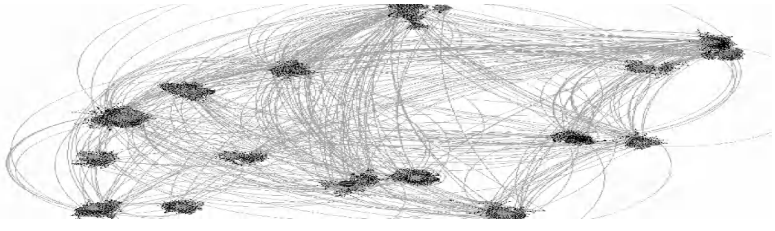


Fig. 5. Entire graphs on clustering result of measured topology

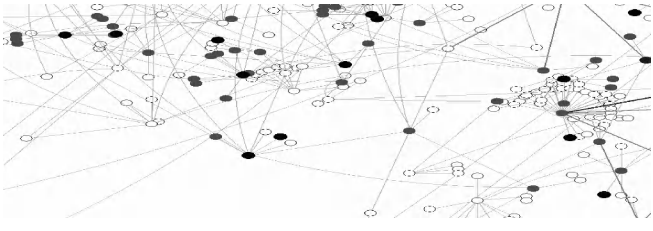


Fig. 6. Effect of reduced control router algorithm

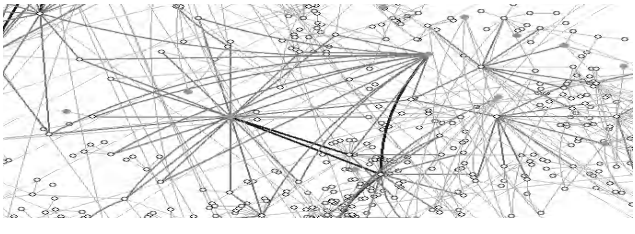


Fig. 7. Effect of key control router selecting algorithm

5 Conclusion

The paper has three contributions. First, we present a network security event discovery and collaboration system. It is novel that the active topology discovering system is integrated with distributed IDS. Second, we propose a control router selection algorithm based on topology router clustering. Regarding as the reduced control routers set and the key control router sets, we put forward three rules for searching the right control routers. Finally, we provide a simple and effective policy to control the events spreading. The experiment is performed on the real network topology measured by auto topology discovering subsystem, which contains 19,847 distinct routers and 24,864 links. And the experimental results proved the control technology is effective. It offers administrators the significant advice to prevent large-scale Network Security Events from overspreading.

References

1. S. Snapp, J.Brentano, et al.: DIDS (Distributed Intrusion Detection System) – Motivation, Architecture and an Early Prototype. In: Proceedings of the National Computer Security Conference, 1991.
2. S. Staniford-Chen, S. Cheung, R. Crawford et al.: GrIDS - A Graph Based Intrusion Detection System for large networks. In Proceedings of the 20th National Information Systems Security Conference, 1(1996) 361-370.
3. Phillip A. Porras and Peter G. Neumann: EMERALD: event monitoring enabling responses to anomalous live disturbances. In 1997 National Information Systems Security Conference, 12(1997)125-131.
4. Rajeev G., Eugene H., Spafford: A Framework for Distributed Intrusion Detection using Interest Driven Cooperating Agents. 2001.
5. Jai S. B., Jose O. G., David I., Eugene S., Diego Z.: architecture for intrusion detection using autonomous agents. Proceedings of Fourteenth Annual Computer Security Applications Conference, IEEE Computer Society, 12(1998) 13-24.
6. K.C.Claffy, D.McRobb: Measurement and visualization of Internet connectivity and performance, http://www.caida.org/tools/skitter/.
7. Jiang, Y., Hu, M. Z., Fang, B. X., Zhang, H. L.: An Internet router level topology automatically discovering system. J. China Institute of Communications. 2002, 23(12), 54-62.
8. George C.Sackett: Cisco router manual (the Second Version), China Machine Press. 2002.
9. LIANG Feng, David Yau: Using Adaptive Router Throttles Against Distributed Denial-of-Sevice Attacks, Journal of Software, 13(2002) 1220-1227.

Adaptive Hiding Scheme Based on VQ-Indices Using Commutable Codewords

Chinchen Chang[1,3], Chiachen Lin[2], and Junbin Yeh[3]

[1] Department of Information Engineering and Computer Science,
Feng Chia University, Taiwan, China
ccc@cs.ccu.edu.tw
[2] Department of Computer Science and Information Management,
Providence University, Taiwan, China
mhlin3@pu.edu.tw
[3] Department of Computer Science and Information Engineering,
National Chung Cheng University, Taiwan, China
ycp92@cs.ccu.edu.tw

Abstract. Hiding secret message in an image has received significant attention that it will not arise attackers' suspicion. Recently, VQ-based compression domain is frequently employed. Although some hiding schemes are VQ-based compression domain, their hiding capacities are limited. To increase the hiding capacity, this paper applies codewords grouping concept to design an adaptive capacity hiding scheme. Based on our scheme, more bits can be embedded into single VQ index only if its degradation caused by replacing other codeword is less than the predefined threshold. The advantages of this proposed scheme are the higher capacity exploration on secret embedding, and acceptable stability on the display of stego-images.

1 Introduction

Over the past years, Internet has become a popular information transmission channel, so secure communication is required. Cryptography can protect the confidential content from being wiretapped and tampered. Information hiding embeds the encrypted/plain secret message into some meaningful media, such as images or compressed codes. Even if intruders wiretap the Internet and intercept those doubtful images, the stego-images are hard to be distinguished from others by human vision system. Furthermore, the hidden secret is also difficultly to be extracted from the stego-images without any critical information that only belongs to the legal senders and receivers.

There are two major techniques developed in information hiding, spatial-domain manner [1] and frequency-domain manner [3, 5]. The most well known solution in spatial-domain is least-significant-bit (LSB) replacement [1]. While in the frequency-domain, it decomposes the cover image into non-overlapping blocks and transforms them into frequency domain. The secret data is then combined with the relative coefficients in the frequency-form stego-image. Generally, hidden secret in frequency domain is hard to be detected, but the hiding capacity is lower than spatial-domain schemes.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 567 – 572, 2005.
© Springer-Verlag Berlin Heidelberg 2005

To conquer the limitation of Internet bandwidth, compression is a general tool to reduce the size of transmitting images. Since VQ is a famous image compression technique, its compression codes are commonly employed to hide secret data. Because these indices indicate the approximate blocks, Jo and Kim proposed a novel hiding scheme [4]. They split a codebook into three groups, to represent the bit valued 1, 0, or none. They selected a suitable codeword from the relative group to represent an index of each block according to its secret bit. In Chang et al.'s scheme [2], each sub-cluster only contains two or four codewords, and each index can hide one or two secret bits.

The rest of our paper is organized as follows. In Section 2, a brief review of the past literatures related to embedding secret data into VQ indices is given. In Section 3, we will depict our proposed scheme. The empirical results will be given in Section 4, and, finally, we will conclude our scheme in Section 5.

2 Related Works

In this section, we will describe Jo and Kim's [4] (being called "VQIW" for short) and Chang et al.'s [2] (being called "VQCQ" for short) schemes of hiding secret data in VQ indices. When embedding, they usually compressed a cover image by VQ, and grouped the codewords in codebook into different groups. Next, they combine the grouping results and secret data to replace index for current block. Finally, the set of indices is sent to a receiver. During the extraction procedure, the codewords were grouped as in embedding procedure. Next, the secret data is extracted according to which the current index belongs to. Finally, the reconstructed cover image is generated by composing of the retrieved blocks.

2.1 VQ-Based Image Watermark Scheme

In VQIW, Jo and Kim split the codebook $C = \{c_0, c_1, c_2, \ldots, c_{N-1}\}$ into three groups, named G_0, G_1, G_{-1}, respectively. Codewords in G_0 and G_1 respectively represent that the embedding secret bit is 0 and 1. On the contrary, any codeword located in G_{-1} means there is nothing embedded. They also assumed that each codeword in codebook only belongs to one of these three groups. Each codeword c_i in G_0 must have a corresponding codeword c_j in G_1, and there exists the minimal squared Euclidean distance between c_i and c_j.

VQIW is simple, but improvement can still be made on their hiding capacity because VQIW now only embeds less than one bit for each index.

2.2 Adaptive Steganography for Index-Based Images Using Codeword Grouping

VQCG also splits the codebook into three groups G_1, G_2, G_4. Here, a codeword only belongs to one sub-cluster. Each sub-cluster in G_4 contains four similar codewords. Each sub-cluster in G_2 contains two similar codewords. Each sub-cluster in G_1 only has one codeword.

Before codewords' grouping, VQCG predefines two thresholds, *TH4* and *TH2*. If codewords belong to the same 4-member sub-cluster, S_{4i}, the distance in between of

them is less than *TH4*. If codewords belong to the same 2-member sub-cluster, S_{2j}, the distance between each other is less than *TH2*. The remainder codewords are grouped individually as 1-member sub-cluster, S_{1k}.

In general, VQCG can embed more bits in an index than Jo and Kim's scheme did. Nevertheless, they used two threshold values when they assigned codewords into three groups. Our scheme only uses one threshold to design codewords' grouping as well as the whole embedding procedure. The detailed description of our proposed scheme is given in the following Subsections.

3 The Proposed Scheme

Unlike VQIW or VQCG, our hiding scheme adaptively group codewords into sub-clusters. In order to maintain image quality of VQ-based stego-images, we still strictly set the distance between each two codewords in the same sub-cluster to be less than our predetermined threshold.

3.1 Codewords' Grouping Phase

In order to select suitable codewords to a sub-cluster, we only collect codewords together when the distance between each other is less than the threshold D. If there are n codewords meet our requirement, we only select the $\lfloor \log_2 n \rfloor$ power of 2 codewords with minimum distances and organize them as a sub-cluster. The remainder codewords will be another sub-cluster, and keep doing the grouping procedure.

Generally speaking, our proposed scheme splits the codebook $C = \{c_0, c_1, c_2, \ldots, c_{N-1}\}$ into $G = \{S_0, S_1, S_2, \ldots, S_{M-1}\}$, where sub-clusters S_i, $0 \le i \le N-1$. Moreover, any codeword only belongs to one sub-cluster. The squared Euclidean distance $SED(c_j, c_k)$ between c_j and c_k must be less than D, where $c_j, c_k \in S_i,\ 1 \le j,k \le |G_i|$, where $|G_i|$ denotes the number of members in G_i.

We do not restrict the number of codewords in each sub-cluster; hence, we can embed as many secret bits as possible into one index. Fig.1 shows the different results after VQIW, VQCG, and the proposed scheme perform codewords' grouping, respectively. Although the total distortion caused in Fig.1(c) is greater than 1(b), we still can make sure that any distance between each two codewords in the same sub-cluster is shorter than the threshold D. We pay the little distortion, but embed more bits.

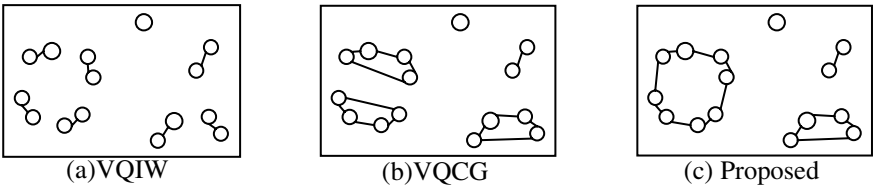


Fig. 1. Different results of grouping schemes

3.2 Embedding Procedure

After codewords' grouping is completed, our proposed scheme then embeds secret bits into VQ-compressed indices. Given VQ-indices x_t, $0 \le t \le T-1$, where T is the number of blocks in a cover image. We find out the sub-cluster S_i to which x_t belongs. Next, we compute $b = \lfloor \log_2 |S_i| \rfloor$, and then we know how many bits b can be embedded into the sub-cluster S_i. If b equals 0, then the current index will be skipped because there is no alternative in this sub-cluster. If b does not equal 0, we translate b secret bits into an unsigned integer S first. Next, we use the S-th member of S_i to replace x_t, and then generate a new index called x_t'. Finally, we can successfully embed b secret bits by transforming x_t into x_t'.

3.3 Extracting Procedure

Our proposed extraction procedure only requires looking up tables. We have the result of codewords' grouping, G, and the embedded VQ-compressed indices, $x_t's$ for $0 \le i \le N-1$, so we can find out which sub-cluster S_i x_t' belongs. Next, we can get the order of index x_t' in the sub-cluster S_i. The order is then translated into binary form and the binary bits are the extracted secret bits.

3.4 Discussions of Threshold Value

In the proposed scheme, we observe the threshold plays a crucial role in the image quality of VQ-based setgo-images and the number of codewords in one sub-cluster. It is a trade-off between image quality and the hiding capacity. Some researchers pointed out that we can modify the least three bits at most in one pixel in LSB to maintain the image quality of setgo-images. It means that one pixel's value may increase or decrease eight at most. In general, each block in VQ process is usually in size of 4×4 pixels. If we do not want to make anyone suspicious of a block's modification, we should keep the squared Euclidean distance of a block less than $(4\times4)\times8^2=1024$. That is why we set threshold as 1024. To prove our inference, more supportive experimental results are given in Section 4.

4 Experiments

In conducting our experiments, we take some gray-level images with size of 512×512 pixels for test. These images are divided into 16384 non-overlapping blocks with size of 4×4 pixels by traditional VQ compression. We randomly generate 16384 bits to be our secret message.

Fig.2 is the embedded results of "Lena" using VQIW [4], VQCG [2] and our proposed scheme with codebook size of 512 and threshold of 1024. Fig.2(a) is a VQ-compressed image before embedding. Figs.2(b), 2(c) and 2(d) are the hiding result by using VQIW, VQCG, and our proposed scheme. We can see that the *PSNR* of VQIW is the highest, and the *PSNR* of our scheme is only 0.37 less than that of VQIW.



(a) VQ-compressed (b) VQIW (31.7) (c) VQCG (31.5) (d) Proposed (31.33)

Fig. 2. VQ-compressed cover image of "Lena" and three stego-images using different schemes with codebook size of 512 and threshold of 1024

Table 1. Hiding capacities and PSNRs of different schemes with the codebook size of 512

	Schemes / Cover Image	VQIW *TH*=1024	VQCG *TH4*=1024 *TH2*=512	Proposed scheme *TH*=1024	VQIW *TH*=2048	VQCG *TH4*=2048 *TH2*=1024	Proposed scheme *TH*=2048
Hiding Capacities	Lena	12283	22036	26620	14437	26401	35243
	Sailboat	10977	17095	19698	13204	23476	27409
	Pepper	12133	21221	25677	13946	25835	32842
	Toys	6558	10106	13702	9139	16632	17706
	Baboon	6777	10530	12899	10318	16155	21077
	Airplane	12717	23191	29605	14057	26107	36897
PPSNR	Lena	31.7	31.5	31.33	31.2	30.97	30.55
	Sailboat	28.91	28.8	28.69	28.59	28.39	28.25
	Pepper	30.94	30.73	30.57	30.58	30.25	30.02
	Toys	30.85	30.57	30.38	30.33	29.86	29.93
	Baboon	24.33	24.29	24.26	24.13	24.09	23.97
	Airplane	31.18	31.04	30.75	30.88	30.64	30.29

Table 1 lists the hiding capacity and *PSNRs* of these three hiding schemes with codebooks all in size of 512. In general, our proposed scheme has superior increment than other schemes. Compared *PSNRs* with VQIW, in the worst case, our proposed scheme only decreases about 0.47 when threshold is 1024, and about 0.65 when threshold is 2048.



(a) *TH*=512, *PSNR*=31.72, *Bits*=19973 (b) *TH*=1024, *PSNR*=31.33, *Bits*=26620 (c) *TH*=2048, *PSNR*=30.55, *Bits*=35243 (d) *TH*=4096, *PSNR*=29.3, *Bits*=46645

Fig. 3. Stego-images of an 512×512 "Lena" generated by the proposed scheme with codebook size 512

Fig.3 shows the results using different thresholds. Although the hiding capacities of Figs.3(c) and 3(d) have great increment, their *PSNRs* decrease rapidly. As discussed in Subsection 3.4, it is a good choice to set the threshold as 1024, because it can keep balance between image quality and hiding capacity.

5 Conclusions

In this paper, we proposed a novel hiding scheme based on codewords' grouping concept to achieve higher hiding capacity. In our scheme, we utilize only one threshold to collect as many suitable codewords as possible together, and then select those feasible ones with minimal distortion to organize as one sub-cluster.

The experimental results confirm that our proposed scheme can embed more bits without huge degradation of image quality. Also, two alternatives are provided to increase hiding capacity, one is codebook size and the other is threshold. The adaptive hiding capacity makes our proposed scheme beneficial to various applications.

References

1. C. C. Chang, J. Y. Hsiao, and C. S. Chan: Finding Optimal Least-significant-bit Substitution in Image Hiding by Dynamic Programming Strategy, Pattern Recognition, Vol. 36(2003) 1583-1595
2. C. C. Chang, P. Y. Tsai, and M. H. Lin: An Adaptive Steganography for Index-based Images Using Codeword Grouping, Pacific Rim Conference on Multimedia, Tokyo, Japan, (2004) 731-738.
3. I. J. Cox, J. Killian, T. Leighton, and T. Shamoon: Secure Spread Spectrum Watermarking for Multimedia, IEEE Transactions on Image Processing, Vol. 6, No. 12(1997) 1673-1687.
4. M. Jo and H. D. Kim: A digital Image Watermarking Scheme Based on Vector Quantisation, IEICE Transactions on Information and Systems, Vol. E85-D, No. 6, Jun. (2002).
5. H. Noda, J. Spaulding, M. N. Shirazi, and E. Kawaguchi: Application of Bit-plane Decomposition Steganography to JPEG2000 Encoded Images, IEEE Signal Processing Letters, Vol. 9, No. 12(2002) 410-413.

Reversible Data Hiding for Image Based on Histogram Modification of Wavelet Coefficients

Xiaoping Liang, Xiaoyun Wu, and Jiwu Huang*

School of Information Science and Technology, Sun Yat-sen University,
Guangzhou, Guangdong 510275, China
Guangdong Province Key Laboratory of Information Security, 510275, P.R. China
isshjw@zsu.edu.cn

Abstract. Reversible data hiding has drawn intensive attention recently. Reversibility means to embed data into digital media and restore the original media from marked media losslessly. Reversible data hiding technique is desired in sensitive applications that require no permanent loss of signal fidelity. In this paper, we exploit statistical property of wavelet detail coefficients and propose a novel reversible data hiding for digital image by modifying integer wavelet coefficients slightly to embed data. Our algorithm doesn't need any lossless compression technique and provides several embedding modes for choice according to capacity and visual quality requirement. Experimental results show that our algorithm achieves high capacity while keeping low distortion between marked image and the original one.

1 Introduction

Traditional data hiding technique embeds data into digital media while distorting the original media permanently. Those permanent distortions may not be allowable in some sensitive applications, such as law enforcement, medical and military image systems. Reversible data hiding technique can solve this problem, because it can restore the original media from marked media losslessly.

Reversible data hiding has drawn intensive attention recently. Some reversible data hiding methods have been reported in the literature in recent years. Most of the reversible data hiding methods rely on lossless compression technique more or less to make space for embedding data. Independent bit-plane compression method proposed by Fridrich et al. [1] achieves very small capacity by compressing bit planes in spatial domain. RS embedding method proposed by Goljan et al. [2] achieves higher performance (capacity versus visual quality measured in PSNR, i.e. peak signal-to-noise-ratio) than the foregoing method by compressing the R/S feature string. Generalized-LSB embedding method proposed by Celik et al. [3] achieves higher performance and greater flexibility than the two foregoing methods by exploiting more efficient compression technique. Difference expansion method proposed by Tian [4] achieves much higher performance than all the above methods by compressing location map of selected expandable difference values, and it is one of the best methods among the existing reversible data hiding techniques. Is there any reversible data hiding method that

* Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 573 – 580, 2005.
© Springer-Verlag Berlin Heidelberg 2005

doesn't count on lossless compression technique? Ni et al. [5] proposed such reversible data hiding by modifying image histogram in spatial domain. It achieves much high visual quality of marked image (PSNR is above 48dB) than all the foregoing methods. However, its max capacity is quite small, much smaller than those of the foregoing methods using lossless compression technique. Moreover, Ni's method needs to transmit side information by extra channel to the receiver for successful restoration.

We discover that the histogram of integer wavelet detail coefficients approximately corresponds to discrete Generalized Gaussian Distribution (GGD). In this paper, we exploit the statistical property of integer wavelet detail coefficients and propose a novel reversible data hiding for digital image by modifying integer wavelet coefficients slightly to embed data. Our algorithm doesn't need any lossless compression technique and provides several embedding modes for choice according to capacity and visual quality requirement. Moreover, our algorithm is easy to implement and doesn't need to transmit side information by extra channel. Experimental results show that our algorithm achieves much higher capacity than Ni's [5], and has similar performance to Tian's [4] which is one of the best methods in all existing reversible data hiding algorithms.

The rest of this paper is organized as follows. A brief introduction to integer wavelet transform (IWT) and statistical property of its detail coefficients is provided in Section 2. Section 3 describes embedding modes for choice to embed data. The proposed algorithm is presented in detail in Section 4. Some experimental results are presented in Section 5. The conclusion is drawn in Section 6.

2 IWT and Statistical Property of Its Coefficients

2.1 Integer Wavelet Transform

Our experimental comparison study shows that CDF 9/7 biorthogonal wavelet is more suitable for our high capacity purpose than other wavelet families, so we choose IWT of CDF 9/7 biorthogonal wavelet based on lifting scheme to realize our algorithm. An example of the lifting of CDF 9/7 biorthogonal wavelet is given in [6]. To one dimensional signal $\{x_l\}_{l\in Z}$, the lifting steps are described as follows.

$$\begin{cases} s_l^{(0)} = x_{2l} \\ d_l^{(0)} = x_{2l+1} \end{cases} \quad \begin{cases} d_l^{(1)} = d_l^{(0)} + \alpha(s_l^{(0)} + s_{l+1}^{(0)}) \\ s_l^{(1)} = s_l^{(0)} + \beta(d_l^{(1)} + d_{l-1}^{(0)}) \end{cases} \quad \begin{cases} d_l^{(2)} = d_l^{(1)} + \gamma(s_l^{(1)} + s_{l+1}^{(1)}) \\ s_l^{(2)} = s_l^{(1)} + \delta(d_l^{(2)} + d_{l-1}^{(2)}) \end{cases} \tag{1}$$

$$\begin{cases} s_l = \zeta s_l^{(2)} \\ d_l = d_l^{(2)} / \zeta \end{cases} \tag{2}$$

$$\begin{gathered} \alpha = -1.586134342;\ \beta = -0.05298011854;\ \gamma = 0.8829110762; \\ \delta = 0.4435068522;\ \zeta = 1.149604398 \end{gathered} \tag{3}$$

where s_l and d_l are generally referred to as lower frequency and detail coefficients, respectively. $s_l^{(i)}$, $d_l^{(i)}$ (i=0, 1, 2) are mid-output.

According to integer wavelet theory [7], we construct integer wavelet transform based on the framework mentioned above. That is:

$$\begin{cases} s_l^{(0)} = x_{2l} \\ d_l^{(0)} = x_{2l+1} \end{cases} \quad \begin{cases} d_l^{(1)} = d_l^{(0)} + Int(\alpha(s_l^{(0)} + s_{l+1}^{(0)})) \\ s_l^{(1)} = s_l^{(0)} + Int(\beta(d_l^{(1)} + d_{l-1}^{(0)})) \end{cases} \quad \begin{cases} d_l^{(2)} = d_l^{(1)} + Int(\gamma(s_l^{(1)} + s_{l+1}^{(1)})) \\ s_l^{(2)} = s_l^{(1)} + Int(\delta(d_l^{(2)} + d_{l-1}^{(2)})) \end{cases} \tag{4}$$

$$\begin{cases} d_l^{(3)} = d_l^{(2)} + Int((\zeta - \zeta^2)s_l^{(2)}) \\ s_l^{(3)} = s_l^{(2)} + Int((-1/\zeta)d_l^{(3)}) \end{cases} \quad \begin{cases} d_l^{(4)} = d_l^{(3)} + Int((\zeta - 1)s_l^{(3)}) \\ s_l^{(4)} = s_l^{(3)} + d_l^{(4)} \end{cases} \quad \begin{cases} s_l = s_l^{(4)} \\ d_l = d_l^{(4)} \end{cases} \tag{5}$$

where $Int(x)$ means integer part of x. The values of parameters $\alpha, \beta, \gamma, \delta, \zeta$ are given in formula (3). Equation (5) is an extra lifting step different from Equation (2). We adopt it here because it can achieve reversible transform according to [7] while Equation (2) cannot.

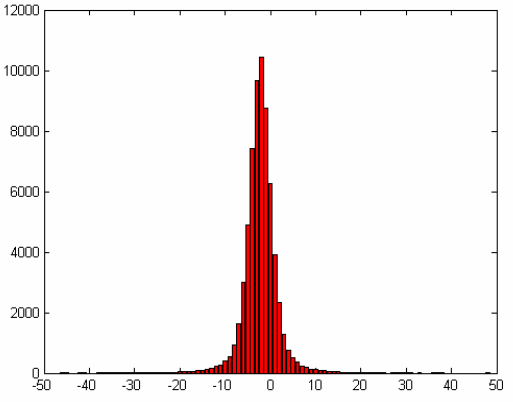


Fig. 1. Histogram of one detail subband of 512×512 grayscale image Lena

2.2 Property of Wavelet Coefficients

Experimental work [8] revealed a particular property of wavelet detail coefficients of almost every natural textured image that they have a histogram corresponding to a Generalized Gaussian Density (GGD). Our experiments on a wide range of natural images show that histogram of IWT detail subband has approximate shape of discrete Gaussian distribution function likewise. Fig. 1 shows histogram of a detail subband of 512×512 grayscale image Lena. Moreover, study on human visual system (HVS) in digital wavelet transform domain indicates that modification on wavelet coefficients under detection threshold can't be detected by human eyes [9].

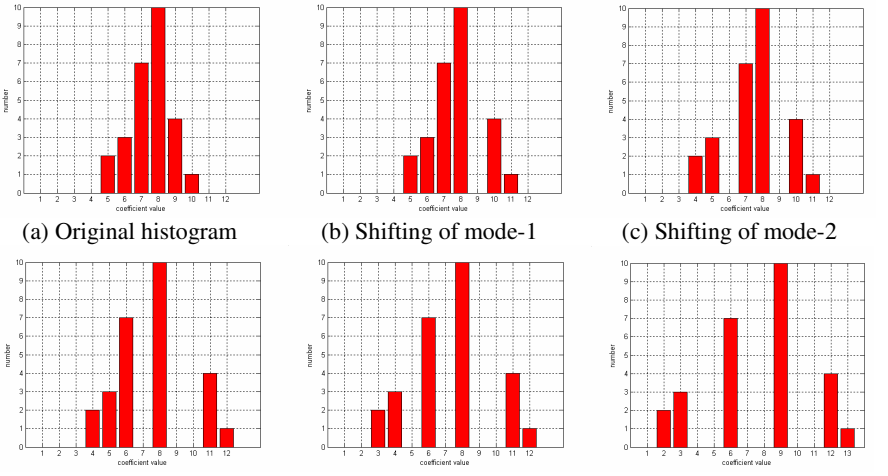
3 Embedding Modes

Our algorithm provides five embedding modes to realize reversible data hiding by modifying histogram of IWT detail coefficients. Five embedding modes can be named mode-1, mode-2, mode-3, mode-4 and mode-5. Embedding mode refers to the

shifting way of histogram of IWT detail coefficients and the moving way of coefficients with the maximum number (i.e. peak point of the histogram). The max capacity of our algorithm depends on the maximum number. In order to obtain the highest capacity, gaps adjacent to peak point are introduced in the histogram by shifting part of the histogram apart from peak point. The shifting ways of the five embedding modes are illustrated in Fig. 2 using artificial examples. Data embedding is carried out by moving the coefficient with the maximum number to the gaps when bit "1" is embedded, or keeping it unchanged when bit "0" is embedded.

Let n_{1st} equal to the maximum number in the histogram, and n_{2nd} the sub-maximum number in the histogram. The max capacity of mode-1, mode-2, mode-3, mode-4 and mode-5 is n_{1st}, $n_{1st} + n_{2nd}$, $2n_{1st}$, $2n_{1st} + n_{2nd}$ and $2(n_{1st} + n_{2nd})$ respectively.

If there are multiple peak points in the histogram, any peak point can be chosen to embed data when using mode-1 and mode-3, and two peak points can be chosen instead of a peak point and a sub-peak point when using other embedding modes, and all choices are on the condition that most of coefficients should be kept unchanged during embedding process.



(a) Original histogram (b) Shifting of mode-1 (c) Shifting of mode-2

(d) Shifting of mode-3 (e) Shifting of mode-4 (f) Shifting of mode-5

Fig. 2. Artificial examples showing histogram shifting of embedding modes

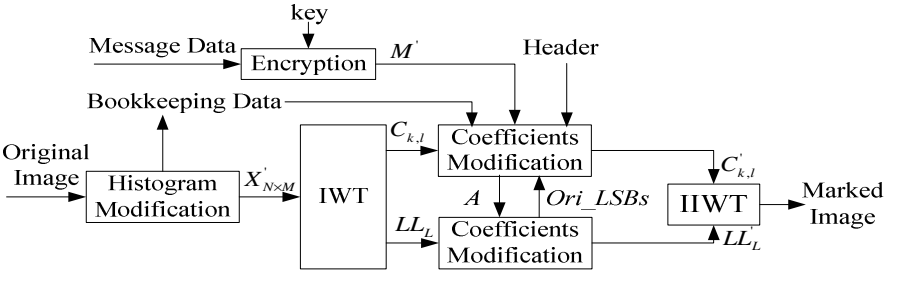


Fig. 3. The block diagram of data hiding

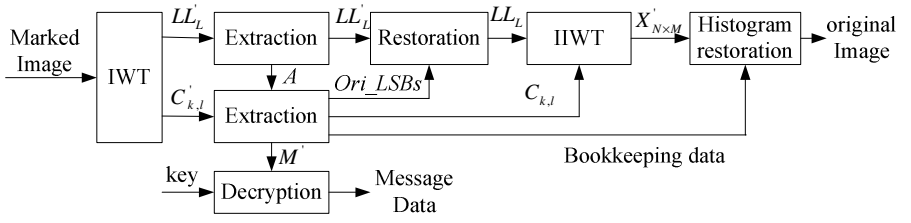


Fig. 4. The block diagram of data extraction and restoration

4 The Proposed Algorithm

To avoid pixel overflow/underflow, we adopt histogram modification in spatial domain to narrow the histogram from both sides before embedding [10]. Let $X_{N\times M}=\{x(i, j),\ 1<i<=N,\ 1<j<=M\}$ be a grayscale image of size $N\times M$. Let $C_{k,l}=\{c_{k,l}(i, j),\ \ k=h(horizontal),\ \ v(vertical),\ \ d(diagonal);\ \ l=1,2,...,L\}$ be IWT detail subband, and LL_L low frequency subband. We embed message data M into $C_{k,l}$ using some selected embedding mode according to capability and visual quality requirement, and embed assistant information used for correct decoding into LL_L by replacing LSBs (least significant bits) of LL_L coefficient. Assistant information denoted as A includes the chosen embedding mode, number of selected $C_{k,l}$, coefficient values with peak points of selected $C_{k,l}$ and histogram moving direction of selected $C_{k,l}$. In this paper, the term capacity means the bit length of M, also called payload in some literature. The embedding procedures are described as follows.

1. Pre-process the original image by carrying out histogram modification to prevent overflow/underflow and record bookkeeping data generated in this procedure.
2. Decompose the preprocessed image into L-level subband images includeing $C_{k,l}$ and LL_L by IWT.
3. Evaluate embedding capability, and decide embedding mode according to M. Select $C_{k,l}$ for embedding in the order from low level to high level, and in diagonal, vertical, horizontal according to HVS. Record peak points of selected $C_{k,l}$.
4. Form A and Embed A into LL_L by replacing LSBs of LL_L coefficients, and record the original LSBs of LL_L coefficients replaced by A as Ori_LSBs. LL_L is denoted as $LL_L^{'}$ after modifying coefficients.
5. Form bit-stream B=header $\cup$ Ori_LSBs $\cup$ bookkeeping data $\cup$ M'. M' comes from M after encryption with secret key.

6. Embed B into selected $C_{k,l}$ using chosen embedding mode. $C_{k,l}$ is denoted as $C'_{k,l}$ after modifying coefficients.
7. Combine LL'_L and $C'_{k,l}$, and perform the inverse IWT to generate a marked image.

The extraction and restoration procedures are the inverse of the embedding. Fig. 3 shows a block diagram of the proposed reversible hiding algorithm and Fig. 4 shows block diagram of the extraction and restoration part.

Table 1. Experimental results of some test images using 3-level IWT decomposition, embedding mode-1 and mode-4

Test images (512 x 512 x 8)	**Capacity(bpp)**		**PSNR(dB)**	
	mode-1	mode-4	mode-1	mode-4
Lena	0.1376	0.4090	43.20	39.44
Baboon	0.0407	0.1417	43.65	39.59
Goldhill	0.0874	0.2611	43.18	39.50
Barbara	0.1153	0.3416	43.13	39.46
Pentagon	0.0829	0.2487	43.34	39.53
Peppers	0.0947	0.2913	43.47	39.52
Airplane	0.1572	0.4622	42.92	39.39
Pills	0.0551	0.2502	43.54	39.54
Boats	0.0831	0.2602	43.29	39.50

5 Experimental Results and Analysis

We applied the proposed reversible data hiding algorithm to grayscale images that are frequently used. Table 1 contains experimental results of some test images using 3-level IWT decomposition, embedding mode-1 and mode-4. The term bpp means bits per pixel, derived from the ratio of total bit length of M to pixel number of image. For a 512 x512 image, 1 bpp means that 512x512 bits are embedded in the image. Data in the table shows that the proposed reversible data hiding algorithm can obtain high capacity while keeping high PSNR of the marked image against the original image. The max capacity of Ni's method [5] on grayscale image Lena is 0.0208 bpp, while the proposed method's is 0.1376 bpp when using 3-level IWT decomposition and embedding mode-1, which is much higher than Ni's. The visual impact of the proposed reversible data embedding on Lena can be seen from Fig. 5.

We compared the performance of the proposed algorithm with that of Celik's [3] and that of Tian's [4] on 512x512 grayscale image Lena on condition that those methods give the best performance. The capacity versus distortion (in PSNR) comparison among the three methods is shown in Fig. 6. It indicates that our algorithm outperforms Celik's, and has similar performance to Tian's at PSNR from 36dB to 42.5dB, but outperforms Tian's at high PSNR above 42.5dB.



(a) PSNR = 43.20 dB, 0.1376 bpp (b) PSNR = 39.44 dB, 0.4090 bpp

Fig. 5. The visual impact of the proposed reversible data embedding on image Lena

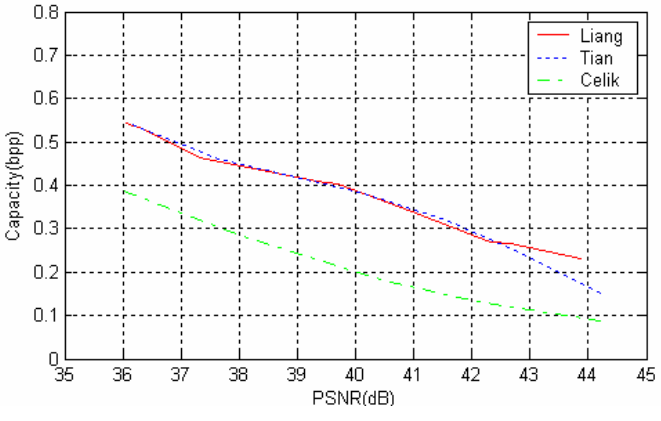


Fig. 6. Capacity versus distortion comparison of Lena of three reversible algorithms

6 Conclusion

A novel reversible data hiding algorithm for digital image by modifying integer wavelet coefficients is presented in this paper. The proposed algorithm doesn't need any lossless compression technique and can be implemented in several embedding modes. Moreover, our algorithm is easy to implement, and doesn't need to transmit side information by extra channel. Experiment results show that the proposed algorithm provides high capacity while keeping low distortion between the marked image and the original one. The proposed reversible data hiding technique can be applied in sensitive image systems, such as law enforcement, e-government, medical and military image systems.

References

1. Fridrich, J., Goljan, M., Du, R.: Invertible Authentication. In: Proc. of SPIE Photonics West, Security and Watermarking of Multimedia Contents III, vol. 3971. San Jose, California, USA (Jan 2001) 197-208
2. Goljan, M., Fridrich, J., Du, R.: Distortion-free Data Embedding. In: 4th Information Hiding Workshop, LNCS, vol. 2137. Springer-Verlag, New York, USA (2001) 27–41
3. Celik, M.U., Sharma, G., Tekalp, M.A., Saber, E.: Lossless generalized-LSB data embedding. In: IEEE Transaction on Image Processing (Feb 2005) 253-266
4. Tian, J.: Reversible Data Embedding Using a Difference Expansion. In: IEEE Transactions on Circuits and Systems for Video Technology (Aug 2003) 890-896
5. Ni, Z., Shi, Y.Q., Ansari, N., Su, W.: Reversible Data Hiding. In: Proceedings of International Symposium on Circuits and Systems (ISCAS 2003), vol. 2. Bangkok, Thailand (25-28 May 2003) 912–915
6. Daubechies, I., Sweldens, W.: Factoring wavelet transform into lifting step. In: Journal of Fourier Analysis, vol. 4 (1998) 245-267
7. Calderbank, R., Daubechies, I., Sweldens, Yeo, W., B. L.: Wavelet transforms that map integers to integers. In: Journal of Applied and Computational Harmonic Analysis, vol. 5 (1998) 332-369
8. Mallat, S.: A theory of multiresolution signal decomposition: the wavelet representation. In: IEEE Transactions Pattern Analysis and Machine Intelligence (1989) 674-693
9. Watson, A. B., Yang, G. Y., Solomon, J. A., Villasenor, J.: Visibility of wavelet quantization noise. In: IEEE transactions on image processing, vol.6 (1997) 1164 – 1175
10. Xuan, G., Zhu, J., Chen, J., Shi, Y. Q., Ni, Z., Su, W.: Distortionless Data Hiding Bsased on Integer Wavelet Transform. In: IEE Electronics Letters, December (2002) 1646-1648

An Image Steganography Using Pixel Characteristics

Young-Ran Park[1], Hyun-Ho Kang[2], Sang-Uk Shin[1], and Ki-Ryong Kwon[3]

[1] Department of Information Security, Pukyong National University, 599-1 Daeyeon-3Dong, Nam-Gu, Busan 608-737, Republic of Korea
young@shannon.pknu.ac.kr, shinsu@pknu.ac.kr
[2] Graduate School of Information Systems, University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, Japan
kang@ice.uec.ac.jp
[3] Department of Electronic and Computer Engineering, Pusan University of Foreign Studies, 55-1 Uam-Dong, Nam-Gu, Busan 608-738, Republic of Korea
krkwon@pufs.ac.kr

Abstract. This paper presents a steganographic algorithm in digital images to embed a hidden message into cover images. This method is able to provide a high quality stego image in spite of the high capacity of the concealed information. In our method, the number of insertion bits into each pixel is different according to each pixel's characteristics. That is, the number of insertion bit is dependent on whether the pixel is an edge area or smooth area. We experimented on various images to demonstrate the effectiveness of the proposed method.

1 Introduction

Steganography enables us to make communication invisible by hiding secret information in innocuous cover objects. Image steganography is a secret communication method used to transmit secret messages that have been embedded into an image. To accommodate a secret message in a digital image, the original cover image is slightly modified by the embedding algorithm. As a result, a stego image is obtained.

Many steganographic algorithms have been proposed to hide secret information in digital images. The most common method is the LSBs (least-significant-bits) method, which utilizes some least bits of pixels in the cover image to embed the hidden data. The advantages of the LSBs method are ease of computation and a large payload of data that can be embedded in the cover image with high visual quality. The LSBs-based scheme is often employed for these reasons [1]-[4].

While human perception is less sensitive to subtle changes in edge areas of a pixel, it is more sensitive to changes in the smooth areas. Several methods exploiting these human visual characteristics have been studied recently [5]-[7]. Our proposed method improves on the Chang *et al.* technique [7] by increasing capacity, and we used Thien *et al.* algorithm [8] to enhance the quality of the stego images. The results have shown that our method increases the amount of secret data that can be stored in a cover image, while at the same time producing a higher quality of stego image. To increase overall capacity our procedure determines the number of bits to be inserted into a target pixel by using the largest difference value between the other four pixels close to the target pixel. When inserting secret data into a target pixel, we used the modular

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 581 – 588, 2005.
© Springer-Verlag Berlin Heidelberg 2005

operation to improve the quality of stego image. We experimented with various images to evaluate the efficiency of the proposed method, and as a result, we found that our method was able to insert much more information into an image than existing methods. Besides this, as stated above, the quality of the image was not degraded.

The structure of the paper is as follows: In Section 2 we present our steganographic algorithm for gray-level images. In Section 3, the results of experiments show that our proposed method is feasible. Finally, we will draw our conclusions in Section 4.

2 Proposed Method

Our system refers to four pixels adjacent to a target pixel in the embedding process.

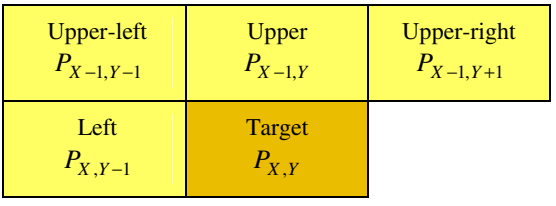


Fig. 1. A target pixel and four neighboring pixels finished insertion process

2.1 The Insertion of Hidden Information

Our method refers to the four neighboring pixels that have already finished insertion process to embed the secret message into the target pixel (refer to Fig. 1). In the cover image: Given target pixel $P_{X,Y}$ with gray value $g_{x,y}$, let $g_{x-1,y-1}$, $g_{x-1,y}$, $g_{x-1,y+1}$, and $g_{x,y-1}$ be the gray values of its upper-left $P_{X-1,Y-1}$ pixel, upper $P_{X-1,Y}$ pixel, upper-right $P_{X-1,Y+1}$ pixel, and left $P_{X,Y-1}$ pixel, respectively. The embedding procedure is as following:

Step 1. Select the maximum and the minimum values among the four pixel values that have already finished the embedding process. Calculate the difference value d between the maximum pixel value and the minimum pixel value using the following formula:

$$d = g_{\max} - g_{\min} . \tag{1}$$

Where: $g_{\max} = \max(g_{x-1,y-1}, g_{x-1,y}, g_{x-1,y+1}, g_{x,y-1})$ and

$g_{\min} = \min(g_{x-1,y-1}, g_{x-1,y}, g_{x-1,y+1}, g_{x,y-1})$.

Using equation (1), we judge whether the target pixel is included in an edge area or a smooth area. This is why the number of bit n, inserted into the target pixel is determined by value d. Also, because we are using the upper-left, upper, upper-right and left pixels that were already handled when calculating difference d the embedding and extracting process has the same value d, making it more accurate.

Step 2. Calculate n: the number of the insertion bits in a target pixel $P_{X,Y}$ using the following formula:

$$n = \lfloor \log_2 d \rfloor - 1, \quad \text{if } d > 3. \tag{2}$$

If the value of d is less or equal to 3, n in $P_{X,Y}$ is determined to be 1, otherwise value of n is taken from the result calculated using equation (2). We appropriately adjust n to enhance both the capacity and the imperceptibility within the cover image.

Step 3. Calculate a temporary value $t_{x,y}$ using:

$$t_{x,y} = b - (g_{x,y} \bmod 2^n). \tag{3}$$

Where b is the decimal representation of the hidden message using n bits.

Step 4. To make the quality of the image higher, we select a value nearest to the target pixel's value within the cover image using:

$$t'_{x,y} = \begin{cases} t_{x,y}, & \text{if } (-\lfloor (2^n - 1)/2 \rfloor) \le t_{x,y} \le \lceil (2^n - 1)/2 \rceil; \\ t_{x,y} + 2^n, & \text{if } (-2^n + 1) \le t_{x,y} < (-\lfloor (2^n - 1)/2 \rfloor); \\ t_{x,y} - 2^n, & \text{if } \lceil (2^n - 1)/2 \rceil < t_{x,y} < 2^n. \end{cases} \tag{4}$$

Step 5. Finally, we calculate the new pixel value $g^*_{x,y}$ for $P_{X,Y}$ using:

$$g^*_{x,y} = g_{x,y} + t'_{x,y}. \tag{5}$$

The search for the best $b + (q \cdot 2^n)$, in fact, can be reduced to the search for the best $g^*_{x,y} = g_{x,y} - (g_{x,y} \bmod 2^n) + b + (q \cdot 2^n)$, where $q \in \{0, 1, -1\}$, so that $g^*_{x,y}$ is the nearest to $g_{x,y}$. Equation (4) helps to determine this q automatically, because equation (3) and (4) imply $g^*_{x,y} = g_{x,y} + t'_{x,y} = g_{x,y} + t_{x,y} + (q \cdot 2^n)$ $= g_{x,y} - (g_{x,y} \bmod 2^n) + b + (q \cdot 2^n)$ with q chosen from $\{0, 1, -1\}$. The effectiveness of equations (3) and (4) has already been proved in Thien *et al.* method. Please refer to Thien *et al.* method [8], and Chan *et al.* method [9] for a more detailed proof.

When the new value of pixel $P_{X,Y}$ is outside the necessary range, it has to be adjusted using the following formula:

$$g^*_{x,y} = \begin{cases} g^*_{x,y} + 2^n, & \text{if } g^*_{x,y} < 0; \\ g^*_{x,y} - 2^n, & \text{if } g^*_{x,y} > 255. \end{cases} \tag{6}$$

Thus we solved the problem that may occur when pixels exceed the boundary from 0 to 255.

The stego image can be created after performing the insertion process on all of the pixels except the first row and column in the cover image. The hidden message is

embedded into the cover image using a raster scan, except for the pixels of the first row and the first column because the upper or the left pixels no longer exist.

2.2 The Extraction of Secret Information

The extracting speed of our method is faster than the Chang *et al.* method because our algorithm is much simpler. In the stego image, given a target pixel $P^*_{X,Y}$ with gray value $g^*_{x,y}$, let $g^*_{x-1,y-1}$, $g^*_{x-1,y}$, $g^*_{x-1,y+1}$ and $g^*_{x,y-1}$ be the gray values of its upper-left $P^*_{X-1,Y-1}$ pixel, upper $P^*_{X-1,Y}$ pixel, upper-right $P^*_{X-1,Y+1}$ pixel, and left pixel $P^*_{X,Y-1}$ respectively. Then the insertion procedure is performed using the following steps:

Step 1. Calculate the difference value d^* between the maximum pixel value and the minimum pixel value among the upper-left pixel $P^*_{X-1,Y-1}$, the upper pixel $P^*_{X-1,Y}$, the upper-right pixel $P^*_{X-1,Y+1}$, and the left pixel $P^*_{X,Y-1}$ in stego image using the following formula:

$$d^* = g^*_{\max} - g^*_{\min}. \tag{7}$$

Where: $g^*_{\max} = \max(g^*_{x-1,y-1},\ g^*_{x-1,y},\ g^*_{x-1,y+1},\ g^*_{x,y-1})$ and

$g^*_{\min} = \min(g^*_{x-1,y-1},\ g^*_{x-1,y},\ g^*_{x-1,y+1},\ g^*_{x,y-1})$.

The d^* equal d, because values $g^*_{\max}$ and $g^*_{\min}$ in equation (7) are equal to the values $g_{\max}$ and $g_{\min}$ in equation (1) of the embedding procedure respectively. So, the hidden data can be extracted precisely.

Step 2. Calculate the value of n as the number of bits inserted into pixel $P^*_{X,Y}$ using:

$$n = \lfloor \log_2 d^* \rfloor - 1, \quad \text{if } d^* > 3. \tag{8}$$

In equation (8), if the value of d^* is less than or equal to 3, the value of n is 1, otherwise n is the value calculated in equation (8) as in the embedding procedure.

Step 3. Calculating the value of b using:

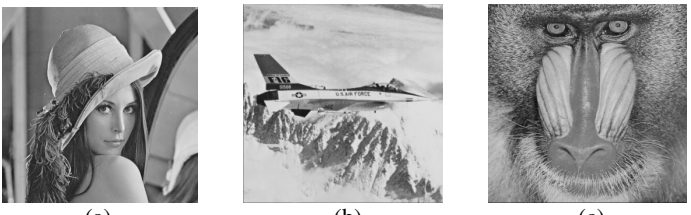
$$b = \mathrm{mod}\,(g^*_{x,y},\ 2^n). \tag{9}$$

The decimal value b is represented into the binary of n bits. We can see the secret message after performing the extraction processing.

3 Experimental Results

This section presents the results of experiments and an analysis of those results. We compared our procedure with the Chang *et al.* system that the number of insertion bits per pixel was different. However we did not compare our results with the Thien *et al.* method because the number of insertion bits per each pixel equals.

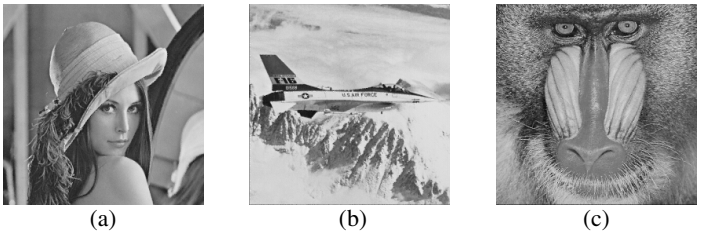
Several experiments were performed to evaluate our method. Test images applied to the experiment are gray-scale images with size of 256 by 256 as shown in Figure 2.



(a) (b) (c)

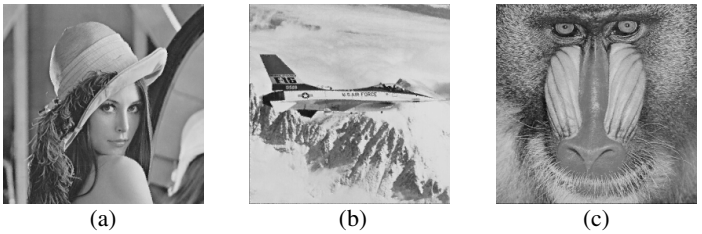
Fig. 2. Cover images with a size of 256 by 256 pixels: (a) Lena, (b) Airplane, and (c) Baboon

Figure 3 shows the stego images obtained using the Chang *et al.* method, and Figure 4 indicates the stego images obtained using our method. When comparing cover images to stego images, it is difficult to distinguish visually between the two images. However, in the computational quality of stego images our scheme is superior to the Chang *et al.* method. In addition, our method is able to insert many more secret bits into the cover image than the Chang *et al.* method. We compared Chang *et al.* method with our proposed method using the amount of insertion and PSNR respectively. As a result, our scheme provides a high quality of stego images, has a large message capacity and has more embedded bits in a pixel.



(a) (b) (c)

Fig. 3. Stego images obtained using the Chang *et al.* method: (a) Lena image with hidden information amounting to 105,493 bits and PSNR of 37.15 dB, (b) Airplane image with 109,673 bits and PSNR of 34.97 dB, and (c) Baboon image with 178,662 bits and PSNR of 31.96 dB



(a) (b) (c)

Fig. 4. Stego images of the proposed method: (a) Lena image with hidden information amounting to 138,839 bits and PSNR of 37.90 dB, (b) Airplane image with 132,960 bits and PSNR of 36.93 dB, and (c) Baboon image with 209,353 bits and PSNR of 35.10 dB

Although the amount of bits inserted is larger, the PSNR is high. This is made possible by using the property of the modular operation. When changing the value of a target pixel in our scheme to insert the secret message, we get a better quality of the stego images because our method selects the value nearest to that of the original pixel.

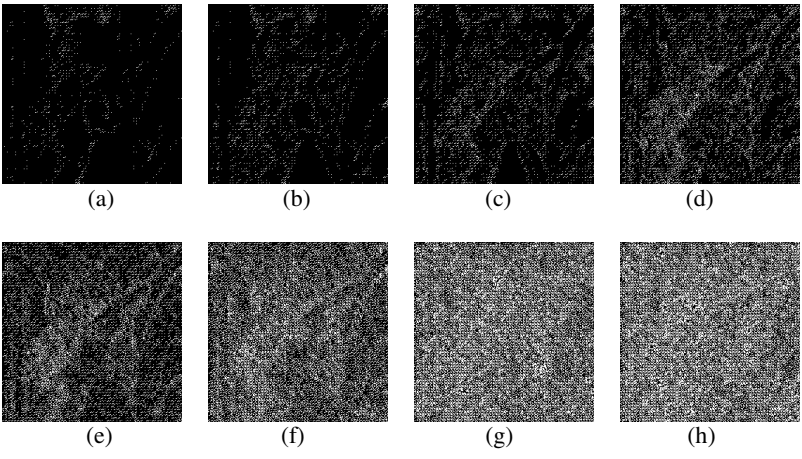


Fig. 5. The difference between the bit-plane of cover image and the bit-plane of stego image using the Chang *et al.* method: (a) MSB (Most Significant Bit) plane, (b) 7^{th} LSB plane, (c) 6^{th} LSB plane, (d) 5^{th} LSB plane, (e) 4^{th} LSB plane, (f) 3^{rd} LSB plane, (g) 2^{nd} LSB plane, and (h) LSB (Least Significant Bit) plane

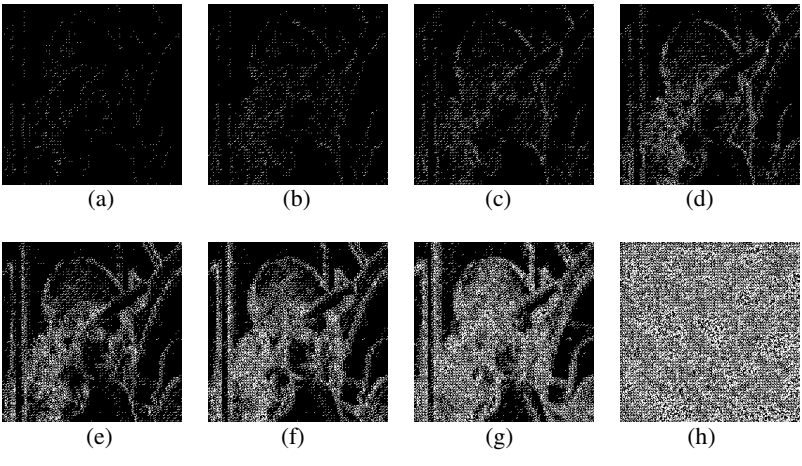
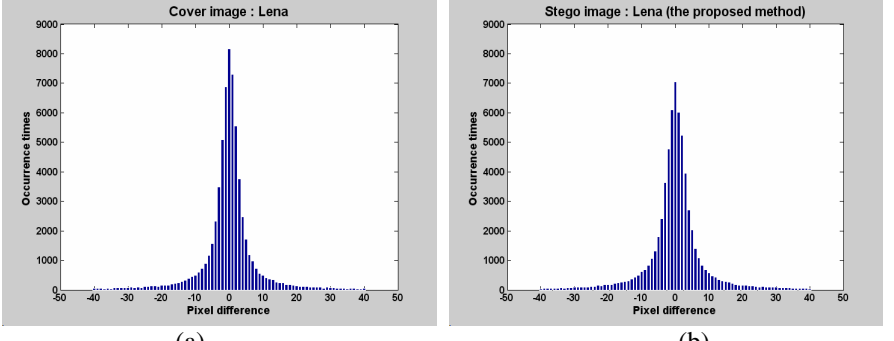


Fig. 6. The difference between the bit-plane of cover image and the bit-plane of stego image using our proposed method: (a) MSB (Most Significant Bit) plane, (b) 7^{th} LSB plane, (c) 6^{th} LSB plane, (d) 5^{th} LSB plane, (e) 4^{th} LSB plane, (f) 3^{rd} LSB plane, (g) 2^{nd} LSB plane, and (h) LSB (Least Significant Bit) plane

Figure 5 and Figure 6 show difference between the bit-planes of cover image and the bit-planes of stego images using the Chang *et al.* method and our method. In Figure 6, we can see that our scheme embedded hidden information into pixels, including the exact edge area. That is, our scheme embeds more bits into the edge areas than the smooth areas.

Figure 7 shows a histogram charting the difference between consecutive pixels of the cover image and our stego image. The histogram of our stego image is very similar to the histogram of cover image. Figure 7 shows the fact that our method is secure against attacks in image processing.



(a) (b)

Fig. 7. Histogram showing the difference between consecutive pixels: (a) cover image and (b) stego image by the proposed method

4 Conclusions

In this paper, we presented a method of image steganography using the difference of value between pixels neighboring the target pixel. In our method, the number of insertion bits is dependent on each pixel according to whether the pixel is an edge area or smooth area. As we can see from Figure 6, our method has judged the edge area and smooth area more accurately for each pixel and chosen the proper number of the bits to be embedded therein. Also, our processing time is short because our embedding and extracting algorithm is simple. Our scheme can improve the quality of stego images by using the modular operation and the capacity of bit rates to be embedded. We have shown that our system is more effective than other methods through conducting experiments utilizing capacity and imperceptibility.

Acknowledgements

This research was supported by the Program for the raining Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce, Industry and Energy of the Korea Government.

References

1. W.N. Lie and L.C. Chang: Data Hiding in Image with Adaptive Numbers of Least Significant Bit Based on the Human Visual System", International Conference on Image Processing, IEEE, Vol. 1, (1999) 286-290
2. R.Z. Wang and C.F. Lin: Image Hiding by Optimal LSB Substitution and Genetic Algorithm, Pattern Recognition, ELSEVIER, Vol. 34, (2001) 671-883,.
3. H. Noda, J. Spaulding, M.N. Shirazi, M. Niimi and E. Kawaguchi: BPCS Steganography Combined with JPEG2000 Compressio, Proceedings of Pacific Rim Workshop on Digital Steganography (STEG), July (2002) 98-107.
4. T. Zhang and X. Ping: A New Approach to Reliable Detection of LSB Steganoraphy in Natural Image, Signal Processing Journal, ELSEVIER, Vol. 83 May (2003) 2085-2093,.
5. D.C. Wu and W.H. Tsai: A Steganographic Method for Images by Pixel-value Differencing, Pattern Recognition Letters, ELSEVIER, Vol. 24 (2003) 1613-1626,.
6. X. Zhang and S. Wang: Vulnerability of pixel-value differencing steganography to histogram analysis and modification for enhanced security, Pattern Recognition Letters, ELSEVIER, Vol. 25(2004) 331-339,.
7. C.C. Chang and H.W. Tseng: A Steganographic Method for Digital Images Using Side Match, Pattern Recognition Letters, ELSEVIER, Vol. 25, June (2004) 1431-1437,.
8. C.C. Thien and J.C. Lin: A Simple and High-Hiding Capacity Method for Hiding Digit-by-Digit Data in Images Based on Modulus Function, The Journal of The Pattern Recognition Society, PERGAMON, Vol. 36,June (2003) 2875-2881,.
9. C.K. Chan and L.M. Cheng: Hiding Data in Images by Simple LSB Substitution, The Journal of The Pattern Recognition Society, PERGAMON, Vol. 37, (2004) 469-474.

Alternatives for Multimedia Messaging System Steganography

Konstantinos Papapanagiotou[1], Emmanouel Kellinis[2], Giannis F. Marias[1], and Panagiotis Georgiadis[1]

[1] Dept. of Informatics and Telecommunications, University of Athens, Panepistimiopolis, Ilissia, Greece, GR15784
conpap@di.uoa.gr, marias@mm.di.uoa.gr, georgiad@di.uoa.gr
[2] KPMG LLP, One Canada Square, London E14 5AG, United Kingdom
emmanouel.kellinis@kpmg.co.uk

Abstract. The Multimedia Messaging System allows a user of a mobile phone to send messages containing multimedia objects, such as images, audio or video clips. MMS has very quickly gained the popularity of SMS among mobile users. Alongside, the need for a secure communication became more imperative. Hiding information, especially in images has been an attractive solution for secret communication. In this paper we examine the possibilities for the use of steganography within a multimedia message. The most widely known algorithms for steganography are presented and discussed. Their application in a mobile environment is analyzed and a theoretical evaluation is given.

1 Introduction

One of the most popular uses of mobile phones has been the exchange of messages between users. The Short Messaging System (SMS) was introduced with GSM mobile phones and it very rapidly became popular among users. The Multimedia Messaging System (MMS) offers the ability to send and receive multimedia content using a mobile phone. Nowadays, most of the mobile phones not only are capable of sending and receiving Multimedia Messages (MM), but also contain an embedded camera and can run customized applications (e.g. using Java 2 Platform Micro Edition, J2ME).

Most research regarding security in mobile environments with limited resources in terms of processing power, memory capacity and energy autonomy, is focused on the implementation of symmetric and asymmetric cryptographic algorithms. Steganography differs from cryptography in the sense that it tries to hide the message instead of transforming it so as to obscure its meaning [17]. In some cases, steganography may actually prove to be more effective. The combination of both may give the best results, as a message can be encrypted before it is hidden into another object. Cryptography has received most attention until now, leaving a great space for research on steganography. Steganography concerns itself with ways of embedding a secret message into a cover object, without altering the properties of the cover object evidently. The embedding procedure is typically related with a key, usually called a stego-key. Without knowledge of this key it will be difficult for a third party to extract the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 589 – 596, 2005.
© Springer-Verlag Berlin Heidelberg 2005

message or even detect its existence. Once the cover object has data embedded in it, it is called a stego object. The amount of data that can be hidden in a cover object is often referred to as embedding capacity. The embedding capacity is directly related with the secrecy of the message. The distortions in the cover object caused by the steganographic algorithm become more obvious as a user tries to add more hidden data. Evidently, there is a point of balance when the embedded data do not alter the cover object significantly enough to arouse suspicion.

In this paper we examine how steganography can be used in the context of MMS. To the best of our knowledge, there are only very few known implementations of steganographic algorithms for mobile devices [15]. We present some widely used algorithms for steganography and explain how they can be applied in MMS. Users can benefit from covert channels in MMS in order to secretly exchange hidden messages and keys, without arousing suspicion of their existence. It could also be used as a mean for hiding the exchange of one-time passwords between a corporate server and a mobile worker [25]. An evaluation of steganographic techniques is also required, as, to the best of our knowledge, there has not been any benchmarking of steganographic algorithms. We propose assessment metrics that can provide us with a sufficient view of their performance. The structure of this paper is as follows: first we examine the architecture of the MMS and the supported media formats. Subsequently, we present the steganographic algorithms that can be used in the MMS. These algorithms can be applied to image, audio or video files, but also to presentation information. In section 4 the presented steganographic techniques are evaluated and some examples of their applications are given. Finally, we provide our concluding remarks and give suggestions for future research.

2 MMS Technology

MMS [5] is a technology that allows a user of a properly enabled mobile phone to create, send, receive and store messages that include text, images, audio and video clips. Figure 1 shows the architecture of the MMS network. A user interacts with an MMS Client which is usually an application in his mobile phone. The Client uses the MMS_M interface in order to communicate with the MMS Proxy-Relay, which is responsible for communication with other messaging systems. Proxy-Relays use the MMS_R interface to communicate among themselves. They also communicate with the MMS Server, which stores messages, through the MMS_S interface. In a typical protocol run, a user will compose a MM and select the address of the receiver. The MMS Client will submit the message to the associated MMS Proxy-Relay, which will forward it to the target Proxy-Relay. The MM will be stored in the associated MMS Server and a notification will be sent to the target MMS Client. Finally, the receiver instructs its MMS Client to retrieve the MM from the Server.

MMS supports a variety of media formats [7]. Concerning audio at least two codecs are supported: aacPlus (Advanced Audio Coding Plus) and AMR-WB (Adaptive Multi-Rate Wideband). Support for MP3, MIDI and WAV formats is also suggested [8]. For still images JPEG with JFIF should be supported by MMS Clients while for bitmap graphics the supported formats are: GIF87a, GIF89a and PNG. Finally, the supported media formats for video are: H.263 Profile 3 Level 45, MPEG-4 Visual

Simple Profile Level 0b and H.264 (AVC) Baseline Profile Level 1b. Moreover, there are some constraints for the size of media [6]: text cannot exceed 30kB, an image must be below 100kB and a video must be smaller than 300kB.

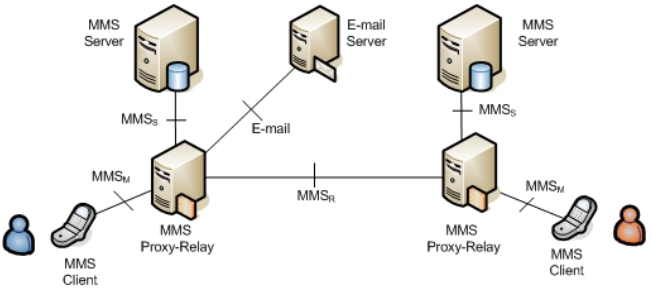


Fig. 1. MMS network architecture

A single MM message may contain many different media elements. An MMS Client should be able to render the multimedia content in a meaningful, for the user, way. Various presentation languages are supported for this purpose, like the Wireless Markup Language (WML) and the Synchronized Multimedia Integration Language (SMIL) [5]. MMS presentation information is transferred along with the multimedia objects. SMIL is a simple XML-based language and currently the most widely used language for MMS presentation.

3 MMS Steganography

MMS-capable devices can send and receive messages containing image, audio or video. Such multimedia objects may contain hidden information, embedded to them using steganographic techniques. To the best of our knowledge, currently there are no widely deployed tools that perform stego-functions. Bond Wireless has created a tool [14] for sending a secure SMS, which can also embed hidden messages inside an image within an MMS. However, only image steganography is supported, even though MMS also supports audio and video. In this section we will present algorithms that can be used for embedding hidden information in a MM.

3.1 Image Audio and Video Steganography

Digital images are the most widely used cover-objects for steganography. Despite the use of compression in many image formats a high degree of redundancy characterizes a digital representation of an image. The most widely known algorithm for image steganography, the LSB algorithm, involves the modification of the LSB layer of images. In this technique, the message is stored in the LSB of the pixels which could be considered as random noise. Thus, altering them does not have any obvious effect to the image [1]. Variations of this algorithm include changing two or three LSBs [4], or adding a weak noise signal [2] in order to make detection harder. One of the first tools to implement the LSB algorithm is Steganos [12]. Masking techniques take advantage of the properties of the human visual system [11], as an observer cannot

detect a signal in the presence of another signal. Algorithms that use this technique analyze an image in order to detect regions in it where a message can be placed [11].

Most research in the category of transform domain embedding algorithms is focused on taking advantage of redundancies in Discrete Cosine Transform (DCT). DCT is mainly used in the JPEG format in order to compress images. Information can be embedded in a JPEG image by altering DCT coefficients, for example by changing their LSB. Even though changing a large number of coefficients does not produce any visible alterations, it has a significant effect in compression rate and it may also produce statistical anomalies. Thus, the embedding capacity of the DCT algorithm is much smaller than LSB's [1]. An algorithms that use DCT is F5 [9], a successor to jsteg. Another transform domain which has been used for embedding information is the frequency domain. Alturki et al. [10] propose quantizing the coefficients in the frequency domain using Fourier transformation in order to embed information which will appear as Gaussian noise. Similarly, the wavelet domain is also suitable for embedding information. The signal is decomposed into a low-pass and a high-pass component using a Discrete Wavelet Transform (DWT) and data is embedded, for example, in the LSB of the wavelet coefficients [13].

Audio steganography can be applied to most audio formats and can be very effective on high compression audio algorithms such as MP3 [21]. The LSB algorithm for audio steganography uses a subset of available audio samples and alters their LSB. However, as the number of the adjusted LSBs increases the introduced noise becomes more audible [13]. Masking techniques can be used for audio steganography in the same concept as with image steganography. A faint sound becomes inaudible in the presence of a louder sound. Similarly, if two signals are close in the frequency domain, the weaker one will be inaudible [15]. A widely known method to hide information inside audio is echo hiding [16], which adds echo to a host audio. The delay between the original signal and the echo depends on the echo kernel that is being used [16]. Cox et al. [19] have proposed the spread spectrum method where the data is spread throughout the maximal range of frequencies so that it becomes very difficult for someone to find the hidden data unless they know the spreading algorithm. Mp3stego [21] is one of the most widely known tools for hiding text in MP3 audio files. In mp3stego the data to be hidden are embedded in the MP3 bit stream as parity during the encoding process. The properties of parity ensure that more data can be hidden if more values are selected as a cover message [17]. Similar methods can be used for other compressed audio formats such as AMR and AAC. However, to the best of our knowledge there are only few proposals for information hiding algorithms specifically designed for AAC-encoded audio [22].

The techniques used to hide information inside video files are not very different from the methods used in image steganography. There are some obvious differences though in the way that data is stored. Someone can store more information inside a video file than in an image file since video is a sequence of images (frames). Therefore in a 30 frames/sec video we can store 30 times the information we could store in a single picture. Moreover, someone can pick frames to store data in such a way that even if someone could detect the existence of modification, he would not be able to put the extracted information in sequence. A technique for embedding data in an MPEG compressed video clip is analyzed in [18].

3.2 Markup Language Steganography

Mark-up languages (ML) such as HTML, XML and SMIL contain presentation (i.e., markup) information which consists of tags. Tag attributes can be reordered freely without causing any visual change [24]. Additionally, if a parser does not recognize an attribute it ignores it. By exploiting these steganography opportunities, someone can hide information inside a document. For example, attribute reordering can be used to embed information into a document. In the following example of SMIL code we have the <region> tag which has five attributes:

```
<region top="0" id="Image" left="0" height="144"
width="176"  />
```

By rearranging them we get 5! possible combinations of these five attributes, i.e. 120 different ways to display the same thing. Using this line alone we can hide a total of $\log_2 120 = 6.9$ bits. Since we don't want to arouse suspicion by changing all attributes' position for each tag, we can split the number of permutations throughout the document in order to achieve increased robustness and a larger storage space. In addition we can introduce a white space after the tag name [24]. Using the same SMIL code we can create hidden binary information by doing the following for binary 0 and 1:

```
<region id="Image" width="176" height="144" left="0"
top="0" />
<region  id="Image" width="176" height="144" left="0"
top="0" />
```

Many tags such as <Image>, <text> or <audio> can also include the end-tag inside the start tag or they can split into two separate tags [24]. For example for binary 0 and 1:

```
<img src = "ThirdImage.jpg" region="Image"/>
<img src = "ThirdImage.jpg" region="Image"></image>
```

Finally someone can hide information inside a SMIL document using simple text based steganography [24], by introducing extra spaces at the end and start of each line or by appending at the end of lines a TAB (binary 0) or SPACE (binary 1) character.

4 Evaluation and Applications of MMS Steganography Methods

Most research concerning benchmarking steganographic algorithms is focused on techniques for steganalysis [23]. However, the time required to embed or extract a hidden message is critical in limited environments like mobile phones. Thus, we need to adopt a parameter, which we will call Speed, that will reflect the time required to embed a certain amount of information to a specific cover object using a steganographic algorithm. The embedding capacity for each algorithm should be examined along with the secrecy level offered, given the message size restrictions that each MMS message sets. We use two different parameters in order to express these aspects: Capacity for the embedding capacity and Secrecy for how easily suspicions are aroused for the given Capacity. We establish here a metric that would give a measurement for the performance of a steganographic algorithm as follows:

$$StegoPerformance = f(Speed, Capacity, Secrecy) \quad (1)$$

Regarding speed, LSB substitution techniques are considered to be the least complex. Methods involving transformations, such as DCT, FFT and DWT, or masking require increased processing power and produce an additional overhead. However, LSB can only be applied to uncompressed data. Practically, images in MMS are sent in JPEG format. Therefore the use of DCT or DWT-based methods for image steganography is suggested. Additional analysis is also required for echo hiding or spread spectrum techniques. Contrarily, SMIL and generally markup language steganography is considered to be very efficient as it only involves text manipulation. As far as embedding capacity and secrecy are concerned, LSB algorithms give the possibility of hiding a large amount of data. Nevertheless, as the number of the substituted LSBs increases, it becomes more and more apparent that steganography may have been used [4]. Conversely, transform embedding techniques are used in applications where robustness is required [11]. The fact that hidden data is inserted in the transform domain makes the message much harder to detect but also limits the embedding capacity of such algorithms. SMIL steganography provides very little room for embedding data. A SMIL file in a MM can be very small and thus, it can be hard to embed a large amount of data into it. Also, text-based steganography lacks secrecy as an observer may easily detect an abnormality in plain text. However, in our case a user will never see a SMIL file as it only provides information so that a mobile device can render a MM. Table 1 summarizes the advantages and disadvantages of each algorithm.

Table 1. Evaluation of Steganographic Algorithms

Algorithm	Object	Speed	Capacity	Secrecy	Perform.
LSB	Image	☺	☹	☹	☹
DCT-DWT	Image	😐	😐	☺	😐
LSB	Video	😐	😐	😐	😐
DCT-DWT	Video	☹	😐	☺	😐
LSB	Audio	☺	😐	☹	😐
Echo	Audio	😐	😐	😐	😐
Spread Spectrum	Audio	☹	😐	☺	😐
Attribute Syntax	ML	☺	😐	☺	☺
White Space	ML	☺	☺	☺	☺

A user could take advantage of the fact that SMIL steganography can be undetectable in order to embed a secret message in the presentation part of a MM. For example, one could embed hidden information in image, audio or video clips using a stego key and then hide this key in the presentation part of the MM using SMIL steganography. Even if a malicious user could realize that hidden information is contained in a multimedia clip contained in a MM, it would be hard to also find the key in order to retrieve this information. Thus, SMIL steganography can be used to hide information on how to extract data hidden in other parts of a MM. Steganography in MM can also be used for various other purposes. In terms of digital rights management, author and copyright information could be hidden in media that are transmitted by MMS (for example, ringtones, wallpapers, etc.). Origin authentication can also be enforced, by secretly embedding information regarding the identity of a sender in order to counter

spoofing attempts. Nowadays, mobile message marketing is becoming more and more popular. For example, premium services are advertised using SMS or even MMS. A user has no way to know if such messages are legitimate and if the advertised cost is real. Steganography could aid in this direction by embedding data (for example a hash) that could verify the validity and integrity of such a message.

5 Conclusions

In this paper we presented and discussed the possibility for using steganography in the MMS. We examined various, widely-used algorithms that could be applied in a mobile environment, in order to hide information in images, audio or video clips. We also suggested the use of SMIL steganography in order to embed a small message (e.g. the stego-key) in the presentation part of the MM. We evaluated the presented algorithms on a theoretical basis, taking into account the fact that they will be applied in a restricted environment, with limited processing power, memory and energy autonomy. For the evaluation of such algorithms, specific metrics were proposed, that can provide a complete view of their performance. Finally, we provided some suggestions for possible uses of steganography in MMS. Currently, we are experimenting with a practical implementation of steganographic algorithms. A real world application, written in J2ME, will enable us to perform thorough tests in order to provide further suggestions for the use of the presented algorithms in mobile devices. Combined use of steganography and cryptography is also examined for optimal results.

Acknowledgements

Part of this work was performed in the context of the project entitled "PERAS: PERvasive and Ad hoc Security" funded by the Greek Ministry of Development, General Secretariat for Research and Technology, under the framework "PENED".

References

1. M. Kharrazi, H. T. Sencar, N. Memon: “Image Steganography: Concepts and Practice”, National University of Singapore (2004)
2. J. Fridrich and M. Goljan: “Digital image steganography using stochastic modulation" SPIE Symposium on Electronic Imaging, San Jose, CA (2003)
3. S. Katzenbeisser, F. A. P. Petitcolas: “Information Hiding Techniques for Steganography and Digital Watermarking”, Artech House (2000)
4. K. Curran, K. Bailey: “An evaluation of image-based steganography methods”. International Journal of Digital Evidence. (2003)
5. Open Mobile Alliance: “Multimedia Messaging Service Architecture Overview Approved Version 1.2”, OMA, March (2005)
6. Open Mobile Alliance: “MMS Conformance Document 1.2 Approved Version 1.2”, OMA, March (2005)
7. 3rd Generation Partnership Project, “Multimedia Messaging Service (MMS): Media formats and codecs (Release 6)”, 3GPP, March (2005)

8. 3rd Generation Partnership Project, “Multimedia Messaging Service (MMS): Functional description; Stage 2 (Release 1999)”, 3GPP, June (2002.)
9. A. Westfeld: “F5a steganographic algorithm: High capacity despite better steganalysis," 4th International Workshop on Information Hiding. (2001)
10. F. Alturki, R. Mersereau: “Secure blind image steganographic technique using discrete fourier transformation," IEEE Intl. Conference on Image Processing, Greece (2001)
11. E. T. Lin and E. J. Delp: "A Review of Data Hiding in Digital Images," Proc. of the Image Processing, Image Quality, Image Capture Systems Conference, Savannah (1999)
12. Steganos Software: http://www.demcom.com/english/steganos/index.htm
13. Cvejic N and Seppänen T: “A wavelet domain LSB insertion algorithm for high capacity audio steganography”. Proc.10th IEEE Digital Signal Processing Workshop, USA (2002)
14. Bond Wireless, SMS AV Software: http://www.bondwireless.com/bwproducts/bwsmssecure.asp
15. T. Moerland: “Steganography and Steganalysis”, Universiteit Leiden, Rhone-Alpes (2003)
16. D Gruhl, A Lu, W Bender: “Echo Hiding", Information Hiding, Springer Lecture Notes in Computer Science v 1174, (1996) 295-315
17. R. J. Anderson and F. A. P. Petitcolas: “On the limits of steganography”, IEEE Journal of Selected Areas in Communications, 16(4):474-481,. Special Issue on Copyright & Privacy Protection, May (1998)
18. J. J. Chae and B. S. Manjunath: "Data hiding in Video" Proc. 6th IEEE International Conference on Image Processing (ICIP'99), Kobe, Japan (1999).
19. I. J. Cox, J. Killian, T. Leighton, and T. Shamoon: “A secure robust watermark for Multimedia,” IEEE Trans. Image Processing, Vol. 6. no. 12, Dec. (1997) 1673-1687
20. I. J. Cox, J. Kilian, T. Leighton and T. Shamoon: “Secure Spread Spectrum Watermarking for Multimedia", IEEE Trans. on Image Processing, 6, 12 (1997) 1673-1687
21. F. A. P. Petitcolas: “MP3Stego.” http://www.petitcolas.net/fabien/steganography/mp3stego, Aug. (1998)
22. Ryuki Tachibana: "Two-Dimensional Audio Watermark for MPEG AAC Audio," in Proc. Security, Steganography and Watermarking of Multimedia Contents VI, SPIE USA (2004)
23. M. Kharrazi, T. H. Sencar, N. Memon: “Benchmarking steganographic and steganalysis techniques”, EI SPIE San Jose, CA, January 16-20, (2005)
24. S. Inoue et al.: “A Proposal on Information Hiding Methods using XML”, 1st NLP and XML Workshop, Japan (2001)
25. T. Morkel, M.S. Olivier, J.H.P. Eloff, H.S. Venter: “One-time Passwords in a Mobile Environment Using Steganography”, Proc of IEEE SecPerU 2005, Greece, (2005)

Error Concealment for Video Transmission Based on Watermarking

Shuai Wan, Yilin Chang, and Fuzheng Yang

National Key Lab. on ISN, XiDian University, Xi'an 710071, China
{shwan, ylchang, fzhyang}@mail.xidian.edu.cn

Abstract. We propose an error concealment method for robust video transmission based on watermarking technique. In this method, the motion vector information is used as the watermark and embedded in the transform coefficients before quantization. A proper embedding algorithm is designed with the characteristics of the video coder taken into consideration, so that the embedding watermark has little influence on the video quality and the coding efficiency. At the decoder, the extracted watermark is used to effectively detect errors and restore the corrupted motion vectors. Then the restored motion vectors are utilized in error concealment. Simulation results demonstrate the advantages of the proposed method in error prone environment. The proposed method is applicable to all the block-transform based video coding schemes.

1 Introduction

Video communication over digital networks has become an important part of current video applications. However, the compressed video stream is very sensitive to transmission errors. To provide a robust video communication, a variety of error control techniques have been proposed such as Automatic Repeat reQuest (ARQ), Forward Error Correction (FEC), and Error Concealment (EC) [1] [2]. The application of ARQ is rather limited because the retransmission of damaged frames using ARQ is not suitable for real-time video communications. Compared with FEC, EC needs a lower overhead, or even no overhead at all, which makes the most of the channel rate. Furthermore, EC can be used as a remediation when FEC fails [1]. So error concealment is an important part in robust video transmission.

Most of the error concealment methods fall into two categories: spatial and temporal. In temporal error concealment, the corrupted macroblock (MB) is replaced by the corresponding MB in the previous frame. A simple way is just to repeat the MB at the same position, and more effective methods resort to protecting or restoring the motion vector of the corrupted MB, and employing motion prediction and error concealment according to the motion vector. Therefore, how to protect and restore the motion vector is a crucial technique in error concealment. Various kinds of error concealment method have been presented to protect the motion vectors, e.g., by giving them higher priority [3]. However, this

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 597–604, 2005.
© Springer-Verlag Berlin Heidelberg 2005

kind of method often introduces more overhead and complexity into the system, and may not be compatible with the video coding standards.

Watermark was first designed to hide information in multimedia for the purpose of ownership certification and tampering detection [4]. It is usually visually imperceptible, and has the ability to carry side information which can be useful for many other purposes. Based on this idea, some error control methods for video transmission using watermarking are also proposed in recent literatures [5] [6]. For example, fragile watermark can be embedded for error detection [5], which is an important step in video error concealment, and [6] proposed a method to protect the motion vectors using data hiding for video transmission over packet-loss channels.

In this paper, we propose a new error concealment method for video transmission based on watermarking. In our method, the motion vector information is considered as the watermark and embedded in the transform coefficients before quantization. A proper embedding algorithm is designed with the characteristics of the video coder taken into consideration, so that the embedding watermark has little influence on the video quality and coding efficiency. In this paper, we describe our algorithm using H.263 video coding as an example, however, the extension to other block-transform based motion compensated video coding schemes is straightforward. The rest of this paper is organized as follows. In Section 2, we present the basic idea and the description of our error concealment method based on watermarking. The experimental results given in Section 3 demonstrate the effectiveness of our method. And conclusions are drawn in Section 4.

2 Error Concealment Based on Watermarking

The basic idea of our method is to use the watermark to convey the motion vector information, which is quite important for error concealment. Fig. 1 shows the framework of an H.263 video coder with an additional watermark embedding module. As shown in Fig. 1, the estimated motion vector is delivered through two separate ways: along one way the motion vector is to be coded using Variable Length Coding (VLC) as usual, and along the other way to be stored as watermark information for further embedding. In our method, the watermark is embedded into the DCT (Discrete Cosine Transform) coefficients to protect the motion vectors, and the embedding algorithm is designed according to the characteristics of the video coder. As a result, the proposed method is compatible with the video coding standards.

2.1 Embedding Algorithm

In H.263 video coder and other block-transform based video coders, the basic unit for coding is an MB. In order to protect the motion vectors, we embed the motion vectors of MBs in the current frame into the MBs at the corresponding position in the next frame. Take the i^{th} MB in frame n–1 ($\mathrm{MV}_{i,n-1}$) for example, whose motion vector is denoted as $\mathrm{MV}_{i,n-1}$. The information of $\mathrm{MV}_{i,n-1}$

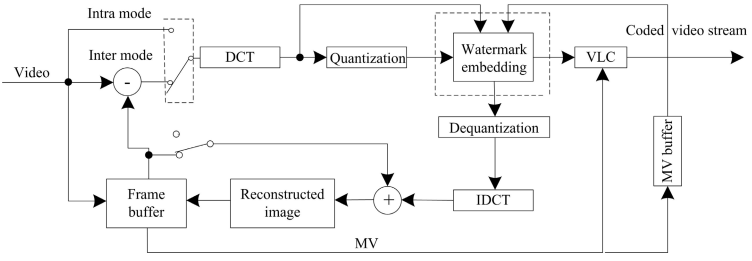


Fig. 1. Block diagram of the H.263 video coder with watermark embedding module

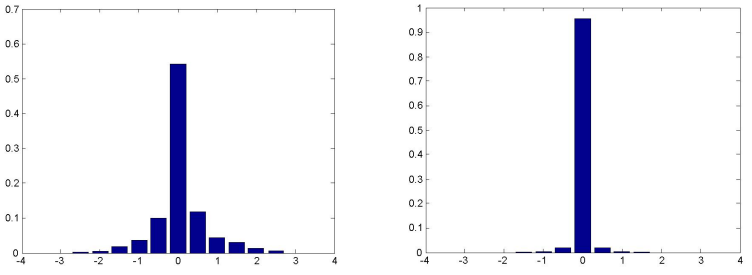


Fig. 2. Motion vector distribution histogram. (a) "Forman" sequence. (b) "Claire" sequence.

will be embedded into the i^{th} MB in frame n ($\mathrm{MB}_{i,n}$) as watermark. When $\mathrm{MB}_{i,n-1}$ has been destroyed, its motion vector can be well restored if $\mathrm{MB}_{i,n}$ is correctly decoded, so $\mathrm{MB}_{i,n-1}$ can be effectively concealed using the restored motion vector.

Usually the motion vector is represented by a pair of components: the horizontal component and the vertical component, respectively. In H.263, the motion vector component is of half-pixel precision and it is coded using variable length code with the maximum length of 13 bits [2]. It is obvious that if all the bits for coding a motion vector are embedded into the coding bit stream, the video quality will be severely impaired. Through analysis of the distribution of the motion vectors in H.263 video coding, we found that the values of a single motion vector component are mainly within the range of –2 to 2. Fig. 2 shows the histogram of the motion vectors in the test sequences of "Forman" and "Claire" with QCIF format.

So we classified the motion vectors into 16 value species: less than - 3, - 3,- 2.5, - 2, - 1.5, - 1, - 0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, larger than 3.5. For the convenience of embedding, the value species are respectively coded using fixed length codes, which are designed according to the distribution of the motion vectors. The codes are shown in Table 1. For the motion vector of an MB, its horizontal and vertical components are respectively represented by 4 bits according to Table 1. These bits representing the motion vector information are used for embedding

Table 1. Motion Vector Codes

MV	< -3.5	-3	-2.5	-2	-1.5	-1	-0.5	0
Code	1111	1101	1110	1100	1010	0001	0010	0000
MV	0.5	1	1.5	2	2.5	3	3.5	> 3.5
Code	1000	0100	1001	0110	0101	0011	1011	1111

as watermark. In H.263, each MB contains 4 luminance blocks (marked as block 1 to block 4) and two chroma blocks (marked as block 5 and block 6), each of which contains 8×8 pixels. Every 8×8 sized block is to be transformed using DCT resulting in 1 DC coefficient and 63 AC coefficients. In this paper, the motion vector information will be hidden in the transformed coefficients.

According to the characteristics of the Human Visual System (HVS), observers are much more sensitive to the distortion in the DC coefficient or the low-frequency AC coefficients than in the high-frequency AC coefficients. So the watermark should not be embedded in the DC coefficient or the comparatively low-frequency AC coefficients. However, after motion compensation and DCT, the residual block has very small DCT coefficients, especially many zero values in the high-frequency AC coefficients. If we use the high-frequency AC coefficients for embedding, there will be an increase in the coding bit rate. And the HVS is more sensitive to the distortion of the chroma component than to the luminance component. In order to avoid a significant impairment of the coding efficiency and the video quality, we choose the 8^{th} and 9^{th} AC coefficients of each luminance block (denoted as $\mathrm{AC}_{i,8}$ and $\mathrm{AC}_{i,9}$ $(i = 1 \ldots 4)$) to embed the watermark.

Then the 4 bits for coding the horizontal component are respectively embedded into the 4 AC_8 coefficients of the luminance blocks in the MB. For example, if the horizontal component of a MV is 3, the bits of 0010 will be respectively embedded into $\mathrm{AC}_{i,8}$ $(i = 1 \ldots 4)$. And the bits for coding the vertical component are respectively embedded into the $\mathrm{AC}_{i,9}$ $(i = 1 \ldots 4)$ in the same way.

Now comes the question of how to embed the corresponding bits. Let C_F be the value of transformed coefficient, C_Q be its quantized value, and C_{QE} be the value of C_Q after embedding. Then C_{QE} can be expressed by

$$C_{QE} = C_Q + \Delta \tag{1}$$

where Δ is the distortion caused by embedding. Let b be the bit for embedding $(b = 0, 1)$. In order to expediently extract b from C_{QE} at the decoder, we expect to obtain

$$|C_{QE}| \bmod 2 = \begin{cases} 0, & b = 0 \\ 1, & b = 1 \end{cases} \tag{2}$$

where "mod" represents the module operation. Therefore b can be embedded as:

(1) When $b = 0$,
(i) if C_Q is an even number

$$\Delta = 0 \tag{3}$$

(ii) if C_Q is an odd number

$$\Delta = f(C_Q) \tag{4}$$

(2) When $b = 1$,
(i) if C_Q is an odd number

$$\Delta = 0 \tag{5}$$

(ii) if C_Q is an even number

$$\Delta = f(C_Q) \tag{6}$$

where $\Delta = f(C_Q)$ is a function depending on C_Q. In the traditional algorithms, $f(C_Q) = -1$ or $f(C_Q) = 1$ are frequently used to implement the embedment. However, when the video coding process is considered, such $f(C_Q)$ functions will introduce large quantization errors when the DCT coefficient is reconstructed at the decoder. For example, suppose C_F is within the range of 39-50 and the quantization parameter is 6. The reconstructed value of C_F can be determined according to the quantization function and the re-quantization function defined in H.263, which are given by

$$C_Q = sign(C_F) \cdot (|C_F| - QP/2)/(2 \cdot QP) \tag{7}$$

$$C_F^{'} = sign(C_Q) \cdot (QP \cdot (2 \cdot |C_Q| + 1)) \tag{8}$$

where QP is the quantization parameter, and $C_F^{'}$ is the reconstructed value of C_F. The function $sign(\cdot)$ is defined as

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{9}$$

We suppose the DCT coefficient to be uniformly distributed within such a value scope, so for the original data the mean squared error of reconstruction is approximate 18. However, if $f(C_Q) = -1$ is used, the mean squared error will respectively increase to 222, which will result in a significant decrease in the reconstructed video quality.

In order to reduce the reconstructed error, we determine Δ according to both the original DCT coefficient C_F and the QP. The detailed function is given by

$$\Delta = f(C_F, QP) = \begin{cases} sign(C_F) & |QP \cdot (2 \cdot (R-1)+1) - |C_F|| > \\ & |QP \cdot (2 \cdot (R+1)+1) - |C_F|| \\ -sign(C_F) & |QP \cdot (2 \cdot (R-1)+1) - |C_F|| < \\ & |QP \cdot (2 \cdot (R+1)+1) - |C_F|| \end{cases} \tag{10}$$

where $R = (|C_F| - QP/2)/(2 \cdot QP)$, and the symbol "/" denotes dividing exactly where the residue is discarded and the integer quotient is remained as the result. When C_F is within the range of 39-50 and the QP is 6, the mean squared error will be 78 using this embedding method, which is much smaller than using the traditional embedding methods.

After embedding, the coefficients embedded with watermark are coded as usual. Then the coded video stream will be transmitted over the error-prone channel.

2.2 Error Concealment

After transmission, the decoder extracts the watermark to detect errors and restore the corrupted motion vectors. Firstly the embedded 8 bits in each MB are extracted by

$$b = \begin{cases} 0, & |C_{QE}| \bmod 2 = 0 \\ 1, & |C_{QE}| \bmod 2 = 1 \end{cases} \tag{11}$$

Then they are arranged to two codes as the order by which they are embedded. Then the motion vector can be found out through looking up Table 1. During the decoding process, the decoder compares the extracted motion vector with the decoded motion vector to detect errors. Take $\mathrm{MB}_{n,m}$ for example, the following error detection algorithm is carried out to determine whether $\mathrm{MB}_{n,m}$ has been impaired by errors. Let $\mathrm{MV}_{n,m}$ be the decoded motion vector for $\mathrm{MB}_{n,m}$ and $\mathrm{MV}'_{n,m}$ be the motion vector extracted from $\mathrm{MB}_{n+1,m}$

$$\mathrm{If}((\mathrm{MV}_{n,m} == \mathrm{MV}'_{n,m}) \cup (\mathrm{MV}_{n-1,m} == \mathrm{MV}'_{n-1,m})) == \mathrm{TRUE},$$

$$\text{then} \quad \mathrm{MB}_{n,m} = \mathrm{TRUE}.$$

$$\begin{aligned} \mathrm{If}(((\mathrm{MV}_{n,m} \neq \mathrm{MV}'_{n,m}) \cap (\mathrm{MB}_{n+1,m} == \mathrm{TRUE})) \cup \\ ((\mathrm{MV}_{n-1,m} \neq \mathrm{MV}'_{n-1,m}) \cap (\mathrm{MB}_{n-1,m} == \mathrm{TRUE}))) == \mathrm{TRUE}, \end{aligned}$$

$$\text{then} \quad \mathrm{MB}_{n,m} = \mathrm{FALSE}.$$

Once $\mathrm{MB}_{n,m}$ is detected to be wrong, the following algorithm will be performed with an attempt to restore its MV.

$$\mathrm{If}\ \mathrm{MB}_{n+1,m} == \mathrm{TRUE} \quad \text{then}\ \ \mathrm{MV}_{n,m} = \mathrm{MV}'_{n,m}$$

$$\text{else}\ \ \mathrm{MV}_{n,m} = (0,0)$$

According to the obtained $\mathrm{MV}_{n,m}$, the corresponding area for $\mathrm{MB}_{n,m}$ is found in the previous frame. Then the corrupted $\mathrm{MB}_{n,m}$ is replaced by the corresponding pixels in the previous frame.

3 Experimental Results

We implement the proposed error concealment method in the test model of H.263 video coder. Two standard test sequences with QCIF format are used in our experiments: "Foreman", and "Claire". Without loss of generality, for each test sequence the first frame is coded as I frame using Intra mode, and the others are coded as P frames using motion compensation. B frame is not used. The frame rate is set to 30 frames per second, and 300 frames are tested for each sequence.

First, the coding efficiency of the proposed method is evaluated. Fig. 3 shows the PSNR curves at the encoder under different coding bit rate of the original H.263 video coding and the proposed method based on watermarking. It can be observed that the influence on the coding efficiency caused by watermarking is almost negligible, because the embedding method is well designed according to the characteristics of the motion vector, the DCT coefficients, and the quantization progress.

In order to evaluate the error concealment performance, we compare our method with the generic error concealment (GEC) proposed in [1]. For both

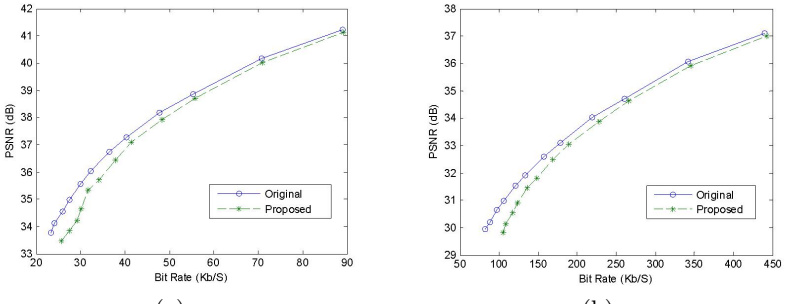


Fig. 3. Coding efficiency. (a) "Foreman" sequence. (b) "Claire" sequence.

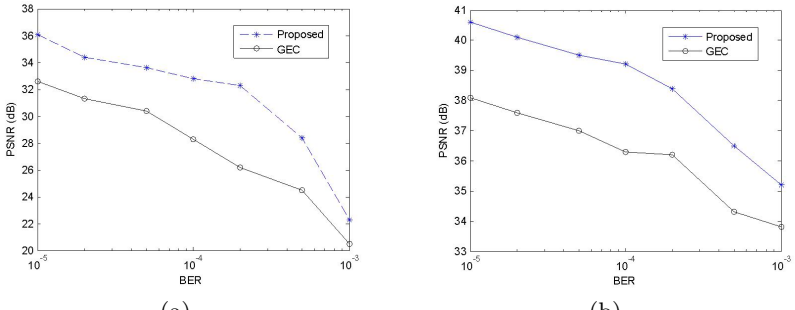


Fig. 4. PSNR versus BER. (a) "Foreman" sequence. (b) "Claire" sequence.

methods, quantization parameter is set to 6, and the first I frame in every sequence is heavy protected so it is subject to no errors because we focus on the temporal error concealment. Stochastic noise with uniform distribution is added to the coded bit rate to simulate the transmission errors. For each error rate condition tested, 30 simulation runs with different error patterns are performed, and the luminance Peak Signal Noise Ratio (PSNR) is averaged over all decoded frames and all channel realizations for objective video quality assessment. Fig. 4 shows the average luminance PSNR at the decoder under different bit-error rates (BER). The results strongly support the claim that the proposed error concealment method yields more consistent and significant gains over GEC. More specifically, 2-3 dB gains can always be observed when the error rate is less than 10-3. The performance improvement can also be perceptually observed on the actual decoder-reconstructed video.

4 Conclusion

In this paper, we have proposed an error conceal method for robust video coding using watermarking. At the encoder, the motion vector information is embedded into the DCT coefficients as watermark. After transmission, the decoder extracts the watermark to detect the errors and perform error concealment according to the restored motion vectors. Since a proper embedding algorithm is designed according to the characteristics of video coding, the influence of the proposed method on the coding efficiency is almost neglectable. Simulation results show that our method substantially and consistently outperforms the traditional error concealment method, which indicates that the error concealment based on watermarking is an effective way to improve the performance of video transmission over error-prone channels.

References

1. Y. Wang and Q. Zhu: Error control and concealment for video communication: a review. Proceedings of the IEEE, Vol. 86. (1998) 974-997
2. ITU Telecom. Standardization Sector of ITU: Video Coding for Low Bit Rate Communication. ITU-T Recommendation H.263 Version2 (1998)
3. L. H. Kieu and K. N. Ngan: Cell-loss concealment techniques for layered video codecs in an ATM network. IEEE Trans. Image Proc., Vol. 3. (1994) 666-677
4. ET Lin and EJ Delp: A Review of Fragile Image Watermarks. Proceedings of the Multimedia and Security Workshop (ACM Multimedia '99), (1999) 25-29
5. Minghua Chen, Yun He and Reginald L. Lagendijk: Error detection by fragile watermarking. Proceedings of PCS2001, Seoul. (2001) 287-290
6. Peng Yin, Min Wu and Bede Liu: A Robust Error Resilient Approach for MPEG Video Transmission over Internet. VCIP, SPIE. (2002) http://www.ee.princeton.edu/ pengyin/ publication/vcip2002-errcon-final.pdf.

Applying the AES and Its Extended Versions in a General Framework for Hiding Information in Digital Images

Tran Minh Triet and Duong Anh Duc

Faculty of Information Technology, University of Natural Sciences, VNU-HCMC,
227 Nguyen Van Cu St, Hochiminh City, Vietnam
{tmtriet, daduc}@fit.hcmuns.edu.vn

Abstract. Watermarking techniques can be applied in many applications, such as copyright protection, authentication, fingerprinting, and data hiding. Each different purpose of usability requires different approaches and schemes. In this paper, we present the framework of combining the Advanced Encryption Standard (AES) with watermarking techniques for hiding secret information in digital images. This framework can also be customized for other type of multimedia objects.

1 Introduction

Nowadays with the increasing development of telecommunication and digital multimedia services, digital resources are available for public accesses. The arising problem is to protect intellectual properties of valuable digital masterpieces while preserving their public accessibility[8]. Watermarking is considered one of the applicable solutions for this problem [9].

Besides applications in copyright protection or authentication which require the robustness of the algorithms, watermarking can also be used for other purposes [8], such as data hiding [14], public watermarking [9]... Any watermarking technique has the common capability that is embedding pieces of information (called watermark) into other information (such as multimedia objects) [9]. Therefore, we can hide secret message in other information, especially multimedia object, such as digital images, audio, or video...

Among millions of digital images, it is almost impossible for anyone to find out exactly what image contain secret message. People can not distinguish between the original and embedded images. Even when the attacker suspects correctly the "special" image, he has to find out the way to exactly extract the secret message embedded in this image. If the embedding process uses the efficient watermarking technique to hide secret message in the image, the attacker will not have enough clue to extract accurately the hidden information.

However in case the attacker knows or can guess the method that has been used to hide secret data together with correct parameters of the embedding process, the content of hidden message will be revealed as soon as the message has been extracted from the image.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 605 – 610, 2005.
© Springer-Verlag Berlin Heidelberg 2005

To strengthen the secrecy of hidden information, we can utilize the high security cryptographic algorithms and processes to encrypt secret message before embedding it into other information, such as digital images. In this paper, we propose the framework of combining the modern conventional algorithm, the Advanced Encryption Standard (AES), together with some of its extended versions, with watermarking techniques for hiding secret information in digital images. This framework can also be used for other types of multimedia objects.

The rest of the paper is organized as follows. Section 2 introduces briefly the AES and some extended versions. Section 3 describes the proposed framework of combination between the AES (and its extended versions) with watermarking techniques for data hiding. We conclude with a summary and directions for future work in Section 4.

2 The Advanced Encryption Standard and Some of Its Extended Versions

2.1 Brief Introduction of the Algorithms

In October 2nd 2000 the National Institute of Standards and Technology (NIST) announced the Rijndael Block Cipher, proposed by Vincent Rijmen and Joan Daemen, to be the Advanced Encryption Standard (AES). This new standard has been utilized in many applications, from personal security and privacy applications to electronic transactions over networks.... For details of the AES, we refer to [1].

We have studied this new block cipher and devised the extended versions of this new cryptography standard to stimulate its strength and resistances against attacks. The 256/384/512-bit extended version ([3]) and the 512/768/1024-bit extended version of the AES ([2]) have been designed basing on the same mathematical bases of the original cipher but can process larger blocks using larger cipher keys in compare with the original algorithm. While the possible length of the block or the cipher key is 128 or 192 or 256 bits for the original algorithm, the 256/384/512-bit extended version can process blocks and cipher keys having 256 or 384 or 512 bits, and the 512/768/1024-bit extended version can process blocks and cipher keys having 512 or 768 or 1024 bits

When designing these extended versions, we try to satisfy all the criteria of the motivation of choices that were used in the original cipher. Adjustments and optimizations have also been made when possible, e.g. in the criteria for choosing the coefficients in the MixColumns transformation. As a result the extended version is expected, for all key and block length defined, to behave as good as can be expected from a block cipher with the given block and key lengths. Besides, the computational cost has also been carefully considered to optimize the speed of the Cipher and the Inverse Cipher with respect to their security strength.

2.2 Strength Against Known Attacks

As each round uses different round constant and the cipher and its inverse use different components, the possibility for weak and semi-weak keys, as existing for DES,

can be eliminated [1]. The fact that all non-linearity in the cipher is in the fixed S-box eliminates the possibility of weak keys as in IDEA [1], [6]. The key schedule, with its high diffusion and non-linearity, makes it very improbable for the related-key attacks [5]. The complicated expression of the S-box in $GF(2^8)$, in combination with the effect of the diffusion layer prohibits interpolation attacks [7] for more than a few rounds [3]. For all block lengths of each extended version, this is sufficient to protest against differential and linear cryptanalysis. For details we refer to [3].

Just like the original AES, the most efficient key-recovery attack for the two extended versions of this algorithm mentioned above is still the exhaustive key search. Obtaining information from given plaintext-cipher text pairs about other plaintext-cipher text pairs cannot be done more efficiently than by determining the key by exhaustive key search.

3 The Proposed Framework

3.1 The Embedding Process

The embedding process is illustrated Fig. 1. This process can be divided into three sequential steps:

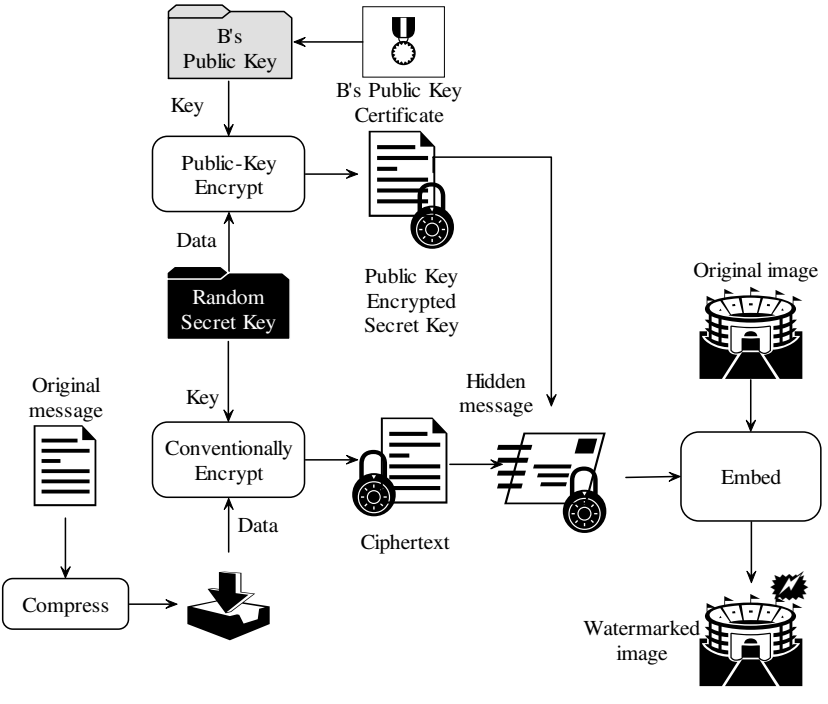


Fig. 1. The embedding process

Step 1: Conventionally encrypt the content of the secret message
In this first step, the original message that we want to hide in the selected image will be compressed to reduce the size of the data to be embedded. The smaller the size of data is, the more secure it will be. Besides, the compression will also remove repeated patterns in the secret message.

The compressed message will then be encrypted using an efficient robust conventional algorithm with a one-time-only secret key generated randomly.

The chosen cipher must be robust to known attacks such as differential cryptanalysis, linear cryptanalysis, truncated differentials, brute force attack… The original AES, together with our new devised extended versions of this cipher, can be applied in this first step of the embedding process.

Step 2: Encrypt the secret key with the public-key of the recipient
This second step ensures that only the legal receiver B can decrypt exactly the content of the secret message. The secret key used to encrypt the secret message will be encrypted using a public key algorithm with the public key of the receiver B.

Step 3: Embed the encrypted message together with the encrypted secret key into the selected image
The hidden message, including the ciphertext and the public-key encrypted secret key, together with the original selected image, will become the inputs of the embed module.

Because the main purpose of our process is only to embed the hidden message into the selected image, we do not have to worry about the robustness of the watermarking techniques in use. As a result, we have wide range of choices to select candidate watermarking techniques, from the very simple scheme that embeds watermark into the least significant bits of the image to the complicated algorithms in spatial domain [10], DCT domain [11] or wavelet domain ([12],[13])...

3.2 The Extracting Process

From the embedding process, we can easily figure out the straightforward extracting process shown in Fig. 2. This process also consists of three sequential steps:

Step 1: Extract the hidden message embedded in the image.

Step 2: The recovered hidden message contains the public-key encrypted secret key and the conventionally encrypted message. Using the receiver B's private key, which is known only by the owner – the receiver, B can easily decrypt the content of the secret key that was used to encrypt the content of the message.

Step 3: Using the secret key, the receiver B can decrypt the content of the secret message.

In fact, the watermarked image might be changed during transmission from the sender A to the receiver B. The changes may be caused by unintentional operations or malicious attacks to reveal or destroy the content of the message hidden in the digital image. If the chosen watermarking technique sustains the malicious attacks, the receiver can recover exactly the content of hidden message. However, the purpose of this framework is not to protect content of hidden message to be changed or destroyed. Its main goal is to keep the content of the message to be absolutely secret.

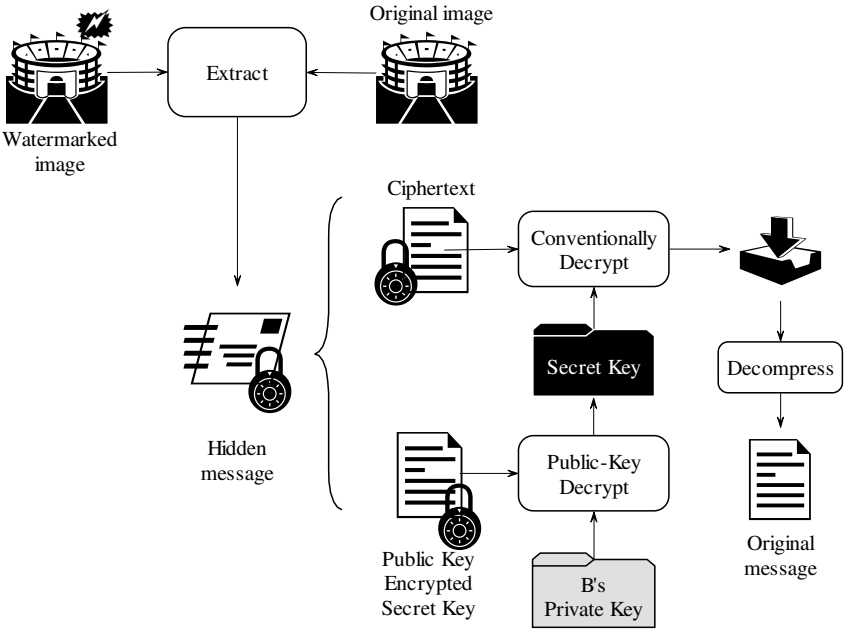


Fig. 2. The extracting process

Even when the attacker can extract the hidden message, he has to face with the challenge of the encrypted message.

4 Conclusions

In the first step of the embedding process, the original message will be compressed. This data compression not only reduces significantly the size of the data to be embedded into the selected image but also strengthens cryptographic security. Most cryptanalysis techniques exploit patterns found in the plaintext to crack the cipher. Compression reduces these patterns in the plaintext, thereby greatly enhancing resistance to cryptanalysis. In practice, the LZW algorithm can be used efficiently because of its high compression ratio and fast speed. Furthermore it is also easy to modify and adjust several steps in this algorithm to devise other LZW-like algorithms that can be used to strengthen the cryptographic security.

Just like the original AES [1], the most efficient key-recovery attack for the two extended versions of this algorithm, including the 256/384/512-bit extended version [3] and 512/768/1024-bit version [2], is still the exhaustive key search. Obtaining information from given plaintext-cipher text pairs about other plaintext-cipher text pairs cannot be done more efficiently than by determining the key by exhaustive key search.

Because the secret key that has been used to encrypt the secret message is encrypted with the public key of each recipient, only this legal recipient can successfully decrypt the public key encrypted secret key using his or her own private key.

In this paper we introduce briefly the characteristics of the AES and some of our proposed extended versions of this cipher. We also present the general framework of combining cryptographic algorithms with watermarking techniques to embed secret and sensitive information into digital images. With high flexibility level, the framework allows users to select their favorite efficient algorithms, including the compression method, the conventional cipher, and the public key technique. This feature supports various practical implementations. The framework can also be customized easily to deal with other type of multimedia objects, such as audio or video.

References

1. J. Daemen, V. Rijmen: AES Proposal: Rijndael, AES Algorithm Submission (1999)
2. Duong Anh Duc, Tran Minh Triet, Luong Han Co: The extended version of the Rijndael Block Cipher. Journal of Institute of Mathematics and Computer Sciences, India (2001).
3. Tran Minh Triet: Research in Some Issues on Information Security and Applications. MSc Thesis, University of Natural Sciences, VNUHCM, Vietnam (2005)
4. Duong Anh Duc, Tran Minh Triet, Luong Han Co: The extended Rijndael-like Block Ciphers. International Conference on Information Technology: Coding and Computing - 2002, The Orleans, Las Vegas, Nevada, USA (2002)
5. J. Kelsey, B. Schneier, D. Wagner: Key-schedule cryptanalysis of IDEA, GDES, GOST, SAFER, and Triple-DES. Advances in Cryptology (1996)
6. J. Daemen: Cipher and hash function design strategies based on linear and differential cryptanalysis. Doctoral Dissertation, K.U.Leuven (1995)
7. T. Jakobsen, L.R. Knudsen: The interpolation attack on block ciphers. Fast Software Encryption, LNCS 1267, E. Biham, Ed., Springer-Verlag (1997) 28–40.
8. Duong Anh Duc, Tran Minh Triet, Dang Tuan, Ho Ngoc Lam: Watermarking – An overview and applications in Intellectual Property Management and Protection. Proc. 3rd Scientific Conference of Hochiminh City University of Natural Sciences, Vietnam (2002)
9. Stefan Katzenbeisser, Fabien A.P. Petitcolas: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston, London (2000)
10. N. Nikolaidis, I. Pitas: Robust image watermarking in the spatial domain. Signal Processing (1998).
11. M. Barni, F. Bartolini, V.Cappellini, A. Piva: A DCT domain system for robust image watermrking. Signal (1998)
12. P. Meerwarld: Digital Image Watermarking in the Wavelet Transfrom Domain. Salburg (2001)
13. Wenwu Zhu, Zixiang Xiong, Yaquin Zhang: Multiresolution watermarking for image and video. Proc. IEEE Inter. Conf. Image Processing (1998)
14. C. Hsu, J. Wu: Hidden signatures in images. Proc. IEEE Inter. Conf. Image Processing (1996)

An Image Hiding Algorithm Based on Bit Plane

Bin Liu, Zhitang Li, and Zhanchun Li

Network Center, Huazhong University of Science and Technology, Wuhan 430074, China
bliu@mail.hust.edu.cn

Abstract. In this paper, an image hiding method which used for hiding one secret image into multiple open images is addressed. The method is roughly divided into three steps. First, based on the correlations analysis, different bit planes of a secret image are hided into different bit planes of those different open images. Second, a group of new hiding images is obtained by image fusion method, and then we get a new image by taking the "Exclusive-NOR" operation on these images. At last, the final hiding result is obtained by hiding the image obtained in above steps into certain open image. The experimental result shows the effectiveness of the method.

1 Introduction

In recent days, all sorts of imaging sensors have been developed rapidly, so it is convenient to obtain the images with different resolutions, flat roofs and spectrums such as Spot series, Landset, Quickbird, and so on. The images, with vivid and intuitionistic characters, are becoming an indispensable means for the transfers of information. Hence, an associated problem --- information security is becoming more and more important with the development of multimedia and network. The problem how to give an effective method for image encryption is a hot topic presently [1,2,3]. Image hiding technique is an important method to realize image encryption; Its rough idea is to hide a secret image into a quotidian image invisibly. In general case, image encryption is used by combing image-scrambling technique [3].

Up to now, several literatures have been reported on the topic of image hiding. In [2], an adaptive algorithm, used to embed data into vector quantization (VQ) compressed images, is presented by Du and Hsu. This method adaptively varies the embedding process according to the amount of hidden data. Using a binary matrix and a weighted matrix as a key, seng, Chen and Pan described a hiding method for binary image by [4]. In [5], Zhang et. al. give an image hiding method based on iterative blending, the result of image hiding is obtained by image fusion method which is performed by a iterative scheme. Because chaos sequence is non-periodicity, ubiquitous, easily to be generated and sensitivity to original value, it has been widely used in image encryption and image hiding [6,7].

In this article, a novel multi-image hiding method for one secret is presented. The proposed algorithm is given based on the correlation analysis at first, which determines how to hide the different bit planes of the secret image in the different bit planes of the different open images. Then, a series new hiding images are obtained by image fusion, and then we get a new image by taking the "Exclusive-NOR" operation on these images. The final hiding result is achieved by hiding the image obtained according to

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 611 – 616, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the above steps into certain open image. The experimental result shows that the proposed method is effect and promising.

2 Image Hiding Algorithm Based on Bit Plane

2.1 The Proposed Algorithm

As we know, for a secret image which is hided into another image, if the carrier image is intercepted and captured by third party and threw doubt upon it, it is possible to incur destroying or decoding [5]. So it is very dangerous that the security of a secret image is depended on one image. In order to enhance the security of information, a natural method is to hide one image into several open images. Hence, we propose a novel image hiding algorithm which is shown as Fig.1.

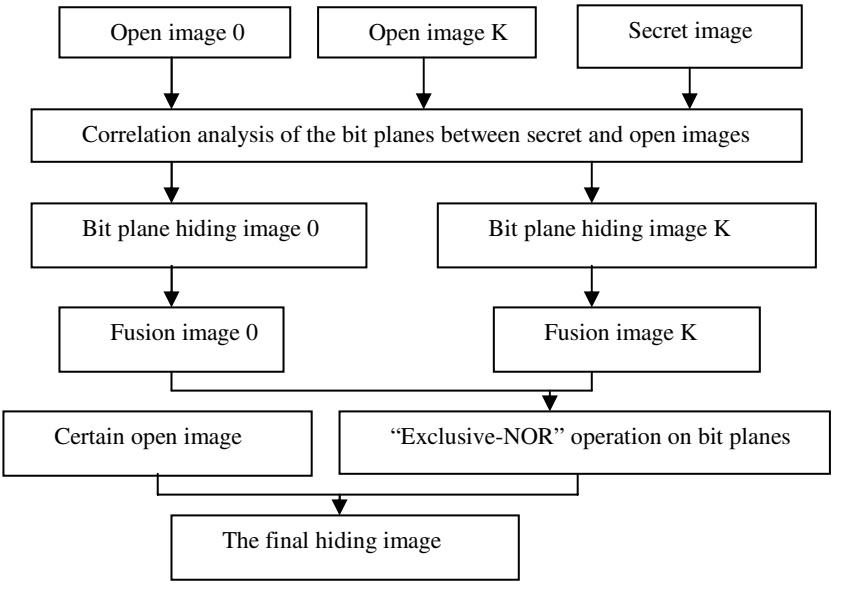


Fig. 1. The flowchart of image hiding

Now, we described the proposed algorithm in detail. According to Fig.1, the algorithm can be divided into several steps as follows (suppose that the secret and open images possess 8 bit planes):

Step 1: Input the secret image $g(x, y)$ and open images $f_i (i = 0,..., K-1, K \le 8)$;

Step 2: Compute the correlation coefficients for the bit planes of the secret image and different open images;

Step 3: Select the bit plane with the largest correlation coefficient as the location which the bit plane of the secret image is embedded (record these bit planes as keys), thus the hiding image $f_i'(i = 0,...K-1)$ is obtained;

Step 4: By image fusion method, the images $f_i'(i=0,...,K-1)$ are hided into the images $f_i(i=0,...,K-1)$, the results are denoted by $\tilde{f}_i(i=0,...,K-1)$;

Step 5: Take the nesting "Exclusive-NOR" operation on the corresponding bit planes of $\tilde{f}_i(i=0,...,K-1)$, this means that the "Exclusive-NOR" operation is carried out between the same bit plane of $\tilde{f}_0$ and $\tilde{f}_1$, the new image is denoted by S_1. Repeating above procedure for S_1 and $\tilde{f}_2$, we get the result S_2;

Step 6: The final hiding result f' is obtained by hiding $\tilde{f}$ into certain open image by image fusion method. That is

$$f' = \lambda \tilde{f} + (1-\lambda) f_{\Xi}, (0 < \lambda < 1) \ . \tag{1}$$

where f_{Ξ} is certain open image among $f_i(i=0,...,K-1)$.

2.2 Illustrations of the Algorithm

Without lost of generality, the secret image $g(x, y)$ is denoted by $g = (g^1,..., g^8)^T$ and the open images f_i are denoted by $f_i = (f_i^1,...f_i^8)^T \quad (i=0,...K-1)$, where g^m and $f_i^m \ (m=1,...,8)$ are the mth bit planes of the secret and open images. We explain the embedding method in the case of g^1. If we calculate the correlation coefficient of g^1 and f_i^m, a matrix can be obtained as follows:

$$C_1 = \begin{bmatrix} c_{01} & ... & c_{08} \\ ... & ... & ... \\ c_{K-1,1} & ... & c_{K-1,8} \end{bmatrix} \tag{2}$$

here $c_{pq} = c(g^1, f_p^q)(p=0,...K-1, q=1,...8)$ are the correlation coefficients between g^1 and f_p^q. The correlation coefficient between two vectors X and Y is calculated by Eq.(3).

$$C(X,Y) = \frac{\sum\sum(X-EX)(Y-EY)}{(\sum\sum(X-EX)^2\sum\sum(Y-EY)^2)^{1/2}} \tag{3}$$

Suppose the largest element of C_1 is $c_{k,l}$, then g^1 should be embedded into f_k^l, the embedding location is the lth bit plane of the kth open image. When we select the embedding bit plane for g^2, the lth bit planes of each open image are not involve the calculation of the correlation coefficient. By the similar procedure, the embedding bit plane for g^2 can be determined. Repeating the same procedure for the rest bit planes, the corresponding embedding locations can be determined. Thus the different bit planes of the secret image can be embedded into the different bit planes of those different open images. In other words, two bit planes of the secret image cannot be embedded into the same bit plane of one image or same bit planes of two images.

According to above method, the bit planes of the secret image may be congregated a few open images (thus, these images may be changed greatly). So we adopt the following amended step: taking out $8-(\lfloor 8/K \rfloor+1)$ bit planes from the images in which the number of embedded bit planes is larger than $\lfloor 8/K \rfloor$ ($[\bullet]$ is taking integer operation) and embed these bit planes into the open images in which the number of embedded bit planes is smaller than $\lfloor 8/K \rfloor$ (such that the number bit planes embedded into each open image is smaller $\lfloor 8/K \rfloor+1$, more over, each open image should be embedded one bit plane of the secret at least).

When the nesting operation is employed, the order of the images $\tilde{f}_i(i=0,...K-1)$ can be changed, this maybe chances the key space. In the last fusion step, the image with the largest correlation coefficient with $\tilde{f}$ is selected as the carrier image, the reason for this is to achieve better hiding effect.

3 Extract Algorithm

The extract algorithm is the contrary procedure of the hiding algorithm. It can be implemented as follows:

Step 1: Using Eq. (1), the image $\tilde{f}$ is resolved by f', $f_{\equiv}$ and λ;

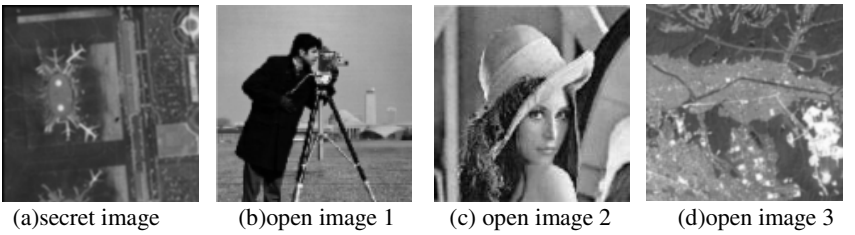
Step 2: By $\tilde{f}$ and the reversibility of the "Exclusive-NOR" operation applied in step5 in section 2, the $\tilde{f}_i(i=0,...K-1)$ is obtained;

Step 3: From step 4 in section 2 and $\tilde{f}_i(i=0,...K-1)$, we have $f_i'(i=0,...K-1)$;

Step 4: Extract g^m from each $f_i'(i=0,...K-1)$ by the embedding key, and then the secret image g is restored.

4 Simulation and Analysis

To verify the effectiveness of the approach proposed in this paper, example will be given in this section. The secret image and open images are shown as Fig.2.



(a)secret image (b)open image 1 (c) open image 2 (d)open image 3

Fig. 2. Secret and open images

By the image hiding and extracting algorithm introduced in section 2 and 3, the hiding image and restoration image are shown as Fig.3.



(a)hiding image (b)restored image

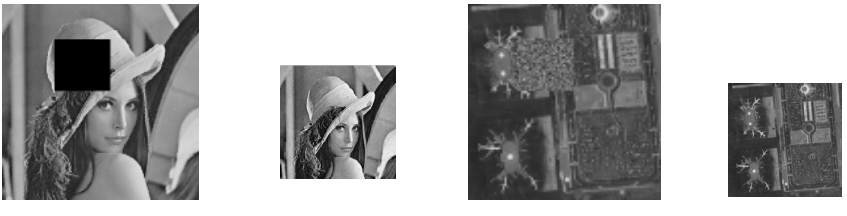
Fig. 3. Hiding and restored images

In the procedure of encryption, based on the correlation analysis and the amended approach, the first, third and fourth bit planes of the secret image are embedded into the first, second and fifths bit planes of Fig.2(d), the second, fifths and eighth bit planes of the secret image are embedded into the third, sixth and seventh bit planes of Fig.2 (c), the sixth and seventh bit planes of the secret image are embedded into the fourth and eighth bit planes of Fig.2(b). The parameter λ used in Eq.(1) is 0.05.

It can be observed that the secret image is hided into Fig.2(c) without visibility. The difference between the restored image and the original secret image is very slight. In order to assess the hiding ability quantitatively, the index root mean square error (RMSE) is employed:

$$RMSE = \sqrt{\frac{1}{M \times N} \sum_{j=0}^{N-1} \sum_{i=0}^{M-1} (\hat{P}(i, j) - P(i, j))^2} \tag{4}$$

here $M \times N$ is the size of the image, $P(i, j)$ and $\hat{P}(i, j)$ express Fig.2(c) and Fig. 3(a) respectively, the value of RMSE in this case is 0.0609, so the method proposed here has better restorative performance. Several common attack measures including corruption, cutting and zooming are considered. Fig.4 shows the restored results that were destroyed by above ways.



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

Fig. 4. The test for anti-attack performance

Fig4. (a) and (b) are the images that Fig.3(a) are cutting and zooming out (zooming ratio is 2:1). Fig4. (c) and (d) are the restored results from Fig4. (a), (b), respectively.

Fig.4 shows that the contents contained in the secret image are destroyed partly, but most of them are still preserved and the influence for image understanding is slight. In other words, the proposed hiding scheme is robust for above destroyed measures.

5 Conclusion

In this paper, a smart method used for hiding one image by multiple images is discussed. Because the correlation will affect the hiding effect heavily, so the different bit planes of the secret image are first hided into the different bit planes of the open images by correlation analysis. Then, the images obtained in above step are hided into the corresponding open images by image fusion method. Using the operation named as nesting "Exclusive-NOR" for these fusion results, a new hiding image is then obtained. The hiding image is finally obtained by using image fusion again. The experimental result shows that the proposed method is effective and promising.

References

1. M. Wu, B. Liu: Data Hiding In Image and Video : Part I---Fundamental Issues and Solutions. IEEE Trans. on Image Processing Vol.12, (2003)685-695.
2. W.C. Du, W.J.Hsu: Adaptive Data Hiding Based on VQ Compressed Images. IEE Proceedings- Vision, Image and Signal Processing Vol.150, (2003) 233-238.
3. Ding Wei, Yan Weiqi, Qi Dongxu : Digital Image Information Hiding Technology and Its Application Based on Scrambling and Amalgamation. Chinese Journal of Computers Vol.10, (2000) 644~649.
4. Y.T. Tseng,Y.Y. Chen, H.K. Pan : A Secure Data Hiding Scheme for Binary Images. IEEE Trans. on Communications (2002), Vol. 50, 1227-1231.
5. Zhang Guicang, Wang Rangding : Digital Image Information Hiding Technology Based on Iterative Blending. Chinese Journal of Computers Vol.26, (2003) 569-574.
6. Sun Xin, Yi Kaixiang, Sun Youxian : New Image Encryption Algorithm Based on Chaos System. Journal of Computer-aided Design and Computer Graphics Vol.4 (2002) 136-139.
7. P. Moulin and M. K. Mihcak : A Framework for Evaluating the Data-hiding Capacity of Image Sources. IEEE Trans. on Image Processing Vol. 11, (2002) 1029-1042.

A Blind Audio Watermarking Algorithm Robust Against Synchronization Attack*

Xiangyang Wang[1,2] and Hong Zhao[1]

[1] School of Computer and Information Technique, Liaoning Normal University, Dalian 116029, China
[2] State Key Laboratory of Information Security, Graduate School of Chinese Academy of Sciences, Beijing 100039, China
wxy37@126.com

Abstract. Synchronization attack is one of the key issues of digital audio watermarking. In this paper, a robust digital audio watermarking algorithm in DWT (Discrete Wavelet Transform) and DCT (Discrete Cosine Transform) domain is presented, which can resist synchronization attack effectively. The features of the proposed algorithm are as follows: (1)More steady synchronization code and new embedded strategy are adopted to resist the synchronization attack effectively. (2)The multi-resolution characteristics of DWT and the energy-compression characteristics of DCT are combined to improve the transparency of digital watermark. (3)The algorithm can extract the watermark without the help of the original digital audio signal.

1 Introduction

With the rapid development of the network and multimedia technique, digital watermarking has received a large deal of attention recently and has been a focus in network information security [1]. The current digital watermarking schemes mainly focus on image and video copyright protection. A few audio watermarking techniques have been reported [2]. Especially, it is hard to find the robust watermarking algorithms that can resist the synchronization attack effectively [3-5].

Up to now, 4 robust audio watermarking strategies are adopted to resist the synchronization attack. They are: All-list-search [2], combination of spread spectrum and spread spectrum code [6], utilizing the important feature of origin digital audio [7-8](or we call it self-synchronization strategy), synchronization code [9-10]. Among them, All-list-search strategy need great calculating amount and has high false positive rate; the second strategy cannot achieve blind detection; the current self-synchronization algorithm cannot extract feature points steadily. By contrast, synchronization code strategy has more obvious technological advantages. Barker code has better self-relativity, so literature [9] and [10] chooses it as synchronization mark. These methods can resist synchronization attack effectively. But it has such defects as follows: (1)it chooses a

* This work was supported by the Natural Science Foundation of Liaoning Province of China under Grant No.20032100 and the Open foundation of State Key Laboratory of Information Security of China under Grant No.03-02.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 617 – 622, 2005.
© Springer-Verlag Berlin Heidelberg 2005

12-bit Barker code which is so short that it is easy to cause false synchronization. (2)it only embeds the synchronization code by modifying individual sample value, which reduces the resisting ability greatly.

Taking the problems mentioned above in mind, we introduce a DWT and DCT-based digital audio blind watermarking algorithm which can resist synchronization attack effectively. We choose 16-bit Barker code as synchronization mark, and embed it by modifying the mean value of several samples.

2 Watermark Embedding Scheme

The diagram of our audio watermarking technique is shown in Fig.1.

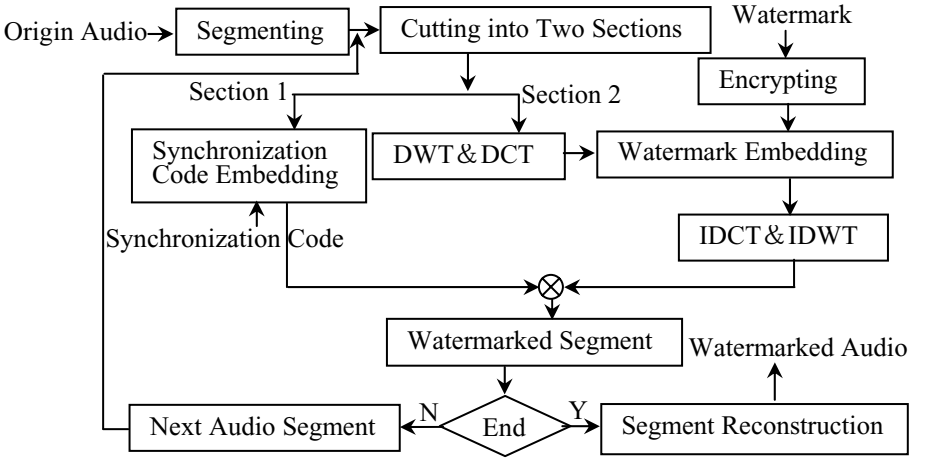


Fig. 1. Watermark embedding scheme

In order to guarantee robustness and transparency of watermarking, the proposed scheme embeds synchronization code in the mean value of several samples.

Let $A = \{a(i), 0 \le i < Length\}$ represent a host digital audio signal with $Length$ samples. $W = \{w(i, j), 0 \le i < M, 0 \le j < N\}$ is a binary image to be embedded within the host audio signal, and $w(i, j) \in \{0,1\}$ is the pixel value at (i, j). $F = \{f(i), 0 \le i < Lsyn\}$ is a synchronization code with $Lsyn$ bits, where $f(i) \in \{0,1\}$. The main steps of the embedding procedure developed can be described as follows.

2.1 Pre-processing

In order to dispel the pixel space relationship of the binary watermark image, and improve the security performance of the whole digital watermark system, watermark scrambling algorithm is used at first. In our watermark embedding scheme, the binary watermark image is scrambled from W to W_1 by using Arnold transform.

In order to improve the robustness of proposed scheme against cropping and make the detector available when it loses synchronization, audio segmenting is used at first, and then, synchronization code and watermark are embedded into each segment. Let A^0 respects each segment, and A^0 is cut into two sections A_1^0 and A_2^0 with L_1 and L_2 samples respectively. Synchronization code and watermark are embedded into A_1^0 and A_2^0 respectively.

2.2 Synchronization Code Embedding

The proposed watermark embedding method proceeds as follows:

1) The audio segment A_1^0 is cut into $Lsyn$ audio segments, and each audio segment $PA_1^0(m)$ having $n = \left\lfloor L_1 / Lsyn \right\rfloor$ samples

$$PA_1^0(m) = \left\{ pa_1^0(m)(i) = a_1^0(i + m \times n), 0 \le i < n, 0 \le m < Lsyn \right\}. \tag{1}$$

2) Calculating the mean value of $PA_1^0(m)$, that is

$$\overline{PA_1^0(m)} = \frac{1}{n} \sum_{i=0}^{n-1} pa_1^0(m)(i), (0 \le m < Lsyn). \tag{2}$$

3) The synchronization code can be embedded into each $PA_1^0(m)$ by quantizing the mean value $\overline{PA_1^0(m)}$, the rule is given by

$$pa_1^{0'}(m)(i) = pa_1^0(m)(i) + (\overline{PA_1^{0'}(m)} - \overline{PA_1^0(m)}), \tag{3}$$

where $PA_1^0(m) = \{ pa_1^0(m)(i), 0 \le i < n \}$ is original sample, $PA_1^{0'} = \{ pa_1^{0'}(m)(i), 0 \le i < n \}$ is modified sample and

$$\overline{PA_1^{0'}(m)} = \begin{cases} \mathrm{IQ}(\overline{PA_1^0(m)}) \times S_1 + S_1/2, & \text{if } \mathrm{Q}(\overline{PA_1^0(m)}) = f(m) \\ \mathrm{IQ}(PA_1^0(m)) \times S_1 - S_1/2, & \text{if } \mathrm{Q}(PA_1^0(m)) \ne f(m) \end{cases}, \tag{4}$$

$$\mathrm{IQ}(\overline{PA_1^0(m)}) = \left\lfloor \overline{PA_1^0(m)} / S_1 \right\rfloor, \tag{5}$$

$$\mathrm{Q}(\overline{PA_1^0(m)}) = \mathrm{mod}(\mathrm{IQ}(\overline{PA_1^0(m)}), 2), \tag{6}$$

where $\mathrm{mod}(x, y)$ returns the remainder of the division of x by y, and S_1 is the quantization step.

2.3 Watermark Embedding

1) DWT: For each audio segment A_2^0, H-level DWT is performed, and we get the wavelet coefficients of $A_2^{0^H}, D_2^{0^H}, D_2^{0^{H-1}}, \cdots, D_2^{0^1}$, where $A_2^{0^H}$ is the coarse signal and the detail signals are $D_2^{0^1}, \cdots, D_2^{0^{H-1}}, D_2^{0^H}$.

2) DCT: In order to take the advantage of low frequency coefficient which has a higher energy value and robustness against various signal processing, the DCT is only performed on low frequency coefficient A_2^{0H}.

$$A_2^{0HC} = \mathrm{DCT}(A_2^{0H}) = \{a_2^0(t)^{HC}, 0 \le t < L_2/2^H\}. \tag{7}$$

3) Watermark Embedding: The proposed scheme embeds watermark signal bit in the magnitude of the DCT-coefficient by quantization.

$$a_2^{0'}(t)^{HC} = \begin{cases} \mathrm{IQ}(a_2^0(t)^{HC}) \times S_2 + S_2/2, & if \ \ \mathrm{Q}(a_2^0(t)^{HC}) = w_1(i,j) \\ \mathrm{IQ}(a_2^0(t)^{HC}) \times S_2 - S_2/2, & if \ \ \mathrm{Q}(a_2^0(t)^{HC}) \ne w_1(i,j) \end{cases}, \ t = (i-1) \times N + j, \tag{8}$$

where $0 \le i < M$, $0 \le j < N$, and S_2 is the quantization step, and

$$\mathrm{IQ}(a_2^0(t)^{HC}) = \left\lfloor (a_2^0(t)^{HC})/S_2 \right\rfloor, \mathrm{Q}(a_2^0(t)^{HC}) = \mathrm{mod}(\mathrm{IQ}(a_2^0(t)^{HC}), 2). \tag{9}$$

4) Inverse DCT: The Inverse DCT is performed on low frequency coefficient $A_2^{0'HC}$ as follows:

$$A_2^{0'HC} = \begin{cases} a_2^{0'}(t)^{HC}, & 0 \le t < M \times N \\ a_2^0(t)^{HC}, & M \times N \le t < L_2/2^H \end{cases}, \ A_2^{0'H} = \mathrm{IDCT}(A_2^{0'HC}). \tag{10}$$

5) Inverse DWT: After substituting the coefficients A_2^{0H} with $A_2^{0'H}$, H-level Inverse DWT is performed, and then the watermarked digital audio signal is $A_2^{0'}$.

2.4 Repeat Embedding

In order to improve the robustness against cropping, the proposed scheme repeats Sect. 2.2 and Sect. 2.3 to embed synchronization code and watermark into other segments.

3 Watermark Detecting Scheme

The watermark detecting procedure in the proposed method neither needs the original audio signal nor any other side information. The watermark detecting procedure is stated as follows.

1) Locating the beginning position B of the watermarked segment is achieved based on the frame synchronization technology of digital communications.

2) H-level DWT is performed on each audio segment $A^*(m)$ after B, and then get the coefficients as follows $A^{*H}, D^{*H}, D^{*H-1}, \cdots, D^{*1}$.

3) The DCT is performed on the low frequency DWT-coefficient A^{*H}

$$A^{*HC} = \mathrm{DCT}(A^{*H}) = \{a^*(t)^{HC}, 0 \le t < L_2^*/2^H\}. \tag{11}$$

4) The extraction rule is

$$W' = w'(i, j) = \left\lfloor a^*(t)^{HC} \Big/ S_2 \right\rfloor \bmod 2, (0 \le i < M, 0 \le j < N). \quad (12)$$

5) Finally, the watermark image $W^* = \{w^*(i, j), 0 \le i < M, 0 \le j < N\}$ can be obtained by descrambling.

4 Experimental Results and Robust Test

In order to illustrate the robust nature of our scheme, the proposed algorithm is applied to two digital audio pieces. These audio signals are music with 16 bits/sample, 44.1kHz sample rates. We use a 64×64 bit binary image as our watermark for all audio signals and a 16-bit Barker code 1111100110101110 as synchronization code. The Daubechies-1 wavelet basis is used, and 3-level DWT is performed.

Several attacks are used to estimate the robustness of our scheme.

Re-quantization. We tested the process of re-quantization of a 16-bit watermarked audio signal to 8-bit and back to 16-bit.

Re-sampling. The watermarked audio are down sampled to several different sampling rates (22.05 kHz, 11.025 kHz, 8 kHz) and then upsampled back to 44.1 kHz.

Additive noise. White noise with 10% of the power of the audio signal is added.

MP3 compression. The coding/decoding was performed using a software implementation of the MP3 coder with several different bit rates (112 kbits/s, 64 kbits/s, 48 kbits/s, 32 kbits/s.)

Table 1 summarizes the proposed watermark detection results comparing with that of scheme [10] against various attacks. In addition, the Normalized Cross-correlation (NC) are given in the Table1.

Table 1. The watermark detection results for various attacks

Attack type	The test sample 1 (NC)		The test sample 2 (NC)	
	Scheme [10]	Proposed scheme	Scheme [10]	Proposed scheme
Origin watermark	1.0000	1.0000	1.0000	1.0000
Re-quantization	1.0000	1.0000	0.9996	0.9996
Re-sampling 22.05kHz	0.4087	0.9935	0.3733	1.0000
Re-sampling 11.025kHz	0.3618	0.6624	0.3773	0.8932
Re-sampling 8kHz	0.3550	0.5595	0.3702	0.7778
Additive noise	0.9619	0.9652	0.9161	0.9212
Cropping (Front) 10%	1.0000	1.0000	1.0000	1.0000
Cropping (Middle) 10%	1.0000	1.0000	1.0000	1.0000
Mp3-112k	1.0000	1.0000	1.0000	1.0000
Mp3-64k	0.3925	0.9948	0.3619	0.9528
Mp3-48k	0.3838	0.9489	0.3638	0.9748
Mp3-32k	0.3722	0.8096	0.3667	0.9104

5 Conclusion

In this paper, we propose a synchronization digital audio watermark algorithm based on the quantization of coefficients. To improve the robustness of audio watermark, the proposed algorithm is constructed by selecting robust Barker code as synchronization code, embedding synchronization code into the mean value of several samples and embedding watermark into DWT and DCT coefficients. The experimental results have illustrated the inaudible and robust nature of our watermarking scheme.

References

1. Cox I J, Miller M L.: The First 50 Years of Electronic Watermarking. Journal of Applied Signal Processing, Vol. 56, No. 2. (2002) 225–230
2. Wei Li, Yi-Qun Yuan, Xiao-Qiang Li, Xiang-Yang Xue, Pei-Zhong Lu: Overview of Digital Audio Watermarking. Journal on Communications, Vol. 26, No. 2. (2005) 100–111
3. Wen-Nung Lie, Li-Chun Chang: Robust and High-Quality Time-Domain Audio Watermarking Subject to Psycho Acoustic Masking. In: Proceedings of IEEE International Symposium on Circuits and Systems, Vol. 2. Arizona, USA, (2002) 45–48
4. Megías D, Herrera-joancomartí J, Minguillón J.: A Robust Audio Watermarking Scheme Based on MPEG 1 Layer Ⅲ Compression. Communications and Multimedia Security - CMS 2003, LNCS 963. Springer-Verlag (2003) 226–238
5. Kim H J.: Audio Watermarking Techniques. Pacific Rim Workshop on Digital Steganography, Kyushu Institute of Technology, Kitakyushu, Japan (2003)
6. Sheng-He Sun, Zhe-Ming Lu, Xia-Mu Niu: Digital Watermarking Technique. Science Press, Beijing (2004)
7. Chung-Ping Wu, Po-Chyi Su, C.-C. Jay Kuo: Robust Audio Watermarking for Copyright Protection. In Proc. SPIE Vol. 3807, (1999) 387–397
8. Wei Li, Xiang-Yang Xue: Audio Watermarking Based on Music Content Analysis: Robust Against Time Scale Modification. In : Proceedings of the Second International Workshop on Digital Watermarking, Korea, (2003) 289–300
9. Yong Wang, Ji-Wu Huang, Yun Q Shi.: Meaningful Watermarking for Audio with Fast Resynchronization. Journal of Computer Research and Development, Vol. 40, No. 20. (2003) 215–220
10. Ji-Wu Huang, Yong Wang, Shi Y. Q.: A Blind Audio Watermarking Algorithm with Self-Synchronization. In: Proceedings of IEEE International Symposium on Circuits and System, Arizona, USA, Vol. 3. (2002) 627–630

Semi-fragile Watermarking Algorithm for Detection and Localization of Temper Using Hybrid Watermarking Method in MPEG-2 Video

Hyun-Mi Kim, Ik-Hwan Cho, A-Young Cho, and Dong-Seok Jeong

Dept. of Electronic Engineering, Inha University,
253 Yonghyun-Dong, Nam-Gu, Incheon 402-751, Republic of Korea
{chaos_th, teddydino, ayoung}@inhaian.net, dsjeong@inha.ac.kr

Abstract. In this paper, a novel semi-fragile watermarking adapted to MPEG-2 video is proposed. It is achieved by using the hybrid watermarking method which is implemented by the combination of the robust watermark applied in DCT and spatial domain and the fragile watermark applied in motion vector. The proposed method can not only detect attacks but also localize them. Besides, it can distinguish malicious attacks such as frame dropping or swapping from non-malicious tempers like re-compression. This method satisfies invisibility and does not need the original information for the watermark detection.

1 Introduction

With the development of digital technology, various digital contents are widely generated, distributed and stored. However, rapid development of digital contents and infra structures also raised many problems such as the protection of copyrights or authentication of contents. The importance of copyright protection or authentication of digital contents has been increased due to the characteristics of digital contents such that it could be copied easily and the copy is identical to the original. Therefore, technologies for copyright protection or authentication are essential and various watermarking algorithms are developed.

Many researchers have studied robust watermarking algorithms. They are mainly focused on copyright protection and identification which requires very high robustness. On the other hand, the fragile or semi-fragile watermarking is developed for authentication purpose. While the use of the fragile watermarking is just for authentication, the semi-fragile watermarking can distinguish malicious attacks from non-malicious manipulations and localize the attacks in addition to the authentication.

There have been studies on semi-fragile watermarking for detection and localization of temper on the image [1] and videos [2, 3]. Their target was mostly object and block in the image or videos for localization. However, localization of temporal attack hasn't been considered even if video has temporal characteristics. Therefore, a new semi-fragile watermarking algorithm that targeted temporal property is proposed by using the hybrid watermarking method which is the combination of various watermarking algorithms in this paper.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 623 – 628, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Watermark Embedding Scheme

The proposed watermarking method is the combination of three watermarking algorithms. Firstly, robust watermarks are embedded into DCT coefficients of intra frame and into spatial domain of inter frame, respectively. This kind of watermarks is used to distinguish the malicious attacks from non-malicious ones and to localize the attacks. Secondly, the fragile watermark is embedded into motion vectors of inter frame. The motion vector is changed so easily that it is proper to a fragile watermarking. This watermarking works for contents authentication.

2.1 Embedding Robust Watermark in DCT Coefficients of Intra Frame

This embedding scheme is for distinguishing intentional attacks from unintentional ones and localizing the attacks at GOP (Group of Pictures)-level. Watermark bits are embedded in quantized AC coefficients of 8×8 luminance blocks within intra frame. To support the robustness against re-compression, two principles are considered. Firstly, watermark signals must be strong enough to survive for the quantization in compression process. Secondly, a watermark bit must be embedded in one of quantized AC coefficients in high frequency along the diagonal positions (i.e., u=v in Eq. (1)) to minimize the watermark detection errors [4].

$$X_q^* = \begin{cases} \max\{X_q(u,v), M_q(u,v)\} & \text{if } w_n = 1 \\ 0 & \text{if } w_n = 0 \end{cases} \quad (1)$$

Where w_n is the nth watermark bit to be embedded, $M_q(u,v)$ is the watermark signal, $X_q(u,v)$ is the quantized AC coefficient in the position (u,v), and X_q^* is the watermarked coefficient. If the embedded bit is '0', the AC coefficient is replaced by '0'. Because most of the AC coefficients have the value of '0', this process may not affect video quality much.

2.2 Embedding Robust Watermark in Spatial Domain of Inter Frame

This embedding scheme is for distinguishing malicious attacks from non-malicious manipulations and localizing the malicious attacks applied in inter frame.

First of all, blocks in which watermark to be embedded must be selected from inter frames. For this, the differences of DC coefficients of the corresponding 8×8 blocks in the two successive frames are calculated and blocks with the difference over predefined threshold are selected. Because the DC coefficient means the average (or mean) value of the block and represents basic characteristics of the block, this difference can be used as a motion-detection criterion [5].

When the watermarks are embedded into selected blocks, a pre-selected bit-plane of every pixel in selected block is replaced with a watermark bit. Therefore, the embedded watermark bit is only one bit for one block. Since pixel values can be changed easily in inter frame, only one watermark bit is inserted in several blocks of one frame for robustness against the lossy compression. In addition, watermark bits are embedded in GOP-level and this scheme localizes malicious attacks of frame-level within GOP. Fig. 1 describes the block selection scheme and the watermark embedding scheme respectively.

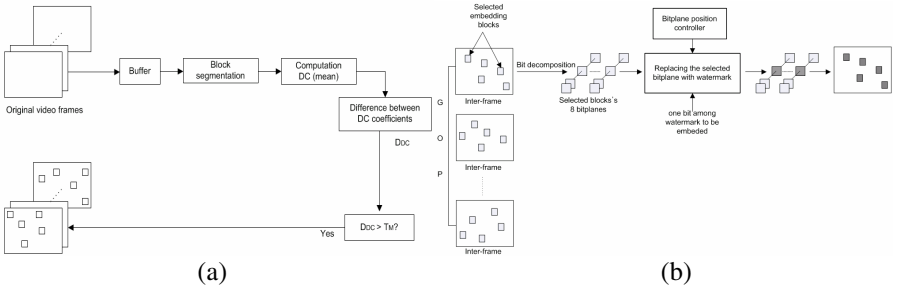


Fig. 1. Diagram of (a) block selection process for embedding the watermark bit and (b) watermark embedding procedure in inter frame

2.3 Embedding Fragile Watermark into Motion Vector of Inter Frame

Because motion vector is changed easily even for trivial attack, it is appropriate target for fragile watermarking. Therefore, the usage of motion vector information in fragile watermarking makes it possible to judge whether the content is authenticated and its integrity is verified.

In this paper, the fragile watermarking using motion vector is based on two principles [6]. One is to embed watermarks in blocks with large motion vector magnitude. The other is to embed watermarks in motion vector components having less change of phase angle after watermark embedding. These principles are for minimizing visual difference between the original video and the watermarked video.

This embedding scheme uses the method proposed in [6]. Therefore, two principles mentioned above is used for selecting of embedding component of motion vector and following method is used for watermark embedding.

If MVC[j] mod 2 ≠ mark[k], then MVC*[j] = MVC[j] + d;
If MVC[j] mod 2 = mark[k], then MVC*[j] = MVC[j].

Where MVC[j] is selected component of motion vector, MVC*[j] is the watermarked component, mark[k] is the kth watermark bit to be embedded, and d is odd integer, and 'mod' means modular operator.

In Addition, one of the important considerations is that motion vector has a search range. After d is added, MVC*[j] must be in search range. Therefore, d is set up properly in boundary of search range.

3 Watermark Detection Scheme

Watermark detection scheme is simply implemented by considering the embedding procedure. So, detection scheme is composed of three sub-schemes.

Firstly, robust watermark in DCT coefficients of intra frame is easily extracted by checking the selected AC coefficient. If the coefficient is '0' then watermark is also '0', '1' otherwise.

Secondly, robust watermark in spatial domain of inter frame is extracted directly from pre-determined bit-plane of pixel values in selected block after selecting blocks

like embedding scheme. Because the watermark is embedded in several blocks in a frame, the watermark is calculated into the mean of extracted watermark bits in a frame. This procedure is repeated in every frame, and the watermarks can be detected as many as GOP size. Also, this process is used to increase the detection rate as mentioned in the embedding scheme.

Finally, fragile watermark in motion vector of inter frame is extracted by selecting motion vector and its component using the same method as for embedding scheme and modular operating the selected motion vector component.

4 Experiments and Results

The proposed watermarking in this paper has semi-fragile watermarking characteristics. Therefore, the proposed method must support functions including authentication, verification of integrity, distinction between malicious and non-malicious attacks, localization of the attacks, etc. Especially, two attack categories are used in this experiment to distinguishing malicious attacks from non-malicious attacks. Recompression for non-malicious attack and frame dropping and swapping for malicious attack are used. Because re-compression is usually used in such cases as transcoding or changing the bit-rate, etc. All experiments have been conducted on the video sequence Foreman (QCIF) and Carphone (QCIF). The total frame numbers used in this experiment is 100 and GOP size is set up as 15 frames. Table 1 shows PSNR of each video when the original video and the watermarked video is compressed.

Table 1. The comparison of PSNR between the original video and watermarked video

	The original video	The watermarked video
Foreman/Carphone	42.96644/45.63685 (dB)	42.48967/45.53592 (dB)

The PSNR between the original video and the watermarked video is represented in Table 1. Consequently, this proposed method satisfies the requirement of invisibility.

Table 2. The detection result of the robust watermarking applied into intra frame (Foreman / Carphone)

	NoAttack (%)	Recompression (%)	FrameDropping (%)	FrameSwapping (%)
Frame 0	100 / 100	100 / 100	100 / 100	100 / 100
Frame 15	100 / 100	100 / 100	58 / 55	100 / 100
Frame 30	100 / 100	91 / 96	53 / 49	89 / 97
Frame 45	100 / 100	96 / 98	49 / 45	95 / 90
Frame 60	100 / 100	98 / 91	47 / 47	98 / 93
Frame 75	100 / 100	100 / 100	54 / 53	100 / 100
Frame 90	100 / 100	96 / 100	49 / 49	89 / 98

The characteristics of the fragile and semi-fragile watermarking are proved through the robustness. Table 2 represents the detection results of the robust watermarking applied into intra frame. Table 2 shows that the watermarking into intra frame is robust for the most of attacks. However, the detection rate of extracted watermark is decreased largely after a frame is dropped (fifth frame in this experiment). Therefore, localization of frame dropping attack is possible in GOP-level.

Fig. 2 illustrates the detection result of the robust watermarking applied into inter frames. Fig. 2 is the graph that presents the ratio of correctness of extracted bits from the (a) no-attacked video, (b) re-compressed video, (c) frame dropped video, and (d) frame swapped video respectively. This graph denotes extracted watermarks from one GOP only.

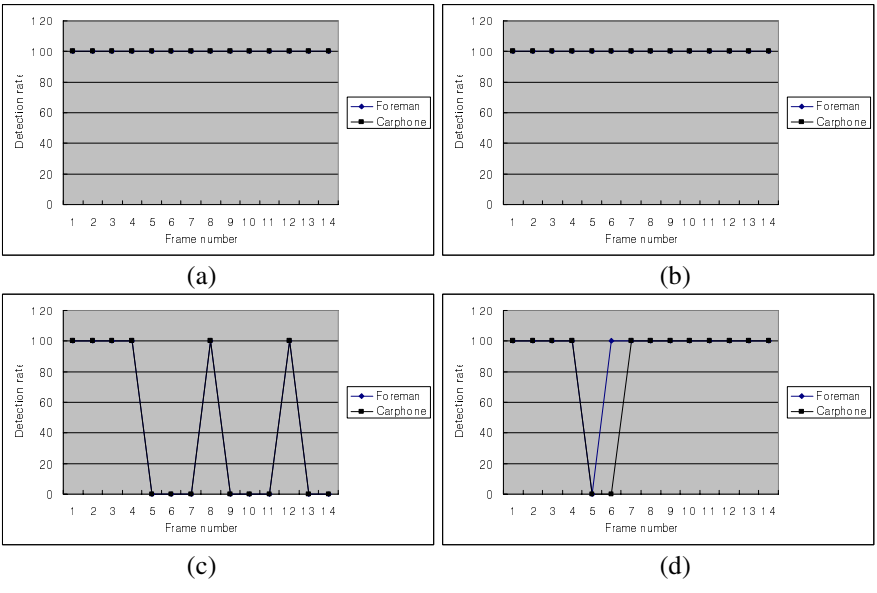


Fig. 2. The detection result of the robust watermark applied into inter frame for videos with (a) no attack, (b) re compressed, (c) frame dropped (frame 5) and (d) frame swapped (frame 5 and 6). In most of the cases, the results of two videos are same or similar.

In Fig. 2, the watermark is robust against re-compression attack, but the watermark isn't extracted from the same position as attacked in frame-level. From this result, it is possible to distinguish the malicious attack from the non-malicious one and localize the attack. Table 3 denotes the detection result of the fragile watermark applied into inter frame.

As shown Table 3, if the original video is changed, the fragile watermark is not ex-racted well. Therefore, this can be used for authentication and verification of integrity.

Table 3. The detection result of the fragile watermark applied into inter frame. This result is the average of detection rate(Foreman/Carphone).

NoAttack (%)	Recompression (%)	FrameDropping (%)	FrameSwapping (%)
100/100	56.25/62.5	68.75/65	68.75/65

5 Conclusion

We present a novel semi-fragile video watermarking scheme using the hybrid watermarking method which is the combination of robust DCT spatial domain watermarking and the fragile motion vector-based watermarking. The experimental results show that the proposed algorithm is capable of doing the authentication and verification of integrity, the differentiation of malicious and non-malicious attack and localization of attack. Although many works have been proposed for localization of attacks, they were mainly adopted in object or block level. However, frame-level attack should be considered because video has temporal characteristic. We consider frame-level attack for localization. Though the proposed watermarking scheme has many advantages, the attacks used in this paper are somewhat restrictive. Therefore, our future works will be focused on the study for the robust algorithm against many other attacks.

Acknowledgement

This work was supported by INHA UNIVERSITY Research Grant.

References

1. Jiri Fridrich : A hybrid watermark for temper detection in digital images. Symposium on Signal Processing and its Applications (1999)
2. Jae Yeon Park, Jae Hyuck Lim, Gyung Soon Kim, Chee Sun Won : Invertible semi-fragile watermarking algorithm distinguishing MPEG-2 compression from malicious manipulation. Digest of Technical Papers of IEEE International Conference on Consumer Electronics (2002)
3. Tsong-Yi Chen, Chien-Hua Huang, Thou-Ho Chen, Cheng-Chieh Liu : Authentication of lossy compressed video data by semi-fragile watermarking. Proceeding of IEEE International Conference on Image Processing (2004)
4. Gang Qiu, Pina Marziliano, Anthony T.S. Ho, Dajun He, Qibin Sun : A Hybrid watermarking scheme for H.264/AVC video. Proceeding of IEEE International Conference on Pattern Rec-ognition (2004)
5. Qing-Ming Ge, Zhe-Ming Lu, Xia-Mu Niu : Oblivious video watermarking scheme with adaptive embedding mechanism. Proceeding of IEEE International Conference on Machine Learning and Cybernetics (2003)
6. Jun Zhang, Jiegu Li, Ling Zhang : Video Watermark Technique in Motion Vector. Proceeding of IEEE Symposium on Computer Graphics and Image Processing (2001)

Public Watermarking Scheme Based on Multiresolution Representation and Double Hilbert Scanning

Zhiqiang Yao, Liping Chen, Rihong Pan, Boxian Zou, and Licong Chen

Department of Computer Science, Fujian Normal University,
Fuzhou 350007, China
{yzq, lpchen, rhpan}@fjnu.edu.cn

Abstract. A novel robust watermarking algorithm is presented using multiresolution representation of quasi-uniform cubic B-spline curves. The 4×4 blocks of host image are reindexed into 1-D in Hilbert scanning order. Every two neighbor blocks from Hilbert sequence are selected, and part pixels in this two blocks are taken as control points by Hilbert scanning order to define a couple of quasi-uniform cubic B-spline curves. The wavelet-based multiresolution representations of quasi-uniform B-spline curves are carried out and the low-resolution control points of quasi-uniform B-spline curves are modified according to a binary watermark. Experimental results show that this scheme is strongly against the attacks of filtering, JPEG compression and translation.

1 Introduction

With the development of the communication networks, more and more digital products are spreading in the Internet. The reproduction and distribution of the digital product are becoming extremely easier and faster. As a result, the problems of protecting copyrighted digital media and information security become more urgent. The solution counts on the digital watermarking techniques that imperceptibly and robustly embed a special symbol into the digital product to mark ownership of it and authenticate it, ensuring the integrity and reliability of the digital information by detecting and analyzing the watermark. Thusly it becomes an effective means of the knowledge copyright protection and warding off the fake.

Current techniques for the image watermarking can be broadly classified into two categories: (i) spatial domain methods which embed the data by directly modifying the pixel values of the original image [1], (ii) transform domain methods which embed the data by modulating the coefficients of a properly chosen transform domain such as DCT [2], DFT [3] ,wavelet [4,5,6] and DWT-DFT composite[7]. Watermarking techniques can be alternatively split into two distinct categories depending on whether the original image is necessary for the watermark extraction or not. Although the existence of the original image facilitates a great extent to watermark extraction ([2,5]), owner of the original image is compelled unsecurely to share his works with anyone who wants to check the existence of the watermark.

Parameteric B-spline curve is one of components of the shape mathematics description and it provides an important math modeling implement for a large amount of geometric modeling problems. Wavelet analysis is widely applied in many fields. The

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 629 – 634, 2005.
© Springer-Verlag Berlin Heidelberg 2005

combination of geometric modeling and wavelet analysis facilitates to curve and surface editing,smoothing and progressive transmission. Naturely applying them in image watermarking, a novel digital image watermarking algorithm is proposed in this paper.

2 The Theoretical Base

Assume a quasi-uniform cubic B-spline curve,denoted as $\gamma(t)$, has 2^j+3 control points. $\gamma(t)$ can be decomposed as a low-resolution part $\gamma_{j-1}(t)$ and a detail part $\beta_{j-1}(t)$. Both $\gamma_{j-1}(t)$ and $\beta_{j-1}(t)$ are quasi-uniform cubic B-spline curves. Let C_j be the control points of $\gamma(t)$ and C_{j-1} be the control points of $\gamma_{j-1}(t)$, the procedure of decomposing $\gamma(t)$ is equivalent to that of splitting C_j into a low-resolution part C_{j-1} and a detail part D_{j-1}(To distinguish D_{j-1} from the control points, we call D_{j-1} the vertex of $\beta_{j-1}(t)$). The relationship between C_j, C_{j-1} and D_{j-1} can be represented as[9]:

$$C_j=P_jC_{j-1}+Q_jD_{j-1}\ . \tag{1}$$

Where P_j and Q_j are the reconstructing matrix.

When we decompose the quasi-uniform cubic B-spline curves with the multiresolution techniques, it is necessary to calculate C_{j-1} from known C_j by switching (1) to

$$[P_j|Q_j][C_{j-1}{}^T|D_{j-1}{}^T]^T= C_j\ . \tag{2}$$

Here ^{T} is matrix transfer. As long as we solve the linear equations, we can obtain the unique solutions C_{j-1} and D_{j-1}, i.e., split C_j into the low-resolution part C_{j-1} and the detail part D_{j-1}. A solution with less complexity has been proposed in literature [9].

Because the chaos pseudo-random set is used in next section, we give a brief introduction. Suppose $U \subset R$(R is a real space), $f: U \to U$ is a chaos mapping

$$x_{k+1} = f(x_k, \lambda), x_k \in U, \lambda \in R\ .$$

Where k=1,2,… denotes the iterative times, λ denotes the parameter of the chaos control system. Given a initial real x_1, produce the chaos sequence by the chaos mapping and change it to the binary sequence, then acquire m bits of the sequence in turn to obtain the decimal chaos pseudo-random set $\{a_i \mid 0 \le a_i < 2^m, i = 0,1,\cdots\}$ until the number of the set comes up to the indispensable number n.

3 Proposed Watermarking Algorithm

According to Cox *et al.* [2], watermark should be embedded in the DC and low frequency AC coefficients in DCT domain due to their large perceptual capacity. The strategy can be extended to DWT domain. We embed informative watermark into the low resolution part of B-spline to make it more robust while still keeping the watermark invisible. The proposed watermark-embedding algorithm is described as follows.

Step1. Take Arnold transform on the $N*N$ binary watermark to get the scramble image,then turn it into one-dimensional vector by the way of progressive scanning:

$$W = \{w(i) \mid w(i) \in \{0,1\}, 0 \le i \le L-1\} \ .$$

where $L=N*N$, denoting the size.

Step2. Because the Hilbert scanning can keep the regional correlation of the image[8], we divide the original image I of size $M_1 \times M_2$ into blocks of size 4×4. After reindexing all pixels of the every 4×4 blocks by the Hilbert scanning order, Hilbert sequence H is obtained as

$$H = \{H(i) \mid i = 0,1,\cdots, M_1 \times M_2 / 16 - 1\} \ .$$

Step3. Given one chaos initial value(key K), by iterating k times we construct a chaos pseudo-random set defined as A containing 11 elements:

$$A = \{a_i \mid 0 \le a_i < 2^4, i = 0,1,\cdots\} \ .$$

where set m=4, n=11.

Furthermore select one positive even number R and a positive integer T_0<R/2 and set T_1 as T_1= T_0+R/2 ; Then determine $\tau = 0$ and denote the certain watermark bit which is to embed as $w(\tau)$.

Step4. Select the neighboring blocks of the size 4×4 such as $H(2\tau)$ and $H(2\tau+1)$ from H and reindex the pixels of these two blocks respectively by the Hilbert scanning into the one-dimensional vectors $\mathrm{C}^{(1)}$ and $\mathrm{C}^{(2)}$, where $\mathrm{C}^{(s)}$ is written as

$$\mathrm{C}^{(s)} = (v_0^{(s)}, v_1^{(s)}, \cdots, v_{15}^{(s)})^T, (s=1,2) \ .$$

Where $v_i^{(s)}$ denotes the gray-level value of the pixel i+1 after the Hilbert scanning.

Step5. From vector $\mathrm{C}^{(s)}$ we select the components whose subscripts belonging to Set A and construct a new vectors by these components, also marked as $\mathrm{C}^{(s)}$:

$$\mathrm{C}^{(s)} = (v_0^{(s)}, v_1^{(s)}, \cdots, v_{10}^{(s)})^T, (s=1,2) \ .$$

Take $\mathrm{C}^{(s)}$ as the control points and let $(0,0,0,0,1/2^j, \cdots, 1-1/2^j, 1,1,1,1)$ be the knot vector(j=3), two quasi-uniform cubic B-spline curves $\gamma^{(s)}(t)(0 \le t \le 1)$ can be defined. After running decomposition algorithm twice, each curve of just 5 control points is reached. We denote the low-resolution parts as $\mathrm{C}_{j-1}^{(s)} = (v_0^{(s)}, v_1^{(s)}, \cdots, v_4^{(s)})^T$ and the detail parts denoting as $\mathrm{D1}_{j-1}^{(s)} = (d_0^{(s)}, d_1^{(s)}, \cdots, d_3^{(s)})^T$ and $\mathrm{D2}_{j-2}^{(s)} = (d_0^{(s)}, d_1^{(s)})^T$ (s=1,2).

Step6. Set $\mu_s = \sum_{i=0}^{2^{j-2}+2} v_i^{(s)}, (s=1,2), \mu_{dif} = (\mu_1 - \mu_2) \bmod R$. If $w(\tau)$=1 then compute σ as

$$\sigma = \begin{cases} (T_1 - \mu_{dif})/2, & if \quad \mu_{dif} \geq T_0 \\ (T_1 - \mu_{dif} - R)/2, & if \quad \mu_{dif} < T_0 \end{cases} . \tag{3}$$

However if $w(\tau)=0$ then compute σ as

$$\sigma = \begin{cases} (T_0 - \mu_{dif})/2, & if \quad \mu_{dif} \leq T_1 \\ (T_0 - \mu_{dif} + R)/2, & if \quad \mu_{dif} > T_1 \end{cases} . \tag{4}$$

Furthermore $v_i^{(1)} = v_i^{(1)} + \sigma, v_i^{(2)} = v_i^{(2)} - \sigma, i = 0,1,\cdots,2^{j-2}+2$ should be set.

Step7. Reconstruct the new control points $C^{(s)}$ with $C_{j-1}^{(s)}$, $D1_{j-1}^{(s)}$ and $D2_{j-1}^{(s)}$,(s=1,2), which means that a bit of watermark has been embedded into the two neighboring blocks $H(2\tau)$ and $H(2\tau+1)$.

Step8. Increment τ by 1, and repeat from step4 to step7 until all bits of the watermark have been embedded. Finally make use of the inverse-map of the Hilbert scanning to map each pixel back to the original position before the Hilbert scanning.

To go further, the wavelet decompose and reconstruct algorithms of Step5 and Step7 are just mentioned as section 2.

In extracting the watermark, the original image is not required in our algorithm. However, the watermarked image might be subject to some intentional or unintentional attacks, and the resulting image after attack is represented by I'. In the watermark extraction, we process from step2 to step5 as mentioned above. Then we set the parameters as: $\mu_s = \sum_{i=0}^{2^{j-2}+2} v_i^{(s)}$ (s=1,2) and $\mu_{dif} = (\mu_1 - \mu_2) \bmod R$, the $\tau - th$ bit of the extracting watermark can be determined by

$$w'(\tau) = \begin{cases} 1 & if \mid \mu_{dif} - T_1 \mid < R/4 \\ 0 & else \end{cases} . \tag{5}$$

Every $w'(\tau)$ can be gained by (5) from every pair of selected neighboring blocks. Finally we take the Arnold transform to acquire the extracted watermark W'.

4 Experimental Results and Conclusions

In our experiments, we take the trademark, shown in Fig.1(a), as the watermark W. It is a binary image with size 64*64. The watermark W has undergone Arnold transform, shown in Fig.1(b), and been formed as a bit sequence. In order to evaluate the performance of watermarking techniques, similarity degree (ρ) and peak signal to noise (PSNR) are taken as two common quantitative indices. Similarity degree is

$$\rho = 1 - \sum_{\tau=0}^{L-1} \mid w(\tau) - w'(\tau) \mid / L .$$

Some necessary parameters used in our watermark algorithm include R=40, T_0=8, T_1=23. The proposed watermarking scheme is tested on the three popular images: 512×512 Lena, Boat, Pepper. The PSNRs are exhibited in Table 1, i.e., the water-

marks are perceptually invisible. Several types of attack are simulated to evaluate the performance of robustness, including motion blurring, filtering, sharpening, JPEG compression, scaling, and translation. Table 2 shows the quantitative measures for our method. Just with Pepper image as an example, the extracted watermarks under different attack are illustrated in Fig.2(a)-(f). The extracted watermarks survive well even under these attacks by using the proposed algorithm.

Generally, the alteration of the coefficients in the frequency domain leads to the alteration of the pixel value in the spatial domain after the inverse wavelet transform. As a result the visual quality of the image may be influenced. From the formula (3) and (4), we can easily draw a conclusion as $\sigma < R/4$, i.e., the alteration between pre-embedding and post-embedding ranges at (-*R*/4,*R*/4). Accordingly the imperception unobtrusiveness of the watermark is closely relation to the value of *R*. The



(a) (b)

Fig. 1. (a)The watermark; (b) The watermark that has undergone Arnold transformation

Table 1. The PSNR of the marked images versus the original images

image	PSNR
Pepper	39.903
Lena	39.898
Boat	40.015

Table 2. The quantitative indices under several attacks

Attacks	image	ρ	Attacks	image	ρ
Attack-free	Pepper	1.0000	2-pixel horizontal translation	Pepper	0.9270
	Lena	1.0000		Lena	0.9565
	Boat	1.0000		Boat	0.9724
3-pixel motion blurring	Pepper	0.9355	JPEG(Q=50%)	Pepper	0.8101
	Lena	0.9365		Lena	0.8000
	Boat	0.9236		Boat	0.8071
Sharpening	Pepper	0.8655	Shrink 25%	Pepper	0.9092
	Lena	0.9241		Lena	0.9287
	Boat	0.9299		Boat	0.9226
2× 2 median filtering	Pepper	0.8982	Amplifying 25%	Pepper	0.9099
	Lena	0.9302		Lena	0.9575
	Boat	0.9351		Boat	0.9387



(a) (b) (c) (d) (e) (f)

Fig. 2. The extracted watermarks with Pepper image under (a) motion blurring; (b) Sharpening; (c) median filtering; (d) horizontal translation; (e) JPEG(Q=50%); (f) Shrink by 25%;

increment of R lessen imperception of watermark at the same time. Nevertheless the increment of R will improve the robustness of the watermark. Consequently by R we would find a reasonable balance between the robustness and imperception.

Above all a robust algorithm for spline-wavelet-based watermarking has been presented in this paper. It provides greater robustness by use of the multiresolution representation of quasi-uniform B-spline curves, which is similar to embedding the watermark after weeding out the noises from the original image. Other advantages of the proposed algorithm include: (1)Taking advantage of the regional correlation of the image. (2)Less algorithm complexity, just concerning with the multiplication of the low-dimension matrix and vector, together with solving band matrix equations. (3)No requiring the original image for watermark extraction.

Acknowledgements

Authors appreciate the support received from NSF of Fujian (A0510012, A0210016), Funding of Fujian Provincial Education Department (JA03028).

References

1. Liu, J.C., Chen S. Y.: Fast Two-layer Image Watermarking Without Referring To The Original Image And Watermark. Image and Vision Computing. 19(2001) 1083～1097
2. Cox, I.J., Kilian, J., Leighton, F.T. Shamoon, T.: Secure Spread Spectrum Watermarking For Multimedia. IEEE Transactions on Image Processing. 12 (1997) 1673–1687
3. Pereira, S. ,Pun, T.: Robust Template Matching For Affine Resistant Image Watermarks. IEEE Transactions on Image Processing. 6 (2000) 1123-1129
4. Barni, M., Bartolini, F., Piva, A.: Improved Wavelet-based Watermarking Through Pixel-wise Masking[J]. IEEE Transaction on Image Processing. 5 (2001) 783-791
5. Chen,L.H., Lin, J.J.: Mean Quantization Based Image Watermarking. Image and Vision Computting. 21 (2003) 717–727
6. Yao, Z.Q.,Pan, R.J., Zou, B.X., *et al*.: Digital Watermarking Scheme Using B-Spline Curves, Journal of Computer-aided Design & Computer graphics. 17(7) (2005) 1463-1469
7. Kang, X.G., Huang, J.W., *et al*.: A DWT-DFT Composite Watermarking Scheme Robust to Both Affine Transform And JPEG Compression. IEEE Transaction on Circuits and Systems For Video Technology. 8 (2003) 776-786.
8. Gostman, C., Lindenbaum, M.: On The Metric Properties Of Discrete Space-Filling Curves. IEEE Transaction On Image Processing. 4(1996) 794-797.
9. Adam Finkelstein, David H. Salesin.: Multiresolution Curves.Computer Graphics Proceedings, Annual Conference Series, ACM Siggraph,Orlando,Florida. (1994) 261-268.

Performance Evaluation of Watermarking Techniques for Secure Multimodal Biometric Systems

Daesung Moon[1], Taehae Kim[2], SeungHwan Jung[2], Yongwha Chung[2], Kiyoung Moon[1], Dosung Ahn[1], and Sang-Kyoon Kim[3]

[1] Biometrics Technology Research Team, ETRI, Daejeon, Korea
{daesung, kymoon, dosung}@etri.re.kr
[2] Department of Computer and Information Science, Korea University, Korea
{taegar, sksghksl, ychungy}@korea.ac.kr
[3] School of Computer Engineering, Inje University, Kimhae, Korea
skkim@cs.inje.ac.kr

Abstract. In this paper, we describe various watermarking techniques for secure user verification in the remote, multimodal biometric systems employing both fingerprint and face information, and compare their effects on user verification and watermark detection accuracies quantitatively. To evaluate the performance of watermarking for multimodal biometric systems, we first consider possible two scenarios – embedding facial features into a fingerprint image and embedding fingerprint features into a facial image. Additionally, to evaluate the performance of dual watermarking for secure biometric systems, we consider another two scenarios – with/without considering the characteristics of the fingerprint in embedding the dual watermark. Based on the experimental results, we confirm that embedding fingerprint features into a facial image can provide superior performance in terms of the user verification accuracy. Also, the dual watermarking with considering the characteristics of the fingerprint in embedding the dual watermark can provide superior performance in terms of the watermark detection accuracy.

1 Introduction

Traditionally, verified users have gained access to secure information systems via multiple PINs, passwords, and so on. However, these security methods have important weakness that can be lost, stolen, or forgotten. In recent years, there is an increasing trend of using **biometrics**, which refers the personal biological or behavioral characteristics used for verification[1].

In general, biometric verification/identification system can be divided into two modes: a **unimodal** biometric system which uses only a single biometric characteristic and a **multimodal** biometric system which uses multiple biometric characteristics. Building multimodal biometric systems has become an important research trend to overcome several limitations of unimodal biometric systems such as unacceptable identification performance in **large-scale** applications[1-3]. In this paper, both **fingerprint** and **face** are chosen as the biometric characteristics for our multimodal biometric system.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 635 – 642, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Another issue related with biometrics for **remote applications** is how to protect the biometric information securely from unauthorized accesses. As mentioned above, biometric techniques have inherent advantages over traditional personal verification techniques. However, if biometric data(biometric raw data or biometric feature data) have been compromised, a critical problem about confidentiality and integrity of the individual biometric data can be raised. For verification systems based on physical tokens, a compromised token can be easily canceled and the user can be assigned a new token. On the contrary, the biometric data cannot be changed easily since the user only has a limited number of biometric features such as one face and ten fingers.

To provide secrecy and privacy of the biometric data in remote applications, several techniques are possible such as cryptography and digital watermarking. The straightforward approach to guarantee the confidentiality/integrity of the biometric data is to employ the standard cryptographic techniques[4]. Unfortunately, encryption does not provide secrecy once data is decrypted.

Digital watermarking[5-6] can be considered as method for complement to cryptography. Digital watermarking places information, called *watermark*, within the content, called *cover work*. Since watermarking involves embedding information into the host data itself, it can provide secrecy even after decryption. The watermark, which resides in the biometric data itself and is not related to encryption-decryption operations, provides another line of defense against illegal utilization of the biometric data[7]. As in [7], we will focus on watermarking only in the remaining of this paper to provide secrecy and privacy of the biometric data.

In general, traditional digital watermarking techniques may degrade the quality of the content, regardless of the difference between the original and watermarked versions of the cover work. Since watermarking techniques used to increase the security of the biometric data affects the verification accuracy of the biometric system due to that quality degradation, the effects on biometric verification accuracy need to be considered in applying the watermarking techniques. Also, watermarking techniques generally can be classified into *robust watermarking* and *fragile watermarking* according to the embedding strength of the watermark. Watermarks embedded by a robust watermarking technique cannot be lost by possible attacks. Watermarks embedded by a fragile watermarking technique, however, can be lost easily by embedding some other information. By using these characteristics, **dual watermarking** techniques can be used to guarantee the integrity of information transmitted with the fragile watermarking, and to hide some important information with the robust watermarking.

In this paper, we first consider possible two scenarios for multimodal biometric systems and evaluate the effects of each scenario on biometric verification accuracy. In the Scenario 1, we use the fingerprint image as the cover work and hide the facial information into it. On the contrary, we hide the fingerprint information into the facial image in the Scenario 2. After implementing each scenario, we measure the effects of watermarking on biometric verification accuracy with various parameters. In addition to this evaluation, we consider another two scenarios to evaluate the performance of the dual watermarking technique in terms of the effect of embedding the fragile watermark on the detection ratio of the robust watermark. For this evaluation, we

select the fingerprint image used in the Scenario 1 as a cover work for the dual watermarking. In embedding the dual watermarks into a cover work, the Scenario 3 considers the characteristics of the cover work to improve the detection ratio of the watermark. On the other hand, the Scenario 4 does not consider the characteristics of the fingerprint.

The rest of the paper is structured as follows. Section 2 explains the overview of previous biometric watermarking techniques. Section 3 describes the two scenarios to evaluate the watermarking techniques for the multimodal biometric system, and another two scenarios to evaluate the dual watermarking techniques for the secure biometric system in the Section 4. The results of performance evaluations are described in Section 5. Finally, conclusions are given in Section 6.

2 Watermarking Techniques for Biometric Systems

Traditionally, digital watermarking is a technique that hides a secret digital pattern in a digital image or data. However, Yeung and Pankanti[8] proposed a fragile invisible digital watermarking of fingerprint images and Jain and Uludag[7] argued that when the feature extractor sends the fingerprint features to the matcher over an insecure link, it may hide the fingerprint features in a cover work whose only purpose is to carry the fingerprint feature data.

As mentioned above, most previous researches about biometric watermarking use fingerprint image as cover work in order to increase integrity of the fingerprint image or hide a fingerprint feature data into unrelated cover work to guarantee the confidentiality of the fingerprint image. In addition to this capability of the watermarking techniques, biometric watermarking must consider a possible degradation of the verification accuracy since embedding the watermark may change the inherent characteristics of the cover work. Ideal solutions of a digital watermark have been proposed in many papers. Although each different watermarking technique has specific algorithm for the watermark embedding, a general form of the watermark embedding algorithm can be summarized by the following equation

$$I_{WM}(x, y) = I(x, y) + k * W \tag{1}$$

where $I_{WM}(x, y)$ and $I(x, y)$ are values of the watermarked and original pixels at location (x, y), respectively. The value of watermark bit is denoted as W, and watermark embedding strength is denoted as k.

3 Watermarking Techniques for Multimodal Biometric Systems

In the multimodal biometric system considered, two biometric data(fingerprint and facial images) are acquired from biometric sensors.

To hide biometric data using a fingerprint and a face, we first consider possible two scenarios. In the Scenario 1, we use a fingerprint image as the cover work and

hide the facial information into the fingerprint image. That is, the facial features(e.g., Eigenface coefficients) are extracted to be used as watermarks. Finally, the embedding site embeds the facial features into the fingerprint image. After detecting the facial features from the received watermarked fingerprint image, the detecting site computes a similarity between the facial features stored and the facial features extracted. The detecting site also executes the fingerprint verification module with the watermarked fingerprint image received. The results of face and fingerprint verification modules can be consolidated by various fusion techniques such as weighted sum to improve the verification accuracy of unimodal biometric systems.

Note that, in the Scenario 1, the accuracy of the face verification module is not affected by watermarking, whereas the accuracy of the fingerprint verification module can be affected by watermarking. To minimize the degradation of the accuracy of the fingerprint verification module, we use the embedding method proposed by Jain and Uludag[7]. That is, possible pixel locations for embedding are determined first by considering either minutiae or ridge information of the fingerprint image. Then, a bit stream obtained from the Eigenface coefficients is embedded into the fingerprint image according to the possible pixel locations. This watermarking method, however, has several disadvantages. Because it is a type of *informed* watermarking, the minutiae or the ridge information is also needed in the detecting site.

In the Scenario 2, the fingerprint features(e.g., minutiae) used as watermarks are extracted at the embedding site. Then, the minutiae are embedded into a facial image(cover work). After extracting the minutiae from the received watermarked image, the detecting site computes a similarity between the minutiae stored and the minutiae extracted. The detecting site also executes the face verification module with the received facial image.

We use our fingerprint verification algorithm[9] to evaluate the effects on fingerprint verification accuracy with the watermarked fingerprint image. Also, to evaluate the effects on face verification accuracy with the watermarked facial image, face verification algorithm used is Principal Component Analysis(PCA)[10]. Results of evaluation that evaluate the effects on face and fingerprint verification accuracy with the watermarked images will be described in the section 5.

4 Dual Watermarking Techniques for Secure Biometric Systems

In this section, we describe two scenarios of dual watermarking to guarantee the integrity of information transmitted with the fragile watermarking and to hide some important information such as facial information. A dual watermarking technique using the fingerprint as cover work needs to satisfy the following conditions: 1) Robust watermarks are embedded first, and then fragile watermarks are embedded, 2) Watermarks are free from interference between embedded information, 3) The effects on fingerprint verification accuracy is minimized.

We chose the multi-resolution watermarking technique of Dugad *et al.*[11] for the robust watermarking, and the watermarking technique of Jain *et al.*[7] for the fragile watermarking. The details of the techniques can be found in [7,11].

In the Scenario 3, we integrate the robust watermarking technique and the fragile watermarking technique with considering the ridge information of the fingerprint image. On the other hand, Scenario 4 integrates straightforwardly both watermarking techniques without considering the characteristics of the fingerprint image. If both watermarking techniques are integrated straightforwardly, the robust watermark may not be detected accurately because of the possible collisions in some embedding locations. Therefore, we first analyze the factors that can influence on detecting the embedded robust watermarks, and then describe a modified fragile watermarking technique that can avoid the possible interferences with the robust watermarking.

Most collisions take place near the ridges because the robust watermarking technique embeds the watermark near the ridges. Therefore, to minimize the interference caused by embedding both watermarks, the fragile watermark needs to be embedded by considering the ridges where the robust watermark may be embedded into.

In the embedding side, we perform the robust watermarking first, and then apply the modified fragile watermarking to check the integrity of the fingerprint image resulted from the robust watermarking. The modified fragile watermarking embeds watermarks for guaranteeing the integrity and avoiding the collision using Eq. (2).

$$P_{wm}(i,j) = P(i,j) + \{(2s-1)P_{AV}(i,j)k(1+\frac{P_{SD}(i,j)}{A})(1+\frac{P_{GM}(i,j)}{B})\beta(i,j)\} \tag{2}$$

where $P_{wm}(i,j)$ and $P(i,j)$ are values of the watermarked and original pixels at location (i,j), respectively. The value of watermark bit is denoted as s and watermark embedding strength is denoted as k. $P_{AV}(i,j)$ and $P_{SD}(i,j)$ denote the average and standard deviation of pixel values in the neighborhood of pixel (i,j), and $P_{GM}(i,j)$ denotes the gradient magnitude at (i,j). The parameters A and B are weights for the standard deviation and gradient magnitude, respectively, and they modulate the effect of these two terms; increasing either of them decreases the overall modulation effect on the amount of change in pixel intensity, while decreasing them has the opposite effect.

Note that β in Eq. (2) means the locations to be embedded and is determined based on a pseudo-random number generator initialized with a shared secret key. As mentioned previously, the ridges should be excluded from β to avoid the collisions of the embedding locations between the robust and the fragile watermarking. Though the embedding location β should be identical at both embedding and detecting sides, the detecting side cannot generate the same embedding location β because the original fingerprint image of the embedding side and the watermarked fingerprint image of the detecting side generate different edge information by the Sobel operation. To solve this problem, we creates an approximated fingerprint image $\hat{P}(i,j)$ using Eq. (3) before selecting the embedding location β at both embedding and detecting sides. Therefore, the same location β can be generated at both sides because the Sobel operation is applied to the approximated fingerprint image.

$$\hat{P}(i,j) = \frac{1}{8}\left(\sum_{k=-2}^{2} P(i+k,j) + \sum_{k=-2}^{2} P(i,j+k) - 2P(i,j)\right) \tag{3}$$

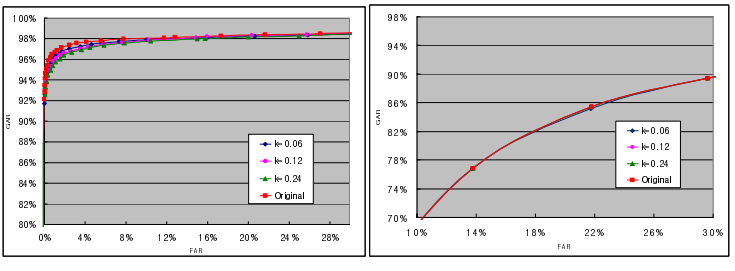
5 Performance Evaluation

We perform experiments with two categories. First, to evaluate the effects on verification accuracy with watermarked images, we measure the accuracy of face and fingerprint verification with watermarked and non-watermarked biometric data. Then, we evaluate the performance of the dual watermarking in terms of watermark detection accuracy.

For the purpose of evaluating of the fingerprint verification accuracy, a data set of 4,272 fingerprint images composed of four fingerprint images per one finger was collected from 1,068 individuals. For the purpose of evaluating the face verification accuracy, a data set was collected from 55 people and composed of 20 facial images per one individual. The images from 20 people were used to construct bases for feature extraction, and the images from the rest were used for training and test.

To evaluate the effects on fingerprint verification accuracy with and without watermarking, we embedded the facial feature into the fingerprint image with various strength k explained in section 2. In the same way, to evaluate the effects on face verification accuracy with and without watermarking, we embedded the fingerprint feature into the facial image with various strength k.

Fig. 1(a) shows ROC curves for fingerprint verification. The effects on fingerprint verification accuracy are degraded according to the distortion of image quality. Fig. 1(b) shows ROC curves for face verification. In spite of serious distortion of image quality, the effects on verification accuracy with the PCA method are negligible. This is because the PCA method uses global information of face images.



(a) Fingerprint Verification (b) Face Verification

Fig. 1. ROC Curves for the Fingerprint and Face Verification: Watermarked Images with Various Strength k

Table 1. Detection Ratio after Attacks

Attack	PSNR	Detection Ratio
Median Filter	60.161461 dB	1.000000
JPEG compression (30%)	25.241369 dB	0.777778
Blur Filter	27.950458 dB	1.000000
Cut	15.169233 dB	0.888889

To evaluate the performance of the dual watermarking technique, we measured the robustness of the dual watermarking technique with various signal attacks. Also, to evaluate the effect of embedding the fragile watermark on the detection ratio of the robust watermark, we measured the performance of the watermarking with considering the characteristics of the fingerprint(Scenario 3) and without it(Scenario 4).

To evaluate the robustness of the dual watermarking technique, we considered four typical types of attacks. As shown in Table 1, the value of PSNR regarding image degradation due to correlated attack. The detection ratio is acceptable even after attacks. The robust watermarking technique can provide an acceptable robustness against the median filter attack compared to other attacks.

Fig. 2 shows the distribution of the detection ratio by using the dual watermarking techniques. The dual watermarking that does not consider the ridge information(represented as "Scenario 4" in Fig. 2) has some influence on the detection ratio of robust watermarks(represented as "Robust Only" in Fig. 2). However, the dual watermarking that considers the ridge information(represented as "Scenario 3" in Fig. 2) has no influence on the detection ratio of robust watermarks.

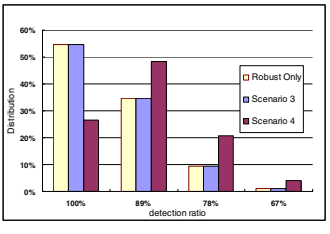


Fig. 2. Comparison of the Dual Watermarking that considers the Ridge Information(Scenario 3) with the Dual Watermarking without Such Consideration(Scenario 4)

6 Conclusions

Biometrics are expected to be widely used in conjunction with other techniques such as the cryptography and digital watermarking on the network. For large-scale, remote user authentication services, the verification accuracy issue as well as the security/privacy issue should be managed. In this paper, the effects of various watermarking techniques on multimodal biometric systems were evaluated where both fingerprint and face characteristics were used simultaneously for accurate verification. Also, we evaluated a dual watermarking technique for the security/privacy issue. Based on the experimental results, we confirm that embedding fingerprint features into a facial image can provide superior performance in terms of the user verification accuracy. Also, the dual watermarking with considering the characteristics of the fingerprint in embedding the dual watermark can provide superior performance in terms of the watermark detection accuracy.

References

1. A. Jain, R. Bole, and S. Panakanti, *Biometrics: Personal Identification in Networked Society*, Kluwer Academic Publishers, (1999).
2. D. Maltoni, et al., *Handbook of Fingerprint Recognition*, Springer, (2003).
3. R. Bolle, J. Connell, and N. Ratha, : Biometric Perils and Patches,*Pattern Recognition*, Vol. 35, (2002) 2727-2738,.
4. W. Stallings, *Cryptography and Network Security*, Pearson Ed. Inc., (2003).
5. H. Lee, et al: An Audio Watermarking By Modifying Tonal Maskers, *ETRI Journal*, Vol.27, No.5, (2005).
6. S. Jung, et al.:An Improved Detection Technique for Spread Spectrum Audio Watermarking with a Spectral Envelope Filter, *ETRI Journal*, Vol. 25, No. 1 (2003) 52-54.
7. A. Jain, U. Uludag, and R. Hsu,: Hiding a Face in a Fingerprint Image, *Proc. of Int. Conf. on Pattern Recognition*, (2002) 52-54.
8. M. Yeung and S. Pankanti: Verification Watermarks on Fingerprint Recognition and Retrieval, *Journal of Electronic Imaging*, Vol. 9, No. 4 (2000) 468-476.
9. S. Pan, et al.: A Memory-Efficient Fingerprint Verification Algorithm using A Multi-Resolution Accumulator Array for Match-on-Card *ETRI Journal*, Vol. 25, No. 3 (2003) 179-186.
10. M. Turk and A. Petland: Face Recognition using Eigenface, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, (1991).
11. R. Dugad, K. Ratakonda, and N. Ahuja: A New Wavelet-based Scheme for Watermarking Images, *Proc. of the ICIP*, (1998).

An Improvement of Auto-correlation Based Video Watermarking Scheme Using Independent Component Analysis

Seong-Whan Kim and Hyun-Sung Sung

Department of Computer Science, University of Seoul, Jeon-Nong-Dong, Seoul, Korea
{swkim7, wigman}@uos.ac.kr

Abstract. Video watermarking hides information (e.g. ownership, recipient information, etc) into video contents. In this paper, we propose an auto-correlation based video watermarking scheme to resist geometric attack (rotation, scaling, translation, and mixed) for H.264 (MPEG-4 Part 10 Advanced Video Coding) compressed video contents. To embed and detect maximal watermark, we use natural image statistics based on independent component analysis. We experiment with the standard images and video sequences, and the result shows that our video watermarking scheme is more robust against geometric attacks (rotation with 0-90 degree, scaling with 75-200%, and 50%~75% cropping) than Wiener based watermarking schemes.

1 Introduction

A digital watermark or watermark in short, is an invisible mark inserted in digital media such as digital images, audio and video so that it can later be detected and used as evidence of copyright infringement. However, insertion of such invisible mark should not alter the perceived quality of the digital media (it is the transparency requirement) while being extremely robust to attack (it is a robust requirement) and being impossible to insert another watermarks for rightful ownership (it is a maximal capacity requirement). Watermark attacks are classified into (1) intentional attacks, and (2) unintentional attacks. Basic requirements for video watermarking are geometric attack robustness (intentional attacks) and H.264 video compression (unintentional attacks). There are four major researches for geometric attack robustness, (1) invariant transform, (2) template based, (3) feature point based, and (4) auto-correlation based [1, 2]. Auto-correlation approach is to insert the mark periodically during the embedding process, and use auto-correlation during the detection process.

We designed an auto-correlation based watermark detection scheme for geometric attack robustness, and present a video watermarking scheme, which is robust on geometric attack (scaling, cropping, rotation, and mixed) and H.264 video compression. In autocorrelation based approaches, separation of Gaussian noise components from raw and watermarked images is a crucial function for watermark performance. To improve performance, we applied Gaussian noise filtering method based on statistical model of natural images. The statistic model is based on ICA (independent component analysis) which models image as factorial distribution of super-Gaussian distribution. Since this model is more relevant for natural image than Gaussian distribution model, ICA based de-noising method shows superior performance for filtering (or

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 643 – 648, 2005.
© Springer-Verlag Berlin Heidelberg 2005

de-noising) Gaussian noise. We implemented a new watermark insertion and detection scheme based on this method and evaluated performance advantages over conventional Wiener filter approach. Result shows that new watermark scheme is more robust and reliable than Wiener filters approach. Lastly our method considers human perception model for video to achieve maximal watermark capacity. Using this perceptual model we can insure the quality of watermarked videos while achieving maximum watermark robustness.

2 Watermark Embedding and Detection Scheme

H.264 is a widely used video compression standard, in which it uses different coding techniques for INTRA (reference) and INTER (motion predicted) frames. We embed the auto-correlated watermark for geometric synchronization in H.264 reference (INTRA) frames, because INTRA frames are used for reference frames of motion predicted (INTER) frames, and they are usually less compressed than INTER frames. In case of geometric attacks for video frames, we can use the watermark in INTRA frames to restore the synchronization misses. We changed the auto-correlation based watermark embedding approach as in [2] with different JND (just noticeable difference) model and different watermark estimation scheme, and also changed the Wiener filter to ICA de-noising based approach.

Watermark embedding for H.264 INTRA frames can be summarized in (1) and (2). In this equation, I' is watermarked frame, E is ICA de-noised frame, N is the ICA estimated noise, λ is perceptual mask as specified in the previous equation, w_p is payload watermark, and w_s is synchronization watermark. We used NVF (noise visibility function) and entropy masking model for perceptual masking model [3, 4]. Previous research uses the Wiener filter to compute E and N. For each 64x64 blocks, we adjusted the watermark strength as image complexity.

$$\text{Embed} \quad I' = I + \lambda N(w_p + w_s), \text{ where } N = I - \text{ICADenoise}(I) \tag{1}$$

$$\lambda = \max(\lambda_N, \lambda_E) \tag{2}$$

$$\lambda_N = (1 - NVF) * A + NVF * B, \text{ where } NVF = \frac{1}{1+\sigma^2}$$

$$\lambda_E = 3.0 * E, \text{ where } E = \sum_{x \in N(X)} p(x) \cdot \log \frac{1}{p(x)}$$

Watermark detection and payload extraction for H.264 INTRA frames can be summarized in (3) and (4). As in the embedding process, we used the same ICA denoising before watermark detection. We compute w_s correlation for watermark synchronization and w_p correlation for payload detection.

$$\text{Detect} \quad \begin{aligned} w_s \cdot w' &= w_s \cdot (I' - E'), \text{ where } E' = \text{ICADenoise}(I') \\ &= w_s \cdot (I + \lambda w_p + \lambda w_s + \delta) \end{aligned} \tag{3}$$

$$\text{Payload} \quad w_p \cdot w' = w_p \cdot (I + \lambda w_p + \lambda w_s + \delta) \tag{4}$$

3 Watermark Estimation Using ICA De-noising

We applied Independent Component Analysis (ICA) based Gaussian noise filtering (de-noising) instead of Wiener filter. From the Bayesian viewpoint Wiener filter is an inference method which computes Maximum Likely (ML) estimates of image signal given the noise variance. Wiener filter assumes Gaussian distribution for both original image and noise. But real image statistics are much different from Gaussian distribution and are better modeled by the ICA model. If the ICA model provides better approximation of real image signals, then it can dramatically enhance noise filtering step in Watermark insertion and detection. ICA is an unsupervised learning method which has found wide range of applications in image processing [5]. ICA finds a linear transform W that maps high dimensional data X into maximally independent source signals S. The source signal S is modeled by a factorial probability density function which is usually super-Gaussian distributions like Laplacian distribution. Super-Gaussian distribution is a distribution which has higher peak at 0 and longer trails than Gaussian distribution.

$$S = WX\ ,\ X = AS \qquad (A = W^{-1})$$
$$P(S) = \prod_{i}^{M} P(s_i)$$

In [6], it is reported that natural image statistics are well described by such ICA model. Also the linear transform filter W learned by ICA algorithm shows much similarity to simple cell filters [6] in human visual system. The filters learned by ICA algorithm are similar to Gabor filters which are multi-orientation, multi-scale edge filters. ICA models complex statistics of natural image well by factorized independent source distributions. Such independent source distributions are usually modeled as generalized Laplacian distribution which is super-Gaussian and symmetric. This simple PDF can be learned from image data and can be used for Bayesian de-noising of images [6]. Let's assume image patch X that follows ICA model (X=AS) and an independent Gaussian noise n with 0 mean and known variance σ. We assume that ICA source signal S has unit variance. Then we can define noisy image Y as $Y = AS + n$. In Bayesian de-noising the goal is to find $\hat{S}$ s.t.

$$\begin{aligned}
\hat{S} &= \underset{S}{argmax} \log P(S \mid Y) \\
&= \underset{S}{argmax} \left[\log P(Y \mid S) + \log P(S)\right] \\
&= \underset{S}{argmax} \left[\log N(Y - AS; \sigma) + \log P(S)\right] \\
&= \underset{S}{argmax} \left[-\frac{|Y - AS|^2}{2\sigma^2} - \sum_{i}^{M} \log P(s_i) \right]
\end{aligned}$$

If we assume unit Laplacian distribution for S and Gaussian noise with signal s, then those equations can be written as.

$$\hat{S} = \underset{S}{argmax}\left[-\frac{|Y - AS|^2}{2\sigma^2} - \sum_i^M |s_i|\right]$$

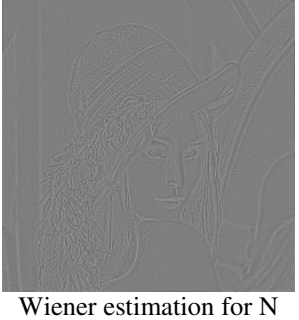
The objective function can be expressed as sum of independent and positive 1D objective functions. So the minimization of the entire objective function is equivalent to minimization of individual source dimensions. The solution to this 1D optimization problem is summarized in [5]. Then using the estimated $\hat{S}$, we can compute denoised image $\hat{X} = A\hat{S}$ as well as the Gaussian noise $n = Y - \hat{X}$. Figure 1 compares Wiener filter and ICA de-noising for separating Gaussian noise N. ICA de-noising result shows the strong edge-like features, whereas Wiener filtered result shows rather weak edge-like features.



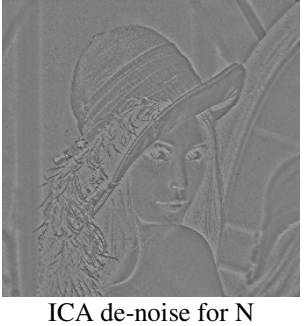
Original image I



Watermarked image I'



Wiener estimation for N



ICA de-noise for N

Fig. 1. Noise estimation comparison between Wiener filter and ICA de-noising

4 Simulation Results

We experimented with the standard test images, and Table 1 shows that we improve the watermark detection performance (increase both correlation value and threshold value) using ICA de-noising based approach over Wiener approach. Threshold value depends on the local characteristics of the resulting auto-correlation image. ICA de-noising increases the threshold more than Wiener filter and it can filter many noisy

auto-correlation peaks, whereby we can make more accurate detection performance. We tested our approach over 20 standard test images, and showed that superior detection performance.

Table 1. ICA driven watermark detection improvements over Wiener

	Using Wiener		Using ICA	
	Correlation value	Threshold	Correlation value	Threshold
aerial.bmp	0.252816	0.074945	0.418893	0.246213
airplane.bmp	0.230755	0.046858	0.353133	0.172529
airport.bmp	0.198668	0.037124	0.411544	0.214705
baboon.bmp	0.192640	0.041524	0.362268	0.204293
barbara.bmp	0.203092	0.036492	0.321886	0.154950
boat.bmp	0.236444	0.041021	0.362430	0.158109
couple.bmp	0.230765	0.041714	0.380771	0.171419
elaine.bmp	0.253149	0.044271	0.350184	0.139988
girl.bmp	0.255930	0.052973	0.388574	0.177383
girls.bmp	0.243981	0.038508	0.346012	0.154674
goldhill.bmp	0.228562	0.039083	0.328786	0.144241
house.bmp	0.219754	0.046751	0.328289	0.161131
lake.bmp	0.213744	0.034296	0.302019	0.138336
lena.bmp	0.236769	0.039555	0.352069	0.156090
man.bmp	0.216027	0.035835	0.326558	0.138451
pepper.bmp	0.254798	0.044070	0.361369	0.157697
splash.bmp	0.250747	0.045205	0.351555	0.166243
stream and bridge.bmp	0.186301	0.037753	0.293211	0.159628
toys.bmp	0.222101	0.058213	0.236835	0.160865

To experiment the geometric attack, we used Stirmark 4.0 [7] geometric attack packages for various geometric attacks (rotation with 0-90 degrees, scaling with 75-200%, cropping with 50-75%, and mixed). We experimented with five cases: (case 1) rotation with 1, 2, and 5 degree clockwise; (case 2) case 1 rotation and scaling to fit original image size; (case 3) cropping with 50% and 75%; (case 4) scaling with 50%, 75%, 150%, and 200%; and (case 5) median, Gaussian, and sharpening filter attacks.

5 Conclusions

In this paper, we presented a robust video watermarking scheme, which uses auto-correlation based scheme for geometric attack recovery, and uses human visual system characteristics for H.264 compression. For watermark insertion and detection our method uses ICA-based filtering method which better utilizes natural image statistics than conventional Wiener filter. Result shows that the proposed scheme enhances robustness while keeping watermark invisible. Our video watermarking scheme is robust against H.264 video compression (average PSNR = 31 dB) and geometric attacks (rotation with 0-90 degree, scaling with 75-200%, and 50%~75% cropping).

References

1. Bas, P., Chassery, JM, and Macq, B.: Geometrically invariant watermarking using feature points. IEEE Trans. Image Proc., vol. 11, no. 9, Sept. (2002) 1014–1028
2. Kutter, M.,: Watermarking resisting to translation, rotation, and scaling. Proc. of SPIE Int. Conf. on Multimedia Systems and Applications, vol. 3528, (1998) 423-431.
3. Voloshynovskiy, S. Herrige, A., Baumgaertner, N., Pun, T.: A stochastic approach to content adaptive digital image watermarking. Lecture Notes in Computer Science: 3rd Int. Workshop on Information Hiding, vol. 1768, Sept. (1999) 211-236
4. Kim, S.W., Suthaharan, S., Lee, H.K., Rao, K.R.: An image watermarking scheme using visual model and BN distribution. IEE Elect. Letter, vol. 35 (3), Feb. (1999)
5. Hinton, G. E., Sejnowski, T. J.: Unsupervised learning: foundations of neural computation (edited), MIT Press, Cambridge, Massachusetts (1999)
6. Hyvarinen, A., Sparse Code Shrinkage: De-noising of non-Gaussian data by maximum likelihood estimation. Neural Computation, 11(7) (1999) 1739-1768
7. http://www.petitcolas.net/fabien/watermarking/stirmark/

A Digital Watermarking Technique Based on Wavelet Packages

Chen Xu[1], Weiqiang Zhang[2], and Francis R. Austin[1]

[1] College of Science, Shenzhen University, Shenzhen 518060, China
xuchen@szu.edu.cn
[2] College of Science, Xidian University, Xi'an 710071, China

Abstract. Digital watermarking is a copyright protection technique that has been developed in response to the rapid growth of networking and internet technology. Possessing also validation capabilities, it is superior to most forms of traditional network safety methods based on encryption and so has attracted widespread attention in recent years. At present, two common forms of watermarking techniques are in use: robust watermarking and fragile watermarking. While the former technique is highly robust, it sorely lacks the capability to recognize forged signals. The latter technique, on the contrary, though being highly sensitive to signal authenticity is overly vulnerable to normal signal processing. This paper introduces a digital watermarking technique (algorithm) that is based on wavelet packages. By embedding the watermarks in the middle-frequency segment of the wavelet package, this new technique possesses all the desirable properties of the previous algorithms, as indicate by experimental results.

Keywords: Wavelet package, digital watermark, hiding, watermark detection, algorithm.

1 The Concept of Digital Watermarks

In general, a digital watermark is a random signal $\{x_1, x_2,...x_n\}$ that is generated from an original signal S_0 using a prescribed rule for assigning the watermark positions. For example, if $\{S_1,S_2,\ldots S_n\}$ is a sequence of positions of S_0, then a watermark can be attached to each of these positions by using a certain pre-determined rule, say $s_i' = f(s_i, x_i)$. Let S denote the signal after it has been watermarked. Then as a result of signal processing and noise contamination during transmission, the received signal, say S', may differ from S. Nevertheless, by applying the pre-determined watermarking rule for S, it is possible to identify the digital watermarks $\{x_1, x_2,...x_n\}$ for S', and consequently determine whether S' contains any digital watermarks $\{x_1, x_2,...x_n\}$ of the original signal.

In order to ensure that the watermarked signal retains the uniqueness of the original signal, it is usual to parametrize both the original message that is being transmitted as well as the protected signal. One way of accomplishing this is as follows. The author

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 649 – 654, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of the message chooses a secret code that is known only to himself; this is the first parameter. At the same time, a second parameter is generated via a uni-directional Hash function that depends only on the protected signal. Then, by using these two parameters as the "seeds' for the generation of a random signal, the signal so generated is uniquely determined by both the author's message and the protected signal. It is then impossible for copyright violators to fake this watermarked message.

2 Robust and Fragile Watermark Algorithms Based on Wavelets

2.1 Robust Watermark Algorithms

Situated at the low-frequency segment of the wavelet spectrum, robust watermarks are resistant to most forms of signal processing such as low-frequency-pass filters, and noise contamination. Watermark hiding algorithms that are based on wavelets can be described as follows:

1) choose a suitable wavelet and apply L-level decomposition; retain the high frequency components of the L-level decomposition; the low-frequency components d_L of L- level decomposition are treated as follows;
2) let N be the length of the watermark that is required to be hidden; choose the N components of d_L with the greatest absolute value, and record their respective positions; the watermark hiding rule is:

$$d_L^{'}(i) = d_L(i)(1+\alpha x(i)); \tag{1}$$

3) Use wavelet inverse transform to recover the signal that carries the watermark;

The corresponding watermark detection procedure can be described as follows:

1) apply the same wavelet transform to the received signal;
2) by using the original speech signal, determine the N hidden random positions in the low-frequency segment of L- level decomposition; evaluate

$$x'(i) = (d_L^{''}(i)/d_L(i) - 1)/\alpha; \tag{2}$$

3) determine the correlation between x' and x, and use this correlation to determine the existence, or non-existence, of watermark signals. The correlation between signal $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ is defined as

$$corr(x, y) = \left(\sum_{i=1}^{N} x_i y_i \right) \Big/ \sqrt{\sum_{i=1}^{N} x_i^2} \,. \tag{3}$$

Speech signals may be subject to various kinds of disruptive processes during transmission. Besides general noise contamination, for example, it may have to go through different kinds of signal filters, speech signal compressors, as well as endure other processes that may alter the sampling frequency of the wave. [131,132,133]. A well –hidden watermark and detection algorithm should therefore remain intact under all adverse circumstances and be received in its original form.

Robust watermarks have the following advantages: 1） they are simple and easy to implement (utilizing the simplicity and speed of wavelet transforms); 2） they are robust against signal processing (this algorithm places the watermark in the high

frequency segment of the speech signal); 3) the embedded watermark is not easily detected (again owing to the fact that the watermark is placed in the high frequency segment of the signal); 4) this watermark is able to recover the original speech signal even after it has been corrupted by processing, provided that the signal is still readable. Robust watermarks are therefore resistant to most forms of vicious attacks and hence are particularly useful in situations where the accuracy of the signals are of paramount importance. However, since this watermark cannot distinguish between an authentic signal and a forged one, its usefulness may be compromised when message safety and reliability are crucial.

2.2 Algorithms for Fragile Watermarks

In contrast to robust watermarks, fragile watermarks are highly sensitive to changes in the contents of the transmitted signal because they are embedded in the high frequency segment of the wavelet. For this reason, fragile watermarks are widely used in authenticating multi-media contents. However, because of its high sensitivity to content changes, it is also common for normal changes to the original signal that are caused by transmission to be mistaken for deliberate attack. Therefore this technique easily creates false alarms which wastes human time and resources. Furthermore, these watermarks are readily destroyed by the slightest amount of signal processing. The theory behind the fragile watermarking technique is similar that of the robust technique, only that the watermarks are placed in different locations of the signal.

3 Digital Watermarking Technique Based on Wavelet Packages

This paper introduces a watermarking technique based on wavelet packages that has none of the disadvantages of the previously discussed techniques. The main difference is to embed the watermarks into the middle segments of the wavelet package. The actual method of embedding the watermark is similar to that of the robust watermarking technique, only that the lengths of the watermarks as well as their locations of insertion are different.

Wavelet packages have more widespread applications because of their finer division of the spectrum. Fig. 1 shows the decomposition of a 3-level wavelet package, where A represents low frequency, D represents high frequency, and the subscripts indicate the levels of the wavelet package (i.e. the level number).

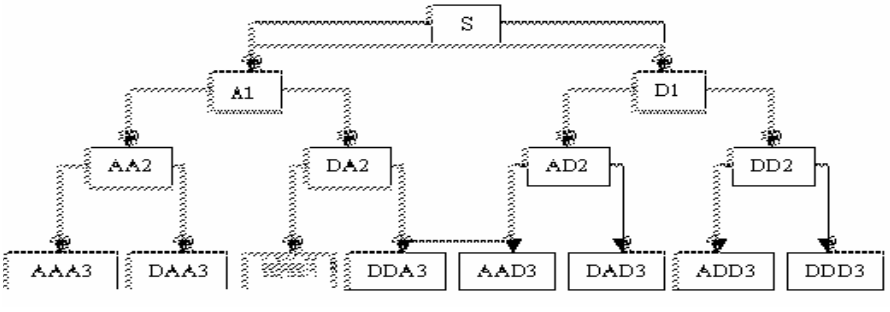


Fig. 1. Decomposition of a 3-level wavelet package

The watermark is embedded in AAD3, which belongs to the middle frequency segment, and so can tolerate most low-frequency-pass filters, high frequency noise, as well as most forms of interference during transmission. It is also sensitive to deliberate attacks on the signal as well as attempted forgery of the original message.

4 Experimental Results and Conclusions

This section details the results of experiments conducted using the authors' own voice signals. The db4 wavelet was used and a 3-level wavelet package decomposition was performed. The watermarked signal used was a white noise with a Gaussian distribution of length N = 200. The length of the speech signal was 2 seconds, and the sampling frequency was 8KHz. The watermark was embedded in the middle-frequency segment AAD3 and Fig. 2 (a) and (b) shows respectively the original speech signal and the watermarked signal. Fig. 2 （c） is the result of using the relative method to extract the watermark.

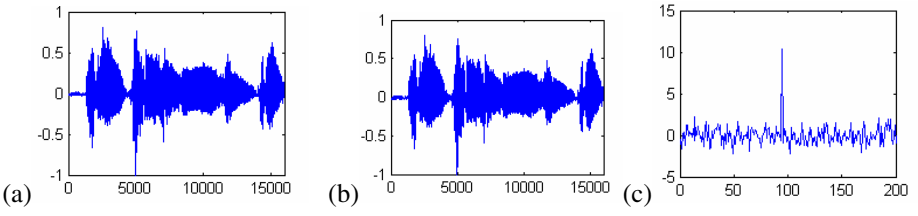


Fig. 2. Embedding watermarks in a speech signal. (a) Speech signal with watermark embedded. (b) Original speech signal (c) The results of using the relative method to extract the watermark.

4.1 Noise Contamination

The effect of noise on a watermark was first examined. A white noise was simulated and then applied to a watermarked signal. The results obtained clearly demonstrates the robustness of this watermark since it is still clearly identifiable in the high-intensity noise regime (see Fig. 3).

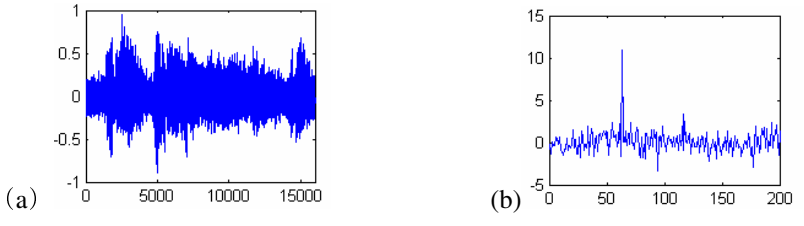


Fig. 3. The effect of white noise. (a) A watermarked signal (b) Detection of the watermark. contaminated by white noise.

4.2 Noise Clearance

The effect of noise clearance processes on a digital watermark was next examined by removing the white noise that was applied to the speech signal in section 4.1. The signal after noise clearance is shown in Fig. 4(a) and it is clear that in spite of a somewhat lowered signal quality, the actual speech message is still largely discernable. This is evidence of the robustness of the watermark against the noise clearance process. Fig. 4(b) shows the watermark extracted using the relative method.

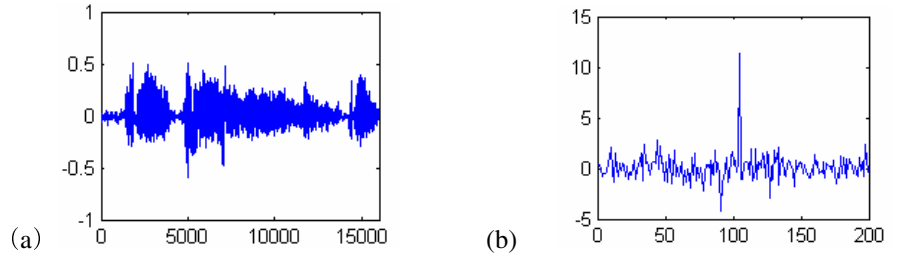


Fig. 4. The effect of noise clearance. (a) Watermarked speech signal after noise removal. (b) Watermark extracted using the relative method.

4.3 Re-sampling

The re-sampling of a digital signal can easily cause unintended changes to the contents of the transmitted message because it is difficult to ensure that the sampling point positions are the same for both the original and the processed message during A/D (or D/A) transformation. A re-sampling simulation was carried out on a watermarked signal and it was observed that the positions of some re-sampled points have indeed been altered (see Fig. 5).

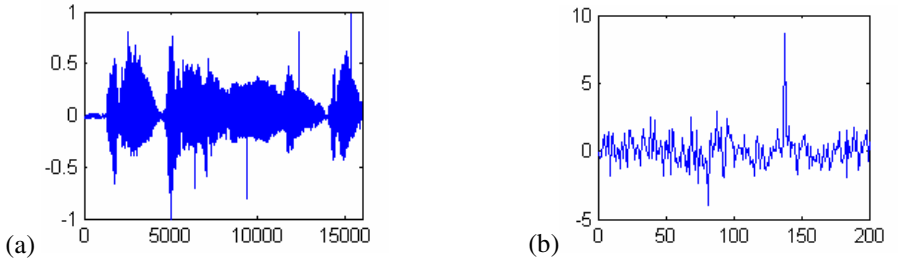


Fig. 5. The effect of re-sampling. (a)A re-sampled watermarked signal. (b) Watermark extracted using the relative method.

4.4 Increased Sensitivity Toward Changes in the Signal Content

In this section, an example is used to illustrate the increased protection afforded by the technique introduced in this paper against signal forgery. First, deliberate changes have been made to a watermarked signal after which an attempt is made to extract the

watermark using the relative method. It was found, however, that the watermark has vanished (see Fig. 6 (b)) as a result of the changes. It is clear, therefore, that by embedding the watermark in the middle segment of the wavelet, it is possible to achieve both robustness and sensitivity in a watermark. As is evident from Fig. 6(a), the wave shape of the speech signal has been severely damaged and the original watermark could no longer be detected (Fig. B (b)).

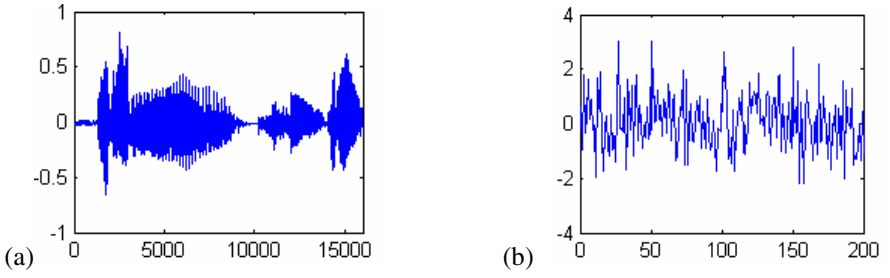


Fig. 6. Sensitivity of the proposed algorithm against forgery. (a) Watermarked signal with deliberate changes. (b) The result of watermark extraction using the relative method.

5 Conclusions

As evidenced by the numerical examples in this paper, a digital watermarking technique based on wavelet packages possesses all the desirable properties of the robust and fragile watermarks and so is a more practical alternative to the watermarks that are presently in use.

References

1. Xu,C.,:Wavelets and Application Algorithms. Science Press Beijing(2004)
2. L,L., Shafer,R., R,W. :Digital Processing of Speech Signals. Science Press Beijing(1983)
3. Niu,X., Yang,Y.: A Digital Watermark Hiding and Detecting Algorithm Based on Wavelet Transforms. Journal of Computers (2000)21~27
4. Martinian,E.,Chen,B.,Wornell,G.:Information Theoretic Approach to the Authentication of Multimedia. SPIE on Security and Watermarking of Multimedia Contents. Vol. 4314. San Jose. the International Society of Optical Engineering(2001)185-196.

A Spectral Images Digital Watermarking Algorithm

Long Ma[1], Changjun Li[2], and Shuni Song[3]

[1] School of Information Science & Engineering, Northeastern University, Shenyang 110004, China
malong1227@sina.com
[2] Department of Color Chemistry, University of Leeds, Leeds LS2 9JT, UK
C.Li@leeds.ac.uk
[3] School of Science, Northeastern University, Shenyang 110004, China
songsn62@@sina.com

Abstract. In this paper we propose a technique to embed a digital watermark containing copyright information into a spectral images. The watermark is embedded by modifying the singular values in PCA/DWT domain of the spectral images. After modification the image is reconstructed by inverse processing, thus containing the watermark. We provide analysis of watermark's robustness against attacks. The attacks include lossy compression and median filtering. Experimental results indicate that the watermark algorithm performs well in robustness.

1 Introduction

Copyright protection is becoming more important in open networks since digital copying maintains the original data and copying can be easily and at high speed. Digital watermarks offer a possibility to protect copyrighted data in the information society. A watermarking algorithm consists of the watermark structure, an embedding algorithm, and an extraction, or a detection algorithm. The watermarking procedure should maintain the following properties: the watermark should be undetectable and secret for an unauthorized user, the watermark should be invisible or inaudible in the information carrying signal and finally, the watermark should be robust on the possible attacks [1, 2].

Watermarking techniques have been developed for spectral images. The gray-scale watermark can be embedded in the DWT transform domain of the spectral image [3,4]. But also the gray-scale watermark can be embedded Principal Component Analysis (PCA) transform domain of the spectral image [5,6].

In this paper we present a new watermarking method for spectral image. The paper is organized as follows: the watermarking procedure is established in section 2. attacks are described in Section 3, experiments with the procedure are in Section 4 and the conclusions are drawn in Section 5.

2 Watermarking Procedure

In this study one band gray-scale image is used as a watermark. Assume the size of watermark is $N \cdot N$, and the size of the spectral image is $N \cdot N \cdot M$. The embedding algorithm is described as follows.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 655 – 660, 2005.
© Springer-Verlag Berlin Heidelberg 2005

1) Apply PCA to spectral domain of the spectral image, the eigenvalues, the eigenvectors, and the eigenimages for the eigenvectors are obtained. The spatial size of the eigenimage is the same as is that of the original spectral image.
2) Using DWT, decompose the eigenimages from one to ten.
3) Apply SVD to the transformed eigenimages E_k:

   $E_k = U_k D_k V_k^T \quad k = 1,2,...,10$, and λ_i^k, $i = 1,2,...,N$ are the singular of E_k.
4) Apply SVD to the visual watermark:

   $W = U_w D_w V_w^T$, where $\lambda_{wi}, \mathrm{i} = 1,...,\mathrm{N}$ are the singular values of W.
5) Modify the singular values of the transformed eigenimages with the singular values of the visual watermark:

$$\lambda_i^{*\max} = \lambda_i^{\max} + \alpha * \lambda_{wi}, i = 1,2...10 \tag{1}$$

   where $\lambda_i^{\max}$ denotes the max of $\{\lambda_i^1,...,\lambda_i^{10}\}$, and the α is the strength.
6) Obtain the modified DWT coefficients:

   $E_k^* = U_k D_k V_k^T \quad k = 1,2,...,10$

The image is reconstructed with the inverse PCA transform using the modified eigenimages and the original spectral eigenvectors.

Naturally the watermarked image differs from the original image. The difference between these images depends on the strength α in embedding of the watermark according to Eq. 1. The strength balances between properties like robustness, invisibility, security and false-positive detection of the watermark.

Since the singular values of the transformed eigenimages and the strength α are known the watermark is extracted from the spectral image. Now the extraction equation is

$$\lambda_{wi} = (\lambda_i^{*\max} - \lambda_i^{\max}) / \alpha, \mathrm{i} = 1,...,\mathrm{N}$$

Finally the watermark is reconstructed using the singulars:

$$W^r = U_w D_w V_w^T.$$

3 Attacks

The watermark must be not only imperceptible but also robust so that it can survive some basic attacks and signal distortions. Since spectral images are often very large in both spectral image and spatial dimensions, then lossy image compression is usually applied to them. lossy compression lowers the quality of the image and of the extracted watermark. 3D-SPIHT is the modern-day benchmark for three-dimensional image compression [7, 8].

Other possible attacks are different kinds of filtering operation, such median filtering. Median filtering remove occasional bit errors or other outliers of a pixel value. It also removes exceptional spectral values even though they would be results of actual measurements.

The quality in embedding was calculated using Peak Signal-to-Noise ratio (PSNR), which is defined for spectral images as

$$PSNR = 10\log_{10}\frac{MN^2 s^2}{E^d}$$

where E^d is the energy of the difference between the original image and the watermarked image, M is the number of bands in the image, N^2 is the number of pixels in the image and s is the peak value of the original image.

The quality of the extracted watermark was measured with the correlation coefficient cc defined as

$$cc = \frac{\sum (W - \overline{W})(W^r - \overline{W}^r)}{\sqrt{\sum (W - \overline{W})^2}\sqrt{\sum (W^r - \overline{W}^r)^2}}$$

where W contain values from the original watermark and W^r contain values from the extracted watermark. $\overline{W}$ and $\overline{W}^r$ are the respective means.

4 Experiments

In the experiments we used one BRISTOL image as a base image. The BRISTOL images are taken in laboratory conditions from plants or they are natural images from

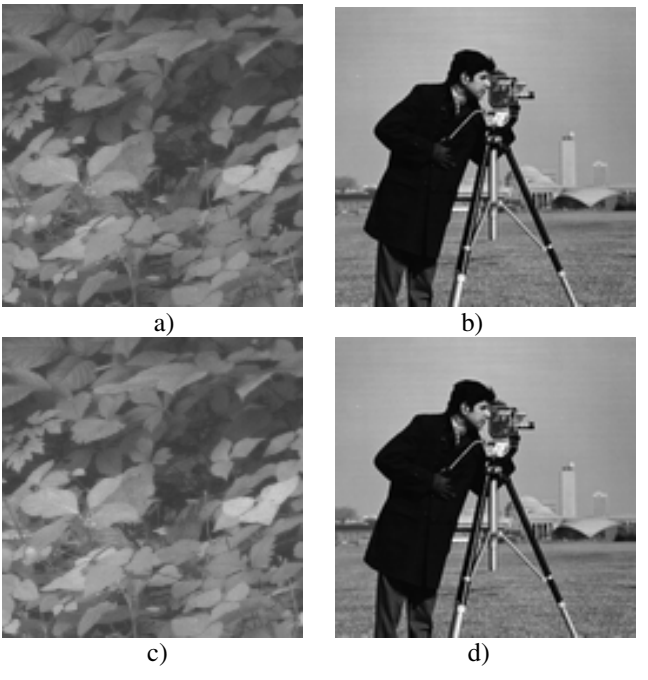


Fig. 1. a) Band 8 from the original image. b) The watermark c) Band 8 of the watermarked image. d) Constructed watermark.

Table I. Parameters in experiments

Parameter	Parameter value
wavelet filter	Daubechies 4-tap
levels in transforms	2
set of eigenimages in embedding	10
The strength α	0.2
quality after embedding	34.74dB

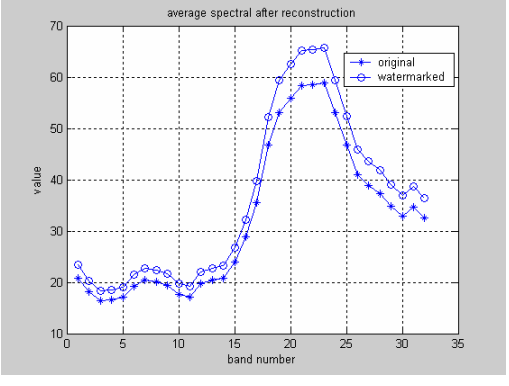


Fig. 2. Average spectra from the watermarked image

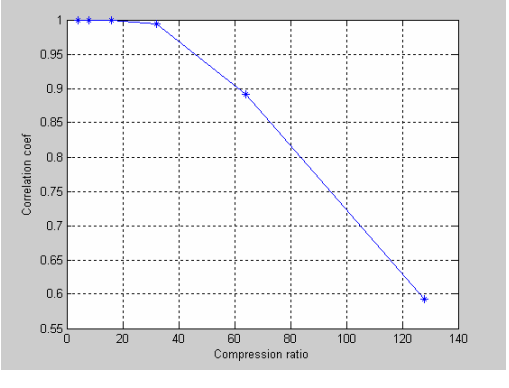
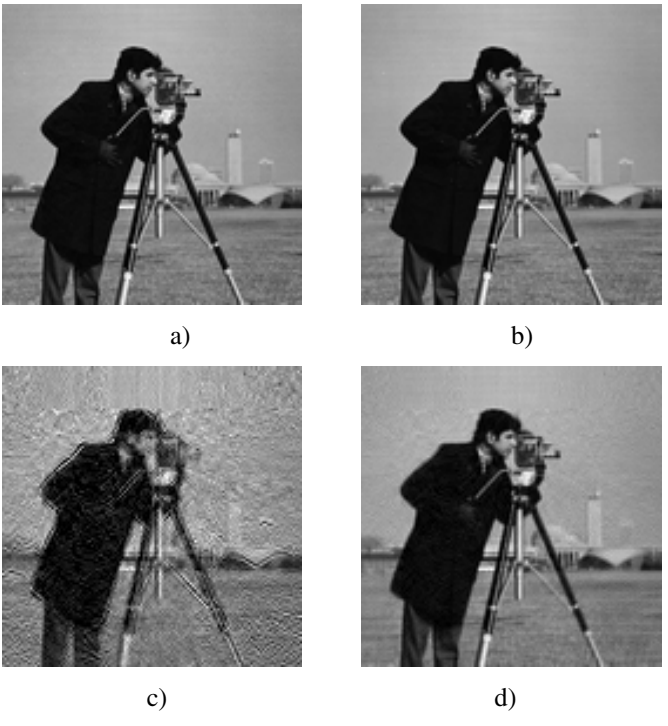


Fig. 3. The qualities of the extracted watermarks

woods, forests, or fields. The spectral range is in 400-700mm. The image was of size $128*128*32$. The resolution of the measurement was 8 bits. The watermark was a visual watermark containing 256 gray-level and it had $128*128$ pixels. In

Fig. 1 we illustrate one band from the spectral images and the watermark applied in the experiments. The parameters used in the experiments are collected in Table I. Fig. 2 show the average spectra of the original image and the watermarked image.



a) b) c) d)

Fig. 4. Watermark extraction quality in lossy compression. Compression ratio a) 8 b) 16 c) 32 d) 64.

Table II. Robustness against filters

attack	PSNR	cc
Median 3×3	32.26	0.6892
Median 5×5	30.10	0.3696

Lossy compression was the first attack that corrupted the watermark. The spectral image was compressed with 3D-SPIHT. Fig. 3 The quality of the extracted watermarks was shown.

The visual degradation of the watermark in lossy compression is shown in Fig. 4.

Median filtering was the second attack applied to the watermarked image: each band of watermarked spectral images was median filtered. The spatial sizes for median filters were 3×3, 5×5. The results are collected in Table II.

5 Conclusions

Watermarking techniques constitute a new area in controlling authorized exploitation of spectral images. We have considered a method for embedding a watermark into a spectral image. The watermark is embedded by modifying singular values. The watermark is a visual watermark, i.e. the extracted watermark can be visually inspected and certified. The robustness of the watermark was tested against attacks, such as 3D-SPIHT and median filtering.

The experiments indicate that the watermark is robust to lossy compression. Against median filtering the watermarking procedure shows higher sensitivity.

References

1. Arta, D., Digital Steganogtaphy: Hiding data within data, IEEE Internet Computing, May/June (2001)75-80
2. Podilchuk, C.I., Delp, E.J., Digital watermarking: algorithms and application, IEEE Signal Processing Magazine, July (2001) 33-46
3. A. Kaarna, J. Parkkinen, digital watermarking of spectral images with three-dimensinal wavelet transform, Proceedings of the Scandinavian Conference on Image Analysis, SCIA 2003, (2003) 320-327
4. A. Kaarna, J. Parkkinen, Multiwavelets in watermarking spectral images, Proceedings of the International Geoscience and Remote Sensing Symposium, IGARSS'04, (2004) 3225-3228
5. A. Kaarna, V. Botchko, P. Galibarov, PCA component mixing for watermarking embedding in spectral images, Proceedings of IS&T's Second European Conference on Color in Graphics, Imaging, and Vision (CGIV'2004), (2004) 494-498
6. A. Kaarna, P. Toivanen, K. Mikkonen, PCA transform in watermarking spectral images, The Journal of Imaging Science and Technology 48(3), (2004) 183-193
7. A. Kaarna, J. Parkkinen, Transform Based Lossy Compression of Multispectral Images, Pattern Analysis & Applications, vol. 4, no 1, (2001) 39-50
8. PL Dragotti, G. Poggi, and ARP Ragozini, Compression of multispectral images by three-dimensional SPIHT algorithm. IEEE Trans. On Geoscience and remote sensing, 38(1) (2000) 416-428

Restoration in Secure Text Document Image Authentication Using Erasable Watermarks

Niladri B. Puhan and Anthony T.S. Ho

Center for Information Security, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798
niladri@pmail.ntu.edu.sg

Abstract. In this paper, we propose a new restoration method for electronic text document images using erasable watermarks. In restoration application, finding sufficient number of low-distortion pixels in a block to embed sufficient information bits and their blind detection is difficult. In the proposed method, an erasable watermark is embedded in each block of a text document image for restoration of the original character sequence. The embedding process introduces some background noise; however the content in the document can be read or understood by the user, because human vision has the inherent capability to recognize various patterns in the presence of noise. After verifying the content of each block, the original image can be restored at the blind detector for further use and analysis. Using the proposed method, it is possible to restore the original character sequence after multiple alterations like character deletion, insertion, substitution and block swapping.

1 Introduction

Watermarking is the art of protecting the multimedia data by inserting the proprietary mark which may be easily retrieved by the owner of the data to verify about its ownership or authenticity. A variety of digital watermarking methods have been developed for such purposes [1]. Authenticity of the multimedia data can be guaranteed with the use of cryptographic signature as the fragile watermark. Lossless format conversion of the watermarked data does not render the embedded signature inauthentic though the representation of the data is changed. Another advantage is that if authentication information is localized, it is possible to achieve tamper localization capability after various attacks. After tamper localization, we are interested to know whether the modified portions of data can be restored. In literature, there are two types of restoration strategies: exact restoration and approximate restoration. In exact restoration, the original copy of the data can be restored without any single bit error. In [2], a Reed-Solomon ECC code was used to generate parity bytes for each row and column of an image. The generated parity bytes were embedded as a watermark in the two least significant bit planes of the image. If there are some changes in the watermarked image and they are within the error correction capability, the changes will be corrected by ECC decoding to restore the original image data. In approximate restoration, the original copy of the data can not be restored; instead an approximate but still valuable version of the data is restored to provide information about the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 661 – 668, 2005.
© Springer-Verlag Berlin Heidelberg 2005

original data. In [3], self-embedding technique was used for embedding a highly compressed version of the image into itself. In this method, the original image was compressed at JPEG 50% quality factor to generate the low-resolution image. For a particular block, the resulting binary sequence is inserted into the LSB plane of a block which is at a minimum distance away and in a randomly chosen direction.

Our present work addresses the issue of restoration for text document images in electronic form whose pixels are either black or white. In text document images, several character from a finite set covey the necessary information. Each character of the English alphanumeric set can be represented by the 7-bit ASCII (American Standard Code for Information Interchange) code. The only previous work for restoration in case of text document images has been reported in [4]. In this method, self-embedding was used for restoration of the original character sequence. The ASCII code of a character was used as its watermark and it was embedded in another character of the document. The particular character for embedding was selected through random permutation or cyclic shift function. This method is effective for restoration against the individual alterations like character substitution, deletion and insertion. However, in this method there exists a possibility of restoration failure after few (even 1) such individual alterations. After multiple alterations, restoration is not possible due to loss of synchronization. To overcome these shortcomings, we propose a new method for restoration of the original character sequence using erasable watermarks in conjunction with error control coding techniques. The proposed method falls under the category of approximate restoration. In binary document images, hiding significant amount of data without causing visible artifacts becomes difficult. In a reasonable block-size, there is insufficient number of low-distortion pixels available. The blind detection requirement of these pixels adds to the difficulty in achieving high watermark capacity. Further, an imperceptible watermark cannot be embedded in white regions of the image. Thus, the white regions are particularly vulnerable to content alteration by the attacker. Due to these shortcomings, it is evident that unless the block size is large, imperceptible watermarking may not be suitable for secure authentication of document images.

So we explore the possibility of embedding the signature in other such pixels that brings visual distortion in watermarked image, but the resulting distortion due to embedding process can be erased at the blind detector. After erasing the embedded watermark, the original image can be restored at the blind detector. This particular concept is known as erasable or invertible watermarking in literature. The algorithm proposed by Fridrich *et al* in [5] for exact authentication of natural images is of particular interest to this paper. Let A represent the information that is altered in the original image when we embed a message of N bits. Fridrich *et al* showed that the erasability is possible provided A is compressible. If A can be losslessly compressed to M bits, N-M additional bits can be erasably embedded in the cover work. In case of document images each pixel is represented by one bit and it can be considered that there is only one bit plane in the image. Within a block if a set of suitable pixels with high correlation can be found, they can be losslessly compressed and an erasable watermark can be constructed. The redundancy achieved by constructing an erasable watermark can be used to embed necessary information for restoration of the original character sequence. This motivation is addressed in the next section by proposing a new restoration method which is particularly useful for text document images.

The paper is organized as follows: in Section 2 the proposed restoration method is described. Simulation results and discussion are presented in Section 3. Finally some conclusions are given in Section 4.

2 Proposed Restoration Method

In this section, we shall propose a restoration method by constructing an erasable watermark in each block of the text document image. In the proposed method we find an ordered set of pixels in each block such that; (1) there exists a high correlation among the pixels and the number of such pixels is high, (2) the same set of pixels can be found at blind detector and (3) the relevant information is preserved after the embedding process. Pixels whose neighborhoods have both white and black pixels are contour pixels and they convey maximum information in the document image. The black pixel whose all neighbor pixels are white is termed as an *isolated* pixel. A white pixel whose all neighbor pixels are white is termed as a *background* pixel. The isolated and background pixels do not convey important information within document images. If these pixels are altered, a background noise will be formed in the image. In text document images, we obtain information by recognizing shapes of different characters and symbols. It is known that human vision has remarkable ability to recognize such shapes even in the presence of noise. So after embedding an erasable watermark in these pixels, the user can still obtain relevant information about the document. The background pixels occur in long sequences and isolated pixels occur in between them with less probability. So such a set of pixels can be significantly compressed using the run-length coding scheme. Flipping of a background pixel creates an isolated pixel and vice-versa in an image; so blind detection of the embedded pixels is possible. We shall outline the proposed method in following steps.

Embedding:

1. The original image of size $M_1 \times N_1$ pixels is divided into non-overlapping blocks of 32×40 pixels. Watermarking is performed for each block independently and in a sequential order starting from left to right and top to bottom of the image. The block-size can be suitably modified if length of the embedded watermark is changed. All characters in the original image are extracted in the sequential order and a binary sequence A_c is obtained after each character is converted into the corresponding ASCII code.
2. Another binary sequence E_s is obtained by encoding A_c with an [*n, k*] BCH encoder [6]. The parameters *n* and *k* are chosen such that the total number of bits in E_s is less than or equal to 108×*B*, where *B* is the total number of blocks in the image.
3. After ECC encoding, if necessary zero-padding is performed in the sequence E_s to make the number of bits exactly equal to 108×*B*. Then E_s is randomly permuted using the secret key *K* and the permuted sequence is denoted as P_s. The errors due to tampering in a localized region of the image often come as bursts. The random permutation converts the burst type of error to be a random type; thus increasing the correction capability of ECC coding.

4. The sequence P_s is divided into B non-overlapping segments of size 108 bits each. Let each such segment be denoted by S_c. A maximum of 108 bits of a particular segment S_c is used in embedding process of the corresponding block in the sequential order.
5. In each block, an ordered set of *insignificant* pixels are searched in the sequential scanning order. Pixels which are in the border with other blocks are not included in this search to maintain block independence. A pixel is defined to be in the insignificant pixel set if the conditions are satisfied in following order.
    a) The pixel is either a background pixel or an isolated pixel.
    b) If there is no other pixel in its 8-neighborhood that has already been decided as an insignificant pixel.
    c) After flipping the current pixel, any new insignificant pixel should not be created among already scanned pixels.
6. The insignificant pixel set is losslessly compressed using the run-length coding scheme. A total of ten bits are used to represent the size of the run-length encoded data. These ten bits along with the run-length encoded data represent the compressed data C_D. The term *redundancy* (R) is defined in Eq. 1 as number of bits available in a block to embed necessary information.

$$R = \text{Size of the insignificant pixel set} - \text{Compressed data size} \quad (1)$$

7. Let the sets S_1 and S_2 be defined such that S_1 contains the first 32 bits and S_2 contains the next n_s bits of the segment S_c. In Eq. 2, n_s is defined to adjust the size of embedded data if the current block has insufficient redundancy.

$$n_s = \min(R - 64, 76) \quad (2)$$

8. The 64-bit authentication signature A_s is computed from the current block according to Eq. 3.

$$A_s = H(C_b, K, I_b, I_K, M_1, N_1, S_2) \quad (3)$$

where H, C_b, I_b *and* I_K denote hash function, current block in the original image, block index and image index, respectively.

The block index is used in the computation of signature to resist block-swapping by the attacker and the image index is necessary to resist the collage attack [7]. The reason for using S_2 in signature computation is to secure the embedded restoration information against hostile attacks.

9. The authentication data A_D of size (n_s + 64) bits consists of three sets. Let the sets be denoted as $A_D^{\,1}, A_D^{\,2}$ and $A_D^{\,3}$. $A_D^{\,1}$ contains the first 32 bits, $A_D^{\,2}$ contains the next 32 bits and $A_D^{\,3}$ contains the remaining n_s bits of A_D. The three sets are computed according to Eq. 4.

$$A_D^{\,1} = A_s^{\,1} \oplus S_1$$
$$A_D^{\,2} = A_s^{\,2} \oplus S_1^{\,c} \quad (4)$$
$$A_D^{\,3} = S_2$$

where $\oplus$ is the exclusive OR operation, $A_s^{\,1}$ contains the first 32 bits and $A_s^{\,2}$ contains the remaining 32 bits of A_s.

10. The compressed data C_D and authentication data A_D are concatenated to create the message *m,* which is embedded in the insignificant pixel set. The embedding is performed pixel-wise; so an insignificant pixel holds one bit of *m* and its pixel value is set equal to the message bit it holds. Likewise all blocks in the image are watermarked.

Detection:

1. The test image of size $M_2 \times N_2$ pixels is divided into non-overlapping blocks of 32×40 pixels. Verification is performed for each block independently and in a sequential order starting from left to right and top to bottom of the image.
2. To verify each block, the message *m* is extracted by finding the insignificant pixel set like in step (5) of the embedding process. The parameters *R* and n_s are computed using Eq. 1 and 2 respectively. The compressed data $C_D^{'}$ and the authentication data $A_D^{'}$ are extracted from the message *m*. The compressed data $C_D^{'}$ together with the current watermarked block is used to reconstruct the block $C_b^{'}$.
3. The authentication data $A_D^{'}$ of size (n_s + 64) bits is separated into three sets. Let the sets be denoted as $A_D^{\,a}, A_D^{\,b}$ and $A_D^{\,c}$. $A_D^{\,a}$ contains the first 32 bits, $A_D^{\,b}$ contains the next 32 bits and $A_D^{\,c}$ contains the remaining n_s bits of $A_D^{'}$.
4. The 64-bit authentication signature $A_s^{'}$ is computed from the reconstructed block $C_b^{'}$ according to Eq. 5.

$$A_s^{'} = H(C_b^{'}, K, I_b, I_K, M_2, N_2, A_D^{\,c}) \quad (5)$$

Two sets $A_s^{\,a}$ and $A_s^{\,b}$ are defined such that $A_s^{\,a}$ contains the first 32 bits and $A_s^{\,b}$ contains the remaining 32 bits of $A_s^{'}$.

5. The segment $S_c^{'}$ of size (n_s +32) bits is computed and it consists of two sets. $S_1^{'}$ contains the first 32 bits, which is computed in Eq. 6. The other set $S_2^{'}$ contains the remaining n_s bits and is equal to $A_D^{\,c}$.

$$S_1^{'} = A_D^{\,a} \oplus A_s^{\,a}$$
$$S_1^{''} = A_D^{\,b} \oplus A_s^{\,b} \quad (6)$$

6. The reconstructed block $C_b^{'}$ is authentic, if the following condition is satisfied.

$$S_1^{'} = \left(S_1^{''}\right)^c \tag{7}$$

7. If necessary, zero-padding is performed such that the size of $S_c^{'}$ becomes equal to 108 bits. During verification of all blocks in the sequential order, the segments $S_c^{'}$ are concatenated to obtain the sequence $P_s^{'}$.
8. The sequence $E_s^{'}$ is obtained after reverse permutation of $P_s^{'}$ by using the secret key *K*. The [*n, k*] BCH decoder is applied in each *n*-bit segment of $E_s^{'}$ to obtain $A_c^{'}$. When the number of error bits in $E_s^{'}$ is within the correction capability of ECC decoder, the original character sequence can be correctly extracted from the ASCII code sequence $A_c^{'}$.

3 Results and Discussion

In this section, we present simulation results by constructing the erasable watermark for the proposed restoration method. The authentication signature to be used in this method is the 64-bit Hashed Message Authentication Code (HMAC) generated using MD5 hash function [8]. The values of *n* and *k* are chosen to be 63 and 18 respectively for BCH error correction coding. Using this scheme, a total of 10 error bits can be corrected in a 63-bit ECC encoded segment. Fig. 1 shows the original image and the watermarked image after pixel-wise embedding of *m* in each block. It is observed that although background noise is present in the watermarked image, the text can still be read and understood by the user. Without any tampering, all blocks in the watermarked images are verified and the original image can be restored at the blind detector. We perform multiple alterations such as character deletion, block swapping, character insertion and character substitution in the watermarked image; (1) the only word '*information.*' in last line is deleted, (2) first three blocks are swapped with the last three blocks, (3) the word 'theory' is inserted in the last line, and (4) the word 'for' in line-5 is substituted by the word 'to'. The resulting attacked image is shown in Fig. 2. Blind detection is performed on the attacked image for restoring the original character sequence after tamper localization. In Fig. 2, the authentic reconstructed blocks are shown after watermark erasing process. The detector correctly localizes the 13 tampered blocks shown in dark regions. A total of 753 error bits are corrected during ECC decoding and the original character sequence is extracted. The number of error bits in each 63-bit segment of the extracted sequence $E_s^{'}$ is shown in Fig. 3. Since the number of the error bits in each segment does not exceed 10, the correct extraction of original character sequence is possible after ECC decoding. The performance of the proposed algorithm can be compared with the previous method that also addresses the restoration issue in [4]. As discussed in Section 1, in this method there is a possibility of false tamper detection and restoration failure. The restoration of a character at any location depends on the watermark which is

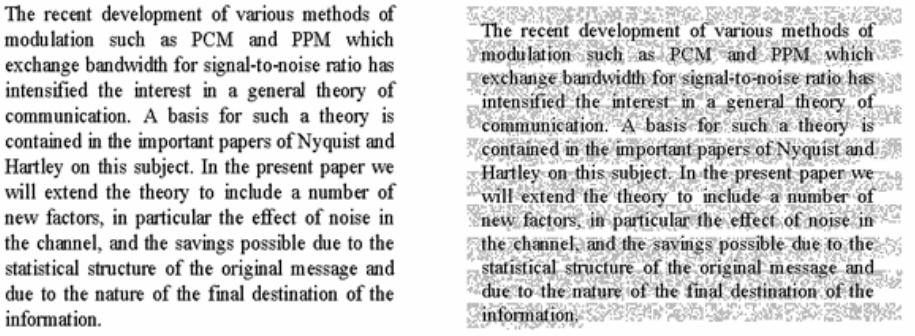


Fig. 1. Left: Original image of 320×440 pixels, right: the watermarked image after embedding the message *m* in each block

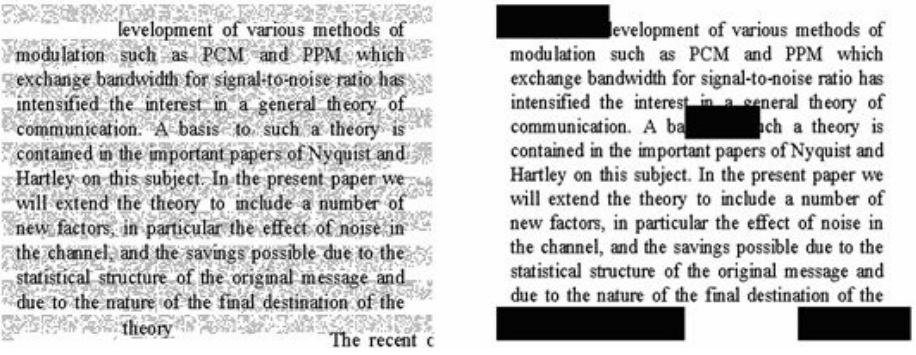


Fig. 2. Left: Attacked image, right: the authentic reconstructed blocks are shown after watermark erasing process and a total of 13 tampered blocks are shown in dark region

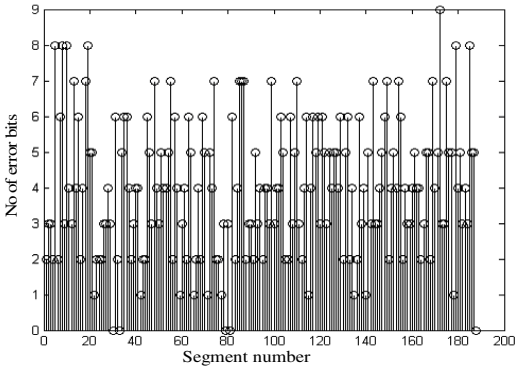


Fig. 3. The number of error bits in each 63-bit segment of the extracted bit sequence $E_s^{'}$

embedded in another character. So the information necessary for restoration is concentrated at certain locations rather than being distributed in the whole image. Thus the restoration capability is reduced in case of certain alterations. After multiple alterations like combined deletion and insertion of characters, restoration is not possible using this method. In the proposed method, the necessary information is embedded in the whole image. So there is no restoration failure even after multiple alterations as demonstrated by the simulation results. A large number of error bits could be corrected by ECC coding since significant amount of information bits are embedded using this method. The restoration capability of the proposed method is limited by the number of error bits in the extracted sequence. If the document image is most edited, then ECC coding could not correct the errors beyond a limit. However, the tampering in the document image can still be localized with high probability. Due to the use of cryptographic hash function, the security of the proposed method is equivalent to the cryptographic authentication.

4 Conclusion

In this paper, we proposed a new watermarking method that is useful in restoration of the original character sequence after tamper localization in text document images. The proposed method can localize various content alterations in the image with high probability and accuracy. It is possible to correct large number of error bits using ECC coding. An erasable watermark was constructed by combining the compressed data, HMAC of the block and ECC encoded ASCII code sequence. After embedding process, the user could easily interpret the text document in the presence of noise due to the inherent ability of human vision. Using the proposed method, the original character sequence can be effectively restored after multiple alterations such as character deletion, block swapping, character insertion and character substitution.

References

1. M. D. Swanson, M. Kobayashi, A. H. Tewfik: Multimedia Data-Embedding and Watermarking Technologies. Proc. of the IEEE, vol. 86, no. 6, (1998)
2. J. Lee, C. S. Won: Authentication and Correction of Digital Watermarking Images. Electronics Letters, vol. 35, no. 11, 886-887, (1999)
3. J. Fridrich, M. Goljan: Images with Self-correcting Capabilities. Proc. IEEE International Conference on Image Processing, vol. 3, 792-796, (1999)
4. A. Makur: Self-embedding and Restoration Algorithms for Document Watermark. Proc. IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2, 1133-1136, (2005)
5. J. Fridrich, M. Goljan, M. Du: Invertible Authentication. Proc. of SPIE, Security and Watermarking of Multimedia Contents, (2001)
6. R. E. Blahut: Theory and Practice of Error Control Codes. Addison-Wesley Publishing Company, (1983)
7. P.W. Wong, N. Memon: Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification. IEEE Trans Image Processing, vol. 10, no. 10, (2001)
8. R. L. Rivest: RFC 1321: The MD5 Message-Digest Algorithm. Internet Activities Board, (1992)

The Study of RED Algorithm Used Multicast Router Based Buffer Management

Won-Hyuck Choi, Doo-Hyun Kim, Kwnag-Jae Lee, and Jung-Sun Kim

School of Electronics, Telecommunication and Computer Engineering,
Hankuk Aviation University, Seoul, Korea
rbooo@korea.com, kdh_r2@yahoo.com, kjlee@seonam.ac.kr,
jskim@mail.hangkong.ac.kr

Abstract. The CBT method that is a dual shared tree enables both a transmitter and a recipient to exchange data through the shortest route to the center core. This method is used to solve the expansion problem of multicast by using a core. However, the current Multimedia Routing Method that involves a packet switching of Best-Effort method tries to transmit only single packet; thus, it causes congestion of data around a core and RP. As for a solution of this problem, the thesis suggests the RED algorithm applied Anycast method. The Anycast Protocol is a way to transfer a packet to an optimal server or a host that is able to redistribute the packet to the most adjacent router or Anycast group members who have Anycast address. Likewise, Anycast redistributes a packet of a core, then again redistributes it to any adjacent router and transmits the packet to an optimal host. In order to redistribute a packet of Core, RED Algorithm is adopted to disperse traffic congestion around the core.

1 Introduction

Due to radical development of internet, various traffic and service have arisen current. Also because of high-speed technology of data transmission, demand to the multimedia used service increases in number rapidly. The multimedia service becomes unconventional and is provided as a synthesized condition of vocal, audio, image, etc. According to this movement, not only data for workstation but data of personal computer gets increasing to be processed. In addition, it shows radical generalization on multimedia communication that uses varieties of data for practical application.

If multimedia communication uses previously used Unicast type technology as its communication method, it would be difficult to improve service of communication because of its limit on transmitting speed and data processing problem. Thus in order to solve such problems in transmission speed and better service, multicast method is highly recommended.

The Multicast Method is a method that multiple transmitters transmit data to more than one recipient who belongs to a same multicast group. The multicast communication is used for video conference, online lecture, renewal and repair of distributed database, software distribution of new version, and so forth and it is considered as an important mechanism in modern technology nowadays.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 669 – 678, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The Multicast Router Protocol transfers users' commonly requesting data through a group, and then a single transmitter provides multicast data to every participant who has subscribed to the group. The Multicast Router Protocol has source-based tree type, such as DVMRP and MOSPF and covalent-based tree type, such as CBT and PIM-SM.

Especially, The CBT method that is a dual shared tree enables both a transmitter and a recipient to exchange data through the shortest route to the center core. This method is used to solve the expansion problem of multicast by using a core. However, the critical problem of CBT is data congestion of Core and Poor Core phenomenon of other routers except for the core.

The current Multimedia Routing Method that involves a packet switching of Best-Effort method tries to transmit only single packet; thus, it causes congestion of data around Core and RP. In order to control such congestion, it is necessary to give feedback regarding the condition of the network congestion to a transmitting station and an adjacent router, which is B-core.

The transmitting station catches such bustling signal and has to reduce the amount of packet to the core router. At this moment, the adjacent router from the core faces Poor phenomenon as data approaches exclusively around the core. The figure 1 and 2 show the problems of CBT that generates data congestion at the core and Poor Core Phenomenon at other router. In order to prevent such problems, packets have to be dispersed and transmitted to B-core router of other side, not to the core router.

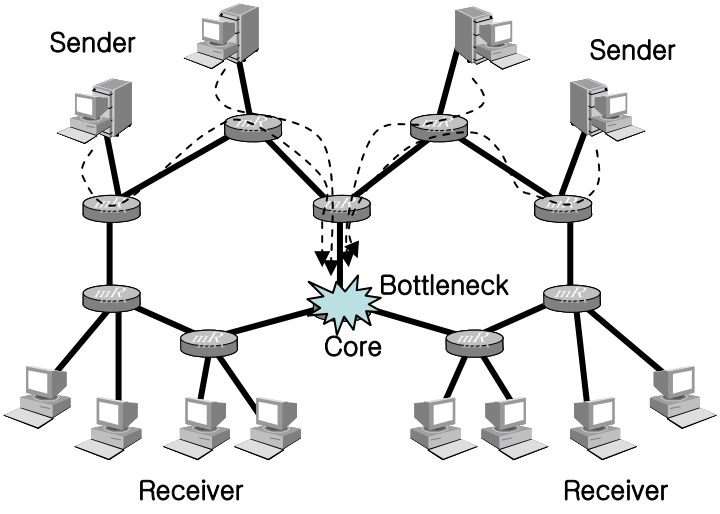


Fig. 1. Bottleneck Core Phenomenon

As for a solution of this problem, the thesis suggests the RED algorithm applied Anycast method. The RED Algorithm is a function to sense ahead and controls the possible congestion on the network by applying average size of cue to a buffer.

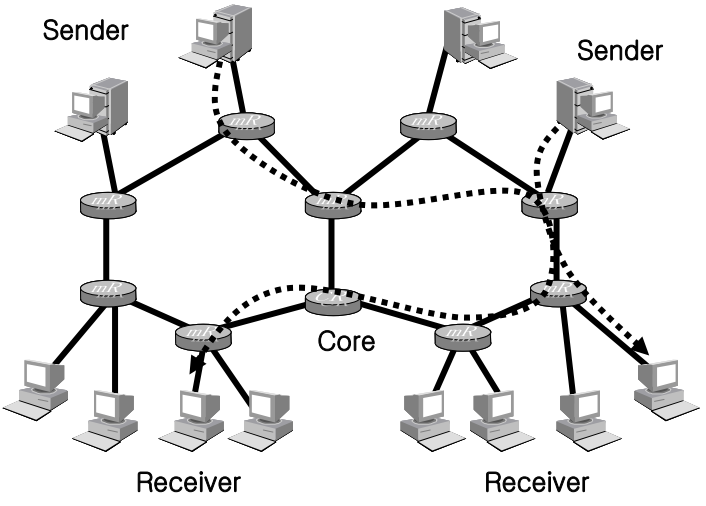


Fig. 2. Poor Core Phenomenon

The Anycast Protocol is a way to transfer a packet to an optimal server or a host that is able to redistribute the packet to the most adjacent router or Anycast group members who have Anycast address. Likewise, Anycast redistributes a packet of Core, then again redistributes it to any adjacent router; thus, it transmits the packet to an optimal host. In order to redistribute a packet of the core, RED Algorithm is adopted to disperse traffic congestion around the core[1].

For the experiment on the subject in the thesis, CBT Method that is classic multicast protocol and Packet dispersed Anycast Method that applies RED Algorithm are actualized with simulation and also their performance are evaluated.

As for the further study in the thesis, the chapter 2 explains CBT Protocol and RED Packet Buffer Management Algorithm among the classic multicast methods. The chapter 3 suggests Characteristics of RED Algorithm utilized Anycast, the chapter 4 analyzes result of the simulation and performance about the proposed methods, and the chapter 5 gives conclusion for the study[2].

2 The Characteristic of Shared Tree Multicast

ShT (Shared Tree) is symbolized as (*, G). The * sign indicates every source and the letter G indicates a group. It is Shared Tree, therefore, the actual size of tree has its value as O(|G|) regardless of number of source. The cost according to the construction of the tree is reasonable; however when the number of source increases, it may cause serious traffic delay. ShT treats fairly small bandwidth and it is suitable when multicast service is started in the network with a large number of transmitters.

Among the Classic Multicast Methods, CBT (Core Based Tree) protocol is relatively recently proposed Multicasting Routing Method. It shares a single tree with all members of a group and uses expandable method that constitutes a bidirectional single tree regardless of a transmitter.

The multicast traffic that was sent to each group is transmitted through same route by having nothing to do with a transmitter. The Shared Tree is connected to the center of the Core Router and the Core Router is selected through multicast address and Hash Function that does mapping to a router. Each router has identical Hash Function, thus, only a core can be selected to each group. For this reason, the routing route of the whole tree is formed around the Core Router.

The multicast packet goes through the core, which causes increase of link cost and as it is described in the Figure 1, there happens Bottleneck due to the traffic congestion around the core. In other words, the bottleneck occurs because of the requested link speed, which happens when bandwidth of the Core link gets exceeded. If such situation becomes more serious, it may lead to deadlock of the Core link and eventually causes the discontinuance of multicast service[3][4].

2.1 RED Algorithm

The RED Algorithm is a function to sense ahead and controls the possible congestion in network by applying average size of cue (q_{ave}) to a buffer. As the RED is a method to let a transmitter know the possible congestion, it uses methods to mark or discard ECN (Explicit Congestion Notification) beat of a packet.

As the [Fig 1] shows if the average size of cue is $(q_{ave})<(min_{th})$, the packet is placed in the router, but if the average size of cue is $(q_{ave})>max_{th}$, every packet is being discarded. In addition, if min_{th} < *the average size of cue* $(q_{ave})max_{th}$, then it marks or discards the packet randomly. The RED Algorithm can be controlled by various intermediating variables as it is demonstrated in the [Fig 1]. The value of w_q is the value of constant that decides level of effect to the average size of cue.

$$q_{ave}=(1-w_q)q_{ave}+w_q q_{ave}$$

Table 1. The Various Intermediating Variables of RED

min_{th}	minimum Threshold of branch off Anycast
max_{th}	maximum Threshold of branch off Anycast
w_q	calculate average queue weight value
max_q	maximum probable value of Anycast determination

Like the formula above, the size of current and past packet amount are up to the value of w_q. The selection of min_{th} and max_{th} value is depended on characteristics of network. If the value of min_{th} becomes too small, the range of cue that administers packets of router gets too larger too, thus, the practical application of the Core router gets decreased. In contrast if the value of min_{th} becomes too large, frequent congestion occurs. Also when max_{th} value is too small, the network utilization gets decreased because the usable size of router becomes too small to be applied. When max_{th} value is big, it operates like Drop Tail.

Likewise, selecting the value of various intermediating variables is very important to RED Algorithm and its performance will be strongly influenced according to the value. The Table 2 indicates recommending value of each intermediating variables to set the value of various RED intermediating variables.

Table 2. The Recommending Value of Various RED Intermediating Variables

min_{th}	5 packet
max_{th}	three times of max_{th}
w_q	0.002
max_q	0.02~0.1

3 RED Algorithm Used Anycast Dispersion Method

The thesis postulates formation of group members by using the following network topology. Based on the formation of given nodes, a data packet is transmitted. With the RED Algorithm , CBT type core method is selected if it is smaller than the critical value and Anycast method is introduced if it is larger than the critical value. Also, CBT floods the adjacent routers during RTT, the fixed round trip time of a router, to inform the current situation in the network to transmitters. Thus, the transmitters can have the shortest access route to routers[7][8].

3.1 RED Applied Anycast

The existing CBT method is a suggested way only to maintain the Share Multicast Method that involves directly service related routers or networks. It causes congestion round a Core router; therefore, the RED Algorithm is introduced to solve this problem.

CBT joins to a group after a host's transmission of a report to the host membership of IGMP. The message received local routers also transmit join_request message to the core router. Such message for the participation is approved by the core router and since then, transmission and reception are made at the core router after the process of member approval is done.

This proposed method introduces RED Algorithm to the Core of CBT. The traffic of core router in CBT is controlled by the recommending value of each intermediating variables. The following is the RED applied traffic control algorithm.

The using Anycast address at this moment is for defying server groups, which provide similar service in the network. The main reason why Anycast is classified as a network service by a recipient in comparison with the existing communication method, the recipient who wants to get service for certain information shares identical characteristics with other recipients that are appointed through the Anycast address.

Thus, any transmitter who is related with such Anycast recipients transmits a data packet to the Anycast address and a router transmits the packet to a nearest recipient from the transmitter. A certain router measures distance which recipient is nearest and it takes charge of the selection of a recipient.

Because of its inner property of Anycast Routing Protocol, the characteristic of Unicast is expanded; therefore, each routing method makes mutual communication internally for the operation process of protocol. The mutual communication of protocol that is made between a transmitter and a recipient of Anycast data is toward a direction from Source to Destination. The following diagram shows CBT conversion to Anycast Method.

```
avg=0
   count=1
   for each packet arrival
       calculate new avg, queue is q_avg:
       if the queue is nonempty
       q_avg=(1-w_q)q_avg+w_q q
   else
       m=f(time q_time)
     q_avg = (1-w_q)m q_avg
   if min_th<q_avg<max_th
    increment count
    calculate probability pa
     p_b: max_p(q_avg=(1-w_q)^m q_avg)
     p_a: pb/(1-count*p_b)
   with probability p_a:
   mark the arriving packet
   count=0
   else count=0
   else count=1
   with queue become empty
   q_time=time
```

Like the above, the RED Algorithm is a technique to predict the transmitting packet of router core once the average packet length of a router core exceeds a certain level. Through such function, the increment of buffer traffic in core router is forecasted.

If avg is bigger than max_{th}, the traffic at the core is increased. At this point, the core transmits non_qu message to a transmitter. The non_qu message received transmitter converts its method to Anycast. The Anycast Protocol operates to transmit a packet to a host or the optimal server that are able to redistribute the packet to a nearest router or local Anycast members, especially who are one of recipient group member and have an Anycast address.

The suggested RED Algorithm applied Anycast Method predicts congestion in CBT with a buffer of router and converts to Anycast. This process enables to transmit data that requires real-time support like multimedia depressively.

In addition, the Poor Phenomenon that is introduced in the introduction can be prevented by the Anycast operation through the appropriate buffer management. As the [Fig 3] shows, the order of CBT conversion to Anycast Method.

4 Simulation

The performance evaluation of RED Algorithm applied Anycast Router is compared and analyzed with the existing multicast based CBT in the thesis. In the experiment, 10 transmitters and 6 groups are presented to vary the number of transmitter and recipient, and each group has maximum three transmitters corresponded to the changes of numbers. The link of each multicast has fixed bandwidth of 105Mbps and 10 ms propagation delay. The simulation measures router reception rate according to the characteristic of each source for 10 minutes. Also the characteristic of data packet is assigned in the simulation. The simulation topology uses a structure as the Figure 4[9].

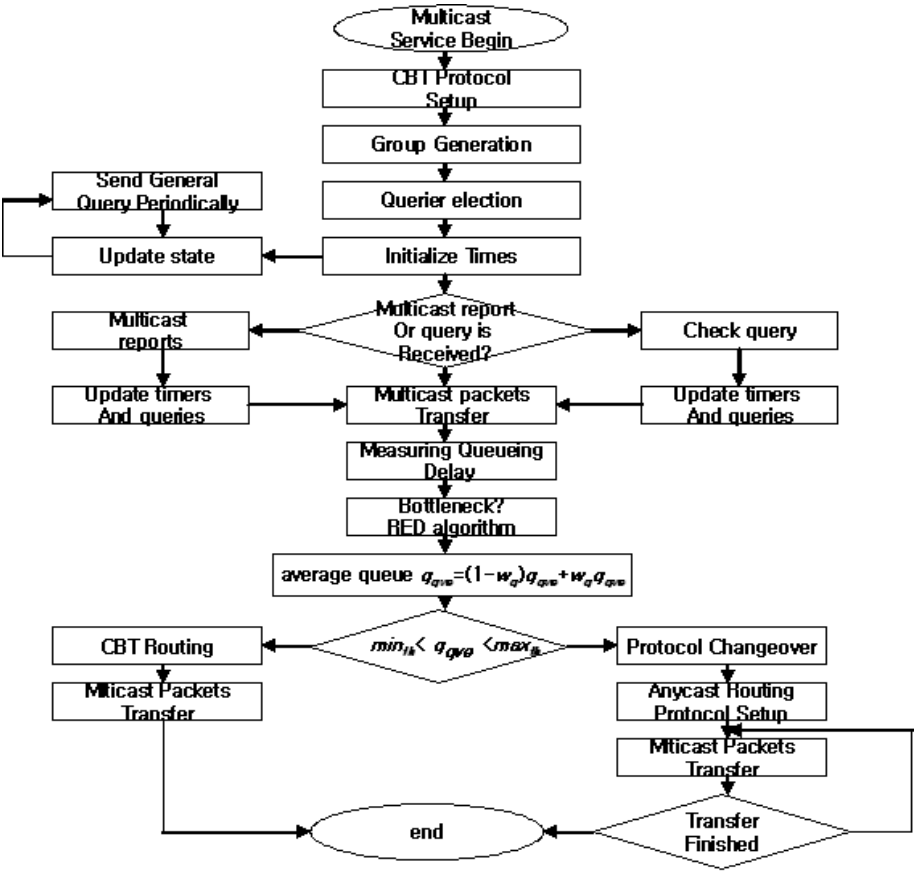


Fig. 3. The Diagram of CBT Conversion to Anycast Method

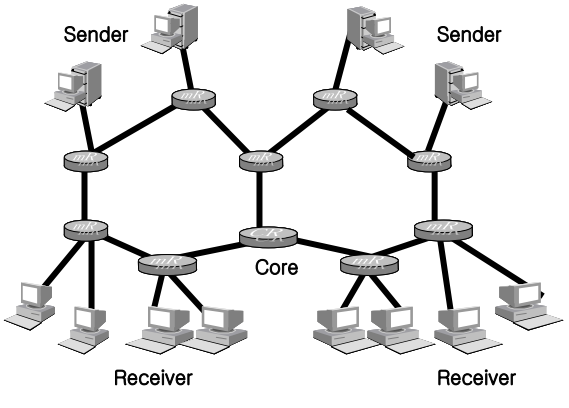


Fig. 4. The Network Topology

For the simulation tool, ns-2 (Network Simulator-Ⅱ) that reflects multicast condition well is used. Also in order to be equitable for the simulation, it randomly picks results after repeating each experiment 100 times and selects an average for the final result.

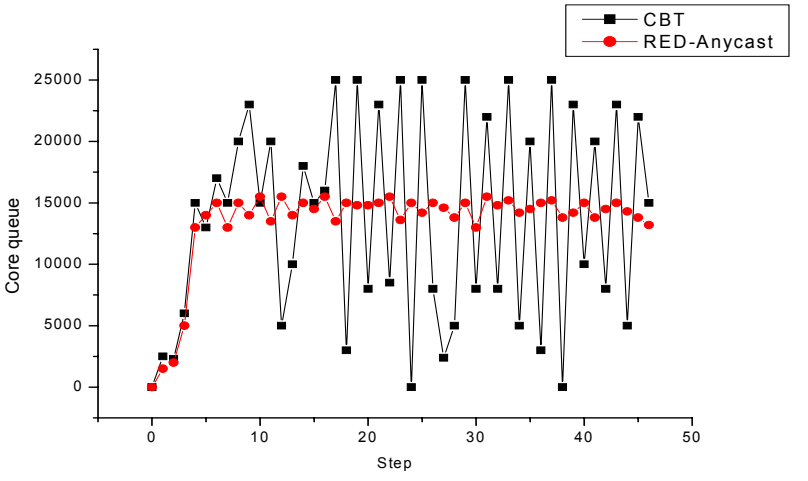


Fig. 5. The Buffer Management in the Multicast Router

The Figure 5 shows the creation and the transmission of a packet from a transmitter to multicast group members. It also measures buffer management according to the packet of multicast router especially for the each routing method.

It is obvious that the characteristic of buffer management for Anycast routing is somewhat superior to CBT routing in buffer management of data transmission. As the data transmission increases, the data processing rate in the CBT routing buffer has radically changed, however in case of Anycast routing, there is no significant change in data increment. For CBT routing, the characteristic of buffer management for the

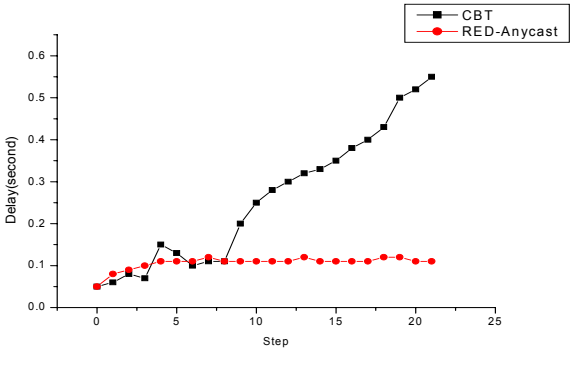


Fig. 6. The Measurement of Average Delay Time at Each terminal when the size of Packet is 512 Byte

data process of Anycast routing shows 38% of improvement and the buffer management rate of Anycast according to the change in size of multimedia data packet is increased to 27%.

The Figure 6 is the result of average delay time at each terminal regarding topology. It is relatively superior to Anycast Routing method; however as time passes, the delay time of CBT routing becomes slow as Anycast routing method does not show any change even in the group number. The delay characteristic of Anycast routing against CBT routing results 14.4% improvement relatively, and the delay time for Anycast according to the size change of multimedia packet gets decreased to average 8.7%.

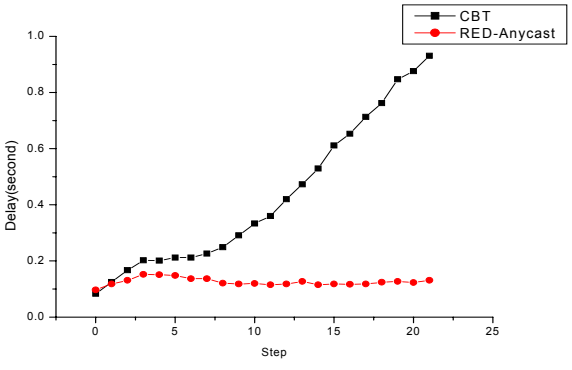


Fig. 7. The Measurement of Average Delay Time at Each terminal when the size of Packet is 1024 Byte

The Figure 7 is result of the average delay time at each terminal when the size of packet is 1024 Byte. The biggest difference between the Figure 6 is that the performance of Anycast Routing has remarkably improved even the number of group is small. Especially when the size of a packet is 1289 Byte, the average delay time at each CBT terminal gets worse because of the congestion around the CBT core router as the size of multicast data packet becomes larger. Eventually, it causes a serious Bottleneck at the links of the core. To summarize the result, Anycast Routing is recommendable routing method that does not be influenced much by the change in packet size and shows stable delay characteristics.

5 Conclusion

The thesis proposed conversion of routing method from CBT Shared Tree Routing method that is stable at the relatively low bandwidth, to Anycast Routing Method that is suitable for traffic dispersion at the high bandwidth depends on the increment of traffic load. In order for the conversion, the study on multicast routing method and the shortcomings of CBT Routing Method are reevaluated. In addition, DEF Cuing Model is applied to analyze the routing method theoretically, and by generalizing the CBT routing method as a cuing model, the cuing delay of CBT core router is measured.

The traffic increment for the core router of CBT Multicast gets increased radically after the tree join of multicast group and along with increment of traffic load in the whole tree, the Bottleneck occurs by the cuing delay at the core router. In order to solve such Bottleneck phenomenon and improve the transmission delay of multicast packet, the conversion of protocol to Anycast Routing Method at the moment when the increment of cuing delay is rapidly made. Along with this process, the routing tables have to be renewed after every multicast group in the simulation topology joined the tree. Through such conversion to Anycast Routing Method, the study confirms the improvement in transmission delay of multicast tree.

References

1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121
2. (a) A. Ballardie, Core Based Trees (CBT) Multicast Routing Architecture, RFC2201, (1997). (b) A. Ballardie, "Core Based Trees (CBT Version 2) Multicast Routing Protocol Specificastion RFC2189, (1997).
3. J.Moy, Multicast Extensions to OSPF, IETF RFC 1584, (1994).
4. K. Ettikan, An Analysis Of Anycast Architecture And Transport Layer Problems, Asia Pacific Regional Internet Conference on Operational Technologies, Kuala Lumpur, Malaysia, Feb.,-March, (2001).
5. J. Lin and S. Paul, RMTP: A Reliable Multicast Transport Protocol, IEEE INFOCOM '96, San Francisco, CA, March (1996).
6. W. Yoon, D. Lee, H.Youn, S. Lee, S. Koh, A Combined Group/Tree Approach for Many-to-Many Reliable Multicast, IEEE INFOCOM'02, June (2002).
7. Christiansen,M.,Jeffay,K.,Ott,D.,and Smith F.D., Tuning RED for Web Traffic, IEEE/ACM Transactions,June, (2001).
8. Floyd, S., Ns Simulator Testrs for Random Early Detection(RED) Gatewqys, Technical Report, October, (1996).
9. "The Network Simulator: ns-2," http://www.isi.edu/nsnam/ns/

Genetic Algorithm Utilized in Cost-Reduction Driven Web Service Selection*

Lei Cao, Jian Cao, and Minglu Li

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 20030, China
{lcao, cao-jian, li-ml}@cs.sjtu.edu.cn

Abstract. A single web service is most likely inadequate to serve the customers' business needs; it takes a selection of various web services composed together to form a business process. The cost is the primary concern of many business processes. In this paper, we propose a new solution using Genetic Algorithm (GA) in cost-reduction driven web service selection. GA is utilized to optimize business process composed of many service agents (SAg), each of which corresponds to a collection of available web services provided by multiple service providers to perform a specific function. Service selection is an optimization process with taking into account the relationships among the services. Better performance has been gotten using GA in the paper than using local service selection strategy.

1 Introduction

Web services are self-describing software applications that can be advertised, located, and used across the Internet using a set of standards such as SOAP, WSDL, and UDDI [1]. A single web service is most likely inadequate to serve the customers' business needs; it takes a selection of various web services composed together to form a business process. Web services composition [2] has been one of the hottest research topics in this new field. However, with the ever increasing number of functional similar web services being made available on the Internet, there is a need to be able to distinguish them using a set of well-defined Quality of Service (QoS) criteria. QoS is a broad concept that encompasses a number of non-functional properties such as cost, response time, availability, reliability, and reputation [3]. These properties apply to both standalone web services and composite web services. Typically, cost and time are two primary factors that customers are concerned about.

The challenge is to select the services that not only satisfy the individual requirements but also best fit the overall composed business process. Therefore, the entire business process needs to be optimized prior to execution. The philosophy of Genetic Algorithm (GA) [4] mimics the evolution process of "the survival of the fittest" in

* This paper has been supported by the 973 project (No.2002CB312002) of China, grand project of the Science and Technology Commission of Shanghai Municipality (No.03dz15027). It is also partly supported by "SEC E-Institute: Shanghai High Institutions Grid", Chinese High Technology Development Plan (No.2004AA104340), Chinese Semantic Grid Project (2003CB317005) and Chinese NSF Project (No.60473092).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 679 – 686, 2005.
© Springer-Verlag Berlin Heidelberg 2005

nature. GA is a parallel global optimization algorithm with high robustness, and it is not restricted by natures of the optimization problem, such as continuity and differentiability. GA is very suited to those complicated optimization problems that cannot be handled efficiently by traditional optimization algorithms (methods of calculus or techniques of exhaustive search). In this paper, we take the cost as the primary concern of many business processes. Using an integer string to represent a web services composition, the best one is the string that leads to the lowest cost. A service selection model using GA is proposed to optimize business process composed of many service agents (SAg), each of which corresponds to a collection of available web services provided by multiple service providers to perform a specific function.

The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 describes GA utilized in our web service selection model. In Section 4, a service selection case is presented. Finally, Section 5 concludes the paper.

2 Related Work

In [2], authors present a framework SELF-SERV for declarative web services composition using state-charts. Their service selection method uses a local selection strategy. The service selection is determined independently to other tasks of the composite services. It is only locally optimal. In [5], authors make research on the end-to-end QoS issues of composite service by utilizing a QoS broker that is responsible for coordinating the individual service component to meet the quality constraint. Authors design the service selection algorithms used by QoS brokers to meet the end-to-end QoS constraints. The service selection problem is modeled as the Multiple Choice Knapsack Problem (MCKP). Authors give three algorithms and recommend the Pisinger's algorithm as the best one. But the solution is usually too complex for run-time decisions. In [6], authors present the Web Services Outsourcing Manager (WSOM) framework via a mathematical model for dynamic business processes configuration using existing web services to meet customers' requirements, and propose a novel mechanism to map a service selection space {0,1} to utilize global optimization algorithms - Genetic Algorithms. The binary encoding is adopted in the algorithm, but it is not human-readable. When the business process is complicated, the chromosome will be too lengthy. Moreover, GA utilization in the service selection is not given in detail, especially about how the relationships among the services will affect the global optimization.

3 GA Utilized in Web Service Selection Model

3.1 Web Service Selection Model

The web service selection model is based on a multi-agent platform, which is composed of multiple components (See Fig. 1).

1. **Service Agent (SAg)** corresponds a list of web services that have the specific function. Its capability depends on the specific function. SAg possesses all useful information about those web services, such as cost, response time, and so on.
2. **UDDI agent (UAg)** is the broker of inner SAg when asking for outer UDDI registry. SAg might get information about required web services via UAg.

3. **Information agent (IAg)** is the information center of the model. All SAg must register themselves with it.
4. **Service composition agent (SCAg)** is responsible for administering the composite business process as a services flow (SF). Using the modeling tool, users can model the SF in advance. The model is composed of many predefined or customized SAg. GA module is the core of SCAg. SCAg will activate GA module to optimize the service selection to get the final executable SF.
5. **ACL channel** [7] is the communication bus among those agents mentioned above.

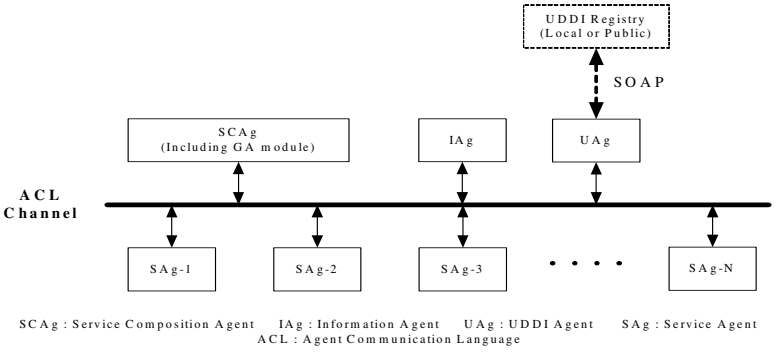


Fig. 1. Model architecture

3.2 Problem Objective Function

When the cost becomes the sole primary factor that customers are concerned about, the service selection is equivalent to a single-objective optimization problem. There are mainly two kinds of service flow:

1. **Pipeline service flow.** For a pipeline services flow that has N steps (N SAg in an execution path) $(S_1, S_2, ..., S_N)$, it only has one type structure: sequence one. The control flow will come through all its SAg, and only one proper web service in each SAg will be chosen to bind to. The overall cost of the SF is equal to the summation cost of all its components.

$$\cos t(SF) = \sum_{i=1}^{N} \cos t(S_{ik}),\ 1 \le k \le size(S_i) \qquad (1)$$
$$size(S_i) = |S_i|,\ i = 1, ..., N$$

   The objective function is: $Min\ \cos t(SF)$

2. **Hybrid service flow.** For a hybrid services flow that has N steps (No more than N SAg in an execution path) $(S_1, S_2, ..., S_N)$, we assume it has three type structures: sequence one, parallel one and conditional branching one. The overall cost of the hybrid SF is described as follows:

$$\cos t(SF) = \sum_{i=1}^{N} \cos t(S_{ik}) * CF_i,\ 1 \le k \le size(S_i)$$
$$size(S_i) = |S_i|,\ i = 1, ..., N$$
$$CF_i \in \{0,1\},\ i = 1, ..., N$$

For some SAg linked by conditional branching structure, only one of them can be passed by the control flow at one execution. Thus the hybrid SF has some different execution path. The overall cost of each execution path can always be represented by the summation cost of its subset components.

$$\begin{aligned} &cost(SF) = \sum_{i=1}^{M} cost(S_{ik}),\ 1 \le k \le size(S_i) \\ &size(S_i) = |S_i|,\ i = 1,...,M \\ &M \le N \end{aligned} \tag{1'}$$

The objective function is also: $Min\ cost(SF)$.

3.3 GA Module

A fitness function is firstly defined to evaluate the quality of an individual, called a chromosome, which represents a solution to the optimization problem. GA starts with a randomly initialized population. The population then evolves through a repeated routine, called a generation, in which GA employs operators such as selection, crossover, and mutation borrowed from natural genetics to improve the fitness of individuals in the population. In each generation, chromosomes are evaluated by the fitness function. After a number of generations, highly fit individuals, which are analogous to good solutions to a given problem, will emerge.

3.3.1 Solution Representation

The representation scheme determines how the problem is structured in GA and also influences the genetic operators that are used. Traditional GA encodes each problem solution into a binary string, called chromosome, which facilities studying GA by schema theorem. However, the binary encoding is not human-readable, which makes it difficult to develop efficient genetic operators that can make good use of the specialized knowledge for the service selection problem. The integer encoding provides a convenient and natural way of expressing the mapping from representation to solution domain. With integer encoding, the interpretation of the solution representation for service selection is straightforward, so the integer encoding is used in this paper. The solution to service selection is encoded into a vector of integers, where the i^{th} element of an individual is k if the k^{th} web service in SAg i is selected. For example, the service flow has four SAg $\{S_1, S_2, S_3, S_4\}$, each of which has four candidate services, and one possible individual is $[1\,4\,2\,3]$. Then the corresponding composite service is $\{S_{11}, S_{24}, S_{32}, S_{43}\}$, which means the 1^{th} service of S_1, the 4^{th} service of S_2, the 2^{th} service of S_3 and the 3^{th} of S_4 are selected concurrently.

3.3.2 Fitness Function

The fitness function is responsible for evaluating a given individual and returning a value that represents its worth as a candidate solution. The returned value is typically used by the selection operator to determine which individual instances will be allowed

to move on to the next round of evolution, and which will instead be eliminated. Highly fit individuals, relative to the whole population, have a higher probability of being selected for mating, whereas less fit individuals have a correspondingly low probability of being selected. To facilitate the selection operation in GA, the global minimization problem is usually changed into a global maximization problem. Through transforming Eq. (1) or Eq. (1'), the proper fitness function for the service selection problem can be obtained**:**

$$F = \begin{cases} U - \sum_{i=1}^{N} \cos t(S_{ik}) & \cos t(SF) < U,\ 1 \le k \le size(S_i) \\ 0 & \cos t(SF) \ge U. \end{cases} \tag{2}$$

Where U should select an appropriate positive number to ensure the fitness of all good individuals are positive in the feasible solution space.

3.3.3 Population Initialization

The population initialization creates the first population and determines the starting point for the GA search routine. An individual is generated by randomly selected web service for each SAg of the services flow, and the newly generated individual is immediately checked to see whether the corresponding solution satisfies the constraints. If any of the constraints is violated, then the generated individual is regarded as invalid and discarded. The process is repeated until enough individuals are generated. The population size $pop_{size} \in [20,100]$ is generally recommended**.**

3.3.4 Genetic Operators

GA uses ***selection operator*** to select the superior and eliminate the inferior. The individuals are selected according to their fitness - the more suitable the more chances they have to reproduce. For the service selection problem, the popular roulette wheel method is used to select individuals to breed. This method selects individuals to reproduce based on their relative fitness, and the expected number of offspring is approximately proportional to that individual's performance. GA uses ***crossover operator*** to breed new individuals by exchanging partial chromosome of the mated parents in stochastic mode. The crossover probability $p_c \in [0.5,1.0]$ is generally recommended. The purpose of crossover is to maintain the qualities of the solution set, while exploring a new region of the feasible solution space. For the service selection problem, the single-point crossover operator is used (See Fig. 2). After each crossover operation, the offspring are immediately checked to see whether the corresponding solutions are valid. GA uses ***mutation operator*** to add diversity to the solution set by replacing an allele of a gene by another in some chromosomes in stochastic mode. The mutation probability $p_m \in [0,0.05]$ is generally recommended. The purpose of mutation operation is to avoid losing useful information in the evolution process. For the service selection problem, the stochastically selected SAg is randomly bound to a web service different from the original one (See Fig. 2). After an offspring is mutated, it is also immediately checked to see whether the corresponding solution is valid.

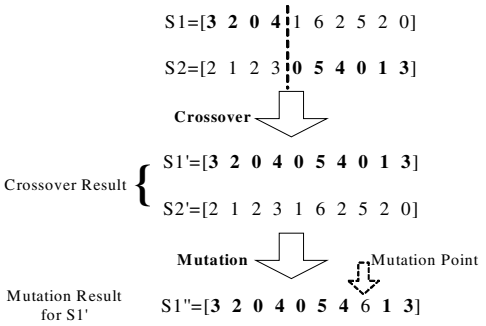


Fig. 2. Crossover operator and Mutation operator

4 A Case Study and the Implementation

There are twenty tasks in a pipeline SF and each of them can be accomplished by a SAg (See Fig. 3). The SF may be a practical "Travel arrangement" composite service. Each kind of services might have multiple providers, each of which could provide several kinds of services. Those business entities may be competitors and do not work with each other. In other way, they would prefer to work with those within the alliance, which claims that there should be some degree discount if customers select some of them. Due to those business rules, there are relationships such as partnerships, alliances, competitors among the services. We must take into account those relationships when select the final services as part of the optimization process.

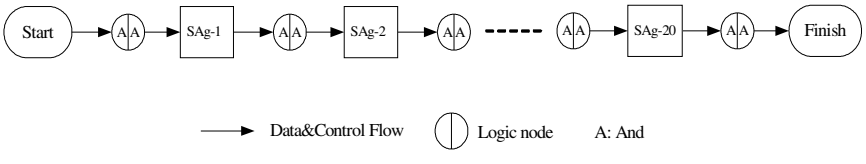


Fig. 3. A pipeline services flow

4.1 Initial Data

Each SAg might have various candidate numbers from 5 to 15 predefined at random. Different choices of web services in each SAg might have different cost (Cost data are omitted here). For Eq. (2), we take the summation of all SAg's maximum cost as the constant U, and get U =**$24385**. The individual fitness can be regarded as how much can be saved after the SF execution if the travel agency has been given the money U to arrange the traveler's trip. If we use the local selection strategy, we might get a final services sequence as follows:

$$[7\ 8\ 5\ 3\ 12\ 15\ 5\ 11\ 1\ 1\ 5\ 11\ 10\ 4\ 13\ 4\ 3\ 9\ 11\ 5]$$

And its overall cost is the summation of all SAg's minimum cost, so it equals to **$22777**. In this case, the individual fitness is **$1608.** According to some given business rules, we define constraints and corresponding actions. For the special individ-

ual, its fitness should be modified with the discount given. And for the invalid individual, it will be discarded. In the paper, some GA parameters are given as follows:

- Solution representation: Integer encoding
- Population size: $pop_{size} = 40$
- Crossover probability: $p_c = 0.8$
- Mutation probability: $p_m = 0.05$
- Maximum generations: $Gen_{\max} = 2000$

4.2 Evolution Procedure

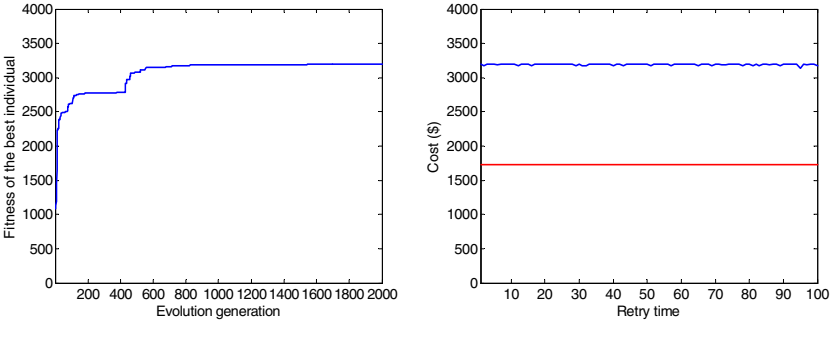
The initial population is randomly generated. After $Gen_{\max}$ generations, we have gotten the final optimal solution:

$$[1\ \ 2\ \ 2\ \ 3\ \ 4\ \ 15\ \ 5\ \ 2\ \ 2\ \ 5\ \ 5\ \ 2\ \ 10\ \ 4\ \ 2\ \ 4\ \ 3\ \ 5\ \ 11\ \ 5]$$

The solution fitness is **$3193**. It is also represented as

{ **SAg1**:WS1, **SAg2**:WS2, **SAg3**:WS2, **SAg4**:WS3, **SAg5**:WS4,
SAg6:WS15, **SAg7**:WS5, **SAg8**:WS2, **SAg9**:WS2, **SAg10**:WS5,
SAg11:WS5, **SAg12**:WS2, **SAg13**:WS10, **SAg14**:WS4, **SAg15**:WS2,
SAg16:WS4, **SAg17**:WS3, **SAg18**:WS5, **SAg19**:WS11, **SAg20**:WS5 }.

We can see the GA evolution process from Fig. 4(a). The best fitness of the population has a rapid increase at the beginning of the evolution process then convergences slowly. It also means the overall cost of the SF is generally decreasing with the evolution process. For better solution, the whole optimization process can be repeated for a number of times (100 times in this paper, and different initial population in each time), and the best one in all final solutions is selected as the ultimate solution to the service



(a) GA evolution process (b) Multi-time GA optimization

Fig. 4. Evolution procedure

selection problem. From Fig. 4(b), we can see the result. The red line represents the solution using local selection strategy. The blue curve represents all best solutions at 100 GA tests. Apparently, all best solutions using GA have better fitness than those using local selection strategy. In addition, the fact about their similar fitness proves that our GA could find the global optimal solution to the service selection problem.

5 Conclusion and Future Work

Web services composition has been one of the hottest research topics. The cost is the primary concern of many business processes. In the paper, we propose a new solution using Genetic Algorithm (GA) in cost-reduction driven web service selection. Using an integer string to represent web services composition, the best one is the string that leads to the lowest cost. Service selection is an optimization process with taking into account the relationships among the services. Seen from the experiment result, better performance can be gotten than that using local service selection strategy. The global optimal solution can be achieved with GA within short time. In future, we will consider more QoS factors besides cost as the objectives of service selection and implement multi-objective global optimization using GA.

References

1. F. Curbera et al.: Unraveling the web services: An introduction to SOAP, WSDL, and UDDI. IEEE Internet Computing, Vol. 6, No. 2, Mar./Apr. (2002) 86-93. IEEE Press.
2. B. Benatallah, M. Dumas, Q. Z. Sheng, and A. H. Ngu.: Declarative composition and peer-to-peer provisioning of dynamic web services. In Proc. of the Int. Conference on Data Engineering (ICDE), San Jose CA, USA, February (2002) 297-308. IEEE Press.
3. J. O'Sullivan, D. Edmond, and A. ter Hofstede.: What's in a Service? Towards accurate description of non-functional service properties. Distributed and Parallel Databases Journal - Special Issue on E-Services, Vol. 12, No.2-3(2002) 117-133,. Kluwer.
4. Stuart Russell, Peter Norvig: Artificial Intelligence: A Modern Approach (Second Edition). Prentice Hall, (2002)
5. Y. Liu , AHH Ngu and L. Zeng: QoS Computation and Policing in Dynamic Web Service Selection. In Proc. 13th Int. Conf. World Wide Web (WWW), New York, NY, USA, May (2004) 66-73. ACM Press.
6. Liang-Jie Zhang, Bing Li.: Requirements Driven Dynamic Services Composition for Web Services and Grid Solutions. Journal of Grid Computing, Vol. 2, No. 2., June (2004) 121-140. Kluwer.
7. FIPA Specifications (http://www.fipa.org)

MacroOS: A Pervasive Computing Platform Supporting Context Awareness and Context Management

Xiaohua Luo, Kougen Zheng, Zhaohui Wu, and Yunhe Pan

College of Computer Science, Zhejiang University,
Hangzhou, Zhejiang Province, China
lxhzju@cs.zju.edu.cn

Abstract. The applications of pervasive computing in smart home constitute a hot research topic. It is time for the concept of the connected home to become a reality. In this paper, we present a pervasive computing application platform, MacroOS. MacroOS is built on the ensemble of multiple technologies and some novel pervasive computing technologies. It consists of micro embedded operating system which supports wireless network sensors, pervasive network which supports various communication protocol and context middleware which meets the requirement of context awareness and context management. Furthermore, based on the MacroOS platform, we implement successful demonstration applications in smart home system.

1 Introduction

Although pervasive computing still falls short of Mark Weiser's "walk in the woods," advances in hardware, software, and networking over the past 12 years have brought his vision close to technical and economic viability [1]. The widespread adoption of digital technologies including real-time capture, storage, processing, display, and analysis of the information generated will truly revolutionize a wide variety of application domains. Furthermore, the developing semantic and adaptive middleware architecture technology provides strong layers of abstraction and isolates applications from the details of embedded device hardware. Naturally, these activities will take place directly in the home and it will take on extraordinary new importance.

The research project, "Research of Operating System to Support Pervasive Computing (MacroOS)", which is supported by 863 National High Technology Program, focuses on this kind of challenge. According to the ensemble of multiple pervasive computing technologies, we research on the key technologies of micro embedded operating system that supports networks of wireless sensors and the network middleware that meets the requirement of context awareness. Furthermore, based on the MacroOS platform, we implement successful demonstration applications in smart home system which supports context awareness and context management.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 687–694, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 MacroOS Overview

Embedded devices applied in smart home range from tiny wireless nodes to special "power assets". The MacroOS model for pervasive computing should integrate various embedded devices into a single platform and provide severe reliability, quality of service, and invisibility for pervasive applications. As shown in figure 1, the MacroOS platform architecture consists of four layers including sensor platform, pervasive network, context middleware and application layer.

The sensor platform layer consists of the hardware, associated drivers, and real-time OS kernel. It is in charge of the various pervasive devices, ranging from tiny motes [2] to relatively big devices like PC.

The pervasive network layer can be divided into two sub layers. The MAC sub layer is responsible for moving data packets to and from one network interface card to another across a shared channel. The later provides transparent transfer of data between end systems, or hosts, and is responsible for end-to-end error recovery and flow control.

The semantic and adaptive context service middleware layer includes the sensor information process service and the context awareness service. The primer service deal with data sorting, data integrating and XML based message enveloping. The later service is in charge of the context data discovery, registry, analysis and binding etc. It is developed as a semanteme-integrated and dynamic reconfigurable middleware.

The application layer is built on the middleware layer. it utilizes the interfaces provided by MacroOS and assembles practical applications. We will describe a typical application in detail later.

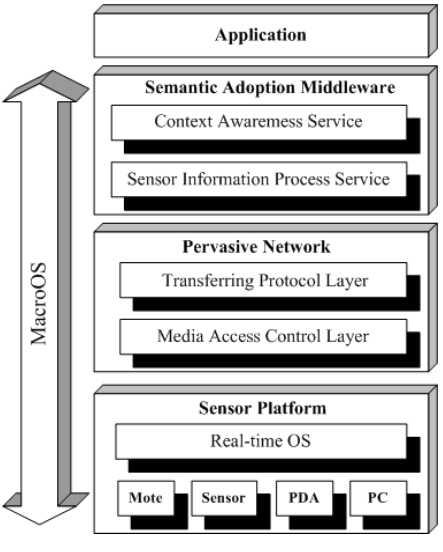


Fig. 1. The architecture of MacroOS

3 Implementation Model

3.1 Embedded Operating System

The sensor platform layer is the hardware/software exchanging system (shown in figure 2). Sensors and other data collecting devices constitute the pervasive computing physical system. They provide for the raw input such as temperature, humidity and image. In theory, pervasive computing should encompass every device worldwide that has built-in active and passive intelligence. Therefore, the corresponding embedded real-time OS should be in duty bound to manage all kinds of physical devices. There is no universal OS can be applied in all of them directly. Traditional operating systems for powerful device such as PDA or PC are well-studied today. The key issue focuses on super-micro sensors.

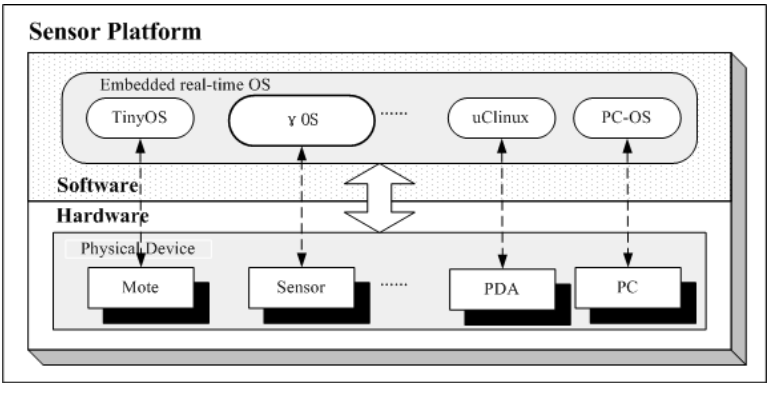


Fig. 2. The Structure of Sensor Platform

The microprocessors used in sensors usually run at several MHz and hold kilobytes of memory. In spite of the resource limitations, software platform for sensors has to meet the requirements of concurrent operation, real-time constraint and reliability. TinyOS [3] is a micro embedded OS designed for wireless sensor network nodes. However, adaptability and fault tolerance has not been considered in its kernel. Since these tiny devices are usually numerous, largely unattended, and expected to be operational for a large fraction of the time, great efforts should be taken to improve their adaptability and reliability.

To provide necessary support for these requirements, we present γOS [4], which provides a flexible scheduling scheme to adapt to the variance of the environment and guarantee the correction of the actions of sensors. In γOS, the primary and subordinate two-level task-scheduling scheme can efficiently isolate the scheduling of time-critical and time-uncritical tasks. Furthermore, its scheduling scheme is extended to tolerate transient faults by re-executing faulty tasks.

3.2 Pervasive Network

Various pervasive computing devices applied in smart home connect to each other and the controlling PC with wired links or wireless RF. Wired networks are well-studied today. However, it is different in wireless communication. Wireless communications solutions have a variety of technical and deployment obstacles. As shown in figure 3, the pervasive network can be divided into two sub layers: MAC layer and transferring protocol layer.

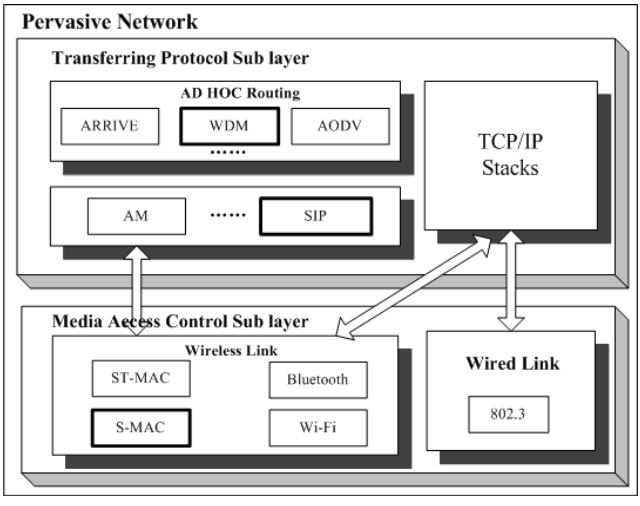


Fig. 3. The Framework of Pervasive Network

The MAC layer frames data for transmission over the network. In tiny mote, for the restrictions of system resource, there is no need of broadband or complex data transfer technologies. Only some energy-efficient CSMA based protocols such as S-MAC [5] can be implemented. Considering the application environment in smart home, we present a TDMA based sensor MAC protocol named ST-MAC [6]. Other devices such as mobile phone and PDA, which are equipped with more powerful communication module, can use standard protocols such as Wi-Fi and Bluetooth.

Above the MAC layer is transferring protocol layer. For powerful devices that support Wi-Fi or Bluetooth, it can easily implement TCP/IP Stacks. However, for tiny devices like network sensors, they must be paid more attention to. AM [7] is used as the basic transferring protocol for mote, but there is an obvious drawback: it's not TCP/IP based. Therefore, we present SIP [8] to implement part of TCP/IP feathers in wireless sensor networks.

Sensor networks are usually ad hoc network. The explosion of short-range ad hoc networking is commensurate with the pervasive device explosion itself. Considering the limited power in tiny sensors, we present an energy-efficient multi-hop routing algorithm named WDM [9].

3.3 Context Awareness Middleware

Like general pervasive computing environments, in smart home applications, people use smart embedded devices interacting seamlessly with space and devices naturally and transparently. Such user-centered ubiquitous computing environments may require a context-aware and lightweight middleware to communicate, operate resources and provide various services. To meet these requirements, we present a context awareness middleware named Semantic Adoption Middleware (SAM) [10]. The traditional middleware is inevitably heavyweight. It has not enough reconfigurable mechanisms and manipulating policies, and also has not specifically semantic interface, which is not suitable for pervasive computing. The overall goal of SAM is to develop a semanteme-integrated and dynamic reconfigurable middleware technology through a synthesis of reflection, component, and knowledge management technologies. The architecture of SAM is shown in figure 4, which the SAM is constituted of semantic virtual agent (SVA), context interface protocol (CIP), and meta objects (MO). Additionally, an ontology-based context representation has been adopted in SAM.

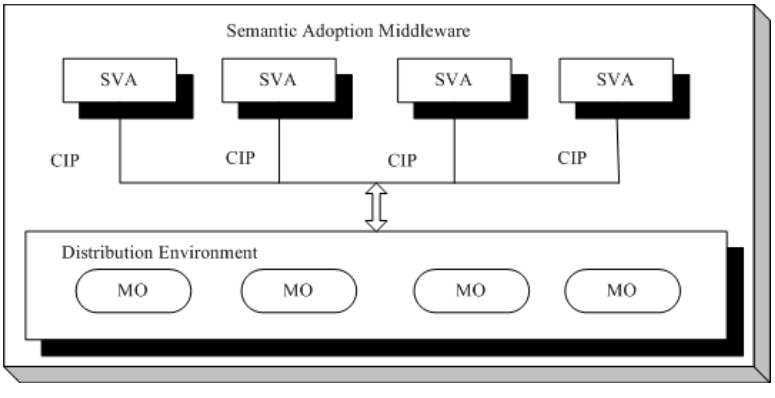


Fig. 4. The Architecture of Semantic Adoption Middleware

Semantic Virtual Agent. SVA is based on distributed components and distributed meta objects. The communications between SVAs are CIP, which consist of SVAs request, response and self-updating protocol. To update SVA's knowledge base and capability repository, a context repository reasoning system is utilized. At the same time, the SVA changes its missions on principle.

In SAM, each task should be transacted by some SVAs. SVA is just a logic abstract union that comprises a number of related meta objects which are kinds of various services. The meta object has direct interface with components. Emphatically, the basic element is the component. The relationships between two components may be mutex or interdependently. In SAM, component manager keeps each component's information and interrelationship repository, which also organizes the meta objects using some strategy.

Context Interface Protocol (CIP). Context interface protocol is the rule of SVA's presence, development, and intercommunication, which includes SVA's naming, request, response, and self-updating protocol. Here, noted a SVA by P=⟨C, M, K⟩, where C is the set of capabilities of this SVA; M is the set of SVA mission; K is the knowledge base of this SVA. For instance, SVA_1 looks up and obtains one performance parameter capability and its missions are providing SVA_2's position for SVA. Considering security and the right of each SVA, the SVA should know who to serve firstly. In particular, the SVA use whose capability depending on the context repository reasoning system and its knowledge base. So, the definitions could be SVA_1=⟨{lookup,getonepara}, {(ProvPos(SVA_2), SVA_3), { k_1, k_1,...}}⟩. Similarly, the protocol for SVA request, response, updating knowledge base, and updating capability can be presented.

Ontology-Based Context Representation. As the middleware is context aware, so the ontology is used to describe the context information. The concept of ontology comes from philosophy and means theory of existence. The new meaning is provided by computer science region, such as AI. The most accepted definition is "ontology is an explicit specification of a conceptualization". The ontology can be understood from two different aspects. Firstly, it is a vocabulary, which uses appropriates terms to describe entities and the relations between entities. Secondly, it is the knowledge base of a specific domain. Ontology is domain-oriented, which describes domain knowledge in a generic way, and provides agreed understanding of the domain.

3.4 Application of MacroOS

In this section, we will introduce a typical smart home application. It is a sophisticate ensemble system above MacroOS. It depends on the home networks both the wireless network and wired network, which connects the terminal device on the dwelling house, such as air-condition, lamp and camera device etc. By the central intelligent management host, the smart home system can receive the multiple sources data, which collect from the sensors in the dwelling houses, process the data, control and monitor both the electric devices and dwelling house. The smart home system can also provide the remote control via the telecommunication devices, such as the phone, computers, PDA etc.

The example of smart home system is shown in figure 5. This system is consisted of the management host, terminal devices, and transfer networks.

The terminal devices in smart home include the electric devices, such as the refrigerator, air-conditioner, lamp, alarm, cameras etc., and the sensors. In the terminal devices, we deploy the embedded real-time OS, the γOS, which controls the electric devices and sensors.

The transfer networks include wired network and wireless network applied ST-MAC and SIP etc. The wireless network will enable the sensors to install in anywhere of the dwelling house without the network topology limitation. And the wired network provides more bandwidth for some devices such as cameras.

The management host deploys the context-aware middleware SVA. It is the core of the smart home, which will process all the data and percept the content

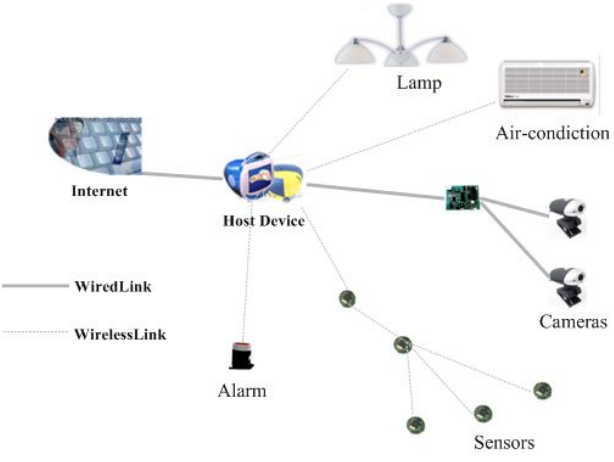


Fig. 5. The smart home application example

of the data in the semantic layer. In our method, we divide the context of smart home space to three parts: device context, environment context, and dweller context. Device context concerns the status and attributes of the appliances inside smart home, such as air-condition and phone. Environment context concerns all the environmental elements, such as temperature, humidity and image. Dweller's context concerns the status of dweller, such as the emotion and status.

For example, the management host receives an ontology-based description from sensors. The semantic adoption middleware in the management host will retrieval the description, and analysis the key words in the description, parser the attributes of these key words, then match the corresponding rules for those key words. We define context by C=(S, P, O). S denotes the center of context. P describes attributes. O is the value of attribute, or the operator, which is logic operator. We also define the interested and important context scenario by SC= (sc_1, sc_2, ..., sc_n) and sc=$\bigcup c_i$. The scenario integrates the situation of system and the entities involved. We define predication c_{n+1}=f(c_1, c_2, ..., c_n) and c_{n+1} denotes the high level context reasoned or the reaction the system should take according to current context scenario. For example: (Room, Property, Temperature)$\bigwedge$(Temperature <15℃)→(Dweller, Status, Cold). It descries if the room temperature is less than 15℃, we conclude that the dweller feels cold. Another, (Environment, Subclass, Light)$\bigwedge$(Light, Property, Brightness) $\bigwedge$ (Brightness <1200Im)→(Appliance, Subclass, Lamp)$\bigwedge$(Lamp, Property, Powerstatus) $\bigwedge$(Powerstatus =highlight). This describes that the room's brightness is too low and the lamp should be set to a highlight work state.

3.5 Conclusion

In this paper, we proposed MacroOS, a pervasive computing platform supporting context awareness and context management. The MacroOS platform has been successfully implemented in smart home application, and it is not limited only

in that system. The future work of applying pervasive computing platform into smart home environment should include providing effective and energy-saving operating system function, supporting nicer network communication protocol and secure routing algorithm, more intelligent semantic and adaptive middleware and so on. Of course, there is still a long way to go before these happen.

References

1. Saha, D., Mukherjee, A.: Pervasive Computing: A Paradigm for the 21st Century. IEEE Computer, 36(3), IEEE Computer Society Press (2003) 25-31
2. Smartdust. (2001) http://robotics.eecs.berkeley.edu/ pister/SmartDust/
3. TinyOS. (2005) http://www.tinyos.net/
4. Zhaohui, W., Lei, W., Kougen, Z.: A reliable OS kernel for smart sensors. 28th Annual International Computer Software and Applications Conference, Hong Kong (2004) 572-577
5. Ye, W., Heidemann, J., Estrin D.: An energy-efficient MAC protocol for wireless sensor networks. The 22nd Conference of the IEEE Communications Society, New York, USA (2002) 1567-1576
6. Xiaohua, L., Kougen, Z., Yunhe, P., Zhaohui, W.: A Sensor Media Access Control Protocol based on TDMA. The First International Conference on Embedded Software and System, Hangzhou, China (2005) 326-333
7. Xiaohua, L., Kougen, Z., Yunhe, P., Zhaohui, W.: A TCP/IP implementation for wireless sensor networks. IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands (2004) 6081-6086
8. Xiaohua, L., Kougen, Z., Yunhe, P., Zhaohui, W.: A TCP/IP implementation for wireless sensor networks. IEEE International Conference on Systems, Man and Cybernetics, Hague, Netherlands (2004) 6081-6086
9. Zengwei, Z., Zhaohui, W.: WDM: An Energy-Efficient Multi-hop Routing Algorithm for Wireless Sensor Networks. 5th International Conference on Computational Science, Atlanta, USA (2005) 461-467
10. Qing, W., Zhaohui, W., Bin, W., Zhou, J.: Semantic and Adaptive Middleware for Data management in Smart Vehicle Space. 5th Advances in Web-Age Information Management, Hangzhou, China (2004) 107-116

A Frame for Selecting Replicated Multicast Servers Using Genetic Algorithm

Qin Liu and Chanle Wu

School of Computer, Wuhan University, 430079 Wuhan, China
liuqin_whu@yahoo.com.cn, wuchl@whu.edu.cn

Abstract. Multicast server replication effectively utilizes the network resources and improves the performance of the clients. Selection of servers in such environments decides the quality of services. In this paper, we provide a frame to select replicated servers with genetic algorithm. A two-level coding scheme is developed to represent a server selection candidate to chromosome efficiently. In the coding process, a method based on Dijkstra's algorithm and random disturbance is designed that ensures generating a valid multicast tree. We discuss two options of genetic operators, namely crossover and mutation only at the first level code (GA1) or two levels (GA2). GA2 offers higher heritability and locality, but it requires more techniques to guarantee the validity of offspring. Extensive simulations demonstrate that both GA1 and GA2 outperform other heuristics. Particularly, GA2 is superior to GA1 in the complex network.

1 Introduction

With the explosive growth of the Internet, multicast is considered to be an effective mechanism to multi-user concurrent applications. However, the service rate is often decided by the bottleneck link of the multicast tree, indicating that even if the clients are of heterogeneous access capabilities, they are served at the same rather low rate. The diversity in access bandwidth calls for new mechanism to meet different needs.

Server replication, widely used in unicast, is introduced in multicast [1], [2]. In this mechanism, multiple servers at different places offer service to different groups of clients simultaneously. So clients in different multicast groups can be served at different rates. The aim of replicated multicast servers selection (RMSS) is to optimize objects such as the average rate and fairness, which is proved to be NP-hard [1].

Existing algorithms of RMSS are divided into special algorithms and general-purpose heuristic algorithms [1], [2]. The former are to handle special network, where RMSS is no longer NP-hard, but strict assumptions limit their wide application. As to the latter, often a near-optimal solution is got, but the quality is hard to be guaranteed.

Since genetic algorithm (GA) promises solutions to complicated problems in various applications successfully, we attempt to employ GA to RMSS. Main issues when applying GA, such as coding, crossover and mutation operators, are discussed. The proposed algorithm consists of several parts, termed GA-RMSS in whole.

The rest of the paper is organized as follows. Section 2 describes the model of RMSS. In section 3, a two-level coding scheme is addressed, and the method to get fitness is given. Section 4 is devoted to the genetic operators, where crossover and

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 695 – 702, 2005.
© Springer-Verlag Berlin Heidelberg 2005

mutation at two levels code are mainly discussed. In section 5, the performance of GA-RMSS is substantiated by extensive simulations. Section 6 concludes the paper.

2 Model

The underlying topology of computer network can be specified by the directed graph $G = (V, E)$, where $V = \{v_0, v_1, ..., v_{N-1}\}$ is a set of nodes (including hosts and routers), and $E = \{e_0, e_1, ..., e_{M-1}\}$ is a set of links. For a link e, $B(e)$ and $D(e)$ represents its bandwidth and delay, respectively. Consider a multicast service with the set of servers $S = \{s_0, s_1, \ldots, s_{n-1}\}$ and the set of clients $C = \{c_0, c_1, \ldots, c_{m-1}\}$, such that $S \subset V$ and $C \subset V$. Clients are assigned to the servers with allocation function φ: $C \rightarrow S$. Denote the set of the clients assigned to server s_i as $\Omega_i = \{c_j | \varphi(c_j) = s_i\}$. We aim at building n multicast trees $T_i = (V_i, E_i)$, $(V_i \subseteq V, E_i \subseteq E, 0 \le i < n)$, that T_i takes s_i as the root and Ω_i as the leaves. Let $route(c_j)$ denote the path from server s_i to c_j in T_i if $\varphi(c_j) = s_i$.

When clients selecting the same server are grouped and a multicast tree is built, the server decides its own service rate, meaning that rates of different groups need not to be same. Let R_i and r_j denote the sending rate of s_i and the receiving rate of c_j, then $(\forall c_j)(\varphi(c_j) = s_i) \Rightarrow r_j = R_i$, i.e. clients in one multicast group have the same receiving rate, equal to the sending rate. If several multicast trees share one link e, the sum of their corresponding service rates should not exceed $B(e)$, i.e.

$$\forall e_j \in E, \sum_{e_j \in T_i} R_i \le B(e_j) \tag{1}$$

The performance of all clients can be evaluated from the average rate and fairness of clients. The fairness performance is defined as

$$P_f = (1/m) \cdot \sum_{j=0}^{m-1} (r_j / r_j^*), \tag{2}$$

where r_j^* is the ideal receiving rate of c_j, i.e. the maximum receiving rate of c_j when no other clients exist. Since $r_j^* \ge r_j$ and $0 \le r_j / r_j^* \le 1$, $0 \le P_f \le 1$.

The average rate performance is originally defined as $(1/m) \cdot \sum_{j=0}^{m-1} r_j$, but in order to make its value in the range [0, 1] for convenience of comparison, it is scaled as:

$$P_r = (m / \sum_{j=0}^{m-1} r_j^*) \cdot (1/m \cdot \sum_{j=0}^{m-1} r_j) = \sum_{j=0}^{m-1} r_j / \sum_{j=0}^{m-1} r_j^*. \tag{3}$$

Since the performances P_r and P_f are expected to be maximized, meeting the requirements of fitness function in GA, they are used as fitness function in GA-RMSS.

3 Genetic Coding

Genetic coding, the primary issue of GA, transforms the solutions of problem into chromosomes. The core principle of coding in GA-RMSS is to generate valid as well as diverse chromosomes to represent various server selection candidates. A server selection candidate answers which server a client selects and the specific route to deliver data. In previous heuristic algorithms [1], the two parts are solved together. For example, the shortest path selection chooses the closest server for a client. Consequently, when a server is selected, data is to be sent via the shortest path. However, in

such schemes, if each client is assigned to a fixed server, the corresponding route is also fixed, although there may be several routes from the server to the client. This limitation decreases the probability to find the optimal solution.

Our server selection method is separated in two phases. In the first phase, server allocation is decided; in the second phase, the multicast tree in each group is built. Correspondingly, we propose a two-level chromosome coding scheme.

Server allocation decides the multicast members of each group. One server can provides service to multiple clients while a client can only select one server. A vector *SA* with a length equal to the number of clients is used to denote the result of server allocation. Initially, the server *id* each client selects is generated in the set {0, ..., $n-1$}.

An example we use in this paper is shown in Fig. 1. There is a multicast service with two servers s_0 (v_0) and s_1 (v_1), and seven clients $c_0 \sim c_6$ ($v_8 \sim v_{14}$). The numbers in the bracket represent the delay and bandwidth of the associated link in two directions. Since there are two servers s_0 and s_1, the value in each element of *SA* is 0 or 1. A possible server allocation is shown in Fig. 1, where clients c_0, c_2, c_3 and c_6 select s_0, and c_1, c_4 and c_5 select s_1, i.e. $\Omega_0 = \{c_0, c_2, c_3, c_6\}$ and $\Omega_1 = \{c_1, c_4, c_5\}$.

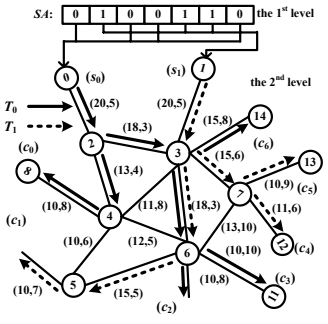


Fig. 1. An integrated chromosome

The second level coding is to build a multicast tree for each multicast group. Tree coding in GA is a hard problem in that effective coding kshould represent all possible trees, and the representation should only express a valid tree [3], [4]. If the representation includes not only tree, but also other invalid representations, it must be examined or repaired. In addition, all leaves in the multicast tree must be client nodes.

We propose a novel algorithm DRMT (Dijkstra-based Random Multicast Tree). It can produce a random tree without examining validity. Dijkstra's algorithm is widely used to get the shortest paths from a sourse *s* to all other nodes in computer networks. If *s* is the server in a multicast group, the sum of all shortest paths of clients in that group form a tree with the server as root and clients as leaves. The multicast tree is valid, but it lacks of diversity, for the multicast tree is fixed under fixed delays.

To bring diversity to the multicast tree, we introduce *disturbances* to the delay of the link. Each time before Dijkstra's algorithm runs, random disturbances are added to the delay of all links to form temporary delays for this time of computation. Consequently, the difference brought by disturbance results in different multicast trees even

in the same server allocation. We use DRMT to get the multicast tree for every group. Fig. 1 shows a random routing result. The solid line and dash line illustrates T_0 and T_1, respectively, which are the second level code. The first and second levels codes constitute an integrated chromosome.

Many structures can be used to store one multicast tree, but we need to store multiple multicast trees. This structure is expected to express one single multicast tree, multiple multicast trees as well as the information of link-sharing. We design a matrix having these properties, and the definition is given below.

Def. 1. $G = (V, E)$ is a directed graph where $V = \{v_0, v_1, \ldots, v_{N-1}\}$. In a multicast service of replicated servers with n multicast trees $T_k = (V_k, E_k)$ $(0 \le k < n)$, the $N \times N$ matrix (r_{ij}), is the exponential adjacency matrix (*EAM*), where $r_{ij} = \begin{cases} \sum_{e_{ij} \in E_k} 2^k \\ 0 \quad (\neg\exists E_k) e_{ij} \in E_k \end{cases}$.

$$EAM = \begin{bmatrix}
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&2&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&3&2&0&0&0&0&0&0&1&\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&2&0&0&0&0&0&\\
0&0&0&0&0&2&0&0&0&0&1&1&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&2&2&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&
\end{bmatrix}$$

Fig. 2. Exponential adjacent matrix of Fig. 1

The *EAM* of Fig. 1 is shown in Fig. 2, where element of "1" or "2" indicates that the related link is only in T_0 or T_1. Element of "3" signifies the link appears in T_0 and T_1 simultaneously. *SA* and *EAM* constitute the structure to represent a chromosome.

After the multicast trees are built, service rates are got by greedy algorithm [1]. It gives priority to the group of the most clients that has not had its rate assigned. Once service rates are known, P_r and P_f can be computed.

4 Genetic Operators

The commonly used genetic operators, selection, crossover and mutation, steer the evolution to produce chromosome of high fitness and get the optimal one eventually. Selection is used to keep the excellent chromosomes and eliminate the bad ones. In GA-RMSS, the best p_s percentage of the chromosomes are reproduced to the next generation. The next two subsections focus on crossover and mutation.

4.1 Crossover

Crossover plays a critical role in the evolution, which rearranges partial structures of two parental chromosomes with the crossover probability p_c to generate two offspring,

improving the searching ability. We adopt the single-point crossover on the first level code. A crossover point is randomly selected in {0, 1, …, m−2}, and the parts of *SA* after the crossover point in parents are exchanged to form two offspring's *SA*s.

As to the offspring's multicast trees, there are two options. The first one is to generate multicast trees using DRMT based on the new formed *SA*, as described in section 3. In this method, crossover only occurs at the first level code, and routes in parents are not inherited. However, the routing information of parents may be decisive to be an excellent chromosome. To provide high heritability, crossover operator should form the offspring's multicast trees that consist entirely of links found in parents. Thus the second option recombines the multicast trees of parents to get those of offspring, which crosses over at two levels. Though obtaining offspring's multicast trees from parents is attractive, it must guarantee the validity of the recombined multicast trees. Decisions should be made that which link is reserved and which is added.

Assume that parents A and B are recombined to get A' and B' .Take the offspring A' as an example. The basic idea is that the multicast trees of A' are initialized to those of A. Then the routes of clients after the crossover point are removed from A', and the routes after the crossover point in B are added to A' to fulfill exchange. We first define the in-degree and out-degree of a node in a multicast tree for later use.

Def. 2 $T_k = (V_k, E_k)$ is a multicast tree of a chromosome A, where V_k is the set of nodes in T_k, and E_k is the set of links in T_k. The in-degree of node v, $v \in V_k$, denoted by $\deg^-(A, v, k)$, is the number of links in E_k with v as their tail node; the out-degree of v, denoted by $\deg^+(A, v, k)$ is the number of links in E_k with v as their head node.

For example, in Fig. 1, there are two links $v_2 \rightarrow v_3$ and $v_2 \rightarrow v_4$ issuing from v_2 in T_0, and therefore $\deg^+(A, v_2, 0) = 2$. There is a link $v_3 \rightarrow v_6$ taking v_6 as its tail node in T_0, so $\deg^-(A, v_6, 0) = 1$. Similarly, $\deg^+(A, v_3, 1) = 2$ and $\deg^-(A, v_4, 1) = 0$.

Deleting Route. As is known, a multicast tree includes the routes of all its clients. If a client is no longer in the multicast group, its route should be deleted. Since a route is made up of links, deleting a route is to remove its links. However, a link may be shared by several routes, so only the non-shared links in the route can be removed. We begin to remove the links of a specific route backward from its last link, for the links closer to the server are more likely to be shared by many routes.

Suppose that we are deleting *route*(c_j) from T_k. Let e be the link considered currently, and *curtail* be the tail node of e. To judge whether e should be removed, we must know if e is shared by any other route. It can be determined by the rule: if $\deg^+(A', curtail, k) > 0$, e is in some other route. The reason is straightforward that $\deg^+(A', curtail, k)$ indicates the number of links taking *curtail* as their head node, and $\deg^+(A', curtail, k) > 0$ means e appears in at least one route. Thus e can not be removed from T_k. Moreover, the previous links to e in *route*(c_j) are also shared by those routes. If $\deg^+(A', curtail, k) = 0$, e is ensured not in any route and can be removed. The immediate link before e is next considered. The procedure continues until it encounters a shared link, or all links in *route*(c_j) are removed.

Adding Route. After deleting all routes after the crossover point, we need to add these corresponding routes in B to A'. However, adding a new route directly to an existing multicast tree may destroy its validity, as two invalid cases shown in Fig. 3.

The black line represents the existing multicast tree, and the gray line denotes the added route (two lines from v_0 to v_2 represent one link). In Fig. 3 (a), there are two parallel paths from v_2 to v_4, namely $v_2 \rightarrow v_4$ and $v_2 \rightarrow v_3 \rightarrow v_4$, not according with the requirement that only one path lies between any two nodes. In Fig. 3 (b), a loop $v_3 \rightarrow v_7 \rightarrow v_6 \rightarrow v_3$ is created, and it is also not allowed. The reason for the two invalid cases is that the new added route makes some node have two paths from the root to it.

We still add *route*(c_j) in B to the multicast tree T_k of A' from its last link, and e is the link currently considered, with *curtail* as tail node. If $\deg^-(A', curtail, k) = 0$, it indicates no link in T_k terminates at *curtail* and it is not in T_k. So e can be added to T_k. The link immediately previous to e in *route*(c_j) is next considered. If $\deg^-(A', curtail, k) > 0$, it suggests that *curTail* is in T_k and there is a path from s_i to *curtail*. Then all links from the first to e in *route*(c_j) should not be added. The procedure stops if all links in *route*(c_j) are added to A', or it encounters a node with in-degree larger than 0.

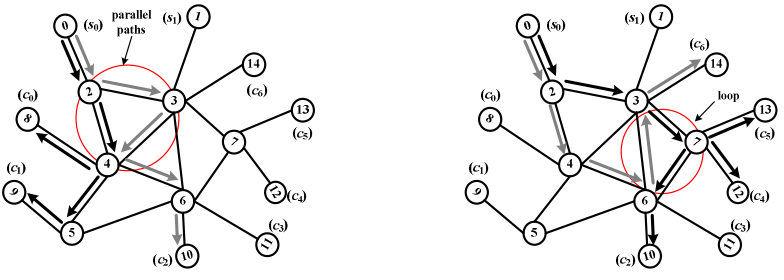


Fig. 3. Invalid cases: (a) parallel paths; (b) loop

4.2 Mutation

Mutation is to change some genes in the chromosome with a small probability p_m to keep the diversity of the chromosomes and avoid seeking slumping into the local optimal solution. In GA-RMSS, the single point mutation is used. In the to-be-mutated chromosome P, an element of SA is randomly selected, and then the value is randomly changed to a server *id* other than the original one. Similar to crossover, there are also two methods in producing the mutated second level code. One is to generate a new chromosome Q using DRMT based on the mutated SA, and take Q directly as the offspring. The other method deletes the route of the mutated client in P, and then adds the corresponding route in Q to P to form the offspring.

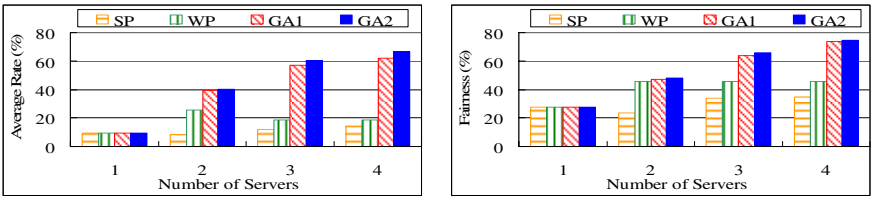
5 Simulations

The performance of GA-RMSS is evaluated by a series of simulations. We compare two versions of GA-RMSS. The first one employs crossover and mutation only at the first level code, referred to as GA1. The second one are at two levels, referred to as GA2. Meanwhile, we also make comparisons among GA-RMSS and two heuristics: shortest path selection and widest path selection (SP and WP for short, respectively)

[1]. In SP, each client selects the nearest server based on the delay. In WP, each client selects a server that has the highest bandwidth from the server to the client.

The network topology used in the simulations is generated by the RouterWaxman model in BRITE [5]. The random network generated has 69 nodes, with the average connecting degree of 6.9. The delay of each link is represented by the distance, and the bandwidth is randomly set in the range [10, 1024]. We assign 4 potential servers.

In the first group of simulations, 8 clients are randomly set, i.e. $m = 8$. The number of servers, n, is increased from 1 to 4. The average rate P_r in (3) or the fairness P_f in (2) is taken as the fitness function, respectively. When there is only one server, GA-RMSS just generates one multicast tree. The average value of the best solutions in 20 independent runs is used for GA1 and GA2 under each network configuration. In each run, these parameters are set: population size is 100, maximum iteration times is 50, $p_s = 0.3$, $p_c = 0.1$ and $p_m = 0.04$. Simulation results are shown in Fig. 4.

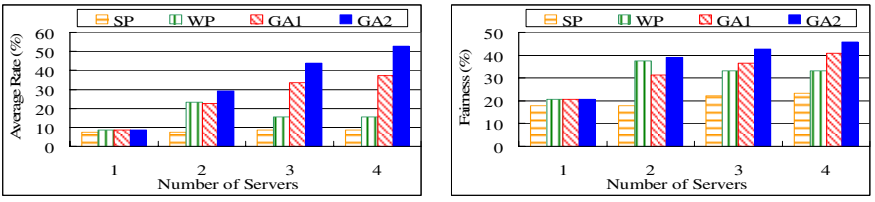


(a) (b)

Fig. 4. Simulations when $m = 8$: (a) average rate; (b) fairness

We can find in all configurations GA1 and GA2 are no worse than SP and WP. It can be proved that WP gets the optimal solution if $n = 1$. In Fig. 3(a) and (b), four algorithms all obtain the optimal solution when $n = 1$. When $n > 1$, GA1 and GA2 are distinctly better than SP and WP. Concerning the rate performance with the varying servers, the average improved percentage of GA1 over SP and WP are 274.16% and 123.73%, while those of GA2 are 293.15% and 136.21%. As to fairness, GA1 is on average 73.95% and 26.22% higher than SP and WP, while GA2 is 77.31% and 28.52%. GA2 exhibits a little superior to GA1.

In another group of comparisons, 7 additional clients are joined, i.e., $m = 15$. The same simulations are conducted as the first group, with results shown in Fig. 5. The similar rules can be observed. It is notable that the average rate and fairness of WP are higher than GA1 when $n = 2$, but GA2 is always the highest. GA2 is on average 16.77% higher than GA1 in terms of the rate performance, and 13.51% higher than GA1 in terms of fairness. Obviously, GA2 is more effective in this group than the pervious. In such configurations, the problem is the more complicated. The superiority of GA2 in the complex environment shows that crossover and mutation at the second level code are of great importance.



(a) (b)

Fig. 5. Simulations when $m = 15$: (a) average rate; (b) fairness

6 Conclusions

This paper presents a genetic algorithm for solving the selection of replicated multicast servers. The main contribution of this paper is twofold. First, we develop a two-level coding scheme to represent a feasible server selection as chromosome in the genetic space efficiently. The multicast trees in a random chromosome are created by the proposed DRMT, which ensures validity of multicast tree. The convenient storing structure of a chromosome also facilitates coding and further processing.

The second contribution is that we investigate the genetic operators in detail. Specifically, crossover and mutation only at the first level code, or at two levels, are discussed. The latter can make the most of the genetic information in good parental chromosomes. Simulations show the excellent performance of GA compared with others. They also confirm that genetic operators at two levels are very important.

References

1. Fei, Z., Ammar, M., Zegura. E.: Multicast Server Selection: Problems, Complexity and Solutions. IEEE Journal on Selected Areas in Communications, Vol. 20 (2002) 1399–1413
2. Fei, Z., Yang, M., Ammar M., Zegura, E.: A Framework for Allocating Clients to Rate-Constrained Multicast Servers. Computer Communications, Vol.26 (2003) 1255–1262
3. Raidl, G.R., Julstrom, B.A.: Edge Set: An Effective Evolutionary Coding of Spanning Trees. IEEE Transations on Evolutionary Computation, Vol. 7, Issue 3 (2003) 225–239
4. Fei, X., Luo, J., Wu, J., Gu, G.: QoS Routing Based on Genetic Algorithm. Computer Communications, Vol. 22 (1999) 1394–1399
5. Medina, A., Lakhina, A., Matta, I., Byers, J.: BRITE: An Approach to Universal Topology Generation. In: Proc of MASCOTS'01, Cincinnati, Ohio (2001)

On a Novel Methodology for Estimating Available Bandwidth Along Network Paths*

Shaohe Lv, Jianping Yin, Zhiping Cai, and Chi Liu

School of Computer, National University of Defense Technology,
Changsha Hunan 410073, China
chi.shaohe@gmail.com, jpyin@nudt.edu.cn, xiaocai@163.net,
3dfx232@sohu.com

Abstract. This paper presents a novel methodology, called COPP, to estimate available bandwidth over a given network path. We first present a rigorous definition of available bandwidth, from which the novel two-steps estimating methodology, e.g. partial and final step, can be derived. In partial step, COPP deploys a particular probe scheme, namely chirp of packet pairs, which is composed of several packet pairs with decremental inter-packet spacing. After each chirp is sent, a partial estimate can be obtained and equal to the weighted average of all turning bandwidth within the chirp. In the second step, the final estimate is the weighted amount of all partial results of all chirps in a measurement episode. Additionally, we develop a heuristic though efficient rule to determine whether a packet pair is turning point. Finally, the evaluation of COPP in various simulations shows that COPP can provide accurate results with relatively less overhead while adapt to network variations rapidly.

1 Introduction

Estimating the bandwidth of an Internet path is a fundamental problem that received considerable attention in the last few years. Knowledge of the bandwidth of a path can be put to good use in various scenarios, such as congestion control and/or avoidance, server selection [8], load balancing, and network security and so on. There have two metrics are commonly associated with bandwidth estimation —capacity and available bandwidth (avail-bw). We discuss the definitions of them in detail in section 3.1. In this paper, we only focus on the measurement of path avail-bw, especially the *average avail-bw* in an interval.

Packet pair technique [1, 3, 5] has been used as one of the primary procedures to measure the capacity of a path. Recently, there has much work [10, 13], including the techniques we will propose, striving to estimate available bandwidth based on packet pair. Current available bandwidth measurement techniques still have some problems [15], including poor accuracy and high overhead etc. We believe that these are mainly due to utilizing the information carried in probe packets and the inherent characters of avail-bw ineffectively. By exploiting that information more efficiently, COPP that we introduce yields better tradeoff between good accuracy and low overhead.

* Supported by National Natural Scientific Foundation of China under Grant No.60373023.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 703 – 710, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The rest of this paper is organized as follows. In section 2, we briefly review the related work in estimating path avail-bw. We describe the rigorous definitions of bandwidth and the measurement process of COPP in section 3. And the simulation results are presented in section 4. Finally, we conclude in section 5.

2 Related Work

We briefly survey the many tools and techniques that have been proposed for estimating avail-bw of network paths. As for capacity estimation, we refer interesting readers to [11], which we omitted here due to space limitations.

Early techniques such as Cprobe measured the asymptotic dispersion rate rather than avail-bw. Many of the recently presented techniques fall into two categories: packet rate model (PRM) and packet gap model (PGM). There are several tools [10, 16], actually, are mixture of PGM and PRM. PGM-based tools [13], in general, send pairs of equal-sized probe packets, and the increasing intra-pair spacing is used to estimate the volume of cross traffic, which is then subtracted from the capacity to yield the avail-bw. Tools such as [2, 6], also based on PGM, which is different for these tools send several trains and the one-way delay (OWD) of every probe packet is used to estimate the utilized bandwidth instead of the pair spacing. PGM assumes that the tight link is also the narrow link, and is susceptible to queue latencies at links except the tight link. Additionally, the errors introduced in capacity estimation would be propagated if the assumption in PGM that the capacity is known is violated. We pick Spruce [13] as the representative PGM-based tool for our experiments.

PRM-based tools [9, 12, 17], are based on the basic observation that probe traffic sent at a lower rate than the avail-bw would be received at the sending rate (on average) —means that the probe packets would experience little interference. In contrast, if the sending rate exceeds the avail-bw, the received rate would be less than the sending rate and the probe packets tend to queue behind previous probe or cross packets—means that the probe packets are likely interfered more seriously. Thus, avail-bw can be measured by searching the turning point at which the interference varied remarkably. The increasing trend of OWD is used to determine this interference in Pathload [12], which is picked as the representative PRM-based tool for our experiments. The technique we present is also PRM-based but uses a pair as a minimum process unit rather than a single packet or train used in previous tools. As follows we will show that COPP can yield accurate estimates with relatively fewer overhead due to exploiting the probe information efficiently.

3 COPP Methodology

In section 3.1, we first discuss the rigorous definition of capacity and available bandwidth. Then we derive the novel methodology, composing of two steps, e.g. the partial and final step, which is shown in detail in section 3.2.

3.1 Definitions of Bandwidth

There have two metrics are commonly associated with bandwidth estimation —capacity and available bandwidth (avail-bw). The capacity of a link is the maximum transmission rate that link can transfer packets in the absence of competing traffic. In contrast, the

avail-bw of a link at a time instance is defined as the maximum transmission rate that link can provide without interfering the existing traffic.

As to a path, the link with the minimum capacity determines the capacity of the path, while the link with the minimum available bandwidth limits the avail-bw, which we refer to as narrow and tight link respectively. In a time scale, the avail-bw of the path is the average of all the avail-bw at every time instance within the scale.

For an end-to-end path *P,* the capacity and avail-bw at time *t* are denoted by C(t), A(t). Then, the avail-bw in a time scale $[t_1, t_2]$ can be defined as equ.1 where the function $\Phi(t)$ defined as equ.2. *P* is idle at time *t* means that there has no packet to be handled at *t* whereas busy means that there has some packets needed to be disposed at that time instance. (Rigorously, the definition of $\Phi(t)$ should distinguish the case of link and path, but this "ambiguous" definition is enough for our discussion.)

$$A(t_1, t_2) = \int_{t_1}^{t_2} \Phi(t)dt /(t_2 - t_1) \tag{1}$$

$$\Phi(t) = \begin{cases} C(t) & if\ \ path\, P\, is\ idle \\ 0 & if\ \ path\, P\, is\ busy \end{cases} \tag{2}$$

Note that the avail-bw is a time-varying metric, though the *average* avail-bw in a time scale may be invariable. Thus, equ.1 can be refined as equ.3, where $t_{i0} = t_1$ and $t_{in} = t_2$. If *n* is large enough, we can further hypothesize that the avail-bw is constant approximately in the short time scale $[t_{ik}, t_{i(k+1)}]$. Then we can obtain equ.4.

$$A(t_1, t_2) = \int_{t_1}^{t_2} \Phi(t)dt /(t_2 - t_1) = \sum_{k=0}^{n} \int_{t_{ik}}^{t_{i(k+1)}} \Phi(t)dt /(t_2 - t_1) \tag{3}$$

$$\begin{aligned} A(t_1, t_2) &= \sum_{k=0}^{n} \left(\frac{\int_{t_{ik}}^{t_{i(k+1)}} \Phi(t)dt}{t_{i(k+1)} - t_{ik}} \bullet \frac{t_{i(k+1)} - t_{ik}}{t_2 - t_1} \right) \\ &= \sum_{k=0}^{n} \left(A(t_{i(k+1)}, t_{ik}) \bullet \frac{t_{i(k+1)} - t_{ik}}{t_2 - t_1} \right) \end{aligned} \tag{4}$$

Notice that the starting and ending point of both the whole and short time scale is already known. Therefore, to obtain an estimation of the avail-bw in a relatively large time scale, what should do is to divide the large scale into several short scales and then acquire a partial estimation in every short scale, and the final estimation of avail-bw in the whole scale can be deduced with the help of equ.4. In other words, equ.4 indicates a possible scheme to estimate the avail-bw in a time scale. In order to obtain an accurate partial estimation, we deploy a particular probe scheme called chirp of packet pairs [4]. We will discuss the whole methodology in the next section.

3.2 The Methodology

In this section, we describe the basic measurement process of COPP. The name COPP is, originally, from the probe scheme, e.g. Chirp of Packet Pairs as illustrated by Fig.1.

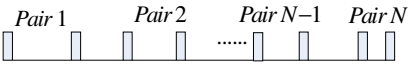


Fig. 1. Chirp of packet pairs: An illustration

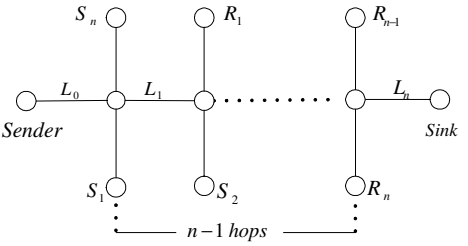


Fig. 2. Experiment Topology and the measuring path from Sender to Sink consist of *n+1* links

Logically, there have two basic steps to obtain the estimation of avail-bw. First, called partial step, we use a chirp of packet pairs to obtain a partial estimation of the avail-bw in the short time scale from the first probe to be sent to the last probe to be received. Second, called final step, after several repeats of the partial step, with the help of equ.4, we use these partial estimations to obtain a final estimation in the targeted relatively large time scale. Hence, in the second step there is no need to send probes. We now describe some details of the two steps.

We use *measurement episode* to indicate the time interval in which a final avail-bw is reported by COPP. In each episode, several chirps of packet pairs are sent and traversed the measured path. For our purpose, a pair chirp as shown in Fig.1 preserves the follow property: the intra-pair spacing is gradually reduced to increase the sending rate of corresponding pair continuously, from the lower estimate bound to the upper bound in a factor of a constant. In addition, a pair can be sent only after the previous pair has been already received. Actually, the interval between the sending time of the first packet of a pair and the received time of the second packet of its previous pair satisfies uniform distribution in [0, 10ms] (only an empirical selection) to avoid the interference between different pairs and ensure that the sample is unbiased.

We use the term *turning point* to denote such packet pair that has been interfered more seriously compared to its previous neighbor. Associated with every turning point, there has a *turning bandwidth* (*TB*) that is the sending rate of the previous pair of the underlying turning point. By exploiting the decision rules depicted in section 4, COPP can discover all turning points and the corresponding *TB*.

Ideally, only the pair sent at rate higher than the avail-bw of which the preceding pair is launched at rate lower than the avail-bw will become the turning point according to the basic PRM observation, which we termed as the expected case. But the burst cross traffic interacted with probes can also trigger some “new” turning point. The short-term burst is a factor contributing to the avail-bw in a relatively long time interval but not the deterministic one.

Thus, we should distinguish the turning point caused by the burst and the ideal and expected case. We give each point a distinct weight according to the interference experienced by the point and its preceding and following neighbors. More details can be found in [4]. Finally, the solution of COPP is: giving each point a distinct weight according to the degree to which the underlying point and its neighbors are disturbed. After each chirp is sent and traversed the measured path, COPP obtain a partial estimate. The follow equ.5 exhibits the derivation of the partial estimate for a chirp

denoted by A, where $TB_1, TB_2, ...TB_k$ and $w_1, w_2 ... w_k$ denote all turning bandwidth and corresponding weight respectively in the chirp.

$$A = \sum_{i=1}^{k} (TB_i * w_i) / \sum_{i=1}^{k} w_i \quad (5)$$

After several partial estimates are obtained, we use equ.6, an variation of equ.4, to compute the final avail-bw estimate in the measurement episode denoted by *Avail-BW*, where n is the amount of chirps launched in the episode and A_k is the partial estimate of the k-th chirp, as well t_k indicates the time length from the first packet to be sent to the last packet to be received in the k-th chirp and T is the total time of the measurement episode or $T = \sum_{k=1}^{n} t_k$.

$$Avail_BW = \sum_{k=1}^{n} (A_k \bullet t_k) / T \quad (6)$$

One should be noted here is that COPP need the cooperation of the two side of the measured path. Sender is in charge of sending probe traffic and calculating bandwidth, whereas receiver should send a special feedback packet indicates that the time instances when a probe packet is received. Thus, all timestamps at the sending and arrival of all probes are available to sender, from which the one-way delay and intra-pair spacing can be derived.

3.3 Related Topics

Obviously, one of the crucial remaining problems is to determine whether a packet pair is turning point or not. Through analyzing the relationship among the one-way delay of a single packet, the inter-packet spacing and the experienced interference of packet pair, we can conjecture that: i) Since there has no interaction of different probe pairs, the first packet of pair fairly reflects the state of the network consists of cross traffic only while the second packet captures the interaction of probe and competing traffic. ii) The one-way delay variation of different packets and the intra-pair spacing variation characterize the difference of interference experienced by the corresponding packets. Based on above conclusion, we develop a rule to determine the turning point: the packet pair can be marked as a turning point only if it satisfies one of the follow conditions where $OWD_{2,k}$ and SV_k denote the one-way delay of second packet and the inter-packet spacing variation of pair k respectively:

I) $OWD_{2,k-1} < AD + \varepsilon \ \ and \ \ OWD_{2,k} \geq AD + \varepsilon$

II) $OWD_{2,k-1} < AD + \varepsilon \ and \ SV_k - SV_{k-1} \geq \varepsilon \ and \ SV_{k-1} > 0$

III) $OWD_{2,k-1} < AD + \varepsilon \ and \ OWD_{2,k} - OWD_{2,k-1} \geq \varepsilon$

There are many parameters should be set before the estimate process is executed, however, such as the size of probe packet, the estimate bound, the amount of pair in a probe, the amount of pair in a episode, the function to generate weight and the

threshold used in decision rules, and so forth. Limited by space, we omit the discussion about the analysis of packet pair and the parameter settings. Some details can be found in [4].

4 Experiments

In this section, we used ns-2 [14] to validate COPP mechanism described in the above section. We first investigate the different ability of pairs with different packet size in terms of reflecting the network dynamics. Then the comparison among Pathload, Spruce and COPP shows that COPP can adapt to network variation and provide accurate estimates with fewer overheads.

4.1 Simulation Setup

Fig.2 shows the experimental topology, where the measured path from Sender to Sink consists of n+1 links. Cross traffic is transmitted over n connections from S_i to $R_i(i=1,2,..n)$. Except the connection from S_n to R_n that has n-1 links overlapped with measured path, all the other connections have only one link overlapped with measured path. The bottleneck link of the measured path is at the center of the path with capacity 10Mbps, and all the remaining links are with capacity 15Mbps.

All link of the measured path has the same amount of cross traffic that consists of several Pareto traffic with shape parameter as 1.6. In addition, The cross traffic over every link of measured path have the same traffic class distribution (95% TCP and 5% UDP) as well as the packet size distribution, e.g. there are three type packet sizes, 1500/700/40Bytes that consumes the 5%, 85% and 10% percentages of total amount of cross traffic respectively.

In simulations, a chirp in COPP consists of 25 packet pairs and the estimate scale is [400K, 10M] which means that the difference between the sending rates of adjacent pairs is 400 Kbps. Additionally, after five chirps are processed, COPP compute and report the time-weighted result as the final estimate.

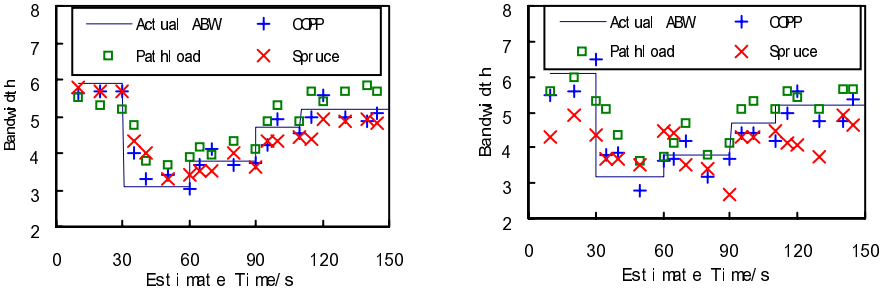
We pick Pathload and Spruce as the references to COPP in experiments. We use the setting of them same as their authors suggest in [12, 13]. As for Pathload, a stream is composed of 100 packets with size of 800Bytes and the amount of stream to yield an estimate is determined by the iterative algorithm in Pathload. To obtain an estimate, whereas, Spruce needs 100 pairs with size of 1500Bytes. We test these tools in several topology but only report the results based on the case of $n=5$ since all conclusions inferred are similar regardless of different simulation setup.

4.2 Experimental Results and Analysis

By controlling cross traffic, we make the available bandwidth varied in time 30s, 60s, 90s and 120s whereas kept stable in the interval between two time points. We conducted the experiments in several cross traffic conditions. We only report the results of the two most candidate cases, called single- and multi-congested cases. The former means that there has a unique link of which the avail-bw is significant less than other links, while the latter refers to the case where there have multi-links of

which the available bandwidth is similar with the tight link. Fig.3 shows the comparison between the estimates of the different tools and actual avail-bw where (a) and (b) refer to the single- and multi- case respectively.

From the picture, several conclusions can be drawn. First, almost all tools can provide approximately correct estimates, but their accuracy is distinct. For Pathload, the estimate can be changed remarkably when the actual bandwidth is invariant which suggests that the stability should be improved. As for Spruce which use probe of size 1500Bytes, the estimates are below the actual value commonly. This confirms our previous observation. Our technique, COPP, provides the best estimates in most case. Second, for agility, COPP is more agile when the variation of network is very acute, as for small variation the performance of all techniques are comparable. Finally, the accuracy of all tools is higher in (a) than (b), means that the multi-bottleneck can effect the estimation significantly, which will be further investigated in future.



(a) The single-congestion case

(b) The multi-congestion case

Fig. 3. The comparison of the actual available bandwidth and the estimate results of COPP, Pathload and Spruce

In almost 90% experiments, the estimate of COPP with five chirps in a episode is fairly accurate. Hence, the total amount of overheads consumed in measurement is trivial for network. Actually, compared to other two techniques, the amount of COPP is lower than Pathload significantly and would also not exceed Spruce. In summary, COPP is an accurate and non-intrusive tool and can reflect the dynamic variation of network quickly in most cases.

5 Conclusions

This paper presents a novel methodology for estimating available bandwidth over a network path, called COPP. Based on our earlier work [4], we refine the original framework of COPP and present a two-step measurement process. We test COPP in several experiments and find that it can provides relatively accurate estimate with less overhead, especially the revised COPP adapts to network variation quickly. The focus of our future work is to evaluating COPP in actual network, i.e. in Internet, and integrating COPP with network applications.

References

1. C. Dovrolis, P. Ramanathan and D. Moore. Packet Dispersion Techniques and A Capacity Estimation Methodology. IEEE/ACM Transactions on Networking Oct. (2004).
2. Liu Min, Shi Jinlin, Li Zhongcheng, Kan Zhigang and Ma Jian. A New End-to-End Measurement Method For Estimating Available Bandwidth. In proceedings of ISCC (2003)
3. X. Liu, K. Ravindran and D. Loguinov. What Signals do Pakcet-Pair Dispersion Carry? In proceedings of IEEE INFOCOM (2005).
4. Shaohe Lv, Jianping Yin, Zhiping Cai and Chi Liu. Towards a New Methodology for Estimating Available Bandwidth on Network Paths. In IFIP/IEEE MMNS, Oct. (2005)
5. V. Jacoboson. Congestion Avoidance and Control. In Proc, SIGCOMM, USA, Sept. (1988).
6. K. Lakshminarayanan, V. Padmanabhan and J. Padhye. Bandwidth Estimation in Broadband Access Networks. IMC, Internet Measurement Conference, Oct. (2004).
7. T. Karagiannis, M. Molle, M. Faloutsos and A. Broido. A Nonstationary Poisson View of Internet Traffic. In proceedings of IEEE INFOCOM, Apr. (2004).
8. R. Carter and M. Crovella. Server Selection Using Dynamic Path Characterization in Wide-area Networks. Proc. of INFOCOM'97, Japan, Apr. (1997).
9. D. Kiwior, J. Kingston and A. Spratt. Pathmon: a Methodology for Determining Available Bandwidth over an Unknown Network. Tech Report (2004).
10. B. Melander, M. Bjorkman, and P. Gunningberg, A New end-to-end Probing and Analysis Method for Estimating Bandwidth Bottlenecks. GLOBECOM (2000).
11. R. Prasad, C. Dovrolis and D. Moore. Bandwidth Estimation: Metrics, Measurement Techniques and Tools. IEEE Network (2003).
12. M. Jain and C. Dovrolis, End-to-End Available Bandwidth: Measurement Methodology, Dynamics, and Relation with TCP Throughput. In proceedings of SIGCOMM, Aug. (2002).
13. J. Strauss, D. Katabi, and F. Kaashoek. A Measurement Study of Available Bandwidth Estimation Tools. In proceedings of ACM IMC (2003).
14. NS version 2. Network Simulator. Http://www.isi.edu/nsnam/ns.
15. M. Jain and C. Dovrolis. Ten Fallacies and Pitfalls on End-to-End Available Bandwidth Estimation. In proceedings of ACM IMC, Internet Measurement Conference (2004).
16. N. Hu L. Li, Z. Mao P. Steenkiste and J. Wang. Locating Internet Bottleneck: Algorithm, Measurement and Implication. In proceedings of SIGCOMM, Aug.(2004)
17. V. Ribeiro, R. Riedi, R. Baraniuk, J. Navrati and R. Cottrell. PathChirp: Efficient Available Bandwidth Estimation for Network Paths. PAM, Passive and Active Measurement Workshops, Apr. (2003).

A New AQM Algorithm for Enhancing Internet Capability Against Unresponsive Flows

Liyuan Zhao[1], Keqin Liu[1], and Jun Zheng[2]

[1] The Department of Computer Science, North China Electric Power University, Beijing 100109, China
{zly_yuan, liu_ke_qin}@163.com
[2] The Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China
zhengjun@hit.edu.cn

Abstract. The unresponsive flows to the network congestion control are dangerous to the equilibrium and the Quality-of-Service (QoS) of the whole Internet. In this paper, a new network queue management algorithm- CCU (Compare and Control Unresponsive flows) is proposed in order to strengthen the robustness of Internet against unresponsive flows. As a sort of active queue management (AQM) algorithm, CCU relies on the detection and punishment of unresponsive flow and gets the elastics control of unresponsive flows, which benefit the buffer queue with the high performance. Via the comparison and evaluation experiments, it has been proved that CCU can detect and restrain unresponsive flows more accurately compared to other AQM algorithms.

1 Introduction

The Internet Quality-of-Service (QoS) requests from network users expand greatly with ever fast development of the network technology as well as new network applications. However, the inherent frailness of the Internet has also become extrusive, especially when the malicious flows flourish these years. For example, DDoS and Worm attacks can bring significant performance degradation and compromise the QoS of the large scale networks, which is also one sort of network congestions.

As unresponsive flows flourish in the Internet, the balance mechanism of network resources is broken by these flows which starving the normal network usage. The scenario mentioned above can result in "Congestion Collapses" [1], where the DDoS and Worm malicious attack flows are just the extreme cases. Accordingly, the common distinct character of malicious attack flows is that these flows are ***unresponsive*** to the Internet congestion control. So, the sticking point of strengthening the Internet immunity to the malicious flows is to enhance the Internet capability against the unresponsive flows. One of the most active research areas to address this problem is Active Queue Management (AQM) [2] which is one of most important approaches to handling "congestion collapse" and mitigating the demolishment of malicious flows.

At present, the traditional queue management technique of router is tail drop (TD) [2]. A lot of AQM algorithms have been proposed to address the inherent problems of TD. Floyd and Jacobson propose a whole new queue management algorithm-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 711 – 718, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Random Early Detection (RED) in their early landmark paper in 1993 [2]. To improve the performance of RED, [3] propose the FRED (Flow Random Early Drop), a modified version of RED. Developed from the AQM algorithm- BLUE [4], [5] propose the SFB (Stochastic Fair Blue) against unresponsive flows directly to protect the TCP-friendly flows and enforce fairness among a large number of flows. To achieve the max-min fairness of bandwidth allocation, the algorithm CHOKe (CHOose and Keep for responsive flows, CHOose and kill for unresponsive flows) is designed as a stateless AQM algorithm [6]. Although the above papers have made significant progress towards improving the fairness of the Internet and suppressing the unresponsive flows, the respective problems of unresponsive flow detections and their impact on the performance of AQM still leave space for further developing in this research direction. At the experiment section of this paper, more details of these AQM algorithms will be presented.

In this paper, a new AQM algorithm- CCU (Compare and Control Unresponsive flows) is proposed. Via the evaluation experiments, it has been proved that CCU can detect and restrain unresponsive flows precisely compared to other AQM algorithms. The rest of paper is organized as follows: Section 2 proposes the new AQM algorithm-CCU, and Section 3 describes the details of our evaluation experiments. Finally, Section 4 gives conclusions.

2 CCU Algorithm

For the purpose of detecting and restraining unresponsive flows effectively, this paper proposed a new AQM algorithm ***CCU*** (***Compare and Control Unresponsive flows***). For remaining low percentage of computation resources usage of the router, CCU only maintains a few parameters recording unresponsive flow states according to the "discrimination principle". CCU also maintains a global drop probability for all other normal flows. The algorithm not only can take various flexible treatment strategies according to different unresponsive flow states, but also guarantee the fairness between the unresponsive flows and normal flows. The following is the main steps of CCU:

For each incoming packet:

```
if (record_src. == unresponsive record)
   to punish the incoming packet belongs to
   unresponsive flows
else
   to detect if the incoming packet belongs to a unre-
   sponsive flow
```

2.1 Detection of Unresponsive Flows

According to the unresponsive flow feature of the high rate of sending data, one incoming packet must have a high probability with same source address as the existing packets in buffers. Therefore, CCU will use the comparison method to find potential unresponsive flows and detect them finally.

Concretely, when current queue length ***qlen*** does not exceed threshold ***qlim*** which set as 90 percent of maximum of buffer, all packets will be permitted to enter the buffer. On the contrary, CCU will compare the incoming packet with several candidate packets that are sampled randomly from the current queue: we set the number of candidate packets ***cand_num*** as 3 in CCU. As long as there is one packet among three candidate packets has the same source address as incoming packet, this candidate packet and incoming packet will both be dropped. The parameter ***drop_num*** is to record the times when candidate packets have the same address as incoming packet. If ***drop_num*** is bigger than 2, i.e. there are two candidate packets have the same address as the incoming packet, we will call this case an **offense** to the router buffer and this flow could be deemed as a potential unresponsive flows. A parameter ***count_off*** is set to record the times when ***drop_num*** is bigger than 2. In others words, ***count_off*** describes the counts of offense of a potential unresponsive flows to router buffer. If ***count_off*** exceeds a threshold, we set it as 5; this flow will be considered as an unresponsive flow.

In fact, not only can this scheme detect unresponsive flows precisely, but also can avoid the case of TCP flows are misclassified as unresponsive flows. It has been proved that this method can really take effect in differentiating unresponsive flows from normal flows. The part of algorithm is showed as following:

For a flow that the incoming pkt belongs to is not a unresponsive flow temporary

```
if (qlen< q_scale* q_lim) {
    all pkts can enter;
} else {
    compare the pkt with three candidate packets ran-
    domly picked form a buffer
    drop_num = the times of candidate packet with the
    same address as pkt
            if (drop_num>1) {
              drop pkt;
            if (drop_num>2)
              count_off ++;
             if (count_off>5)
              the flow is an unresponsive flow;
          }
else {
          if succeed in dropping pkt with p_mark
             p_mark=p_mark-p_decr;
          else
             pkt is permitted to enqueued;
             p_mark=p_mark+p_incr;
           }
}
```

2.2 Punishment of Unresponsive Flows

The punishment scheme for the unresponsive flows includes two parts: "imprisonment" and "parole". We will describe these two different punishment phases for unresponsive flows in detail.

2.2.1 Imprisonment of Unresponsive Flows

Once a certain flow is considered as an unresponsive flow, it will be limited at the sending rate of data packets immediately. In other words, this unresponsive flow is "imprisoned" or it is being in the state of "**imprisonment**". Two Parameters are used to describe the state of a specified unresponsive flow when it is imprisoned. One is ***freeze_time*** which describes the durance of imprisonment for an unresponsive flow; the other is ***p_drop*** which potentially describes the drop probability of this unresponsive flow in the later phase. Usually we set ***freeze_time*** as 50ms in CCU. It has been proved that the unresponsive flows will be depressed excessively and can not obtain fair share of bandwidth if the value of ***freeze_time*** is too big.

2.2.2 Parole of Unresponsive Flows

If the imprisonment is over, the unresponsive flow will be paroled. When every packet of this flow comes into the buffer, CCU will process ***p_drop*** at first, and then the incoming packet will be dropped according the ***p_drop***. If the packet is dropped successfully, CCU will process the next incoming packet, otherwise CCU will estimate if the queue length ***qlen*** has beyond a threshold, which we set it as (***pscale*** * the biggest capacity of buffer). ***pscale*** is set 0.9 in CCU. If it is true, we compare one packet randomly sampled from the current buffer with incoming packet in order to decide if the incoming packet is dropped (and imprisoned again). Only when the comparison fails, this incoming packet could be permitted to enter the buffer. In fact, due to the re-comparison, there is potential circumstance. If the comparison succeeds, we could testify the case that the flow that the incoming packet belongs must have strong relations to the current large length of queue.

2.3 Process of Normal Flows

CCU also maintains a drop probability ***p_mark*** for the normal flows. ***p_mark*** will be increased with the step of ***p_incr*** gradually if the buffer exceeds its capacity, and it will be decreased by ***p_decr*** when buffer come back to below the capacity.

3 Simulation Experiments and Performance Evaluations

3.1 Experiment Environment

The performance of CCU is evaluated by the NS2 [7]. The AQM algorithms-CHOKe, RED, FRED and SFB are also evaluated and compared to CCU. In the experiments, the unresponsive flows are simulated by the constant bit rate (CBR) flows of UDP which send the packets with constant speed, just like the DDoS which pumps packets into the Internet with the near constant speed to aim computers.

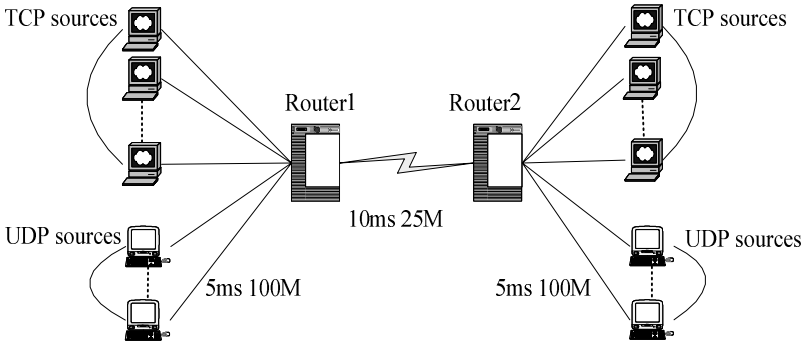


Fig. 1. Network Topology of Experiments

The network topology configuration is configured as Fig. 1. The queue capacity of the congested link between the R1 and R2 is set as 150 packets, and it is shared by 100 TCP sources and 10 UDP sources. The bottleneck link bandwidth is 25 Mbps and its transfer delay is 10ms. Every end host is connected to the routers with 100 Mbps link. All links from hosts to routers have a transfer delay of 5ms.

The TCP flows are derived from FTP sessions which are used to transmit large size files, and all TCP sources are enabled with ECN support The UDP hosts send packets with CBR flows 2 Mbps. In addition, all sources are enabled with ECN support, and randomly started within the first 1 s of simulation, and used 1 KB packets.

3.2 Experiment Result

Fig. 2-6 show the bandwidth allocation among all 100 TCP flows and 10 CBR flows for the steady period of 50ms. In Fig.2, it is clear that CHOKe could not restrain CBR flows effectively. Just as shown in the Table 1, the average bandwidth of every TCP flow is only 0.1187 Mbps, only 53.67 percent of fair share (avg TCP bandwidth) [4]; but the bandwidth of CBR flows reaches 1.2464 Mbps, nearly six times to the fair share. The detailed comparisons between different AQM algorithms are shown in Table 1.

Table 1. Comparisons between AQM algorithms

Item	CHOKe	CCU	FRED	RED	SFB
Real Bandwidth Usage(Mbps)	24.327	24.327	24.327	24.327	21.762
Bandwidth Usage Percent (%)	97.30	97.30	97.30	97.30	87
Avg TCP(Mbps)	0.119	0.220	0.200	0.071	0.216
Avg TCP/Fair Share (%)	53.67	99.27	90.43	31.87	109.20
Avg UDP(Mbps)	1.246	0.237	0.430	1.728	0.015
Avg UDP/Fair Share (%)	563.59	107.25	194.43	781.13	7.58

As Fig. 3 shows, CCU can make the bandwidth of every TCP flow reach fair share, and the bandwidth of CBR flows can also be limited in the range of fair share.

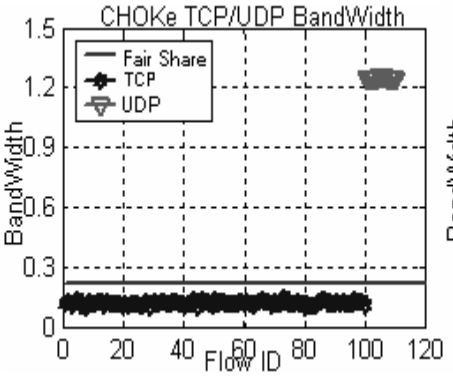


Fig. 2. Bandwidth allocation of CHOKe

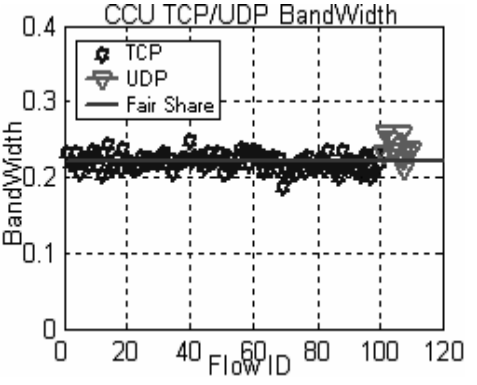


Fig. 3. Bandwidth allocation of CCU

As shown in Fig. 4, although the bandwidth of every TCP flow almost reaches fair share, the bandwidth of all CBR flows are far from fair share. Based on the experiment results, the bandwidth of every TCP flow is 0.2 Mbps, although it has 0.02 Mbps gap to the fair share 0.22 Mbps, the total difference ($0.02Mbps \times 100 = 2Mbps$) is allocated to the other 10 CBR flows. In this way, every CBR flow acquired 0.2 Mbps quotient on the base of obtaining fair share bandwidth. At last, every CBR flow gets 0.43 Mbps bandwidth. In addition, FRED can guarantee the fair bandwidth allocation of TCP flows by dividing current buffer into n sub-queue, which is decided by the number of current active flows in the network, remaining every active flow states in the network. FRED adds additional burden to router buffers. Compared with FRED, CCU can achieve more ideal results by only maintaining the state of unresponsive flows.

As Fig. 5 shows, because of RED absence of protecting TCP flows and TCP-friendly flows in nature, the router could not resist the unresponsive flow congestion which leads to falls of TCP flows and bandwidth "starvations" to TCP flows.

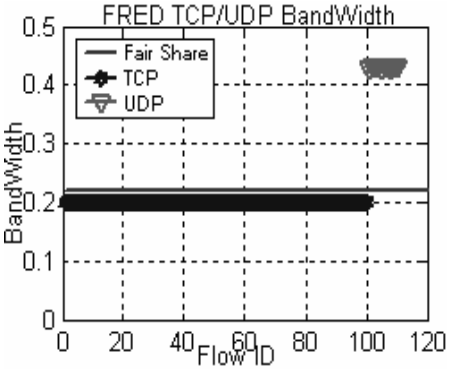


Fig. 4. Bandwidth allocation of FRED

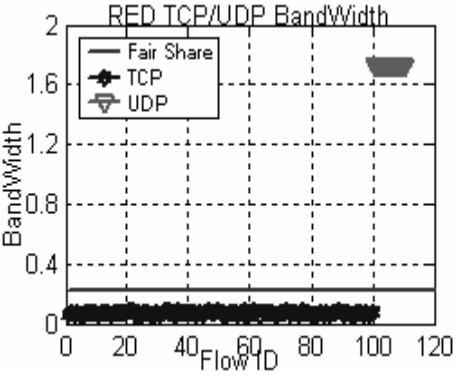


Fig. 5. Bandwidth allocation of RED

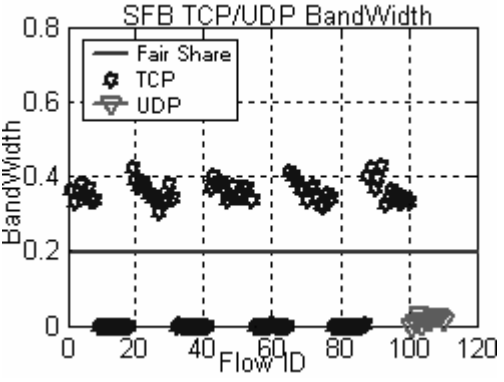


Fig. 6. Bandwidth allocation of SFB

Fig 6 shows the bandwidth allocation among all flows in SFB. Notice that SFB can really restrain CBR flows, however, because of excessive restraint, CBR flows can not achieve fair share bandwidth. On the other hand, nearly half of TCP flows are misclassified as unresponsive flows. At last, like unresponsive flows, these misclassified TCP flows could not get any bandwidth. Moreover, the mentioned above can result in the case that a few TCP flows which are not misclassified will monopolize the whole bandwidth. Eventually, the average bandwidth of these TCP flows is nearly 0.4 Mbps.

4 Conclusions

Unresponsive flows disturb the balance of the Internet and even the extreme unresponsive flows are utilized for attacking the network and steal the network resources. The paper proposed the new AQM algorithm- CCU to address the problems of unresponsive flows. Because CCU only maintains state of the detected unresponsive flows, the router computation load can not be increased. CCU can detect the unresponsive flows with high detection rate. Furthermore, according to the character of complexity of Internet flows, CCU maintains the different elastic control strategies to different flows using various packet dropping policies. The paper analyses the CCU performance compared to other main AQM algorithms by the three detailed experiments, which validate the efficiency of CCU detection and control of unresponsive flows.

References

1. S. Floyd, K. Fall: Promoting the Use of the End-to-End Congestion Control in the Internet. IEEE/ACM Transactions on Networking, Vol. 7(4), (1999) 458–72
2. S. Floyd, V. Jacobson: Random Early Detection Gateways for Congestion Avoidance. IEEE/ACM Transactions on Networking, Vol. 1(4), (1993) 397–413
3. D. Lin, R. Morris: Dynamics of Random Early Detection. In Proceedings of the ACM SIGCOMM, (1997) 127–137

4. Wu-chang Feng, Kang G. Shin: The BLUE Active Queue Management Algorithms. IEEE/ACM Transactions on Networking, Vol. 10(4), (2002) 513–527
5. W. Feng, D. D. Kandlur, S. Debanjan, K. Shin: Fair Blue - A Queue Management Algorithm for Enforcing Fairness. In Proceedings of IEEE INFOCOM, (2002) 1520–1529
6. Ao Tao, Jantao Wang: Understanding CHOKe: Throughput and Spatial Characteristics IEEE/ACM Transactions on Networking, Vol. 12(4) (2004) 694–709
7. S. McCanne, S. Floyd: NS-LBNL Network Simulator. [Online]. Available: http://www-nrg.ee.lbl.gov/ns/

Client Server Access: Wired vs. Wireless LEO Satellite-ATM Connectivity; A (MS-Ro-BAC) Experiment

Terry C. House

Nova Southeastern University 3301 College Avenue Fort Lauderdale
Davie Florida 33314
hterry@nsu.nova.edu

Abstract. This research tested a different approach to Secure RBAC in a post 911 environment. In order to defend the United States against enemies foreign and domestic; it is crucial that combat forces are equipped with the best communication equipment available. 21 years technology experience in a military environment supports this research and contributes an ideology that could revolutionize the government communication process. The "Mobile Secure Role Base Access Control" (MS-Ro-BAC) Network is a system designed to access secure global databases via wireless communication platforms [2]. This research involved the development of a (MS-Ro-BAC) Database, for wireless vs. wired access connectivity. Proprietary software was developed for authenticating with the server system. 40 database access scenarios were tested during day and night hours, for performance. The Iridium Satellite gateway was ineffective as a communication service to efficiently access the database in a global strategic environment.

1 Introduction

This project provided answers that were crucial to testing and analyzing network response time, between the Role Base Access Control (RBAC) Database Server and a client machine, in a wired vs. a wireless environment. The tasks listed below, were accomplished by the end of the research period. (1) Establish a testing strategy and topology for measuring the accessibility of the database information that is requested from the client. (2) implemented an Iridium SATCOM data transmission topology that is coupled with a 'laptop MS-Ro-BAC' computer, to communicate with the databases. (3) Established a data transmission analysis process for LEO Satellite connectivity. (4) Compiled and analyzed statistical data and presented it in a clear concise manner in the final report. (5) established a standard procedure for mobile satellite configuration using: Hardware, software, and database connectivity that are reusable with any satellite-ATM system to send and receive data in a global framework [6].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 719 – 724, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Problem Statement

The current communication approach used in the military and DoD sector falls short to send and receive database information in a global distributed network, throughout the battlefield. Current strategic information systems allow for Secure SATCOM voice transmission and burst data from one point to another; however, the response is static and specific for that individual's request and doesn't pertain to the overall situation in the city, country or world environment [3]. Individuals conducting covert and overt operations around the world are not receiving strategic, actionable information that is crucial to combat missions, as they develop on the ground. In the global war on terrorism, there is a need for a secure communication process that transmits and receives secure information in a timely manner, as it develops anywhere in the world [7].

3 Relevance and Significance

The relevance of this project was to establish a proven standardized framework to access and submit information via Iridium LEO Satellite-ATM TCP/IP protocol, in a global environment. As stated in the former text, the significance and importance to the field of information technology is the implementation of a proven methodology to access global or local information through a push-pull approach and access a wireless or wired database system [4]. Second, establish a dynamic assessment and analysis process to measure the accessibility of the database from a wireless or wired platform with an Iridium SATCOM data system. Information that is useful to other combat forces or government agencies can be entered into the database and accessed by other authorized users from any location on the globe. Each individual will have to submit their clearance level, username, pass-word and access role, before gaining access to their authorized databases [1].

4 Methodology and Approach

Listed below, are several tasks that were accomplished during the experimental process. An in-depth study and analysis of different *data measurement and assessment tools,* for satellite and network monitoring were tested and documented. The research initiated the framework for sending and receiving data by building a complete database network of servers, networked together using an Internet satellite-ATM over TCP/IP gateway [3]. To ensure layered security and data encryption, the network traffic passed through secure IDS monitoring systems, then to a series 2500 CISCO router that controlled the various network connections. A small switch was positioned on one network, hosting two wireless machines: A HP laptop and a workstation, both containing the basic client software and hardware architecture. One of the servers is a standalone RBAC Database machines that contains the authentication data for authorized users. Statistical analysis software was built into the server program, which analyzed the speed and proficiency of the data traffic between

the client and server. All experiments and analysis were conducted and funded by the researcher, to include the testing laboratory and computing machines. The actual Iridium satellite equipment that included airtime, hardware and software, were purchased through an Iridium satellite ISP provider [7].

4.1 Resources and Activities

(1) Establish Lab Facility: The laboratory was established and functional throughout the entire experimental process, it consists of 3 Windows 2000 NT servers supporting the network connectivity of the MS-Ro-BAC database [2]. In addition to the servers, there are three client machines that are connected by wireless Linksys Router, using TCP/IP data transmission. (2) *Establish Lab Network/SATCOM:* A Motorola Iridium Satellite Phone was used as a wireless modem, over LEO Satellite technology; this was possible by coupling the phone with a mobile laptop system. This configuration was representative of what the battlefield operators will use to receive and disseminate information in a permissive or non permissive environment. (3) *Install RBAC database*: The Role Based Access Control Database is installed on "Server 2" The database was designed for the basic user to easily enter or search data through the login interface website. The database prototype was designed in access for simple .asp and CGI scripting over multiple network interfaces. However, the database can easily be transformed to SQL, anytime in the future to allow for a more dynamic approach to storing and retrieving secure data [1].

5 Experimental Process

10 scenarios were executed on the MS-Ro-BAC database server during the daytime: (0710-1900) and night hours: (2000-0400), using wired connectivity and a wireless LEO data communication gateway. Each scenario was designed to test the operating abilities of the servers, network connectivity and database engines that receive information and return requested data. The ten areas of the database that were tested during day light hours, using wired and wireless connectivity, were also tested during hours of limited visibility [2].

Each scenario was timed in minutes, seconds and milliseconds, from the subject accessing the keyboard, to the information requested being returned successfully. An ability rating of 1-10, was assigned to each task within a given scenario to rate the overall actions and response from the user [4].

5.1 Database Scenario Testing

Each test executed 10 identical scenarios, which are listed below; beginning with the initial boot process of the computing system. (1) Initiate a complete cold boot process and login to the RBAC Server. (2) Access the database system and *Add* data to the '*Improvised Explosive Device* (*IED)'* database in client mode: (Add one record). (3) Execute an *Add* to the '*Personnel' database* in client mode. (Add one Record).

(4)S*earch* the *'IED' database*: (Keyword "Car Bombs" in Administrator Mode). (5) S*earch* the *'Personnel' database*. (Key Word "Terrorist" (6) Access the *'IED administrator'* area and '*Add*' an IED record, then confirm the addition. (7) Access the *'Personnel administrator'* area and '*Add'* a personnel record, and then confirm the addition. (8) Access the *'IED administrator'* area and '*Delete*' an IED record, then confirm the deletion. (9) Access the *'Personnel administrator'* area and '*Delete'* a personnel record, and then confirm the deletion. (10) Access the *'News & Updates'* page and '*open both links'*, browse both links and '*Back browse'* to the main page [3].

6 Data Analysis

The data collection methodology, consisted of 10 different test scenarios that were ran 4 times under different conditions, using wired and wireless connectivity. Forty separate forms contained statistical data for each distinct scenario. For example, there were ten wired night test, ten wired day ten wireless day and ten wireless night trials as well. After the 40 tests were conducted, information from each task was recorded, labeled and processed and summarized in the final report [6].

6.1 Data Collection Worksheet

A scenario-form was used to collect data during each experiment. To provide the reader with a view of the data collection and recording process, an example form is presented and explained below. This worksheet was used to record the actual times and performance of the individual tasks within the 10 scenarios [1].

Table 1. The example below, illustrates the task in the left hand column, such as: Cold boot the computer, network connection established, server connection established, login to the authentication server, then login in to role based access control server [2]. All times are recorded in minutes, seconds and milliseconds when appropriate. The ability rating is assessed as: 1-10, where 1 is poor and 10 is excellent. The 'attempts' directly effects the ability score at the end. The 'Total' column contains the final score of: (ability rating – (errors + attempts)) = Total effectiveness [7].

		Scenario 1 Initialization Process			
MIN.SEC.MS		Ability	errors	Attempts	total
27.2.983	time	1- 10	1 - 10	1- 10	
cold boot	0	0	0	0	0
netw.Conn.	0	0	0	0	0
serverConn.	0	0	0	0	0
login	0	0	0	0	0
GUI load	0	0	0	0	0
home	0	0	0	0	0
AVG		0	0	0	0

Table 2. The form below illustrates the statistical labeling process, for all ten scenarios of the: '*Wired Access Day'* Testing. The first column displays the raw time in seconds and milliseconds of how long it took to complete each task from start to finish. While connecting to the network and maneuvering through the various databases and pages, the quality rating score is measured by how well the process works and the ease of negotiation from one area to another. The error ratings are compiled for each scenario and provided as an overall average error rate. For each attempt made during the task, it is considered a negative value in the performance rating. The average, mode, median and range, supportted the final analysis process in the summary report [2].

Example Form for 10 Scenarios of: WIRED ACCESS DAY

	Sec.Mis	quality	Errors	Attempts	Rating
Scenario 1	00.000	00.000	00.000	00.000	00.000
Scenario 10	00.000	00.000	00.000	00.000	00.000
AVERAGE	00.000	00.000	00.000	00.000	00.000
MODE	00.000	00.000	00.000	00.000	00.000
MEDIAN	00.000	00.000	00.000	00.000	00.000
RANGE	00.000	00.000	00.000	00.000	00.000

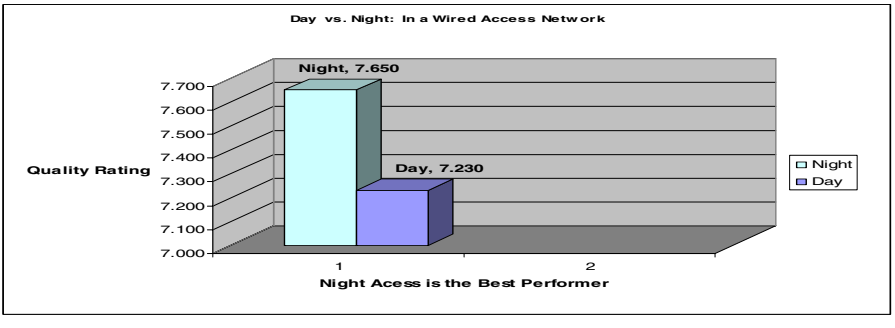


Fig. 1. According to the research results displayed in the above chart*; 'Wireless night access*,' consistently connected at a higher QOS than day access, to the MS-Ro-BAC database network [2], [5]

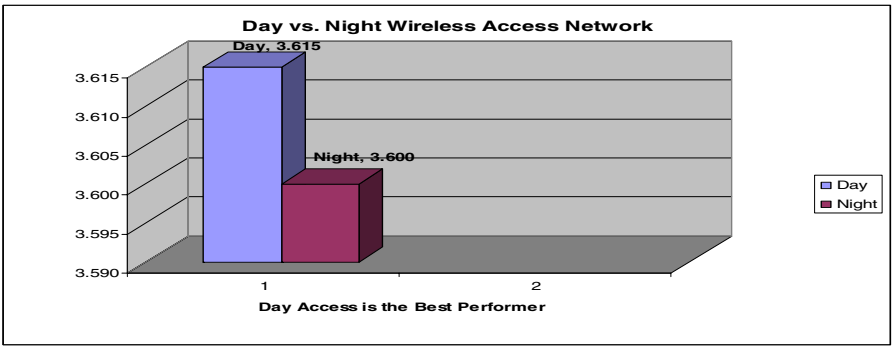


Fig. 2. According to the research results displayed above*; 'Wired night* access,' consistently performed at a higher quality rate than '*Wireless day access,'* to the MS-Ro-BAC database network [3]

7 Summary

This research proved the Iridium Gateway was inefficient as a wireless data connection service; therefore, not effectively accessing the MS-Ro-BAC database. The Iridium satellite data was being transmitted at: 3800 – 8000 kilobytes per second, where the wired T1 connection was transmitting data at: 655,000 – 1.2 megabytes per second; a noticeable difference in speed between the two approaches. This project established that civilian and military forces, could use Iridium Data Phones to send and receive secure database information, via low earth orbit satellite technology; however, the Iridium QOS is inadequate. Future MS-Ro-BAC research will test mobile high bandwidth satellite technology, with non-repudiation attributes.

References

1. Allman, M., Glover, D. “Enhancing TCP Over Satellite Channels using Standard Mechanisms,” IETF draft, February (2003).
2. House, T. “Mobile Instant Secure Role Base Control”; MS-Ro-BAC Device, SoutheastCon 2005, April 8-10, (2005).
3. ITU-R 4B Preliminary Draft Recommendation, “Transmission of Asynchronous Transfer Mode (ATM) Traffic via Satellite,” Geneva, January (2002).
4. Kalyanaraman, S., R. Jain, et. al., “Performance of TCP over ABR with self-similar VBR video background traffic over terrestrial and satellite ATM networks,” (2000).
5. Ponzoni, C. “High Data Rate On-board Switching,” 3rd Ka-band Utilization Conference, September 13-18, (2000).
6. Stallings, W. High-Speed Networks. TCP/IP and ATM Design Principles, Prentice Hall, New Jersey, (2001).
7. Stevens, W. "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms," Internet RFC 2002, January (2002).

An Algorithm for Automatic Inference of Referential Integrities During Translation from Relational Database to XML Schema

Jinhyung Kim[1], Dongwon Jeong[2], and Doo-Kwon Baik[1]

[1] Dept. of Computer Science and Engineering, Korea University, Korea
{koolmania,Baik}@swsys2.korea.ac.kr
[2] Dept. of Informatics & Statistics, Kunsan National University, Korea
djeong@kunsan.ac.kr

Abstract. XML is rapidly becoming one of the most widely adopted technologies for information exchange and representation on the World Wide Web. However, the large part of data is still stored in a relational database and we need to convert those relational data into an XML document. There are existing approaches such as NeT and CoT to convert relational models to XML models but those approaches only consider explicit referential integrities. In this paper, we suggest an algorithm to reflect implicit referential integrities to the conversion. The proposed algorithm provides several good points such as improving semantic information extraction and conversion, securing sufficient referential integrity of the target databases, and so on.

1 Introduction

With XML emerging as the data format of the Internet era, there is a considerable increase in the amount of data encoded in XML. However, the majority of everyday data is still stored and maintained in a relational database [3]. Therefore, we need to translate such relational data into XML documents.

At present, there exist several conversion approaches that enable the composition of XML documents from relational data such as Nesting-based Translation (NeT) and Constraints-based Translation (CoT) [5,8,9].

However, NeT does not support any referential integrity issue. CoT considers explicit referential integrities and reflects them to conversion. Therefore, if referential integrities are defined implicitly, the previous methods cannot secure referential integrities during translation. Moreover, anomalies can arise in target databases because the implicit referential integrity problem does not be guaranteed. To resolve the problems, the implicit referential integrity issue should be considered over the conversion.

In this paper, we study this problem of translation from a relational data model to a XML schema model. For automatic inference of implicit referential integrities, we propose an automatic inference algorithm of referential integrities as a new approach. However, we focus on automatic inference of referential integrities instead of the whole translation from the relational database to the XML document.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 725 – 730, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Translation Model

In this section, we define translation models for conversion from a relational schema to a XML schema. Initial relational schema model is defined in [5,8,9]. However, there are a few elements which have duplicate and unnecessary elements. Therefore, we revise and define new models. We also define relational schema model with implicit referential integrities which includes implicit referential integrity.

Definition 1 (Initial Relational Schema Model). A initial relational schema model is denoted by 5-tuple R_i = (T, C, P, RI_e, K), where

- T is a finite set of table names
- C is a function which represents a set of column names in each table
- P is a function which represents properties of each column and the result of P consists of 3-tuple (t, u, n); t represents data type of column such as integer, string, and so on; u represents unique or not of value of column by u (unique), ~u (not unique); n represents nullable of not of value of column by n(nullable), !n (not nullable).
- RI_e represents explicit referential integrities information
- K is a function which represents primary key information

Definition 2 (Relational Schema Model with Implicit Referential Integrities). A relational schema model with implicit referential integrities is denoted by a 6-tuple R_r = (T, C, P, K, RI_e, RI_i), where

- RI_i represents implicit referential integrities information.

3 Translation Process

3.1 Suggested Algorithm

A metadata of relational database includes the foreign key constraints and the relational data can be converted to the XML data model considering the referential integrity. However, if the database has no foreign key constraints explicitly we must infer the referential integrity and reflect it to the conversion. The inference algorithm consists of 2 steps, comparison step and extraction-refinement step.

First, we compare values of a primary key column in the one table to those of every other column in other tables. At this time, we don't need to compare values of a primary key column in the one table to values of a primary key column in the other table because value of a primary key in the one table cannot be equal to that of a primary key in the other table.

After comparison, we extract candidate of implicit referential integrities based on comparison results. Finally, we check 1:n relation of candidate of implicit referential integrities. In array of RI_{can}, if same column pair is shown up many times, those two columns have referential integrity. Fig. 1 describes the proposed algorithm.

Input: An array of values (V_k) in all columns (C_j) of every table (T_i) which are categorized by tables and attributes such as $t_1.c_1.v_1$ or $t_3.c_1.v_3$. All C_1 is primary key which is represented (PK)
Mid Output : An array of candidate or implicit referential integrities (Ri_{can})
Output : An array of implicit referential integrities (RI_i)
Procedure :
1. Initialize i=1, j=1, k=1, n=1, m=1 /********** I. First Step **********/
2. **Do while** $i<i_{(max)}$
3. **For** $j=2$ to $j_{(max)}$
4. **For** $k=1$ to $k_{(max)}$
5. **For** $n=1$ to $n_{(max)}$
6. **if**($T_i.PK.v_k = (T-T_i).C_j.v_n$)
7. $RI_{can}[m] = (T_i.PK, (T-T_i).C_j)$
8. T is a set of tables, $(T-T_i)$ is a set of table except for T_i
9. **Increment** m
10. **End if**
11. **Next** n
12. **Next** k
13. **Next** j
14. **Increment** i
15. Initialize m=2, p=1 /********** II. Second Step **********/
16. $RI[1] = RI_{can}[1]$
17. **Do while** $m<m_{(max)}$
18. **For** $p=1$ to $p_{(max)}$
19. **if** $RI[m] = RI_{can}[P]$ **Increment** p
20. **else** $RI_{can}[p] = RI[m]$
21. **End if**
22. **Increment** m
End procedure

Fig. 1. Pseudo code for the suggested algorithm

3.2 Translation Example

We show the translation process which is presented from the relational database to the XML document through example. A relational database as initial input is as in Fig. 2.

SID	Sname	PID	Cname
S01	Tom	P01	Database
S02	John	P02	Automata
S03	Cathy	P02	Database
S04	Brown	P03	Simulation
S05	Cabin	P04	Automata

Cname	Room	Time
Database	#701	#2
Automata	#702	#1
Simulation	#703	#2
Algorithm	#704	#3

PID	Pname	Office
P01	Prof. Kim	#217
P02	Prof. Lee	#613
P02	Prof. Park	#121
P03	Prof. Jean	#222

Projname	PID	SID
Data integration in sensor network	P01	S01
Wireless sensor network designation	P02	S02
Ontology system for data integration	P03	S01
Integration system based on XML	P01	S03
Simulation research for network	P02	S02
E-Government roadmap designation	P04	S04
SSL component implementation	P03	S03

Fig. 2. Example of relational database. A professor can tutor one or more students and there canbe one or more students in a class. The column Office of Professor table can be null. A project can be related to one or more students and professors.

First, we change relational database to initial relational schema model as follows:

T = {Student, Professor, Class, Project} P(SID) = {string, u, !n}, P(Sname)={string, ~u, !n}
C(Student)={SID, Sname, PID, Cname} P(PID)={string, ~u, !n}
C(Professor)={PID, Pname, Office} P(Cname)={string, ~u, !n}
C(Class)={Cname, Room, Time} P(PID)={string, u, !n}
C(Project)={Projname, PID, SID}P(Pname)={string, ~u, !n}, P(Office)={integer, u, n}

K(Student)={SID} P(Cname)={string, u, !n};
K(Professor)={PID} P(Room)={integer, u, !n}
K(Class)={Cname} P(Time)={integer, ~u, !n}
K(Project)={Projname} P(Projname)={string, u, !n}, P(PID)={string, ~u, !n}, P(SID)={string, ~u, n}

Because the initial relational schema model does not include implicit referential integrities, we must infer them through inference algorithm. Implicit referential integrities extracted by suggested algorithm are as follows: RI_i = *{(Student.PID, Professor.PID), (Student.Cname, Class.Cname), (Project.PId, Professor.PID), Professor.SID, Student.SID)}.*

Finally, we translate relational schema model with referential integrities into the XML schema model. However, in this paper, such translation part is out of scope. Therefore, we just use NeT and CoT algorithm for this translation. Resulting XML document which is the final outcome of translation is as follows:

<!ELEMENT Student (SID, Sname)>
<!ATTLIST Student Ref_Class IDREF> <!ATTLIST Student ID_Student ID>
<!ELEMENT Professor (PID, Pname, Office?, Student, Project*)*
<!ELEMENT Class (Cname, Room, Time)> <!ATTLIST Class ID_Class ID>*
<!ELEMENT Project (Projname)> <!ATTLIST Project Ref_Student IDREF>

4 Comparison Evaluation

4.1 Comparison with Exchange Scenario

Fig. 3 shows relational database exchanging scenario without our algorithm. Relational database of Side A is translated to the XML document as exchanging format. XML document is converted into the relational database for storing on Side B. If one wants to add new value to some column of relational database in Side B and that column is related to implicit referential integrity, referential integrity information error can break out. Therefore, for overcoming this problem, we apply our algorithm to this exchanging scenario.

Fig. 4 illustrates relational database exchanging scenario with our algorithm. When relational database is translated into a XML document, we can infer referential integrities through referential integrity information by using our algorithm. We can check that value by using referential integrity information received from Side A and then we can decide to allow whether a new value can be added or not. Therefore, we can prevent referential integrity problem.

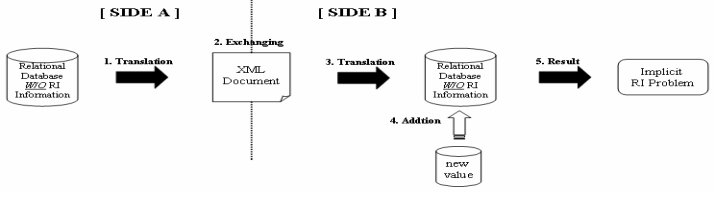


Fig. 3. Relational database exchanging scenario without the proposed algorithm

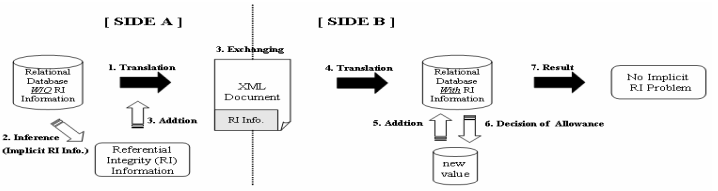


Fig. 4. Relational database exchanging scenario with the proposed algorithm

4.2 Comparison with Translated XML Document

In Fig. 2, suppose to R_e = {(Student.PID, Professor.PID), (Student.Cname, Class.Cname)}. Translated XML documents of NeT, CoT, our algorithm+NeT+CoT are as in Fig 5.

```
<!ELEMENT Student      (SID,Sname,PID+,Cname*)>
<!ELEMENT Professor    (PID,Pname,Office?)>
<!ELEMENT Class        (Cname,Room*,Time)>
<!ELEMENT Project      (Projname, PID*,SID*)>
```

(1) Translated XML document by NeT

```
<!ELEMENT Student      (SID,Sname,PID,Cname)>
<!ATTLIST Student      Ref_Class IDFEF>
<!ELEMENT Professor    (PID,Pname,Office, Student*)>
<!ELEMENT Class        (Cname,Room,Time)>
<ATTLIST Class         ID_Class ID>
<!ELEMENT Project      (Projname, PID,SID)>
```

(2) Translated XML document by CoT

```
<!ELEMENT Student      (SID,Sname)>
<!ATTLIST Student      Ref_Class IDREF>
<!ATTLIST Student      ID_Student ID>
<!ELEMENT Professor (PID,Pname,Office?,Student*,Project*)>
<!ELEMENT Class        (Cname,Room*,Time)>
<!ATTLIST Class        ID_Class ID>
<!ELEMENT Project      (Project)>
<!ATTLIST Project      Ref_Student IDREF>
```

(3) Translated XML document by CoT+NeT+Our algorithm

Fig. 5. The translated XML documents. This figure shows the translations of all methods.

In Net, we can only remove redundancy by using nesting operator such as '*', '+'. We cannot consider referential integrity information. In CoT, we can only reflect explicit referential integrities. Translated XML document by using CoT includes referential integrity information of Re. That is, referential integrities which are defined implicitly are not represented in translated XML document. The translated XML document by using our algorithm + NeT + CoT describes not only explicit referential integrity information which is illustrated in translated XML document by using CoT but also implicit referential integrity information.

5 Conclusion and Future Work

In this paper, we have defined three models: initial relational schema model, relational schema model with referential integrities, and XML schema model. We also illustrated automatic inference algorithm of referential integrities for more exact and effective translation from relational database to XML document. Our approach has the following properties. First, we can automatically infer a more exact and better XML Schema, which includes to referential integrities, from a given relational database. Second, we can avoid the insertion and deletion error.

For future works, we must consider the method for distinguishing whether two columns are same or not. However, our algorithm cannot recognize difference of them because our algorithm only checks values of columns. Therefore, we will research about the method for making a distinction of columns which have same values.

References

1. Bray, T., Paoli, J., and Sperberg-McQueen, C.M.: Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation, October (2000).
2. ISO / IEC JTC 1 SC 34, ISO / IEC 8839:1986: Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML), August (2001).
3. Elmasri, R. and Navathe, S.: Fundamental of Database Systems. Addison-Wesley (2003).
4. Fan, W. and Simeon, J.: Integrity Constraints for XML. In ACM PODS, May (2000).
5. Lee, D., Mani, M., Chiu, F., and Chu, W.W.: Nesting-based Relational-to-XML Schema Translation. In Int'l Workshop on the Web and Databases (WebDB), May (2001).
6. Kurt, C., David, H.: Beginning XML. John Wiley & Sons Inc (2001).
7. Jaeschke, G. and Schek, H.J.: Remakrs on the Algevra of Non First Normal Form Relations. In ACM PODS, Los Angeles, CA, March (1982).
8. Lee, D., Mani, M., Chit, F., and Chu, W.W.: Effective Schema Conversions between XML and Relational Models. In European Converence on Artificial Intelligence (ECAI), Knowledge Transformation Workshop (ECAI-OT), Lyon, France, July (2002).
9. Lee, D., Mani, M., Chit, F., and Chu, W.W.: NeT&CoT: Translating Relational Schemas to XML Schemas using Semantic Constraints. In the 11th ACM Int'l Conference on Information and Knowledge Management (CIKM). McLean, VA, USA, November (2002).
10. Goodson, J.: Using XML with Existing Data Access Standards. In Enterprise Application Integration Knowledge Base (EAI) Journal, March (2002) 43-45.
11. Widom, J.: Data Management for XML: Research Directions. In IEEE Data Engineering Bulletin, September (1999) 44-52.
12. Seligman, L. and Rosenthal, A.: XML's Impact on Databases and Data Sharing. IEEE Computer, Vol. 34, No. 6, June (2001) 59-67.
13. Witkowski, A., Bellamkonda, S., Bozkaya, T. Folkert, N., Gupta, A., Haydu, J., Sheng, L., and Subramanian, S.: Advanced SQL Modeling in RDBMS. ACM Transactions on Database Systems, Vol. 30, No. 1, March (2005) 83–121.
14. Duta, A.C., Barker, K., and Alhajj, R.: ConvRel: Relationship Conversion to XML Nested Structures. In SAC '04, Nicosia, Cyprus, March (2004).

A Fuzzy Integral Method to Merge Search Engine Results on Web

Shuning Cui and Boqin Feng

School of Electronic and Information Engineering, Xi'an JiaoTong University, Xi'an 710049, China
{veini, bqfeng}@mail.xjtu.edu.cn

Abstract. Distributed information retrieval searches information among many disjoint databases or search engine results and merge of retrieved results into a single result list that a person can browse easily. How to merge the results returned by selected search engine is an important subproblem of the distributed information retrieval task, because every search engine has its own calculation or definition about relevance of documents and has different overlap range. This article presents a fuzzy integral algorithm to solve the merging results problem. We have also a procedure for adjusting fuzzy measure parameters by training. Compared to the method of relevance scores fusion and Borda count fusion, our approach has the excellent ability to balance between chore effects and dark horse effects. The experiments on web show that our approach gets better ranked results (more useful documents on top ranked).

1 Introduction

Distributed information retrieval [1] search information from several information collections based on user query. This procedure divided into three stages, (1) choosing information collection, (2) issuing query and (3) merging results.

In stages 3, we should merge or fuse results to a single list. There are general two effects in fusion of multiple information sources: (1) Chorus effect, that is the more numbers of information collection think a document be important, the document must be important. (2) Dark horse effect, if an information collection with high weight thinks a document is important, the document is maybe important. Chorus effect pays an attention to most people attitude, while dark horse effect advocates individualism. The key of fusion is that get better balance of chorus effect and dark horse effect. This article presents a fuzzy integral solution to fuse results for getting better balance.

The rest of the paper is organized as follows. Section 2 discussed others known fusion algorithm. The fuzzy integral algorithm is discussed at section 3. We present our experience results and compare to two others algorithm relevance scores fusion and Borda count fusion at section 4. Section 5 concludes the paper and discussed how to train the fuzzy integral in future work.

2 Related Work / Fusion Algorithm

Different search system gives a different relevance value with a same document. When we fuse results, we need only uniform relevance. Suppose a document is

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 731 – 736, 2005.
© Springer-Verlag Berlin Heidelberg 2005

retrieved by ith systems and has ith relevance value reli. *Relevance Scores Fusion* normalizes the value to [0, 1], and give the value from one of the six algorithms followed [2]:

$$CombMIN = \min(rel_i) \quad CombMAX = \max(rel_i) \quad CombMED = med(rel_i)$$
$$CombSUM = \sum_i (rel_i) \qquad CombANZ = CombSUM / N \qquad CombMNZ = CobmSUM * N$$

Where med is median compute and N is total number of documents.

It has recently been shown [3] that the Borda Count is optimal in the sense that only the Borda Count satisfies all of the symmetry properties that one would expect of any reasonable election strategy. Each voter ranks a fixed set of c candidates in order of preference. If there are c candidates, the top ranked candidate is given c points; the second ranked candidate is given c−1 points, and so on. The candidates are ranked in order of total points, and the candidate with the most points wins the election. In *Borda Count Fusion*, the documents are candidates as well as the search engines are voters. The method is simple and efficient while no relevance scores and training data required. If a weight is associated with a voter, the voters can be distinguished. The weight should be given a fixed value or determined by training [4].

Bayes Fusion is another simple fusion algorithm. Let ri(d) is a rank of document in ith retrieval systems (if the document did not retrieved, this value is infinite), then from bayes formula, the relevant and irrelevant probability of document d are:

$$P_{rel} = \frac{p(r_1, r_2, ..., r_n \,|\, rel) P(rel)}{p(r_1, r_2, ..., r_n)} \tag{1}$$

$$P_{irr} = \frac{p(r_1, r_2, ..., r_n \,|\, irr) P(irr)}{p(r_1, r_2, ..., r_n)} \tag{2}$$

The documents order by the value of relevant probability to over irrelevant probability. Abbreviate the const items and suppose documents are independent, the last expression is [5]:

$$\sum_i \log \frac{P(r_i \,|\, rel)}{P(r_i \,|\, irr)} \tag{3}$$

The better way maybe is re-rank all documents by union systems. *Semi-supervised fusion* [6] system has its own sample database. When users issue a query to search engine, they issue same query to semi-supervised fusion system meanwhile. The documents which retrieved by both fusion system and search engine are labeled sample. Using these samples train the system, and give all documents a new rank.

Suppose that total number of independent retrieval engine is s. Every retrieval engine given a relevant scores pi=pi (d, q) of document d. The document d relevant scores is sum of all relevance. *Linear Fusion* is [7],

$$p(w, d, q) = \sum_{i=1}^{s} w_i p_i(d, q) \tag{4}$$

While w is weight and p is non-negative real number. Notice that we need training data to determine the weight w.

3 Fuzzy Integral Fusion

A fuzzy measure is a set function with monotonicity but not always additivity, and a fuzzy integral is a function with monotonicity which is used for aggregating information from multiple sources with respect to the fuzzy measure.

Let X is finite set, define as $X=\{x_1,x_2,...x_n\}$. And B be a Borel field of X, μ is function from B to interval [0, 1], if μ satisfies the following conditions,

$$\mu(\varnothing)=0,\ \ \mu(X)=1 \text{ And if } A\subset B \text{ then } \mu(A)\le\mu(B)$$

Then μ is a fuzzy measure.

Let μ is a fuzzy measure, and h: X→[0,1]. The fuzzy integral of the function h with respect to a fuzzy measure μ is defined by [8]

$$\int_A h(x)d\mu=\sup\{\alpha\wedge\mu(A\bigcap H_\alpha);\ \ 0\le\alpha\le 1\} \tag{5}$$

While $H_\alpha=\{x;h(x)\ge\alpha\}$ and A is subset of X.

Let all information collections form set $X=\{x_1,x_2,\cdots,x_n\}$, x_i is the ith information collection. $\mu(x_i)$ is the ith weight of information collection, and $\sum_i^n \mu(x_i)=1$, then μ is fuzzy measure. Suppose documents returned by all information collections are $D=\{d_1,d_2,\cdots,d_m\}$, $D_i\subseteq D$ is a document set returned by the ith information collection. Define

$$\begin{cases} r_{ij}=R_i(d_j) & d_j\in D_i \ \ 0\le r_{ij}\le|D_i|-1 \\ r_{ij}=-1 & d_j\notin D_i \end{cases} \tag{6}$$

be the rank of the jth document returned by the ith information collection. Let

$$\begin{cases} h_{ij}=h_i(r_{ij}) & r_{ij}\ge 0 \\ h_{ij}=H_i(r_{ij}) & r_{ij}=-1 \end{cases} \tag{7}$$

be membership of a document j with respect to information collection i. Compute fuzzy integral of every document, that is

$$F(d_j)=\int_A h(x)d\mu=\max_k\{\min(h_{ik},\sum_{h_{ij}\ge h_{ik}}\mu(x_i))\} \tag{8}$$

The fusion of results is ranked in order of fuzzy integral value.

4 Experiment

We do experiments on web. Four selected information collections are Google, Yahoo, AllTheWeb and Ask Jeeves[1]. Our meta-search engine named MySearch does this task. It issues key words to every search engine, and gets top one hundred results. In our experiment, given $h_{ij}=1-\frac{r_{ij}}{2(N-1)}$ $r_{ij}=0,1,2...N-1$, while N=100,and if $r_{ij}=-1$, we have $H_i=0.5$ $(r_{kj}\ge 0\ \wedge\ r_{ij}=-1)$.

According to four information collections, four fixed weight applied in our experiment, that is:

$$\mu(x_1)=0.39\ \ \mu(x_2)=0.22\ \ \mu(x_3)=0.18\ \ \mu(x_4)=0.21$$

Google has the highest weight 0.39 and AllTheWeb has the lowest weight 0.18.

[1] http://www.google.com; http://www.yahoo.com; http://www.alltheweb.com; http://www.ask.com

The results are evaluated based on the following expression [10]:

$$u = u_r * R - u_n * N \tag{9}$$

While u is utility, u_r is value of relevant document, R is total number of relevant document, u_n is cost of irrelevant document and N is total number of irrelevant documents. Here, u_r=1 and u_n=2. The bigger the value of u is, the more effect the search engine is.

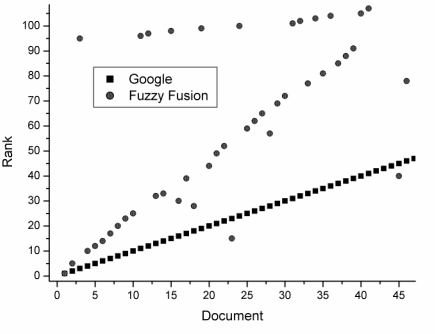


Fig. 1. Google and fuzzy fusion rank

The documents fusion show in figure 1, we only draw top 50 documents rank. From figure 1, the documents of order by Google scatter in the all range of rank. The documents not retrieved by Google but retrieved by others almost evenly insert into the all range of rank (from 1-247). The rank of a few documents was changed greatly, for example the document marked rank 3 became rank 95 after fusion.

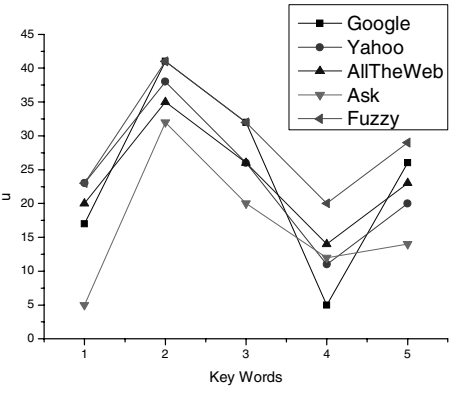


Fig. 2. Evaluate Fuzzy fusion results and other four independent search engine results

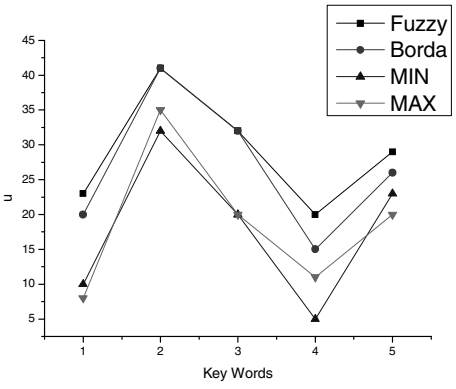


Fig. 3. Compare fuzzy fusion with Borda Count, CombMIN and CombMAX fusion

We issues 5 group key words to four search engines, and then get from 50 top of documents which returned by these search engines. According to Expression 10, we calculate u respectively and show it in figure 2. Obviously, fuzzy fusion has more effect and Google has slimily effect with fuzzy fusion on two key words. Further more, Google has more effect than other search engines.

With the same approach, we calculate u based on CombMIN fusion, CombMAX fusion Borda Count fusion and fuzzy fusion respectively. As figure 3 shown, fuzzy fusion has similarly effect with Borda Count. They both have more effect than MIN and MAX fusion since MIN and MAX do not consider weight.

5 Conclusion and Discussion

Fuzzy integral fusion can get more effect results, because this fusion do not simply ignore many but little "sound", meanwhile, it can weaken dark horse effect when individualism stand out more. It gets more balance with chorus effect and dark horse effect.

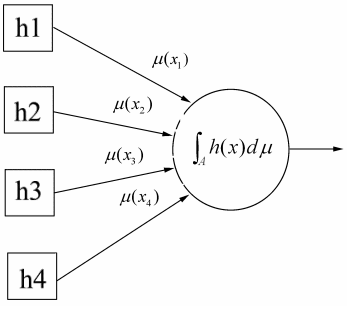


Fig. 4. A Simple fuzzy neuron model. Notice we do not care a rank of the document must be some fixed value, but we want a relevant document to be a top rank.

The weight should be trained to get more effect, one training approach describe as followed (we will consummate this method in future work):

Train data. We random select 10 group key words and get top 100 results from every search engine as training data. All data manually labeled.

Train system. We use a simply fuzzy neuron model (fig 4) presented by [9] to train the weight $\mu(x)$.Once the weights are determined, we did not change it when we merge results, even the weights are not optimal value. We know that is not an optimal weight scheme and leave more sophisticated weighting techniques for future work.

Train Result. $\mu(x_1) = 0.39 \;\; \mu(x_2) = 0.22 \;\; \mu(x_3) = 0.18 \;\; \mu(x_4) = 0.21.$

References

1. Callan J.: Distributed information retrieval. In: Croft W. B. (ed.): Advances in Information Retrieval. Kluwer Academic Pub. (2000) 127–150
2. Fox E.A., Shaw J.A.: Combination of multiple searches. In: Harman D. (ed.): The Second Text Retrieval Conference (TREC-2). Gaithersburg MD USA Mar (1994) 243–249.
3. Saari D.G.: Explaining all three-alternative voting outcomes. Journal of Economic Theory, 87(2) Aug. (1999) 313-355.
4. Javed A. Aslam, Mark Montague: Models for Metasearch. SIGIR'01 New Orleans Louisiana USA. September 9-12 (2001) 276-284.
5. Javed A. Aslam, Mark Montague: Bayes Optimal Metasearch: A Probabilistic Model for Combining the Results of Multiple Retrieval Systems. SIGIR 2000, Athens Greece July (2000) 379-381.
6. Luo Si, Jamie Callan: A Semisupervised Learning Method to Merge Search Engine Results. ACM Transactions on Information Systems, Vol. 21. No. 4. October (2003) 457-491.
7. Christopher C. Vogt, Garrison W. Cottrell: Fusion Via a Linear Combination of Scores. Information Retrieval, 1 (1999) 151–173.
8. M. Sugeno: Fuzzy mesures and fuzzy integrals. In: Gupta M. M., Saridis G. N., Gaines B. R. (eds): A survey, in Fuzzy Automata and Decision Processes. Amsterdam North-Holland (1977) 89-102.
9. James M. Keller, Jeffery Osborn: Training the Fuzzy Integral. International Journal of Approximate Reasoning, vol. 15. (1996) 1-24.
10. Lewis, D. D.: The TREC-4 filtering track. In: Harman, D. (ed.): The Third Text REtrieual Conference (TREC-4). Washington DC. U.S. Department of Commerce (1996) 165-180.

The Next Generation PARLAY X with QoS/QoE*

Sungjune Hong[1] and Sunyoung Han[2,**]

[1] Department of Information and Communication, Yeojoo Institute of Technology,
454-5 Yeojoo-goon, Kyungki-do 469-800, Korea
sjhong@mail.yeojoo.ac.kr
[2] Department of Computer Science and Engineering, Konkuk University,
1, Whayang-Dong, Kwagjin-Gu, Seoul 143-701, Korea
syhan@konkuk.ac.kr

Abstract. This paper describes the Next Generation PARLAY X with QoS / QoE in Next Generation Network (NGN). PARLAY has introduced the architecture for the development and deployment of services by service providers over 3G network. But the existing PARLAY X does not provide the open Application Programming Interface (API) for QoS / QoE. Therefore, to solve this issue, this paper suggests the PARLAY X with QoS / QoE. The object of this paper is to support the architecture and the API of the network service for QoS / QoE in NGN. The PARLAY X can provide users with QoS / QoE in network according to the detected context such as location and speed and user's preference. The architecture of the Next Generation PARLAY X is comprised of the functions for context-awareness, adaptation, and personalization.

1 Introduction

There is increasing interest in Next Generation Network (NGN). NGN needs the provision of seamless applications in the face of changing value chains and business models, requiring the ongoing replacement and extension of service delivery platform enabled by new information technology and software tools.

PARLAY [1][2] has introduced the architecture for the development and deployment of services by service providers over 3G network. However, the existing PARLAY does not provide the open Application Programming Interface (API) for Quality of Service (QoS) [3][4] / Quality of Experience (QoE) [5][6] in Next Generation Network (NGN). It can be expected that QoS / QoE for the customized network service in NGN will be deployed. QoE is defined as the totality of QoS mechanisms, provided to ensure smooth transmission of audio and video over IP networks. These QoS mechanism can be further distinguished as application-based QoS (AQoS) and network-based QoS (NQoS). AQoS includes those services provided by voice and video applications to enhance the desired end-to-end performance, while NQoS includes those services provided by the network and networking device to

* This research was supported by the Ministry of Information and Communication (MIC), Korea, under the Information Technology Research Center (ITRC) support program supervised by the Institute of Information Technology Assessment (IITA).

** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 737 – 742, 2005.
© Springer-Verlag Berlin Heidelberg 2005

enhance end-to-end QoS. QoE is the user-perceived QoS. The objective QoS is measured by QoS metrics. The subject QoS is rated by humans. The mapping between these types of QoS can be achieved by Mean Opinion Score (MOS) [7]. MOS is a measurement of quality of audio heard by the listener on a phone. Most scores break down into the following categories: bad, poor, fair, good, and excellent. The QoE parameter is currently defined for the only voice quality. The QoE parameter of non-voice quality for the multimedia is being defined by ITU-T SG 12 [8].

Therefore, this paper suggests the Next Generation PARLAY X with QoS / QoE in NGN.

The objective of this paper is as follows:

- To support QoS / QoE for the Next Generation PARLAY.

The Next Generation PARLAY X provides users with the QoS / QoE according to the changing context constraints and the user's preference. The existing PARLAY X is the open API to converse telecommunication, Information Technology (IT), the Internet and new programming paradigm. PARLAY Group, a group of operators, vendors, and IT companies, started in 1998 with the definition of an open network Parlay API. This API is inherently based on object-oriented technology and the idea is to allow third party application providers to make use of the network or in other words, have value added service interfaces.

This paper describes the design and implementation of the Next Generation PARLAY X for the next generation PARLAY X in NGN, and is organized as follows: Section 2 illustrates the design of the Next Generation PARLAY X; section 3 describes the implementation of the Next Generation PARLAY X; Section 4 compares the features and performance of the Next Generation PARLAY X. Finally, section 5 presents the concluding remarks.

2 Design of Next Generation PARLAY X

A scenario using the Next Generation PARLAY X is depicted below. We assume that there are the contexts such as location and speed in the surroundings of a wireless device that is detected from sensors called motes [9] or Global Positioning System (GPS). Moreover, we assume that the Wireless LAN region is enough of network resource and the CDMA region is short of network resource. In the middle of an on-line game, a user in the WLAN region decides to go outside to the CDMA region while continuing to play the game on a wireless device. The wireless device is continually serviced with degraded application quality on the screen, albeit there is a shortage of network resources in CDMA region. Therefore, the user can seamlessly enjoy the game on the wireless device.

2.1 Next Generation PARLAY X

Fig.1 shows the architecture of the Next Generation PARLAY X which consists of functions including *Context-awareness*, *Personalization*, and *Adaptation control* for QoS / QoE in the IT area. *Context-awareness* has a role to interpret contexts that comes from mote or GPS. *Context-awareness* is to get the context information such as location and speed, translating the context information to XML format.

Personalization has a role to process the user's personal information such as user's preference and the device type. *Personalization* is to get the personal information, translating the information to XML format. *Adaptation Control* is to reconfigure the protocol for adaptation according to the context. *Adaptation Control* is to re-composite the network protocol modules according the ISP's rule and the policy.

The wireless sensor network called a mote can detect the contexts. The Next Generation PARLAY X is to interpret the context for context-awareness and personalization according to the changing context and the user's preference and to reconfigure the protocol for adaptation.

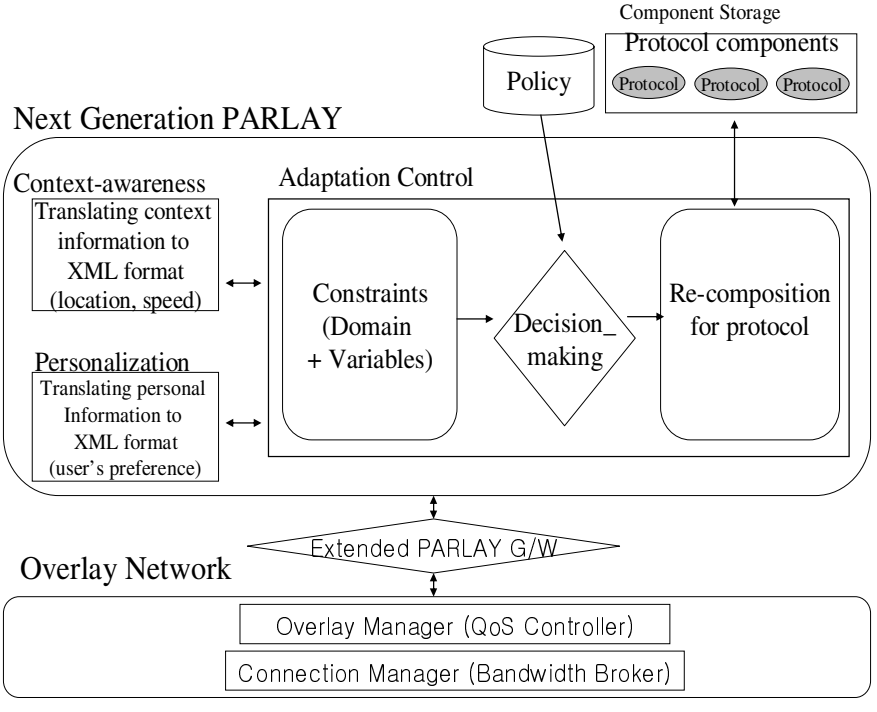


Fig. 1. Mechanism of the Next Generation PARLAY X

The Next Generation PARLAY X is for PARLAY X with QoS / QoE and to support context-awareness, personalization, and adaptation in the service layer. The role of the Next Generation PARLAY X is to obtain the context such as location and speed, to make an interpretation for context-awareness, and to re-composite each protocol for adaptation and personalization according to the context. The network can support QoS / QoE according to the request of the Next Generation PARLAY X in the control layer. The Next Generation PARLAY X uses the XML-based web services technology. The overlay network comprises of functions including the *Overlay manager* for the QoS controller role and the *Connection manager* for the bandwidth broker role.

2.2 Definition of Context for Next Generation PARLAY X

Fig.2 shows the context, profile, and policy of the Next Generation PARLAY X. The context consists of location, speed, weather, temperature, time of day, and presence. The defined variables of context are as follow: l is location, s is speed, wt is weather, tm is temperature, tod is time of day, and p is presence. The profile consists of the user's preference and the device type. The policy is as follows: if (current_location = 'WLAN region') then call RTP protocol else if (current_location = 'CDMA region') then call WAP protocol means to call a function for RTP protocol in the case that the current location is in a Wireless LAN region where resources of a network in the surroundings are enough, and to call a function for WAP protocol in case that the current location is in a CDMA region where the resource of network is scarce.

Fig.2 shows the sequence diagram for the Next Generation PARLAY X. The sensor or GPS detects context information such as the location and speed and it informs context information of the Next Generation PARLAY X. The Next Generation PARLAY X can adaptively choose the optimized protocol, analyzing the context information and policy of the ISP. For instance, as the current location of mobile device is in the WLAN region, users can get the high quality of service through Real Time Protocol (RTP), whereas as if the current location of mobile device is in the CDMA region, users can get the low quality of service through Wireless Application Protocol (WAP).

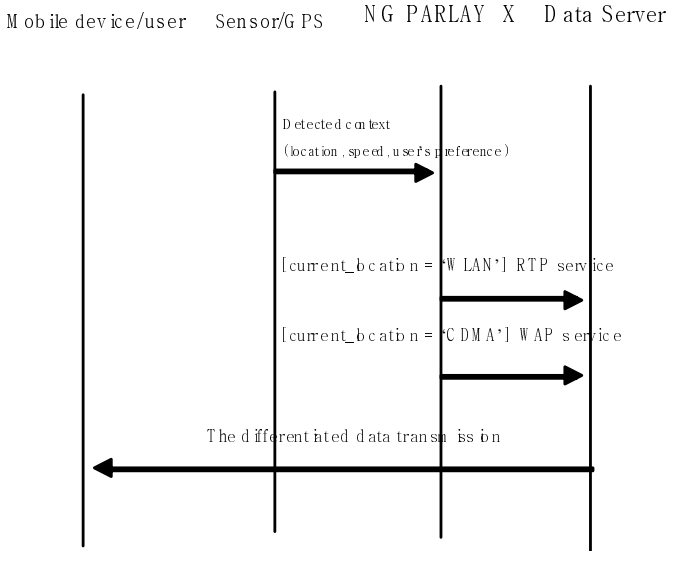


Fig. 2. Sequence diagram for the Next Generation PARLAY X

3 Implementation of Next Generation PARLAY X

The implementation of the Next Generation PARLAY X is based on Windows 2000 server. The Next Generation PARLAY X uses XML-based web services and

intelligent agent called JADE [10]. We have four steps in the implementation. First, UML/OCL notation is defined. The context constraint is the expression using OCL. Second, UML/OCL is translated into XML. Third, XML is processed by the Next Generation PARLAY X. Fourth, the Next Generation PARLAY X provides users with the network service with QoS / QoE. The Next Generation PARLAY X includes the new defined PARLAY X API such as getContextAwareness(), getPersonalization, and adaptiveProtocolType(). We develop the prototype using PARLAY X SDK called GBox [11].

We assume that there are WLAN region and CDMA region according to the horizontal (x) and vertical (y) axes of PARLAY simulator. The Next Generation PARLAY X within PARLAY X simulator can provide RTP protocol or WAP protocol according to the context information such as location.

In Fig.1, if the Next Generation PARLAY gets the location of wireless device, the Next Generation PARLAY can provide network service through the suitable protocol. For instance, if the current location of wireless device is in the WLAN region, the Next Generation PARLAY provides the high quality service through RTP. If the current location of wireless device is in the CDMA region, the Next Generation PARLAY provides the low quality service through WAP.

4 Comparison of Features of Existing PARLAY and Next Generation PARLAY X

4.1 Comparison of Main Features

Table 1 shows the comparison of main features of the existing PARLAY X and the Next Generation PARLAY X. The Next Generation PARLAY X has more features, such as supporting QoS / QoE than the existing PARLAY X. The Next Generation PARLAY X can support QoS including bandwidth and latency and QoE including MOS in the network. Conversely, PARLAY X does not consider QoS / QoE in the network.

Table 1. Comparison of main features

	Existing PARLAY X	Next Generation PARLAY X
QoS (Bandwidth,Latency)	-	X
QoE (MOS)	-	X

4.2 Comparison of Performance

We evaluate performance using ns-2 simulator. There are four nodes for the performance evaluation in ns-2 like Fig.2. The node 0 is for the mobile device. The node 1 is for GPS. The node 2 is for the Next Generation PARLAY X. The node 3 is for the data information server. The node 1 informs the node 2, which is the Next Generation PARLAY X, of the location of user, detecting it from the sensor or GPS.

The node 2 is to re-composite the network protocol according the network resource. We evaluate the packet size of data that is sent to the user. We define the ChangingContext() method using C++ programming language in ns-2 for evaluation in case that the context is changed. The existing PARLAY is stopped in case that the current location in WLAN is changed in CDMA region. Conversely, the Next Generation PARLAY X can keep the service because the RTP protocol service is changed to the WAP protocol service in case that the current location in WLAN is changed in CDMA region. This is attributed to the fact that the Next Generation PARLAY X supports QoS / QoE, whereas the existing PARLAY does not have QoS / QoE.

5 Conclusion and Future Works

This paper suggests the PARLAY X to support QoS / QoE in NGN. This paper shows the Next Generation PARLAY X can provide users with QoS / QoE by detecting the context information such as the location and speed. We believe that the Next Generation PARLAY X addresses new service mechanism on delivery network platform to support more QoS / QoE on the network than the existing PARLAY X. We expect the Next Generation PARLAY X to comply with the industry standard such as PARLAY. We think NGN needs our approach to support the intelligence such as QoS / QoS in the network. Our future work will involve more studies on intelligent network service as well as QoS / QoE in NGN.

References

1. PARLAY home page. www.parlay.org
2. O. Kath, T. Magedanz, R. Wechselberger, "MDA-based Service Creation for OSA/Parlay within 3Gbeyond Environments," First European Workshop on Model Driven Architecture with Emphasis on Industrial Application, University of Twente, Enschede, Netherlands. (2004)
3. ITU-T Recommandation G.1000, “Communication Quality of Service : A Framework and Definitions,” Recommendation, ITU-T. (2001)
4. 3GPP. TS 23 107 V5.12.0 3rd Generation partnership Project, “Technical Specification Group Services and System Aspect : Quality of Service (QoS) Concept and architecture (Release 5),” Technical Specification, 3GPP. (2004)
5. ITU-T Recommendation G 107. The E-Model, “a Computational Model for Use in Transmission Planning,” Recommendation, ITU-T. (2003)
6. Timothy M.O’Neil, “Quality of Experience and Quality of Service, For IP video conferencing,” White paper by Poly com. http://www.h323forum.org/papers/polycom/QualityOfExperience&ServiceForIPVideo.pdf
7. ITU-T. Recommendation P.800.1, “Mean Opinion Score (MOS) terminology,” Technical report, ITU-T. (2003)
8. ITU-T SG 12 document. http://www.itu.int/itudoc/itu-t/ifs/072003/pres_org/tsg12.pdf
9. L. Girod, J. Elson, A. Cerpa, T. Stathopoulos, N. Ramanathan, D. Estrin, "EmStar: a Software Environment for Developing and Deploying Wireless Sensor Networks," in the Proceedings of USENIX General Track. (2004)
10. JADE home page. http://jade.cselt.it/
11. GBox home page - PARLAY X Service Creation Environment. http://www.appium.com

A Design of Platform for QoS-Guaranteed Multimedia Services Provisioning on IP-Based Convergence Network

Seong-Woo Kim[1,*], Young-Chul Jung[2], and Young-Tak Kim[2]

[1] Institute of Information and Communication, Yeungnam University, Korea
[2] Dept. of Info. & Comm. Eng., Graduate School, Yeungnam University, Korea
{swkim, ytkim}@yu.ac.kr, callaguy@yumail.ac.kr

Abstract. In order to provide QoS-guaranteed real-time multimedia services, establishment of QoS-guaranteed per-class-type end-to-end session and connection is essential. Although the many researches for QoS-guarantee have been applying to the provider's network, these have not been completely guaranteeing the end-to-end QoS yet. The topic of the end-to-end QoS-guarantee is still an open issue. In this paper, we propose a platform supporting real-time multimedia services of a higher quality between end-users. To guarantee the QoS between end-users, we use the SDP/SIP, RSVP-TE and CAC. The SDP/SIP establishes end-to-end sessions that guarantee the users' demanded QoS. The RSVP-TE and CAC establish the QoS-guaranteed path between edge nodes on network for the established session through SDP/SIP. The proposed platform can apply to not only the existing IP-based network but also the wired/wireless convergence network of near future.

1 Introduction

Although the many researches for QoS-guarantee have been applying to the provider's network, these have not been completely guaranteeing the end-to-end QoS yet. The topic of the end-to-end QoS-guarantee is still an open issue. In order to provide QoS-guaranteed real-time multimedia services on IP networks, two signaling functions must be prepared: (i) end-to-end signaling to initialize a session, (ii) UNI/NNI signaling to establish QoS and bandwidth-guaranteed virtual circuit for media packet flow.

In this paper, we propose a platform supporting real-time multimedia services of a higher quality between end-users. To guarantee the QoS between end-users, we use the SDP/SIP [1,2], RSVP-TE and CAC (Call Admission Control). We use the SDP/SIP to initialize a end-to-end multimedia session that guarantees the QoS. And we use the RSVP-TE and CAC to establish the QoS-guaranteed connection (or path) between edge nodes on network for the established session through SDP/SIP. These session and connection (or path) must be established among participant's terminals.

The rest of this paper is organized as follows. In section 2, related works are briefly introduced. In section 3, it explains the functional model of the proposed platform such as QoS-guaranteed session, connection establishment, and CAC function. And

* Corresponding author. Tel.: +82-53-810-3939, Fax: +82-53-810-4742.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 743 – 748, 2005.
© Springer-Verlag Berlin Heidelberg 2005

we also implement the proposed platform and experiment it on the test-bed network. Finally, we conclude this paper in section 4.

2 Related Works

2.1 Session Description Protocol (SDP) and Session Initiation Protocol (SIP)

SIP (Session Initiation Protocol) is a signaling protocol on application-layer that was developed by IETF (Internet Engineering Task Force) [1]. SIP executes establishment, modification, and release of sessions among one or more participants [1]. SIP may run on top of several different transport protocols, such as UDP and TCP. SIP invitations used to create sessions carry session descriptions that allow participants to agree on a set of compatible media types. SIP is composed by three main elements as follows:

- *User Agents (UAs)* originate SIP requests to establish media sessions and to exchange media. A user agent can be a SIP phone, or SIP client software running on a PC, laptop, PDA, or other available devices.
- Servers are intermediary devices that are located within the network and assist user agents in session establishment and other functions. There are three types of SIP servers defined in [1]: *proxy*, *redirect*, and *registrar servers*.
- *Location servers* are general term used in [1] for a database. The database may contain information about users such as URLs, IP addresses, scripts features and other preferences.

SDP [2] has been designed to convey session description information, such as session name and purpose, the type of media (video, audio, etc), the transport protocol (RTP/UDP/IP, H.320, etc), and the media encoding scheme (H.261 video, MPEG video, etc). SDP is not intended to support negotiation of session content or media encodings.

2.2 UNI Signaling with Resource Reservation Protocol Traffic Engineering (RSVP-TE)

After determination of QoS and traffic parameters for a multimedia session, QoS-guaranteed per-class-type connections for the session must be established among the participant terminals. In network, connection establishment is accomplished by UNI (user-network interface) and NNI (network node interface) signaling. For UNI signaling between user terminal and ingress edge router, RSVP-TE can be used to carry the connection request [3]. In order to support per-class-type DiffServ provisioning, RSVP-TE must provide traffic engineering extensions so as to deliver the traffic and QoS parameters. The user agent in multimedia terminal must provide RSVP-TE client function, while the ingress edge router must support the RSVP-TE server function. Since RSVP-TE establishes only unidirectional connection, two *PATH-RESV* message exchanges should be implemented to establish bidirectional path between user terminal and ingress router.

3 Platform for QoS-Guaranteed Real-Time Multimedia Services Provisioning

3.1 QoS-Guaranteed Session and Connection Management Using SDP/SIP, RSVP-TE and CAC

Figure 1 shows the overall interactions among session and connection management functions. The user terminal will firstly discover required service from database (service directory) that provides service profiles. If an appropriate application service is selected and agreed via SLA (Service Level Agreement), session setup and control for the multimedia application will be initiated. SDP/SIP will be used to find the location of destination and to determine the availability and capability of the terminal. The SIP proxy server may utilize location server to provide some value-added service based on presence and availability management functions.

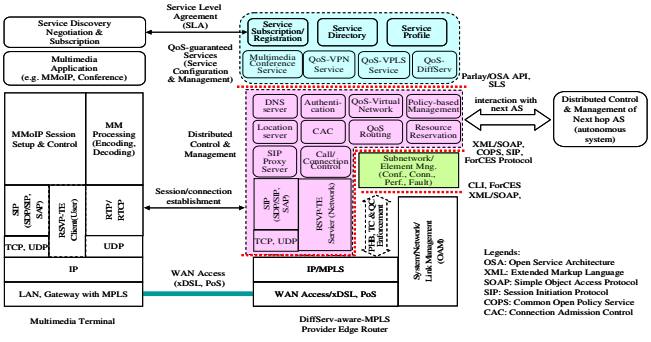


Fig. 1. Session and Connection Management Functional Modules

Once session establishment is agreed by participants, QoS-guaranteed per-class-type end-to-end connection or packet flow establishment will be requested through UNI signaling. RSVP-TE may be used in network. The edge node (router) of ingress provider network will contain connection control and management functions for on-demand connection establishments. When the customer premises network (CPN) or participant's terminal does not support RSVP-TE signaling function, an end-to-end connection will not be established: instead, per-class-type packet flow must be registered and controlled by the connection management function of ingress PE (Provider Edge) node with CAC. The customer network management (CNM) system may support the procedure of per-class-type packet flow registration [5, 6].

Figure 2 depicts the overall flow of session and connection establishment for QoS-guaranteed multimedia service provisioning. User agent A (caller) invites user agent B (callee) through *INVITE* request. SIP proxy server delivers SIP messages. When *INVITE* message created by caller, message body includes information of services that caller can provide. These information are SDP contents related to media, QoS, security, and etc. When callee receives *INVITE* request from caller, callee analyzes SDP information and checks the requested QoS. And then callee sends response message (*183 Session Progress Response*) to caller. Here, callee specifies available services of self corresponding to information of *INVITE* message, in 183 session pro-

gress response message. When caller receives the 183 message, if caller can accommodate callee's request, caller sends *PRACK* (Provisional Response ACKnowledgment) message to callee as response of 183 message. If not, caller notifies failure of session establishment to callee. Callee receives *PRACK* message from caller and sends *200 OK* message to caller as response of *PRACK* message. After caller receives *200 OK* response message, it executes connection establishment through RSVP-TE. To establish QoS-guaranteed connection, *Coordinator* of caller provides the related traffic/QoS parameters to RSVP-TE client. These parameters are based on session parameters that negotiated at previously procedure (show from flow 1 to flow 15 in figure 2). However, RSVP-TE client establishes bidirectional path between caller and callee. Bidirectional path is made by two unidirectional paths between caller and callee actually. For the resource allocation, the RSVP-TE client sends *PATH* message to the RSVP-TE server. The RSVP-TE server calls NMS[6]'s CAC to check the availability of the resources. The CAC checks with the available resources and notifies its result to RSVP-TE server. When requested resource is unavailable, RSVP-TE server notifies error message to the caller's RSVP-TE client. If resource is available, RSVP-TE server forwards the *PATH* message to the callee. The callee responds with *RESV* message to the RSVP-TE server, which will make the resource allocated.

After bidirectional path is established, coordinator creates RTP (Real-time Transport Protocol) stream (media session) connection using *socket API* on bidirectional path. And then caller and callee exchange the rest SIP messages (*UPDATE, Ringing, 200 OK, ACK*) so as to use multimedia session. The real-time Video/Audio data streams are exchanged through this multimedia session. Of course, this guarantees a negotiated QoS.

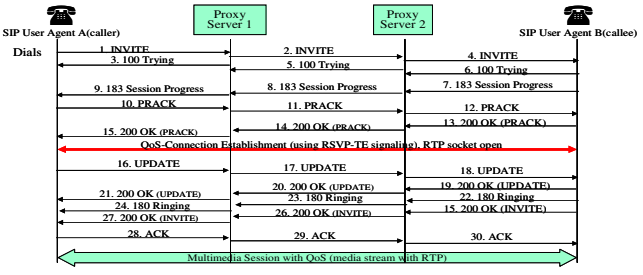


Fig. 2. Overall flow of QoS-guaranteed Multimedia Service Platform

3.2 Implementation and Analysis

Figure 3 shows architecture of proposed QoS-guaranteed multimedia service platform. It is composed of multimedia terminal platform, SIP proxy/registrar server, RSVP-TE server/CAC, and NMS (Network Management System). We use VOCAL system[7] without any modification in it as a SIP server.

In figure 3, multimedia service coordinator manages every functional module (SDP/SIP, RSVP-TE client, Real-time multimedia transport module) on multimedia terminal. SDP/SIP module achieves end-to-end session establishment through interaction with SIP servers. RSVP-TE client module requests QoS-guaranteed connection (path) to RSVP-TE server that interacts with NMS. Real-time multimedia transport module exchanges real-time audio/video data stream between end-users.

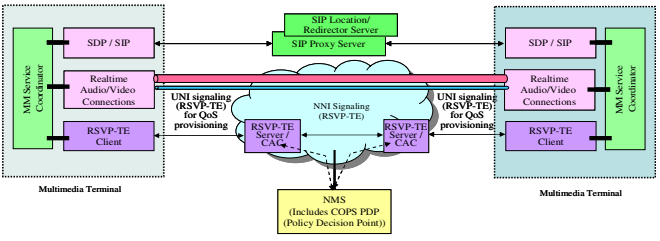


Fig. 3. Proposed QoS-guaranteed Multimedia Service Platform

We implemented proposed platform and analyzed its performance on test-bed network environment as Figure 4. In figure 4, backbone links among PEs (Provider Edges) are 155Mbps OC-3 link. We implemented multimedia terminal function on four laptops and PDAs. The halves of multimedia terminals support QoS-guaranteed service and the rest support best-effort service. The IEEE 802.11b wireless links are used to communicate between PDAs or between laptop and PDA. QoS-guaranteed connection and best-effort connection are integrated into one edge-to-edge tunnel that had been established through RSVP-TE and NMS. We generated background traffic of 155Mbps on physical link among users' terminals. The background traffic is generated after end-to-end session and connection establishment. We allocate bandwidth of 20Mbps to EF (Expedited Forwarding) tunnel (end-to-end LSP) between end-to-end edges. In order to provide QoS-guaranteed multimedia service (EF-service), we also provide bandwidth of 2Mbps to QoS-guaranteed media connection between users' terminals. The rest users' terminals provide best-effort service.

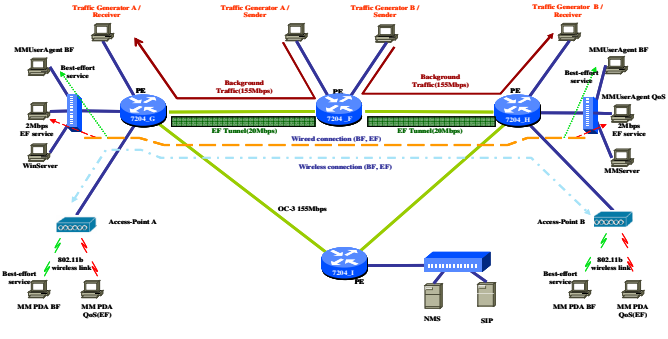
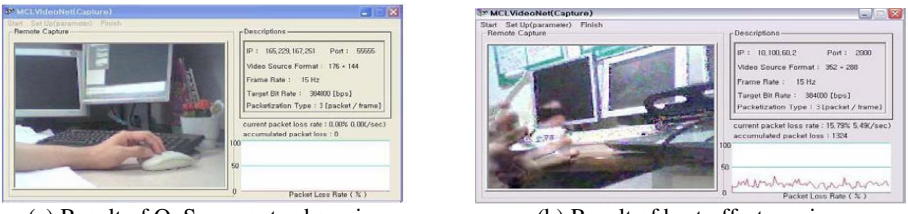


Fig. 4. Test-bed network to test proposed platform

Figure 5 shows results of our experimentation. The experimentation was a real-time video conference. For this experimentation, we implemented video/audio codec (H.263, G.723.1) in users' multimedia terminal platform. We ran EF service and best-effort service on the established EF tunnel simultaneously. In the congestion environment occurred because of background traffic, the users' terminal supporting EF-service satisfied the requested QoS. But users' terminal supporting best-effort service showed performance degradation. With result of this experiment, the proposed

platform demonstrated end-to-end QoS guarantee for real-time multimedia service on wired/wireless convergence network environment.



(a) Result of QoS-guaranteed service (b) Result of best-effort service

Fig. 5. Results of experimentation

4 Conclusions

In order to provide QoS-guaranteed real-time multimedia services, establishment of QoS-guaranteed per-class-type end-to-end session and connection is essential. It is required tightly coupled interactions of session and connection management and CAC.

In this paper, we proposed QoS-guaranteed multimedia service platform. The proposed platform also provides a functional model of interactions among SDP/SIP, RSVP-TE/CAC, and NMS. We implemented the proposed platform and analyzed its performance on test-bed network environment. With the result of our experimentation, the proposed platform demonstrated end-to-end QoS guarantee for real-time multimedia service on wired/wireless convergence network environment. And based on the designed platform architecture, we expect that the proposed platform can be applying to not only the existing IP-based network but also the wired/wireless convergence network of near future.

Acknowledgement. This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

References

1. J. Rosenberg et. al., “SIP: Session Initiation Protocol,” IETF RFC 3261, June 2002.
2. M. Handley and V. Jacobson, “SDP: Session Description Protocol,” IETF RFC 2327, April 1998.
3. D. Awduche, et. al., “RSVP-TE: Extensions to RSVP for LSP Tunnels,” IETF RFC 3209, December 2001.
4. F. Le Faucheur, et. al., “Multiprotocol Label Switching (MPLS) support or Differentiated Services,” May 2002.
5. Young-Tak Kim, Hae-Sun Kim, Hyun-Ho Shin, “Session and connection management for QoS-guaranteed multimedia service provisioning on IP/MPLS networks,” ICSA2005, May 2005.
6. Young-Tak Kim, “DoumiMan (DiffServ-over-universal-MPLS Internet Manager) for Guaranteed QoS Provisioning in Next Generation Internet,” NOMS2004, April 2004.
7. VOCAL SIP server system, “http://vovida.org”.

Introduction of Knowledge Management System for Technical Support in Construction Industries

Tai Sik Lee[1], Dong Wook Lee[2], and Jeong Hyun Kim[3]

[1] Professor, Dept of Civil & Env Engineering, Hanyang University, Ansan, Kyunggi, Korea
cmtsl@hanyang.ac.kr
[2] Research Professor, Dept of Civil & Env Engineering, Hanyang University, Ansan, Kyunggi, Korea
dwlee@cmnet.hanyang.ac.kr
[3] Senior Researcher, Korea Land Corp., Sungnam, Kyunggi, Korea
jeonghkim@iklc.co.kr

Abstract. The "Knowledge Management System" has been introduced for the necessity of convenient communication and productivity improvement; the existing legacy systems have limitations such as the low effectiveness in information sharing and such functions. This study developed an enhanced construction information management system which improved the functions of storing, searching, and sharing of the information. The proposed "Knowledge Document Management (KDM) Portal" can perform Knowledge management through various access methods. Personal files can be managed with a file viewer and advanced viewer functions. The system also enables a 'quick search' using a highlighting system within the text-file search.

1 Introduction

Recently, much attention has been focused towards construction information sharing for easy communication and the improvement of effective business. The construction information sharing between the head office and the field offices since the 1990's is a representative case. This includes the information sharing through enterprise resource planning, construction information sharing through the project management information system, and sharing 'Know-how' through the knowledge management system (KMS). KMS has expanded the range of information sharing not only between the head office and field offices, but also between the owner and the subcontractors. However, less effort has been given to the improvement of the effectiveness in information sharing and its functions. The legacy systems have some limitations in storing, searching and sharing of the information, and these may be an obstacle in the use of the system.

This study is to develop a construction information system which overcomes the prevailing limitations such as storing, searching and sharing the information. The enhanced system should be capable of operating simultaneously with the legacy system.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 749 – 756, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Knowledge Management System Implementation in the Construction System

2.1 Current Application of Technical Information Management System

The server systems are only for the storage of technical information without KMS in most of the construction industries, so that the multiple functions in the expensive server system are not used to full capacity. The technical information is managed by a department and/or persons in the type of a folder, and can hardly be effective knowledge assets for the companies.

The existing systems have spatial and time limitations in technical information management (see **Fig. 1. & Fig. 2.**) and search because it is hardly possible to integrate and manage all of the accumulated technical information. In addition, the standard operation has not been provided and the security management is hardly satisfactory, possibly leading to the degradation of productivity, or a drain in knowledge assets. Other problems include experts for KMS operation and management, as well as budget limitations.

Most of the companies implementing knowledge management fall short in the knowledge asset value sharing due to the lack of KMS recognition. The knowledge sharing is not based on business type and/or the process. It may cause conflicts with existing business; therefore the knowledge sharing is executed separately.

In terms of the function, the knowledge management theories are fully implemented to the information system with various functions, but it may be too complicated for the users. In addition, most KMS functions overlap with those of the existing information systems, so that the usage is not satisfactory. According to the results of a survey, the cost for system construction is regarded to be too high in comparison to business effectiveness, and the lack of system operators/managers is an additional obstacle.

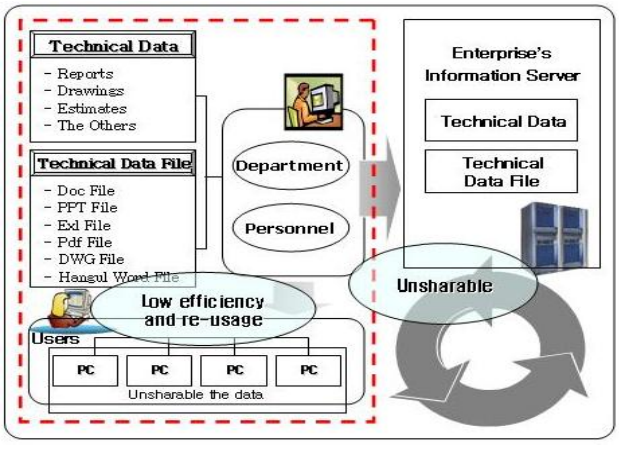


Fig. 1. Limitations of Technical Information Management Systems

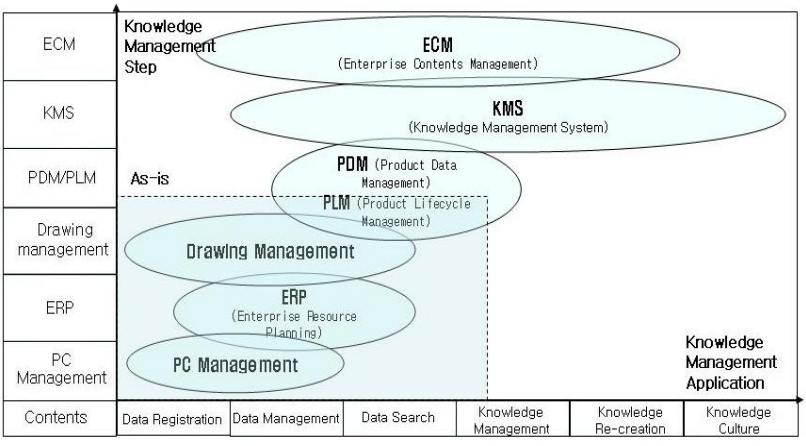


Fig. 2. Current Applications of KMS

2.2 Considerations in KMS Manifestation

The major functions required for knowledge sharing are shown in **Fig. 3.** In order to solve the problems in technical information management and existing KMS, establishment of the standard system for management is required. The standard system can improve the existing technical information management system by developing the system further, based on the business type and process. The quick-searching and full-text searching functions are also necessary. The search function with highlighting for the meta-data and full-text based search is then required. The breakdown structure should be prepared corresponding to the characteristics of each business, and the system should be made flexible for the management of personal technical documents by providing an integrated viewer for various types of technical information. In order to encourage the use of KMS, a specialized function for knowledge sharing should be implemented, as well as security coding for core information by a security system.

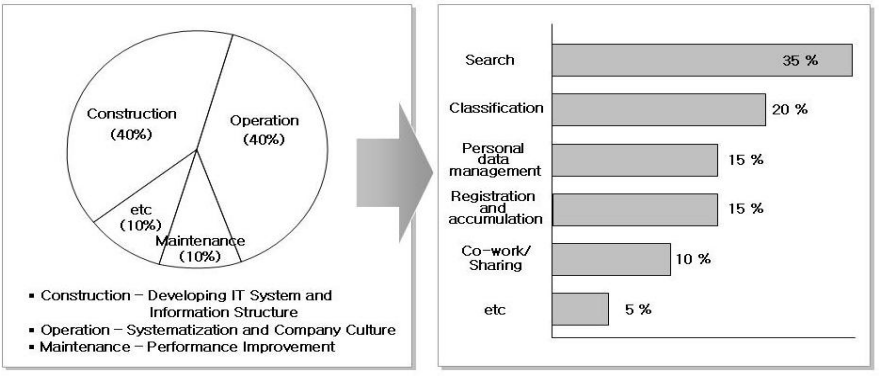


Fig. 3. Required Functions for KMS

3 Procedure and Architecture for KMS Development

In developing KMS, the three main problems are as follows:

- How to find and manage the scattered knowledge in the organization?
- How to use the knowledge better than through the legacy system?
- How to devise an approach for the enhancement of the system?

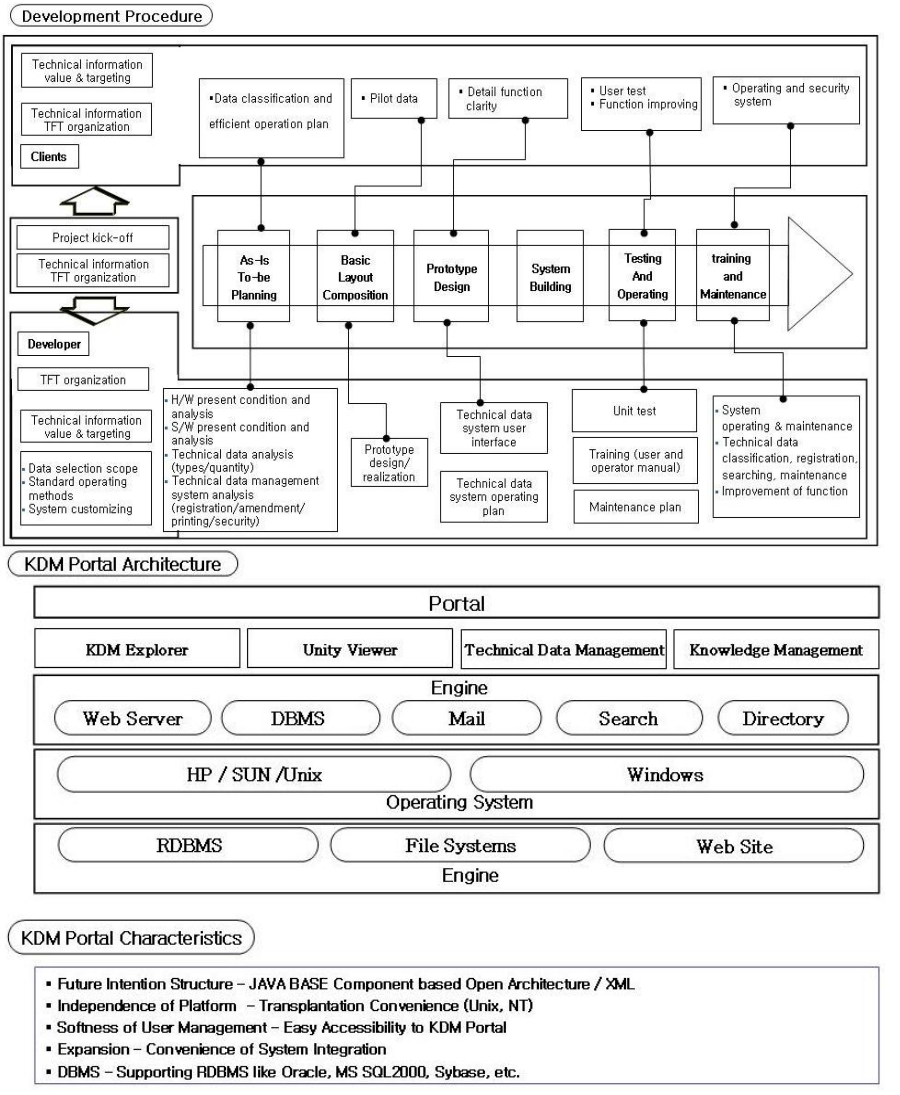


Fig. 4. Procedure for KDM Development and Architecture

The proposed system developed in this study (see **Fig. 4**) is trying to overcome limitations in the existing KMS, named the "Knowledge Document Management System (KDMS)".

4 Characteristics and Function of the KDM Portal

4.1 KDM Portal

The KDM portal (see **Fig. 5**) has various access methods. The KDM explorer and the file browser can manage personal files; and the viewer of text in files for the technical data and the advanced viewer can make flexible knowledge management possible. The portal also has the full-text search function with highlighting support.

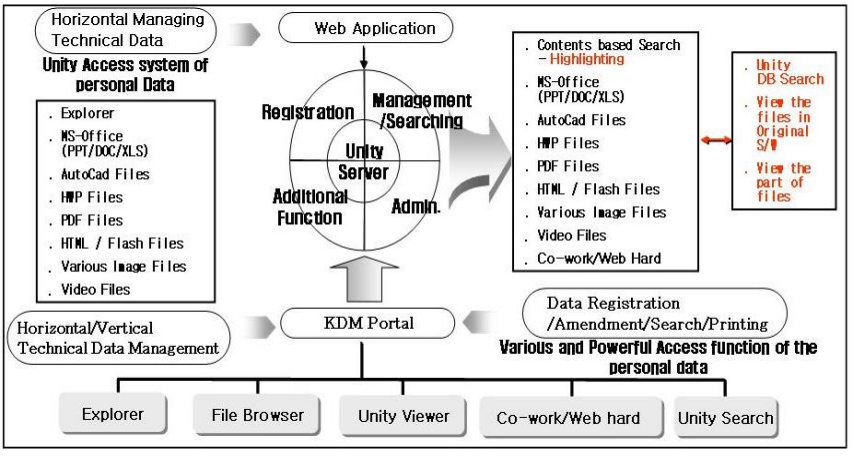


Fig. 5. Characteristic of KDM portal

4.1.1 Process

The process of the KDM portal (see **Fig. 6**) enhanced the storage of technical data by project in technical information, management/searching, and the output of drawings. The KDM portal is designed for the information and the system in a company to be accessible at a single point. Especially the browsing, sharing, and knowledge-searching functions can be a tool for facilitating work between persons or departments. It can improve the ability of competition against other companies by maximizing the use of technical information accumulated in the company.

4.1.2 Browser Function

The browser function of KDM (see **Fig. 7**) can be applied for personal information management systems through a rapid search for all the personal and organizational documents. It is then possible to improve the progress of business, accumulation of know-how, and business processes.

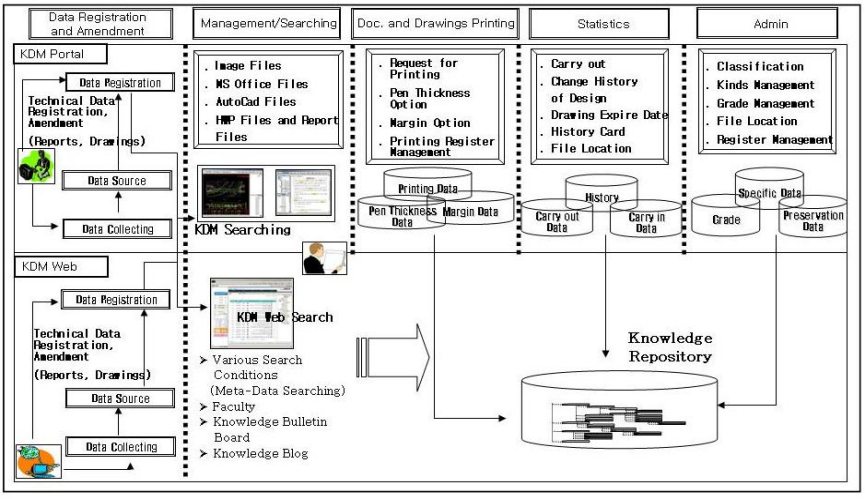


Fig. 6. Process of KDM Portal

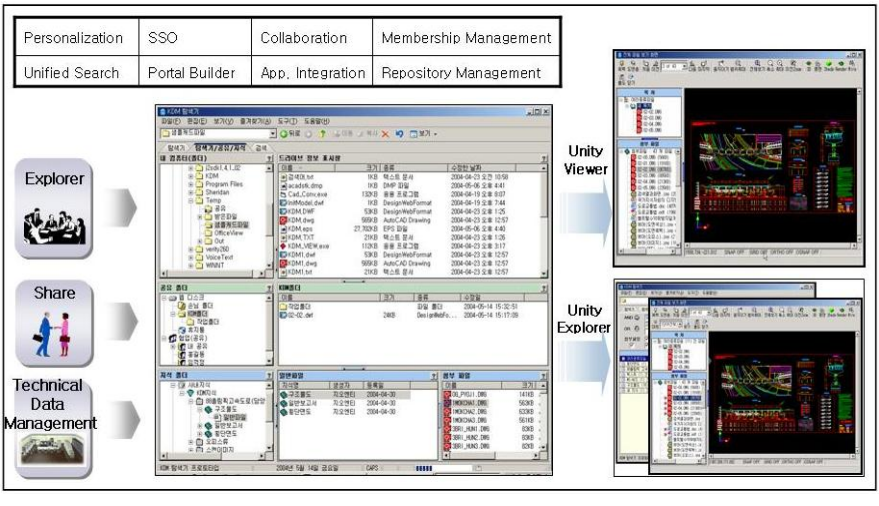


Fig. 7. Portal Function of KDM

4.1.3 Searching Function

The KDM searching function (see **Fig. 8**) integrates the scattered information in/out of the head office and provides a fast search system for business. The most important factor in the function is the co-relations with the other systems.

4.2 KDM Web

The KDM portal is designed for the information and system in a company to access at a single point, providing convenience in system extension and integration. This allows for the effective management of the information system.

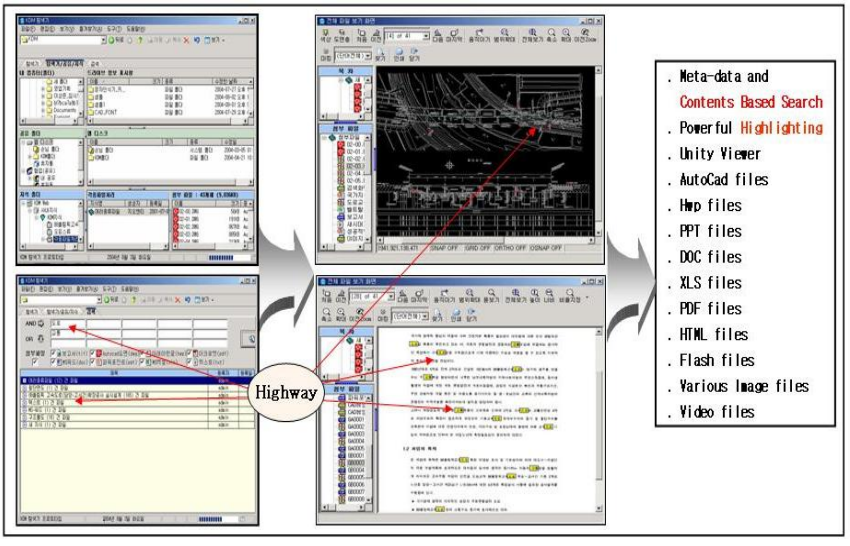


Fig. 8. Portal Searching Function of KDM

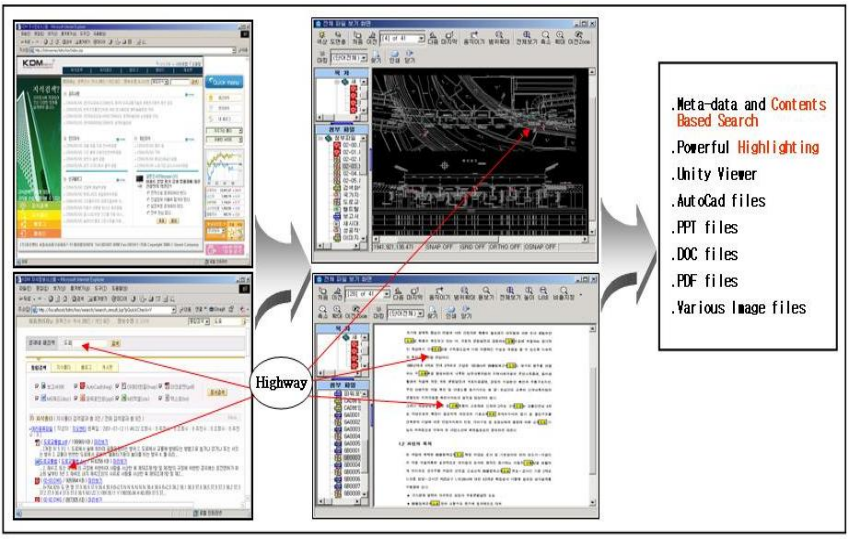


Fig. 9. Web Searching Function of KDM

The KDM Web searching function (see **Fig. 9**) enables fast searches anywhere and anytime by integrating the knowledge materials in/out of the main office. This distinguished system may improve the business process, meanwhile maximizing the use of information.

5 Conclusion

The current legacy systems have various worthwhile functions, but it should be capable of managing extensive amounts of technical information by project and have a quick-search function. This study developed a system which has aforementioned enhanced functions, as well as storing/managing technical information, searching, a systematic classification structure, managing personal information, and printing. The following is to be expected with the proposed system.

- Low-cost system for small and medium-sized construction companies,
- Development of a technical information system corresponding with the legacy system architecture,
- Improvement of business effectiveness and productivity,
- Enhancement of information technological bases and increase in knowledge assets by the accumulation of technical information and security.

This study also suggests further study tasks as follows

- Training construction experts for recognition on value-sharing by KMS, Establishment of a standardized ERD classification system for knowledge sharing throughout the entire construction industry.
- Enhancement of a accessibility in technical information by developing the KMS engine (searching/Web) and publicly releasing the engine Expanding the system by business characteristics from the standardized knowledge engine.

References

1. Lee, T.S., and Lee, D.W., Methodology for development of Knowledge Management System, Proceedings of KSCE Annual Conference, Korean Society of Civil Engineers, (2001).
2. Lee, T.S., and Lee, D.W., Survey for success and failure Factors in Knowledge Management, Proceedings of the 3rd KICEM Annual Conference, Korean Institute of Construction Engineering and Management, pp. 261-264, (2002).
3. PMnCM, Supporting Plan for IT of small and Medium-Sized Construction Companies, The Korean Federation of Construction Industry Societies, (2001).

An Event Correlation Approach Based on the Combination of IHU and Codebook*

Qiuhua Zheng and Yuntao Qian

Computational Intelligence Research Laboratory, College of Computer Science,
Zhejiang University, Hangzhou, Zhejiang Province, P.R. China
State key Laboratory of Information Security,
Institute of Software of Chinese Academy of Sciences, Beijing, P.R. China
zheng_qiuhua@163.com, ytqian@zju.edu.cn

Abstract. This paper proposes a new event correlation technique, which integrates the increment hypothesis updating (IHU) technique with the codebook approach. The technique allows multiple simultaneous independent faults to be identified when the system's codebook only includes the codes of the single fault and lacks the information of prior fault probability and the conditional probability of fault lead to symptoms occur. The method utilizes the refined IHU technique to create and update fault hypotheses that can explain these events, and ranks these hypotheses by the codebook approach. The result of event correlation is the hypothesis with maximum hamming distance to the code of the received events. Simulation shows that this approach can get a high accuracy and a fast speed of correlation even if the network has event loss and spuriousness.

1 Introduction

Event correlation, a central aspect of network fault diagnosis, is a process of analyzing received alarms to isolate possible root causes responsible for network's symptoms occurrences. Since a single fault often caused a large number of alarms in outer related resources, these alarms must be correlated to pinpoint their root causes so that problems can be handled effectively. Traditionally, event correlations have been performed with direct human involvement. However, these activities are becoming more demanding and intensive due to the heterogeneous nature and growing size of today's network. For these reasons, automated network event correlation becomes a necessity. Because failures are unavoidable in large and complex network, an effective event correlation can make network system more robust, and their operation more reliable. Now network event correlation technique has been a focus of research activity since the advent of modern communication systems, which produces a number of event correlation techniques[1-8].

Among these techniques, codebook approach[6, 9] and IHU technique[5] are the two of the most important methods. The codebook approach uses the causality graph model to represent the causal relationship between faults and symptoms, and the complete set of symptoms caused by a problem is represented by a "code" that identifies the problem. In this technique, event correlation is simply the process of

* This research is supported in part by Huawei Technologies.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 757 – 763, 2005.
© Springer-Verlag Berlin Heidelberg 2005

"decoding" the set of observed symptoms by determining which problem matches its code. The problem with minimum distance to the received events code is the optimum candidate. Owing to the inherent error correction capability, the code-based algorithm has a certain degree of tolerance to lost and corrupted event information. Many code-based systems have been shown to perform very well in terms of speed against their more conventional cousins. However, the codebook technique cannot deal with these cases in which more than one fault occurs simultaneously and generate overlapping sets of events. The IHU[4, 5] algorithm can treat with multiple problems simultaneously, and the introduction of the heuristic approach can deduce enormously the number of fault hypotheses that can explain the events occurred in the system. These IHU-based algorithms usually use belief degree to measure those fault hypotheses' likelihood. Since those belief degree are usually gained according to the fault's priori probability $P(f_i)$ and the conditional probability $p(s_j|f_i)$, those IHU-based algorithms' fault-symptom models usually are probabilistic. However, these probabilities are difficult to get in real network system. So the limitation degrades the feasibility of these IHU-based algorithms to a certain extent.

This paper proposes a new event correlation technique that integrates the IHU algorithm with the codebook approach. This method utilizes the codebook approach to encode the network's fault-symptom model. Its codebook only needs to include codes of the single problem. When faults occur in the network, the event correlation engine uses the refined IHU algorithm to create the set of fault hypotheses, and then calculate these fault hypotheses' likelihood through the codebook approach. The hypotheses with the hamming distance to the code of received events are chosen as the result of the event correlation.

The rest of this paper is organized as follows. In section 2, we introduce the codebook approach and the IHU technique in brief. In section 3, we propose our algorithm, and illustrate it in detail with some cases. In section 4, we describe how to design simulation for evaluation the proposed technique. Finally, we conclude our work and present the direction for future research in this area in section 5.

2 The Relative Works

2.1 The Codebook Techniques

The code-based techniques employ information theory to facilitate the process of event correlation. In code-based techniques, fault-symptom relationship are represented by a codebook[6, 9]. For every problem, a code is generated that distinguish one problem from other problems. In the deterministic code-based technique, a code is a sequence of values from $\{0, 1\}$. Problem codes are generated based on the available system model and a fault information model. InCharge uses a causality graph as an intermediate fault-symptom model to generate the codebook. The causality graph is pruned to remove cycles, unobservable events, and indirect symptoms so that the causality graph only contains direct cause-effect relationships between problems and symptoms. A matrix can be generated whose columns are indexed by problems and rows are indexed by symptoms. The matrix cell indexed by (s_i, p_j) contains a probability that problem $p_{i\psi}$causes symptom s_i. In the deterministic model, the probability is either 0 or 1. The correlation matrix is then optimized to minimize the number of symptoms that have to be analyzed but still ensure that the symptom patterns corresponding to different problems allow the problems to be

distinguished. The optimized correlation matrix constitutes a codebook whose columns are problem codes. The observed code is a sequence of symbols from {0,1}; 1 means the appearance of a particular symptom, and 0 means that the symptom has not been observed. The task of event correlation is to find a fault in the correlation matrix whose code is the closest match to the observed coded. For the coding phase is performed only once, the InCharge correlation algorithm is very efficient. Its computational complexity is bounded by $(k+1)log(p)$, where k is the number of errors that the decoding phase may correct, and $p$is the number of problems[10].

2.2 The IHU Technique

The IHU technique is first proposed in the literature[5]. When an event e_i is received, this algorithm creates a set of fault hypothesis FHS_i by updating FHS_{i-1} with an explanation of the event e_i. The fault hypothesis set is a set in which each hypothesis is a subset of Ψ that explains all events in E_O. The meaning that hypothesis h_k can explain event $e_i \in E_O$ is hypothesis h_k includes at least one fault which can lead to the event e_i occurrence. At the worst case, there may be $2^{|\Psi|}$ fault hypotheses can responsible for fault events E_O, so fault hypothesis set don't contain all subsets that can explain events E_O. The way for appending the explanation of the event e_i is that: to analyze each fault hypothesis h_k of FHS_{i-1}, if h_k can explain the event e_i, h_k can be appended into fault hypothesis FHS_i, otherwise, h_k need to be extended with a fault, which can explain the event e_i. Literature[5] proposed a heuristics approach which uses a function $u(f_l)$ to determinate whether fault f_l can be added into hypothesis $h_k \in FHS_{i-1}$. Fault $f_l \in F_{ei}$ can be appended into $h_k \in FHS_{i-1}$ only if the size of h_k, $|h_k|$, is smaller than $u(f_l)$, where function $u(f_l)$ is defined as the minimal size of a hypothesis in FHS_{i-1} that contains fault f_l. The usage of this heuristic comes from the following assumption. In most event correlation problems, the probability of multiple simultaneous faults is smaller than the probability of any single fault. Thus, in these hypotheses containing fault f_l, the fewest size of hypothesis is the one most likely to be the optimal event explanation.

3 The Proposed Algorithm

Our proposed algorithm uses the IHU approach to create and update possible fault hypotheses, and then compares the belief of fault hypotheses with the code-based technique. In the algorithm, the event correlation can be divided into two phases: (1) the phase for creation the set of fault hypotheses, (2) the phase for ranking fault hypotheses. In the following parts, we discuss the event correlation algorithm in detail.

3.1 The Improved IHU Technique

Just as described above, to limit the number of fault hypothesis, a heuristic approach is presented in literate[5]. However, when network very becomes large, the number of hypothesis still grows rapidly. To solve this problem, we improved this heuristic approach by adding a constraint fn_{max}, where fn_{max} is defined as the maximum faults' number occurred in the network. Since in most of problems, the number of faults is usually small, and the number of faults simultaneous beyond a certain value is also very small. This limitation is reasonable to make the size of fault hypothesis in an acceptable range even in a large scale network.

3.2 The Codebook-Based Measure Technique for Fault Hypotheses

These proposed codebook-based techniques are mainly suitable for the single fault cases. They compare the observed events with every problem's code in the codebook, and choose the fault with the minimal hamming distance to code of the received events as the result of the correlation. However, these techniques are not able to deal with the multiple faults simultaneously. If these methods want to correlate the multiple faults cases, they need encode all cases of the multiple faults. To overcome this disadvantage, we propose a new event correlation technique, which integrates the IHU technique with the codebook approach. This technique can treat with multiple faults while it does not increase the codebook's size. It is described as following.

(1) When faults occur in the network, the event correlation engine will receive many events. After receiving these events, the engine creates a set of possible fault hypotheses with the IHU technique, each of which is a complete explanation of the set of observed events.
(2) After creating fault hypotheses, the engine encodes these hypotheses according to the codebook that only include the codes of the single fault-symptom.
(3) The engine compares observed events with these codes of fault hypotheses built in step (2), and chooses these fault hypotheses with the minimal hamming distance to the observed events as the result. In our algorithm, the hamming distance between two number is calculated as the following: $m \oplus n = 0$, $n \oplus 0 = n$, $0 \oplus n = n$, $1 \oplus 0 = 1$, $0 \oplus 1 = 1$, $0 \oplus 0 = 0$, where $m, n > 0$.

The algorithm is defined by the following pseudo-code.

Algorithm 1(The Refine IHU & The Codebook ECA)

load codebook CB_{PN*SN}, *let* $FHS_0 = \{\phi\}$, set $fn_{\max}$.
for every observed events e_i
    compute F_{ei}
    let $FHS_i = \{\phi\}$
    for all $f_l \in F_{ei}$
        let $u(f_l) = fn_{\max}$
    for all $h_j \in FHS_{i-1}$
        for all $f_k \in h_j$ *such that* $f_k \in F_{ei}$
            set $u(f_k) = min(u(f_k), |h_j|)$, *add* h_j *to* FHS_i
    for all $h_j \in FHS_i \backslash FHS_{i-1}$
        for all $f_k \in F_{ei}$ *such that* $u(f_k) > |h_j|$
            add $h_j \cup \{f_k\}$ *to* FHS_{i-1}
    set $minHD = SN$
    for all $h_j \in FHS_i$
        set $C_{h_j} = \underbrace{0 \ldots 0}_{S_N}$
        for all $f_k \in h_j$
            $C_{hj} = C_{hj} + s_{fk}$

HD_j=*HamDis*(E,C_{hj})
if HD_j<*minHD*
    minHD = HD_j, $BFHS_i$={ Φ }, *add* h_j *to* $BFHS_i$
else if HD_j==*minHD*
    add h_j *to* $BFHS_i$
$BFHS_i$ *are the best candidates responsible for the received events.*

4 Simulation Study

In this section, we describe the simulation study performed to evaluate the technique presented in this paper. In real network, it is common that the fault events may be loss or spurious. Many event correlation algorithms do not distinguish the two cases. However, event loss is different from event spurious. In this paper, we consider that: (a) Event loss is that the event engine does not receive these events resulting from those faults occurred in the network. (b) Event spurious is that the event engine receives some events that do not happen in the network. In simulation, we use *LR* to represent the event loss ratio, *SR* to represent the event spurious ratio, *CB* to represent the codebook, *FNP* to denote the fault number distribution, and *FP* to denote the distribution of fault occurrence. Since it is impossible to exhaust all codebook of a given size, we only test limited codebooks in each size.

Given parameters of LR_l, LR_h, SR_l, SR_h and *FNP*, for given parameters of the number *FN* of faults and the number *SN* of symptoms of codebook size, we design *K* simulation cases as follows: (1)Randomly create the codebook $CB_i(1 \leqslant i \leqslant K)$; (2) Randomly generate prior fault probability distribution FP_i, which is a uniformly distributed and their sum is 1. (3)Randomly generate the $LR_j^i(1 \leqslant j \leqslant SN)$ for every event in the codebook which accords with the uniform distribution [LR_l, LR_h]. (4) Randomly generate the $SR_j^i(1 \leqslant j \leqslant SN)$ for every event in the codebook which accords with the uniform distribution [SR_l, SR_h].

For *i*-th simulation case ($1 \leqslant i \leqslant K$), we create *M* simulation scenarios as follows. (1) Randomly generate the set F_i^k ($1 \leqslant k \leqslant M$) of faults according to the fault distribution FP_i and fault number distribution *FNP*. (2) Generate the code C_{Fi} of F_i^k ($1 \leqslant k \leqslant M$) according to the codebook CB_i. (3) Randomly generate the loss events with LR_i and the spurious events with SR_i, then generate the observed events E_i of by adding the noisy events to C_{Fi}. (4) Correlate the observed events $E_i^k(1 \leqslant k \leqslant M)$ and gain the result of the correlation using the proposed algorithm. (5) Calculate the detection rate $(DR_i^k)(1 \leqslant k \leqslant M)$ and the false detection rate (FDR_i^k) $(1 \leqslant k \leqslant M)$ with the following equations.

$$DR_i^k = \left|F_{iD}^k \cap F_{iC}^k\right| / \left|F_{iC}^k\right|, FDR_i^k = \left|F_{iD}^k \setminus F_{iC}^k\right| / \left|F_{iD}^k\right|$$

We calculate the $DR_i = 1/M \sum_{k=1}^{M} DR_i^k$ and $FDR_i = 1/M \sum_{k=1}^{M} FDR_i^k$. Then, we calculate the expected values of detection rate and false detection rate denoted by $DR_{FN,SN}$, and $FDR_{FN,SN}$, respectively. In simulation, we set fn_{max}=7, and the maximum number of faults is 4. We varied *FN* from 20 to 50, and fixed *SN* at 20. The parameter *K* is 100, and *M* is 5000. The results of the simulation is shown as follows:

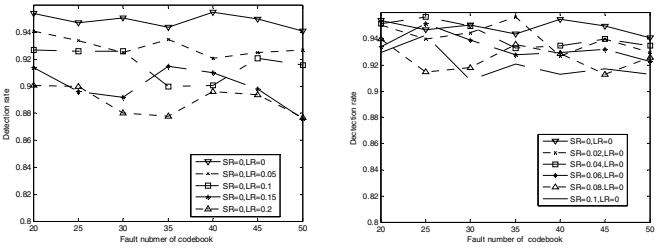


Fig. 1. The impact of LR and SR on detection rate

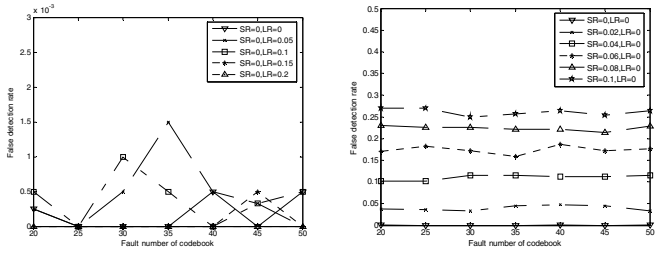


Fig. 2. The impact of LR and SR on false detection rate

From fig.1, we observe that there is no statistically difference in the detection rate on the fault number of codebook, event loss ratio *LR* and event spurious ratio *SR*. However, when the *LR* or *SR* is high, the large codebooks trend to get a worse accuracy than those codebooks with smaller size. Fig.2 presents that the event spurious ratio *SR* has a strong impact on the false detection rate and the event loss ratio *LR* has weak effect on the false detection. The reason is that: (a) The proposed algorithm deals with the spurious event by adding an explain of the event to the fault hypotheses, which does not decrease the detection rate but increase the false detection; (b) Compared with the spurious event cases, our algorithm does not treat with the loss of events in the fault hypotheses creating and updating phase.

5 Conclusion

In this paper, we propose a new event correlation technique, which combines the IHU approach with the Codebook method. The technique can perform the diagnosis of multiple simultaneous independent faults when the codebook only includes the codes of the single fault. The result of experiment shows that our proposed algorithm has a satisfied result on the accuracy of detection and time of correlation. However, when the codebook's size becomes very large but its radius remains small, the accuracy of the proposed algorithm is not satisfied. How to select the codebook and to propose a better distance to distinguish between problems are two problems that need further research. In the near future, we will focus on the two aspects and plan to do more simulations on large size codebooks.

References

1. A. T. Bouloutas, Modeling Fault Management In Communication Networks, in Graduate School of Arts and Sciences, PHD. Columbia University, (1990).
2. A. T. Bouloutas, S. Calo, and A.Finkel, Alarm Correlation and Fault Identification in Communication Networks, IEEE Trans. on Communications, vol. 42(1994) 523 - 533.
3. I. Katzela and M. Schwartz, Schemes for Fault Identification in Communication Networks, IEEE/ACM Trans. on Networking, vol. 3(1995) 753-764.
4. M. Steinder and A. S. Sethi, Increasing Robustness of Fault Localization through Analysis of Lost, Spurious, and Positive Symptoms, presented at INFOCOM2002, New York, (2002).
5. M. Steinder and A. S. Sethi, Probabilistic Fault Diagnosis in Communication Systems through Incremental Hypothesis Updating, Computer Networks, vol. 45(2004) 537-562.
6. S. A. Yemini, S. Kliger, E. Mozes, Y. Yemini, and D. Ohsie. High Speed and Robust Event Correlation, IEEE Communications Magazine, vol. 34 (1996) 82-90.
7. G. Liu, A. K. Mok, and E. J. Yang, Composite events for network event correlation, IM99, (1999) 247-260.
8. Lewis.L, A Case-Based Reasoning Approach to the Management of Faults in Communication Networks, presented at INFOCOM'93, San Francisco, (1993).
9. S. Kliger, S. Yemini, Y. Yemini, D. Ohsie, and S. Stolfo, A coding approach to event correlation, presented at IM95, Santa Barbara, CA, (1995).
10. M. Steinder and A. S. Sethi, A survey of fault localization techniques in computer networks, Science of Computer Programming, vol. 53 (2004) 165-194.

Face Recognition Based on Support Vector Machine Fusion and Wavelet Transform

Bicheng Li[1,2] and Hujun Yin[1,*]

[1] School of Electrical and Electronic Engineering, The University of Manchester, Manchester, M60 1QD, UK
lbclm@263.net
[2] Information Engineering Institute, Information Engineering University, Zhengzhou 450002, China
h.yin@manchester.ac.uk

Abstract. Recently, wavelet transform and information fusion have been used in face recognition to improve the performance. In this paper, we propose a new face recognition method based on wavelet transform and support vector machine-based fusion scheme. Firstly, an image is decomposed with wavelet transform to three levels. Then, Fisherface method is applied to three low-frequency sub-images respectively. Finally, the individual classifiers are fused using the support vector machines. Experimental results show that the proposed method outperforms the best individual classifiers and the direct Fisherface method on original images.

1 Introduction

Within the last decade, face recognition (FR) has become one of the most challenging areas in computer vision, pattern recognition and biometrics. The popular appearance-based approaches are the eigenface method [1] and Fisherface method [2]. The eigenface method is built on the principal component analysis (PCA). The Fisher linear discriminant analysis (FLD or LDA) maximises the ratio of the determinant of between-class scatter matrix and within-class scatter matrix. The Fisherface method is used in conjunction with PCA and LDA.

Recently, wavelet transform has been widely used in image processing and face recognition [3-5]. Lai [4] and Chien [5] exploited an approximation image in wavelet transform. Kwak [3] applied the Fisherface method to four subimages of the same resolution respectively, and then fused the individual classifiers with the fuzzy integral [6]. High-frequency subimages usually provide little valuable information for recognition and give poor performances, while low-frequency subimages can achieve good performance. Furthermore, different resolution images contain auxiliary information. In this paper, we propose a fusion algorithm based on the support vector machines (SVMs) [7-8]. Firstly, an image is decomposed with wavelet transform (WT) to three levels. Secondly, the Fisherface method is applied to these low-frequency subimages separately. Finally, the individual classifiers are fused with the SVMs.

* Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 764 – 771, 2005.
© Springer-Verlag Berlin Heidelberg 2005

To test the algorithm we used the FERET face database [9]. Experimental results show that the proposed method outperforms both the individual classifiers and the Fisherface method. Section 2, 3 and 4 give brief overviews on the SVM, wavelet transform and the Fisherface method. The proposed face recognition method is presented in Section 5, followed by the results and conclusions in Section 6.

2 Support Vector Machine

The SVM [7], a powerful tool for classification developed by Vapnik, is based on the Structural Risk Minimization (SRM) principle. It has good performance in tackling small sample size, high dimension and good generalization.

SVM was originally proposed for linearly separable two-class problems. Consider the training sample $\{(X_i, d_i)\}_{i=1}^{N}$, where X_i is the input pattern for the *i*th example and $d_i \in \{+1, -1\}$ is the corresponding desired response (target output). The equation of a decision surface in the form of a hyperplane that produces the separation is

$$W^T X + b = 0 \tag{1}$$

where X is an input vector, W is an adjustable weight vector, and b is a bias. We may thus write

$$W^T X_i + b \geq 0 \qquad \text{for } d_i = +1 \tag{2}$$

$$W^T X_i + b < 0 \qquad \text{for } d_i = -1 \tag{3}$$

Given the training sample $\{(X_i, d_i)\}_{i=1}^{N}$, the SVM tries to find the parameters W_o and b_o for the optimal hyperplane that maximizes the margin of separation that is the separation between the hyperplane defined in Eq. (1) and the closest data point in the training sample, denoted by ρ. Furthermore, the pair (W_o, b_o) must satisfy the constraints:

$$W_o^T X_i + b_o \geq +1 \qquad \text{for } d_i = +1 \tag{4}$$

$$W_o^T X_i + b_o \leq -1 \qquad \text{for } d_i = -1 \tag{5}$$

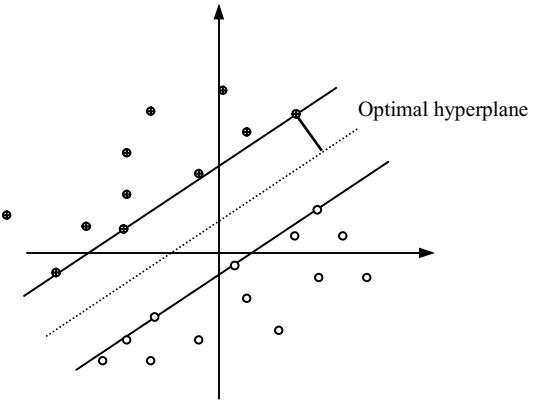


Fig. 1. Illustration of an optimal hyperplane for linearly separable patterns

The idea of an optimal hyperplane for linearly separable patterns is shown in Fig. 1. The particular data points (X_i, d_i) for which the line of Eq. (4) or Eq. (5) is satisfied with the equality sign are called support vectors (SV). For nonlinearly separable problem, kernel machines can be applied to map the date onto linearly separable feature space.

SVM is formed as solving the quadratic programming problem:

Given the training sample $\{(X_i, d_i)\}_{i=1}^N$, find the Lagrange multipliers $\{\alpha_i\}_{i=1}^N$ that maximize the objective function

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(X_i, X_j), \tag{6}$$

subject to the constraints

$$\sum_{i=1}^{N} \alpha_i d_i = 0, \tag{7}$$

$$0 \le \alpha_i \le c, \tag{8}$$

where c is a user-specific positive parameter, and the function $K(\bullet,\bullet)$ is called kernel function. The common used kernel functions are given as following

Radial-basis function: $$K(X, X_i) = \exp\left(-\frac{1}{2\sigma^2} \left\| X - X_i \right\|^2 \right) \tag{9}$$

Polynomial function: $$K(X, X_i) = \left(X^T X_i + 1 \right)^p \tag{10}$$

Multi-layer perceptron: $$K(X, X_i) = \tanh\left(\beta_0 X^T X_i + \beta_1 \right) \tag{11}$$

where power p and width σ^2 are specified priories by the user, and Mercer's theorem [7] is satisfied only for some values of β_0 and β_1.For an input X, the decision rule is

$$f(X) = \operatorname{sgn}(\sum_{i=1}^{N} d_i \alpha_i K(X, X_i) + b) \tag{12}$$

For multiclass problems, there are two basic strategies (where n is the number of class):

(1) The one-against-all approach, where the n SVMs are trained and each SVM separates one class from the rest.

(2) The pairwise approach, where the $n(n-1)/2$ SVMs are trained, and each SVM separates a pair of classes.

3 Wavelet Decomposition

Wavelet transform (WT) [10-12] is a multiresolution analysis tool, and it decomposes a signal into different frequency bands. We use two-dimensional (2-D) WT for decomposing images, generating an approximate subimage and three high frequency subimages via the high-pass and low-pass filtering realised with respect to the column vectors and the row vectors of array pixels. Fig. 2 shows the structure of the filter bank that performs the decomposition at level 1. Here, H and L represent the high-

pass and low-pass filter respectively. In this way we can obtain four subimages (LL1, LH1, HL1, and HH1). In this manner, we can repeatedly decompose the approximation subimages obtained, such as two-level (LL2, LH2, HL2, and HH2), three-level (LL3, LH3, HL3, and HH3), and so forth.

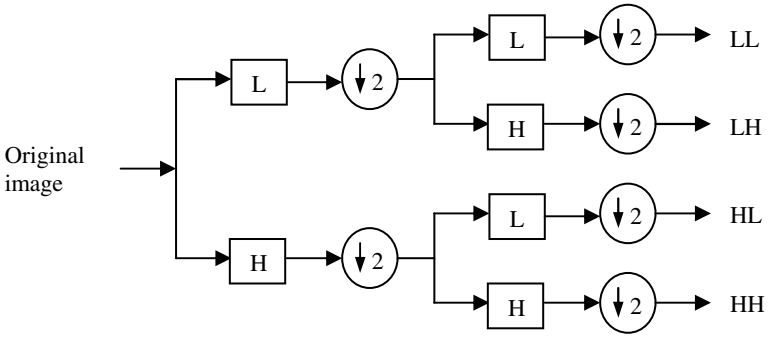


Fig. 2. Structure of 2-D wavelet decomposition occurring at level 1

4 Fisher Method

The Fisherface method [2] is a combination of PCA and FLD, and consists of two stages. First, PCA is used to project face pattern from a high-dimensional image space into a lower-dimensional space. Then, FLD is applied to the feature-vectors.

Suppose a face image is a 2-D $p \times q$ array. An image Z_i is considered as a column vector of dimension $M = p \times q$. A training set consisting of N face images is denoted by $Z = \{Z_1, Z_2, \cdots, Z_N\}$. The covariance matrix is defined as, $R = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \bar{Z})(Z_i - \bar{Z})^T$, where $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$. Let $E = (E_1, E_2, \cdots, E_r)$ consists of the r eigenvectors corresponding to the r largest eigenvalues of R. We project the original face images into a PCA-transformed subspace, and obtain their corresponding reduced feature vectors $Y = \{Y_1, Y_2, \cdots, Y_N\}$ as follows:

$$Y_i = E^T Z_i . \tag{13}$$

In the second stage, FLD is applied to feature vectors, $Y = \{Y_1, Y_2, \cdots, Y_N\}$. Suppose the training set of N face images consists of C classes. The between-class scatter matrix is defined by

$$S_B = \sum_{i=1}^{C} N_i (\bar{Y}^{(i)} - \bar{Y})(\bar{Y}^{(i)} - \bar{Y})^T . \tag{14}$$

where N_i is the number of samples in ith class C_i, $\bar{Y}$ is the mean of all samples, and $\bar{Y}^{(i)}$ is the mean of ith class C_i. The within-class scatter matrix is defined as,

$$S_W = \sum_{i=1}^{C} \sum_{Y_k \in C_i} (Y_k - \bar{Y}^{(i)})(Y_k - \bar{Y}^{(i)})^T . \tag{15}$$

The optimal projection matrix is chosen as a matrix with orthonormal columns that minimizes the determinant of the within-class scatter matrix while maximizing the between-class scatter matrix, i.e.

$$W_{FLD} = \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} = [W_1, W_2, \cdots, W_{C-1}]. \tag{16}$$

where $\{W_1, W_2, \cdots, W_{C-1}\}$ is the set of generalised eigenvectors of S_B and S_W corresponding to the $C-1$ largest generalised eigenvalues $\{\eta_1, \eta_2, \cdots, \eta_{C-1}\}$, that is

$$S_B W_i = \eta_i S_W W_i, \quad i = 1, 2, \cdots, C-1. \tag{17}$$

Because S_B is the sum of C matrices of rank one or less, the rank of S_B is $C-1$ or less. The feature vectors for any face images $V = (V_1, V_2, \cdots, V_N)$ can be obtained by projecting Y onto FLD-transformed space:

$$V_i = W_{FLD}^T Y_i = W_{FLD}^T E^T Z_i, \qquad i = 1, 2, \cdots, N. \tag{18}$$

For a new face image, the classification is realised in FLD-transformed space; that is, the image is transformed into its feature vector using Eq. (18), and then is compared with the feature vectors of the training images.

In the Fisherface method, the upper bound on the number of eigenvectors r selected is $N-C$, which corresponds to the rank of the within-class scatter matrix S_B. To determine r, the eigenvalues of the covariance matrix R, λ_i, $i = 1, 2, \cdots, M$, are ranked in non-increasing order. We choose r ($\leq N-C$) as the minimum of m such that

$$\sum_{i=1}^{m} \lambda_i \Big/ \sum_{i=1}^{M} \lambda_i \geq P. \tag{19}$$

where P is a given constant.

5 Face Recognition Using Wavelet Transform and SVM Fusion

Suppose the training set of N face images $Z = \{Z_1, Z_2, \cdots, Z_N\}$ consists of C classes. The proposed face recognition algorithm consists of three stages.

Firstly, an image is decomposed with WT of 3 levels, obtaining 3 approximation images LL1, LL2, and LL3. Hence we obtain three subsets of images. In this paper, we use the Daubechies family [11] wavelet filters db*N*, where *N* stands for the order of the wavelet. The choice of 3 levels in WT is determined experimentally, and it is a compromise between recognition performance and computational complexity. More levels results in slightly better results, while require more computation.

Secondly, the Fisherface method based on PCA and FLD is applied to the decomposed three subsets respectively, obtaining three classifiers: classifier1, classifier2, and classifier3. In PCA, we choose the number of eigenvectors using Eq. (10) with $P = 95\%$, instead of the commonly used $N-C$ [3]. We obtain the feature vectors of the training images and a given test image using Eq. (18) and calculate Euclidean distances between the feature vector produced from a given test image and those from the training image set. These Euclidean distances are transformed into membership grades as follows:

$$\eta_{ij} = \frac{1}{1+(d_{ij}/\bar{d}_i)},\ i=1,2,3,\ j=1,2,\cdots,N. \tag{20}$$

where i is the index of the classifiers and j is the index of the training face image. $\bar{d}_i$ denotes the average distance between all distance values in ith classifier, d_{ij} is the Euclidean distance between the feature vector of a given test image and that of jth training image in ith classifier. Furthermore, the membership in which a given test image belongs to the kth class in ith classifier is computed by

$$\mu_{ik} = \frac{1}{N_k}\sum_{X_j \in C_k} \eta_{ij},\ i=1,2,3,\ k=1,2,\cdots,C. \tag{21}$$

where C_k is the set of samples in kth class, N_k is the number of samples in C_k. For ith classifier, if the memberships of a given test image satisfy:

$$\mu_{ik_0} = \max_{k=1}^{C} \mu_{ik} \tag{22}$$

Then the given test image is recognised to belong to k_0 th class. Eq.(22) is called rule of the largest membership.

Finally, the individual classifiers are fused by the SVMs. For each image, C memberships μ_{ik} ($k=1,2,\cdots,C$) obtained from each classifier i ($i=1,2,3$) using Eq.(21) constitute a membership vector. The three membership vectors derived from three classifiers are concatenated into a feature vector of size $3C$. The SVMs are designed to use these feature vectors as the inputs and to fuse individual classifiers to generate improved classification results. The output is the label number of the class of the input image.

6 Experimental Results and Conclusions

To test our method, we selected 40 subjects from the "b" series of the FERET database [9], and the indexes of the subjects are from 1141 to1180. For every subject, we select 7 images whose letter codes are ba, bj, bk, bd, be, bf, and bg. They are eight-bit greyscale images of human heads with views ranging from frontal to left and right profiles (pose angle $\le 25^0$) and different illuminations. The original images of size 384×256 are reduced to 192×128. Some sample images are shown in Fig. 3.

In our experiments, 6 images of each subject are chosen randomly as the training set and the rest for testing in each run. The recognition process is executed for 100 runs, and the mean and standard deviation (Std.) of the performance are calculated.

We used db4 [11] to decompose images into 3 levels with boundary effect being considered. Each approximation subimage of different levels, i.e. LL1, LL2, and LL3, is extracted. The sizes of LL1, LL2, and LL3 are 99×67, 53×37 and 30×22, respectively. The Fisherface method with P=95% is applied to the three approximation subimages. Table1 lists the results of Fisherface method with different classification rules, i.e. largest membership (LM), nearest mean (NM), and SVM, for original image, LL1, LL2, and LL3. The result of the proposed method is also given, where polynomial function is applied and the pairwise approach for multiclass problem is used.

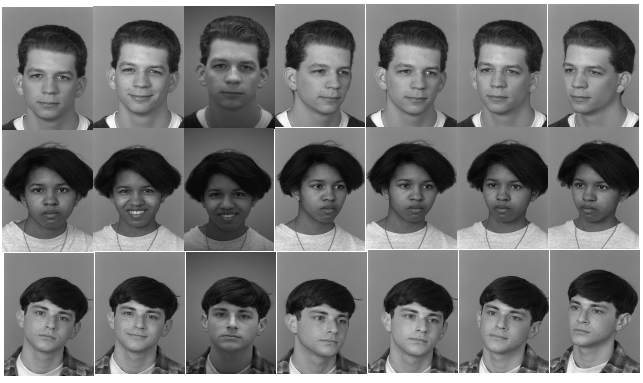


Fig. 3. Samples of three subjects

Table 1. Results (%)of Fisherface method with different classification rules for original iamge, LL1, LL2, and LL3, and the proposed method

Classification rules	Statistics	Original image	LL1	LL2	LL3
LM	Mean	94.90	95.93	96.78	96.10
	Std	3.08	2.99	2.35	2.65
NM	Mean	94.80	96.03	97.08	96.90
	Std	2.99	3.00	2.35	2.67
SVM	Mean	94.63	96.53	97.58	97.55
	Std	3.15	2.67	2.11	2.15
Proposed SVM-based fusion method	Mean	98.42			
	Std	1.68			

Experimental results show that the proposed method outperforms the Fisherface method with different classification rules, including largest membership (LM), nearest mean (NM), and SVM, for original image, LL1, LL2, and LL3. The accuracy rate of the proposed method is 98.42 %, while the highest accuracy rate of Fisherface method for original image, LL1, LL2, and LL3 are 94.90%, 96.53%, 97.58%, and 97.55%, respectively. Usually a fusion method such as fuzzy integrals and radial basis functions can boost the performance by 1-2%. In this case, a simple SVM-based fusion scheme is able to improve the performance by significant 2-4%, indicating the SVM can be used effectively as a fusion tool as well.

References

1. Turk, M.A. and Pentland, A.P.: Eigenfaces for Recognition. *J. Cogn. Neurosci.*, vol.3, 1 (1991) 71–86
2. Belhumeur, P.N., Hespanha, J.P., and Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, (1997) 711–720

3. Kwak, K.-C., Pedrycz, W. : Face Recognition Using Fuzzy Integral and Wavelet Decomposition Method. *IEEE Trans System, Man, and Cybernetics -B*, Vol. 34, (2004) 1666–1675
4. Lai, J.H., Yuen, P.C., and Feng, G.C. : Face recognition using holistic Fourier invariant features. *Pattern Recognit.,* vol. 34, (2002) 95–109
5. Chien, J.T. and Wu, C.C.: Discriminant wavelet faces and nearest feature classifiers for face recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 24, (2002) 1644–1649
6. Sugeno, M.: Fuzzy measures and fuzzy integrals—A survey. *in Fuzzy Automata and Decision Processes,* Gupta, M.M., Saridis, G. N. and Gaines, B.R., Eds. Amsterdam, The Netherlands: North Holland, (1977) 89–102
7. Vapnik, V. : Statistical Learning Theory. New York Wiley, 1998
8. Haykin, S. : Neural Networks: A Comprehensive Foundation. Prentice Hall International, Inc. New Jerey, 1999
9. FERET face database, < http://www.itl.nist.gov/iad/humanid/feret/>
10. Mallat, S.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.11, 7 (1989) 674–693
11. Daubechies, I.: Orthonormal Bases of Compactly Supported Wavelets. *Commun Pure Appl Math*, 41 (1988) 909–996
12. Cohen, A., Daubechies, I. and Feauveau, J. : Biorthogonal Bases of Compactly Supported Wavelets. *Commun. Pure Appl. Math*., Vol.XLV, (1992) 485–560
13. Dai, G., Zhou, C. : Face Recognition Using Support Vector Machines with the Robust Feature. *Proceedings of the 2003 IEEE International Workshop on Robot and Human interactive Communication*, pp:49-53

A Dynamic Face and Fingerprint Fusion System for Identity Authentication

Jun Zhou, Guangda Su, Yafeng Deng, Kai Meng, and Congcong Li

Electronic and Engineering Department of Tsinghua University, Beijing 100084, China
zhoujun@mails.tsinghua.edu.cn

Abstract. This paper presents a novel dynamic face and fingerprint fusion system for identity authentication. To solve the face pose problem in dynamic authentication system, multi-route detection and parallel processing technology are used in this system. A multimodal part face recognition method based on principal component analysis (MMP-PCA) algorithm is adopted to perform the face recognition task. Fusion of face and fingerprint by SVM (Support Vector Machine) fusion strategy which introduced a new normalization method improved the accuracy of identity authentication system. Furthermore, key techniques such as fast and robust face detection algorithm and dynamic fingerprint detection and recognition method based on gray-Level histogram statistic are accepted to guarantee the fast and normal running. Practical results on real database proved that this authentication system can achieve better results compared with face-only or fingerprint-only system. Consequently, this system indeed increases the performance and robustness of identity authentication systems and has more practicability.

1 Introduction

With the development of the information society, personal identity authentication has gained more and more attention. Traditional personal identification approaches based on passwords or ID cards are not sufficiently reliable to satisfy the security requirement because the systems are not able to make the difference between a client and the impostor who fraudulently acquires the password or the ID card. Biometrics has been suggested and investigated by many researchers to recognize users by characteristics which are difficult to impose and biometric characteristics inherently have the capability to distinguish between the client and the impostor by means of physiological features (face, fingerprints, iris, voice, etc.) **[1]**.

However, almost no single biometric can meet the all performance required in practical systems. Face recognition is superior in speed but has poor validity while the fingerprint authentication is reliable but inefficient in database retrieval **[2]**. In order to build an efficiently practical identity authentication, an integration method which combines two or more biometric was proposed. Reference **[2-4]** presents fusion face and fingerprint or speech or iris to build multi-biometric identification or authentication system, and some significant results had been achieved. But these researches paid more attentions to fusion methods or algorithms, and little to actual application identity authentication system.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 772 – 779, 2005.
© Springer-Verlag Berlin Heidelberg 2005

So this paper has introduced a new dynamic face and fingerprint fusion system for identity authentication. This system first processes image acquisition via four parallel threads including three CCD cameras and a fingerprint sensor. In order to increase the robust of dynamic face image acquisition, the three CCD cameras are placed in different angles with respect to the frontal plane to process dynamic face detection respectively which is mainly based on the fast and robust face detection algorithm and get a number of face images with different angles. After image preprocessing, the system chooses the optimal face image which has the minimum angle and uses the MMP-PCA algorithm to implement the face recognition. At the same time, the fingerprint sensor thread checks the fingerprints dynamically and saves the usable one as the fingerprint stored data file to process fingerprint recognition. Finally, the system integrates the two different biometric recognition results (face and fingerprint) and uses the SVM fusion strategy according to self-learning results to improve the authentication performance.

The rest of the paper is organized as follows: Section 2 addresses the system structure. Section 3 presents four key techniques being employed in this system and all the techniques details are depicted. Experiment results on TH face and fingerprint database are described in section 4. Finally, the conclusion is given in section 5.

2 The System Structure

The overall system includes two main modules- preprocessing module and authentication module. Figure 1 shows the architecture of the system.

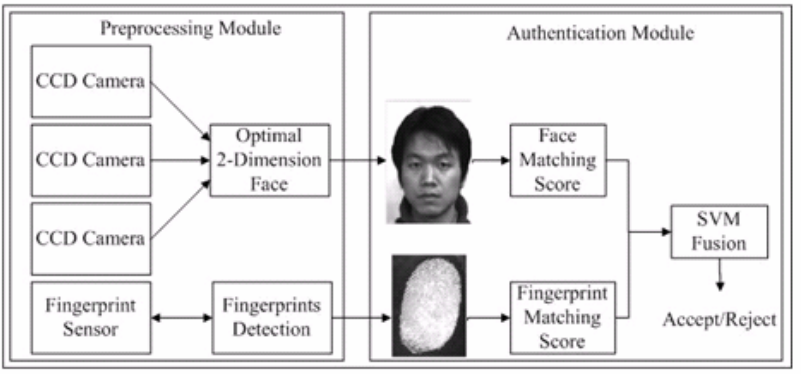


Fig. 1. Architecture of the system

The preprocessing module processes image acquisition via four parallel threads including three CCD cameras and a fingerprint sensor and chooses the optimal face image with minimum angle and a usable fingerprint (see section 3.3) as the authentication files. The authentication module integrates the two different biometric recognition results (face and fingerprint) based on the authentication object from the system database and uses the SVM fusion strategy according to self-learning results to give the final authentication decision.

The whole system processing procedure includes five steps, and the proposed procedures are list as follows:

Step1: Image detection: The system uses the fast and robust face detection algorithm (depicted in section 3.1) with a face tracker to detect the user's face image and the Gray-Level Histogram Statistic method (depicted in section 3.3) to detect fingerprint image automatically at the same time.

Step2: Image acquisition: The dynamic face images are detected and collected from three different directions via multi-route detection in a period of limited time and the all face images form a face image queue. At the same time the fingerprint thread check these fingerprint images and save a useable one as the data file.

Step3: Image preprocessing: The face pose information will be estimated via eyes and nose location, eigenspace analysis and fuzzy clustering **[5]**. The optimal face image with minimum angle from the face image queue will be stored as the data file.

Step4: Image recognition: After reading the stored data files, the face matching threads uses the MMP-PCA algorithm (depicted in section 3.2) to do recognition process. At the same time, the fingerprint threads also do the matching process based on comparing the local feature and global feature minutia with the fingerprint template (depicted in section 3.3).

Step5: Identity authentication: After getting the two matching score, the SVM fusion strategy (see section 3.4) will receive a sequence of strokes and give the final fusion authentication decision according to the self-learning results.

3 Key Techniques

There are four key techniques referred in this system, which have made this system achieve more accuracy and more practicability. The fast and robust face detection algorithm, MMP-PCA face recognition algorithm, dynamic fingerprint detection and recognition algorithm and the SVM fusion strategy are all described in this section.

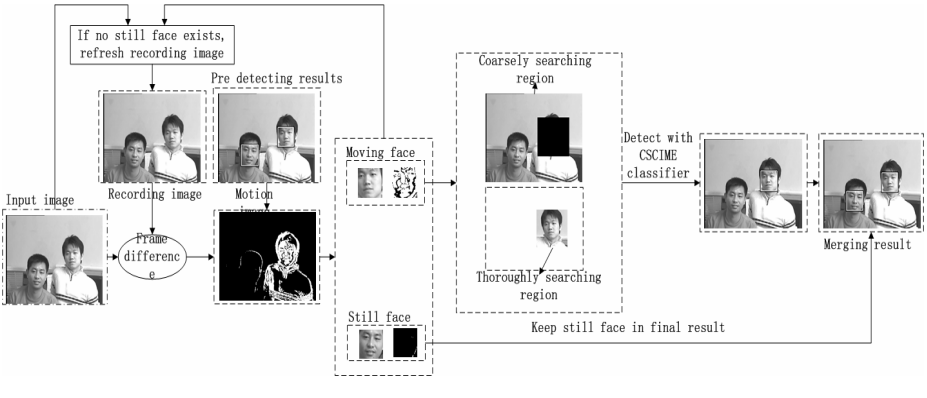


Fig. 2. Procedure of processing face presenting state by dynamic detector

3.1 Fast and Robust Face Detection Algorithm

In our authentication system, a fast and robust face detection algorithm is adopted. By dividing state of face detection system into three and processing input images according to current state with different modes, we treat with special situations such as still face existence, continuously failing to detect face and input without evident change to get a robust face detection system suitable for real application. The procedures of dynamic face detection are shown in Figure 2.

This algorithm is very robust in real situations, and in this authentication system, running on a P4 2.53 GHz PC, with a face presenting, the three-route face detection speed is about 40~80 ms, costing 50~80% of the CPU computational resource **[6]**.

3.2 MMP-PCA Face Recognition Algorithm

MMP-PCA is an improved multimodal part face recognition method based on PCA (principal component analysis) algorithm. We first detach the face parts and divide the face into five parts: bare face, eyebrow, eye, nose and mouth. The positions of these facial parts are all located in Figure 3. Then principal component analysis (PCA) is performed on the five facial parts. We calculate the eigenvalues of each facial part, choose d largest eigenvalues (we use $d=60$) and preserve corresponding eigenvectors, then the eigenvector of the facial part can be calculated.

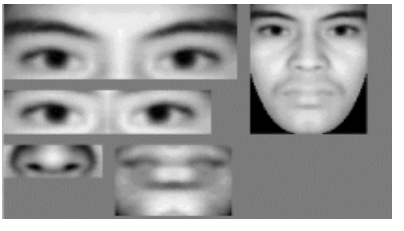


Fig. 3. Five different facial parts

In face recognition, we first calculate the projection eigenvalue of the authentication object from system database and then use the formula 1 to compute the similitude degree between the human face to be recognized and the authentication object human face. In formula 1, $\vec{X}$ means the projection eigenvalue of the human face and $\vec{Y}$ means the projection eigenvalue of the authentication object human face from database.

$$S(\vec{X},\vec{Y}) = \sqrt{1-\frac{\left\|\vec{X}-\vec{Y}\right\|}{\left\|\vec{X}\right\|+\left\|\vec{Y}\right\|}} \tag{1}$$

Using a combination of facial parts to recognize, there are K kinds of combinations:

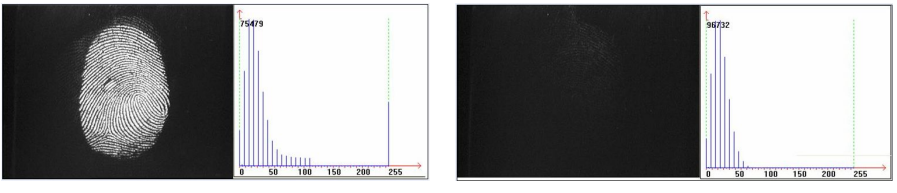
$$K = C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5 = 31 \tag{2}$$

In all these 31 kinds of combinations, the recognition pattern which uses all facial parts is called global recognition pattern (GRP) and the other recognitions are called

local recognition pattern (LRP). When using the GRP, the projection eigenvalues of bare face, eyebrow, eye, nose and mouth of a human face are weighted by *5:6:4:3:2*. In this authentication system, the GRP which performs best in these combinations **[7]** is chosen to process the face recognition part.

3.3 Dynamic Fingerprint Detection and Recognition

In the dynamic fingerprint detection, we use the Gray-Level Histogram Statistic method to detect fingerprint image automatically. Fingerprint images from the fingerprint sensor have two states: fingerprint exists or no fingerprint. We first use the Sobel Edge Detector to reduce the image noise and then do the gray-level histogram. See Figure 4, the highest gray-level (**255**) of the fingerprint image is a large value (15000~20000) and the highest gray-level (**255**) of the image with no fingerprint usually is a very small one (0~200). So we can easily choose the fingerprint image based on the results.



(a) Fingerprint and gray-level histogram (b) No fingerprint and gray-level histogram

Fig. 4. Samples in the dynamic fingerprint detection

In fingerprint recognition, see Figure 5, we use an automatic algorithm to locate the core point and extracted the local structure (direction, position relationship with the neighbor minutiaes) and global structure (position in whole fingerprint) of all the minutiaes **[8].** The matching algorithm uses local and also global structures of every minutiae.

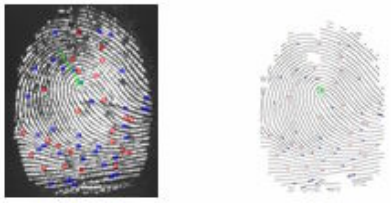


Fig. 5. (a) Fingerprint and it's minutiae points (b) the ridge

3.4 SVM Fusion Strategy

Support Vector Machine (SVM) is an optimal two-class fusion strategy based on minimum structural risk. The goal of the SVM is to find the separating hyper-plane with maximum distance to the closest points of the training set and minimum error.

Therefore, the SVM has more advantages in popularization than some fusion strategies which have the local minimum **[9]**. The classification can be performed as follows:

$$g(x)=\sum_{i=1}^{n_s}\lambda_i y_i K(x_i,x)+w_0 \begin{cases} > \\ < \end{cases} 0 \rightarrow x\in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \tag{3}$$

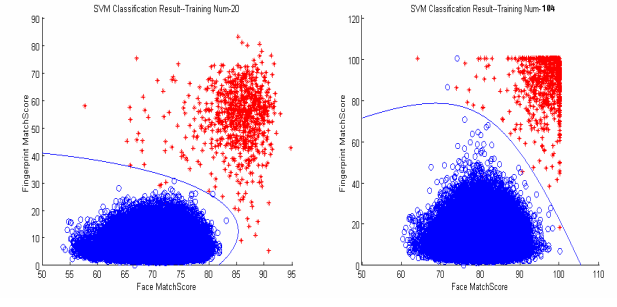
Where $K(x_i,x)$ is a kernel function, and in this authentication system, Polynomial Kernel Function (SVM-Pnm) has been chosen as the authentication fusion strategy:

$$K(x,z)=(x^t z+1)^d, d>0 \tag{4}$$

A new normalization method is proposed in this paper, i.e. constructed a weight vector $W(w_{face},w_{finger})$ for every subject in the database respectively. The W vector is calculated by the following formulates.

$$\begin{aligned} w_{face} &= \frac{100}{\max\limits_{genuine\cup imposter}(s_{face})} \\ w_{finger} &= \frac{100}{\max\limits_{genuine\cup imposter}(s_{finger})} \end{aligned} \tag{5}$$

This weight vector attempts to adjust the variation among the genuine and to emphasize the distance between genuine and impostor. Figure 6 shows the 2 degree SVM-Pnm result to separate genuine and impostor before and after normalization. We can see that SVM-Pnm can separate the two classes almost correctly.



(a) initial (b) normalized

Fig. 6. SVM-Pnm Classification Result

Figure 7 shows that the FAR and FRR curves of SVM-Pnm form a flat and broad vale, this means that the authentication system has a good performance in a large threshold range. And it also proves that the proposed normalization method is effective and better performance can be gained after adopting normalization. In practical authentication system, the decision will be given according to the SVM-Pnm self-learning curve (see figure 6 (b)) after getting the face match score and fingerprint match score.

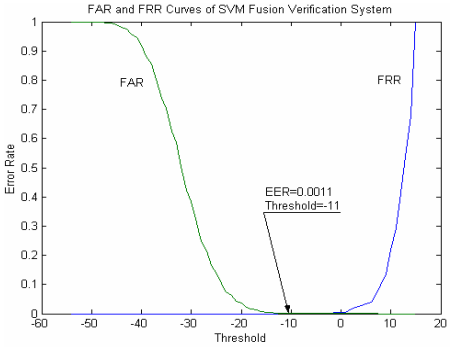


Fig. 7. FAR and FRR curves of SVM-Pnm

4 Experiment Results

This system is tested on TH database. For comparison, we also implement two other dynamic identity authentication systems - a dynamic face identity authentication system and a dynamic fingerprint identity authentication system. The face authentication system uses a common method – face detection and face location and the fingerprint authentication system processes the fingerprint recognition and classification.

The 186 volunteers for the SVM self-learning are selected to be the clients and another 100 volunteers from the TH database are selected to be the impostors to perform the test. Each volunteers tests four times. The condition is based on good illumination and complicated background, see Figure 8.



Fig. 8. Devices and software interface of the system

The results of our system and the other two's are shown in table1. The min TER means minimum of the total error rate where TER=FAR+FRR. From the table we can see that the fusion identity authentication system has more accuracy than the other two.

Table 1. Results of the three systems

System	FAR	FRR	Min TER
Face-only	0.075	0.0925	0.1675
Fingerprint-only	0.0288	0.0124	0.0412
Fusion	0.00362	0.00625	0.00987

5 Conclusion

This paper presented a dynamic fusion biometric system which integrates face and fingerprint in authenticating a personal identification. It combines the advantages of different techniques and performs better than face-only or fingerprint-only systems. The fast and robust face detection algorithm, the optimal face image which has minimum angle based on multi-route detection and the MMP-PCA face recognition indeed improved the face detection and recognition rate. The Gray-Level Histogram Statistic method solved the fingerprint automatic detection and the minutiae match process gets the fingerprints matching score. The SVM authentication fusion strategy performs the fusion authentication decision to separate the client and impostor based on the SVM self-learning results. Experiment results demonstrate that our system performs very well. It does meet the practical response time as well as the accuracy for identity authentication. The real system now is used in our lab.

References

1. A.K.Jain, R.Bolle, and S.Pamkanti, (eds.). Biometrics: Personal Identification in Networked Society. Norwell, Mass.: Kluwer Academic Publishers, (1999).
2. Lin Hong and Anil Jain.: Integrating Faces and Fingerprints for Personal Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.20(12),(1998)1295-1307.
3. Souheil Ben-Yacoub, etc.: Fusion of Face and Speech Data for Person Identity Verification. IEEE Transactions on Neural Networks, Vol.10(5) , (1999) 1065-1074.
4. Wang Yunhong, Tan Tieniu.: Combining Face and Iris Biometrics for Identity Verification. Proceedings of the Conference on Audio- and Video-Based Biometric Person Authentication, (2003).805-813
5. Cheng Du, Guangda Su.: Face Pose Estimation Based on Eigenspace Analysis and Fuzzy Clustering. IEEE International Symposium on Neural Networks (2004).
6. Deng Yafeng, Su Guangda, Zhou Jun, Fu Bo.: Fast and Robust Face Detection in Video. Proceeding of International Conference on Machine Learning and Cybernetics, Aug (2005)
7. Guangda Su, Cuiping Zhang, Rong Ding, Cheng Du.: MMP-PCA face recognition method. Electronic Letters, Vol.38 No.25 December 5, (2002)
8. Huang Ruke.: Research on the Multi-Hierarchical Algorithm for Fast Fingerprint Recognition. Bachelor thesis of Tsinghua University, (2002).
9. Sergios Theodoridis, Konstantinos Koutroumbas.: Pattern Recognition. Elsevier Science, (2003).

Image Recognition for Security Verification Using Real-Time Joint Transform Correlation with Scanning Technique

Kyu B. Doh[1], Jungho Ohn[1], and Ting-C Poon[2]

[1] Department of Telecommunication Engineering, Hankuk Aviation University, 200-1 Hwajeon-dong, Goyang-city 412-791, Korea
kdoh@hangkong.ac.kr
[2] Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA

Abstract. We describe a technique for image verification by use of real-time joint transform correlation with scanning technique. The described method is independent of a spatial light modulator (SLM) in the Fourier plane which has limitations for applying to robust systems. The system is a hybrid processing combined of optical and electronic methods. The system also can perform real-time correlation without SLM at Fourier plane. We develop the theory of the technique and evaluate a performance of the method by estimating the correlation between an authentic fingerprint image and the distorted fingerprints image. Also, we present computer simulation result to demonstrate the validity of the idea.

1 Introduction

Optical image processors have attracted considerable attention because of their ability to perform a large number of valuable operations at extremely high speeds. The correlation of two images is one of the most important mathematical operations in image processing and pattern recognition [1-2]. Optical correlation-based recognition algorithms have been proposed for use in a wide range of application [3-8]. Most real-time coherent optical correlators are derived from two commonly used techniques; the complex matched-filtering operation and the Joint-fourier Transform Correlation (JTC) technique [2,9]. The implementation of a real-time joint transform image correlator generally requires two spatial light modulators. One is employed at the input plane through which the input and reference image are entered. The other is inserted at the Fourier plane to read out the intensity of the interference between the Fourier transform of two signals. In this paper, we describe a technique by use of real-time joint transform correlation with optical scanning. This technique is independent of a spatial light modulator (SLM) in the Fourier plane which has limitations for applying to robust systems. In section 2 we describe the standard architecture of a joint Fourier transform correlator and concept of the real-time JTC with the application of scanning technique. In section 3 we briefly review the theory

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 780–787, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of scanning technique and develop mathematical model of a real-time joint correlation with scanning technique. Our simulation results are shown in section 4, followed by concluding remarks in section 5.

2 Background

The real-time joint transform correlation systems have been accomplished by directly writing the reference object and the target object on SLM, and the detection of the joint transform power spectrum is subsequently performed by another SLM, which may be coherently read out for correlation operation. The conventional real-time joint transform image correlator system is shown in Fig. 1. In real-time JTC systems, the joint transform power spectrum (JPS) is detected by a SLM, a CCD camera and the output of the CCD camera is fed to another SLM for coherent display as shown in Fig.1, where L2 is a Fourier transform lens with the same focal length as lens L1. The spatial resolution requirement of the CCD and SLM should be in general, that is, the CCD camera and the subsequent SLM must be able to resolve these fine details to have the real-time system work. In addition, phase uniformity and contrast ratio of the SLM also are important consideration. The standard architecture of a joint-Fourier transform correlator is a two-stage system. In the first stage, an input image $f(x-a, y)$ and the reference image $r(x+a, y)$ are Fourier transformed and detected by a square-law device, giving a recorded interference intensity $I(u, v)$ as

$$\begin{aligned} I(u,v) &= |\Im\{f(x-a,y) + r(x+a,y)\}|^2 \\ &= |F(u,v)|^2 + |R(u,v)|^2 + F(u,v)^* R(u,v) e^{j2ua} \\ &\quad + F(u,v) R^*(u,v) e^{-j2ua} \end{aligned} \tag{1}$$

where $F(u, v)$ and $R(u, v)$ are the Fourier transform of $f(x, y)$ and $r(x, y)$, respectively, and u and v are the Fourier transform coordinates. In the second stage, the resulting joint-Fourier transform intensity given by (1) is Fourier transformed to give the output at the plane x' and y':

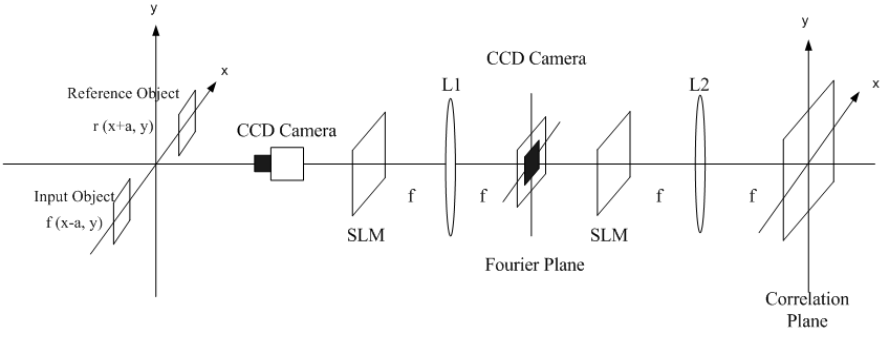


Fig. 1. The conventional Joint-transform Correlation system

$$g(x',y') = f(x',y') \otimes f(x',y') + r(x',y') \otimes r(x',y') + f(x',y') \otimes r(x'-2a,y') + r(x',y') \otimes f(x'+2a,y') \quad (2)$$

where $\otimes$ denotes the correlation operation. Conventionally, $f(x,y)$ and $r(x,y)$ are first recorded on film and the recording of the interference intensity $I(u,v)$ of the two transformed image is subsequently accomplished by a photographic plate. Due to the advent of spatial light modulator, this film recording aspect has been eliminated. Real-time pattern recognition system is accomplished by directly writing $f(x,y)$ and $r(x,y)$ on the spatial modulators, and the detection of the interference intensity is performed by another spatial light modulator which may be coherently read out for correlation operation. The use of spatial light modulators provides the system for real-time correlation. There are, however, major objects to using this scheme. The spatial light modulator in the Fourier plane must be able to resolve the spatial carrier which carries the correlation information. In addition to this, phase uniformity and contrast ration are critical considerations in the design of the spatial light modulator in order to make the real-time JTC practical.

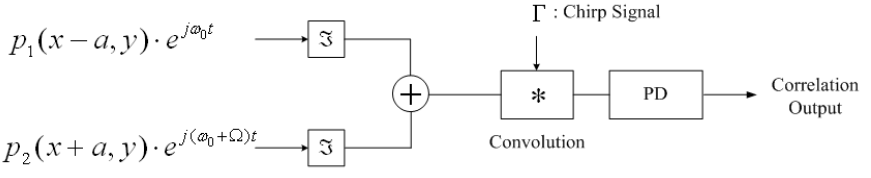


Fig. 2. The System Block diagram of Real-time Joint-Transform Correlation

Realizing the problems of using a spatial light modulator in the Fourier plane, we perform the real-time Joint Fourier transform correlation with the application of optical scanning technique. This approach to JTC does not require a spatial light modulator in the Fourier plane and yet the system can achieve joint transform correlation in real time. The system block diagram is shown in Fig. 2. $p_1(x,y)$ and $p_2(x,y)$ are offset in temporal frequency by Ω. The square-law detection and the subsequent Fourier transform is replaced by a convolution process and an optical heterodyne detection. As it turns out, Fourier transform of a function $h(t)$ is equivalent to the function convolved by a chirp signal under certain approximations. Mathematically, it is given as:

$$h(t) * e^{j2\pi bt^2} \sim \Im\{h(t/2b)\} \quad (3)$$

Where b is some scale constant and $*$ denotes convolution. The approximation employed in (3) is mathematically equivalent to the process for which the Fresnel convolution integral for optical diffraction reduces to a Fourier transform. Convolution can be directly accomplished by using scanning technique, and heterodyne detection is done by optically mixing the light with a photodiode. Note

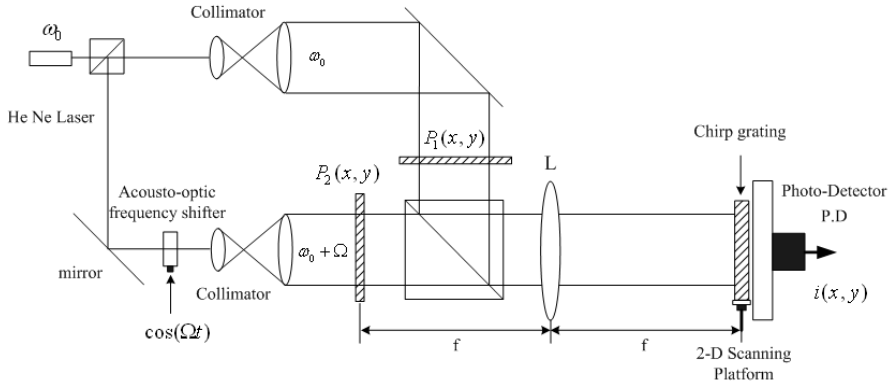


Fig. 3. Implementation of Real-time Joint Transform Correlation system with scanning technique

that no spatial light modulator is required to record the interference intensity in the system. Fig 3. shows an implementation of the system block diagram shown in Fig. 2. $p_1(x, y)$ and $p_2(x, y)$ are illuminated by optical beams with different temporal frequencies, which can be accomplished by using the frequency shifting capability of an acousto-optic modulator. This shift in frequency allows the interaction terms to be carried on a temporal carrier at Ω. By scanning this interaction term with a chirp grating, one effectively takes the Fourier transform of the terms and thus obtains the desired correlation operation. Note that scanning can be achieved by mechanically moving a chirp grating in the Fourier plane. For high speed applications, one could employ an optical scanner such as an acousto-optical scanner placed between the lens L and the chirp grating. The tuned photodiode eventually picks up the mixing current at Ω and sends its output to a display. A peak current output indicates correlation between $p_1(x, y)$ and $p_2(x, y)$.

3 Mathematical Model and Realization of Real-Time Correlation with Optical Scanning

We briefly review the theory of optical scanning technique [10] and its system used to perform a real-time correlation operation. A typical optical scanning system is shown in Fig. 3. Two pupils, $p_1(x, y)$ and $p_2(x, y)$, located in the front focal plane of the lens with focal length f, are illuminated by two broad laser beams of temporal frequencies ω_0 and $\omega_0 + \Omega$ respectively. The transparency function is usually called the pupil function, in our case, input image and the reference image. The laser's temporal frequency offset of Ω can be introduced, for example, by an acousto-optic modulator shown as in Fig. 3. The two beams are then combined by a beam splitter, BS, and used for scanning of the chirp grating, $G_0(x, y)$, located at the back focal plane of lens. The combined scanning field distribution on the back focal plane is given by:

$$P_1(k_x, k_y)\exp(j\omega_0 t) + P_2(k_x, k_y)\exp(j(\omega_0 + \Omega)t) \tag{4}$$

where $P(k_x, k_y)$ is the Fourier transform of the function $p(x, y)$. The combined optical field is used to scan a target (chirp grating) with amplitude distribution given by $|G_0(x, y)|^2 = 1 + cos(ax^2)$ located on the back focal plane of lens L. Alternatively, the target can be placed on a x-y scanner platform, as shown in Fig. 3, while the scanning beam is stationary. The photo-detector, which responds to the intensity of the optical transmitted field, generates a current given by

$$\begin{aligned} i(x, y) = \iint_A |\{P_1(k_x, k_y)\exp(j\omega_0 t) + P_2(k_x, k_y)\exp[j(\omega_0 + \Omega)t]\} \\ \times G_0(x + x', y + y')|^2 dx' dy' \end{aligned} \tag{5}$$

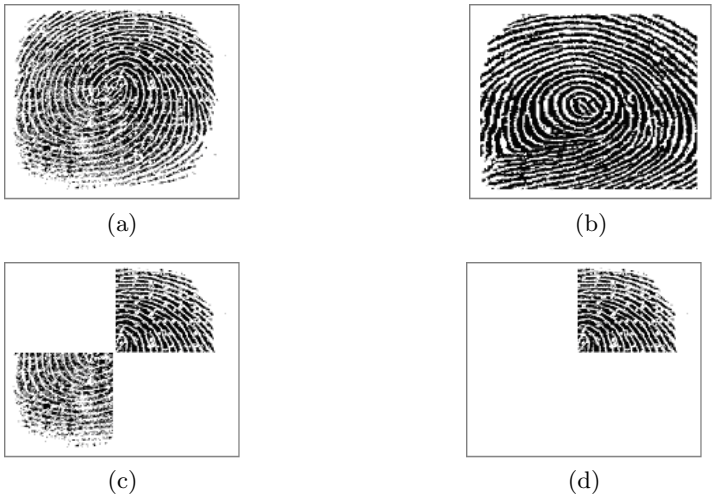
Note that the integration is over photo-detector area A; $x = x(t)$ and $y = y(t)$ represent the instantaneous position of the scanning pattern, and the shifted coordinates of G_0 represent the action of scanning. We can write Equation (5) as

$$\begin{aligned} i(x, y) = \iint_A (|P_1(k_x, k_y)|^2 + |P_2(k_x, k_y)|^2 + P_1^*(k_x, k_y)P_2(k_x, k_y)\exp[j\Omega t] \\ + P_1(k_x, k_y)P_2^*(k_x, k_y)\exp[-j\Omega t]) \times |G_0(x + x', y + y')|^2 dx' dy' \end{aligned} \tag{6}$$

The first and second terms of Equation (6) produce the autocorrelation terms $p_1(x, y) \otimes p_1^*(x, y)$ and $p_2(x, y) \otimes p_2^*(x, y)$ respectively. The third and forth terms of Equation (6) produce the cross-correlation terms between the input signal and the reference signal, that is, $p_1^*(x, y) \otimes p_2(x, y)$ and $p_1(x, y) \otimes p_2^*(x, y)$. The difference between the joint power spectrum of the conventional JTC expressed in Equation (1) and the joint power spectrum of Equation (6) is the quadratic phase function, $\exp(\pm jax^2)$, of chirp grating. The quadratic phase has the form of a chirp signal and we call this modulation chirp encoding. Here we show that owing to the chirp grating, the correlation of the two patterns is achieved.

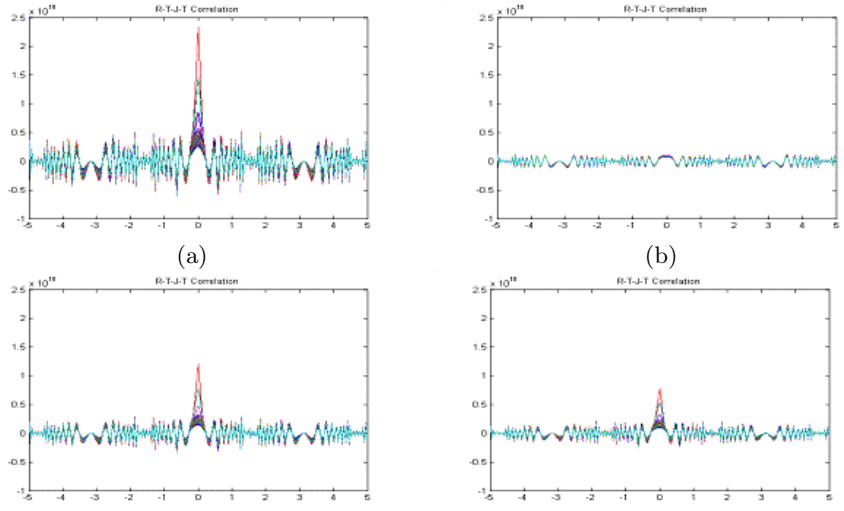
4 Computer Simulation Results

A numerical analysis of the optical correlator is provided to study the effects of the joint power spectrum chirp grating encoding on the correlation signals. To study the performance of the JTC system with scanning technique, we use 2−D fast Fourier transform (FFT). Figure 3 shows the practical implementation of system block diagram shown in Fig. 2. The computer simulation is performed according to the block diagram. The acousto-optic modulator operates in the Bragg regime at sound frequency 40 MHz. The two beams are at frequencies ω_0 and $\omega_0 + \Omega$, respectively, where ω_0 is the frequency of the laser light. The upper path beam at frequency ω_0 is incident on the input pattern p_1 and the lower beam at frequency $\omega_0 + \Omega$ is incident on the reference pattern p_2. The two patterns are located at the front focal plane of lens L. They are combined by the beam splitter and the combined data are then convolved with the chirp grating to achieve



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

Fig. 4. Fingerprints used in the tests: (a) The authorized fingerprint. (b) The unauthorized finger print. (c) Input pattern with missing data when 50% of the authorized fingerprint is blocked. (d) Input pattern with missing data when 75% of the authorized fingerprint is blocked.



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

Fig. 5. Correlation results for the verification. (a) Correlation result for authorized input. (b) Correlation result for unauthorized input. (c) Correlation result for authorized input with 50% missing data. (d) Correlation result for authorized input with 75% missing data.

correlation output. Fig. 4 shows fingerprints of dimensions 30mm × 30mm to be used in the system tests. The fingerprint shown in Fig. 4(a) is chosen to be authentic, and the fingerprint shown in Fig. 4(b) is considered to be unauthorized biometric image to be rejected. Figure 4(c) and 4(d) show an example of an input pattern with missing data when 50% and 75% of the authorized fingerprint is blocked respectively. Simulations with different input images and reference image have been performed. To compare the results between different simulations, all the system parameters, such as the power of the laser source, the time constant, and the value scale are set to the same values throughout these simulations. In Figure 5, the cross correlation between input images and reference image are presented. Figure 5 (a) illustrates the output correlation intensity for verification when an authentic reference fingerprint and an authorized fingerprint are used. A sharp and strong output peak for the input is obtained. Figure 5 (b) shows the output correlation for verification when an authentic reference fingerprint and an unauthorized fingerprint are used. The correlation peak is nearly flat as expected. Figure 5 (c) and 5 (d) show the results of correlation test with missing data when 50% and 75% of the authorized fingerprint is blocked respectively. The correlation peaks are lower than the one in Fig. 5 (a) as expected but they are much stronger than the one in Fig. 5 (b).

5 Concluding Remarks

We have described a JTC with scanning technique for security verification of objects by using hybrid (optical/digital) pattern recognition and the system produces the correlation function in output plane. For the images presented the system can identify an authorized input by production of well-defined output peaks or it rejects unauthorized input. Current methods for real-time optical correlation employing a joint-Fourier transform approach use a spatial light modulator (SLM) in the focal plane to store the interference intensity which is subsequently read out by coherent light. Because the resolution required for accurate representation of the image transform is considerably high, the spatial light modulator which meet the specifications are expensive. The described system does not suffer from the two major drawbacks with the conventional real-time system: the existence of zeroth-order spectra (because of scanning technique) and the use of a high quality spatial light modulator for coherent display (because of the use of a chirp grating). This departure from the conventional scheme is extremely important as the proposed approach does not depend on SLM issues. Consequently, the system should become more readily adapted to industry as well as military pattern-recognition applications. We would expect the correlation peak to be as good as the one obtained from classical JTC systems. To obtain a higher correlation peak in our system, one can preprocess the input image (fingerprint) such as the extraction of its edge information. As final remark, we have described a real-time JTC system for security verification which does not require a spatial light modulator in the Fourier transform plane as in other real-time JTC systems. We will further investigate the performance of the system in terms of input noise,

biometric rotation, and other types of distortions such as missing data caused by wet fingerprint to study the robustness of the system while applying additive factors such as the nonlinear JTC and discrimination ration.

Acknowledgement. This research was supported by the Internet information Retrieval Research Center (IRC). IRC is Regional Research Center of Gyeonggi Province, designated by ITEP and Ministry of Commerce, Industry and Energy.

References

1. VanderLugt, A.: Signal detection by complex spatial filter, IEEE Trans. Inf. Theory IT-10, 139-146(1964)
2. Weaver, C. S., Goodman, J.W.: A technique for optically convolving two functions, Appl. Opt. 5, 1248-1249(1966)
3. Horner, J.: Metrics for assessing pattern-recognition performance, Appl. Opt. 31, 165-166 (1992)
4. Alkanhal M., Kumar B.: Polynomial distance classifier correlation filter for pattern recognition, Appl. Opt. 42, 4688-4708 (2003)
5. Kumar, B., Mahalanobis, A., Takessian, A.: Optimal tradeoff circular harmonic function correlation filter methods providing controlled in-plane rotation, IEEE Trans. Image Process. 9, 1025-1034 (2000)
6. Yu, F. T., Jutamulia, S. , Lin, T., Gregory, D.: Adaptive real-time pattern recognition using a liquid crystal TV based joint transform correlator, Appl. Opt. 26, 1370-1372 (1987)
7. Lu, G., Zang, Z., Wu, S., Yu, F. T.: Implementation of a non-zero-order joint-transform correlator by use of phase-shifting techniques, Appl. Opt. 38, 470-483 (1995)
8. Matwyschuk, A., Ambs, P., Christnacher, F.: Targer tracking correlator assisted by a snake-based optical segmentation method, Opt. Commun. 219, 125-137 (2003)
9. Goodman, J.W.: Introduction to Fourier Optics, McGraw-Hill, New York, 1967
10. Poon, T.-C.: Scanning holography and two-dimensional image processing by acousto-optic two-pupil synthesis, J. Opt. Soc. Am. A2, 621-627 (1985)

Binarized Revocable Biometrics in Face Recognition

Ying-Han Pang, Andrew Teoh Beng Jin, and David Ngo Chek Ling

Faculty of Information Science and Technology, Multimedia University,
Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia
{yhpang, bjteoh, david.ngo}@mmu.edu.my

Abstract. This paper proposes a novel revocable two-factor authentication approach which combines user-specific tokenized pseudo-random bit sequence with biometrics data via a logic operation. Through the process, a distinct binary code per person, coined as bio-Bit, is formed. There is no deterministic way to acquire bio-Bit without having both factors. This feature offers an extra protection layer against biometrics fabrication since bio-Bit authenticator is replaceable via token replacement. The proposed method also presents functional advantages of obtaining zero equal error rate and yielding a clean separation between the genuine and imposter populations. Thereby, false accept rate can be eradicated without suffering from the increased occurrence of false reject rate.

1 Introduction

Biometrics authentication technology is becoming popular. Biometrics characteristics cannot be lost and difficult to duplicate and share. However, biometrics signals from the same user acquired at different times vary dramatically due to the inconsistent acquisition environment and user's interaction with the acquisition device. Besides, it cannot be easily revoked. If a biometrics is compromised, it is compromised forever and rendered unusable. This concern is exacerbated by the fact that a person has limited numbers of biometrics characteristics. Users may run out of biometrics for authentication if the biometrics characteristics keep being compromised.

Bolle et. al. [1] and Davida et al. [2] have introduced the terms *cancelable biometrics* and *private biometrics* to rectify this issue. These terms are used to denote biometrics data that can be cancelled and replaced, as well as is unique to every application. The cancelability issue of biometrics was also addressed by Andrew et al. [3]. They introduced the freshness into the authenticator via a randomized token. The revocation process is the inner-product of a tokenized pseudo-random pattern and the biometrics information iteratively. Savvides et al. [4] proposed a cancelable biometrics scheme. In the scheme, different templates can be obtained from the same biometrics by varying the convolution kernels thus enabling the template cancelability.

This paper proposes a revocable two-factor face authentication approach which combines user-specific tokenized pseudo-random bit sequence with biometrics data to generate a distinct binary code, coined as *bio-Bit*. The overview of the proposed approach is depicted in Fig. 1. Firstly, a compact facial vector $\{\boldsymbol{x}\}$ is derived from the raw biometrics image $\{\boldsymbol{I}\}$ by using moment analysis. A high-dimensional facial data is transformed into a compact feature vector. Secondly, Chang's Multi-bit Scheme

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 788 – 795, 2005.
© Springer-Verlag Berlin Heidelberg 2005

(MbS) [5] is utilized to derive a stable bitstring $\{y\}$ from the features. According to the degree of distingushability, each moment feature in the vector may contribute more than one bits to the bitstream generation. Lastly, a token-generated random bit sequence $\{r\}$ is combined with $\{y\}$ via a logic operation to derive *bio-Bit* $\{B\}$.

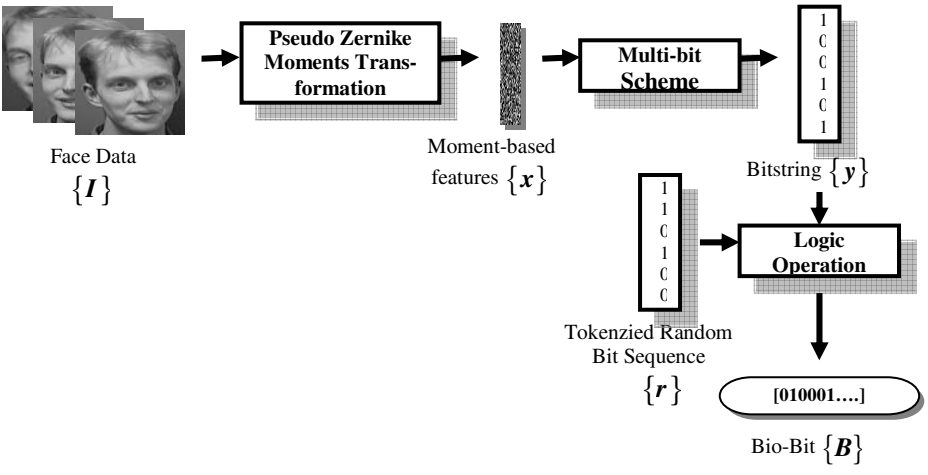


Fig. 1. The block diagram of the proposed approach

The incorporation of the factors improves the accuracy of the bio-Bit by incorporating "who you are" with "something you have". The mixing step even provides higher level security, in which users must possess both authentication factors. Besides, this method also solves the non-reproducible issue of biometrics. Biometric data is not exactly reproducible as it varies from time to time due to different environments (e.g. illumination, pose, facial expressions) or physical changes (e.g. glasses). Implementation of Chang's MbS into our approach outcomes a stable bitstring by considering features that fall within a range determined by the mean and standard deviation of the genuine feature distribution [5]. The irrevocability of biometrics is also resolved by integrating the biometrics with a random bit sequence. This enables bio-Bit to be reissued if it is compromised, by replacing another bit sequence via token replacement.

2 Pseudo Zernike Moments

We adopt pseudo Zernike moments (PZMs) for facial feature extraction [6]. PZM features are rotational invariant, through their magnitude representation [7][8]. Since PZMs are computed using scaled Geometric moments, they are able to produce a set of scale invariant features [9]. Translation invariance is achieved by transferring the image into one whose origin is the image centroid [8][9].

The two-dimensional pseudo Zernike moments of order p with repetition q of an image intensity function $f(r,)$ are defined as [7][8],

$$\boldsymbol{PZ}_{\boldsymbol{pq}} = \lambda(p,N)\int_0^{2\pi}\int_0^{1} V_{pq}(r,\theta) f(r,\theta) r dr d\theta \tag{1}$$

$\lambda(p,N) = \dfrac{4(p+1)}{(N-1)^2\pi}$ and pseudo Zernike polynomials $V_{pq}(r,\theta)$ are defined as,

$$V_{pq}(r,\theta) = R_{pq}(r)e^{-\hat{j}q\theta}; \quad \hat{j} = \sqrt{-1} \tag{2}$$

and $r = \sqrt{x^2 + y^2}$, $\theta = \tan^{-1}\left(\dfrac{y}{x}\right)$, $-1 < x, y < 1$

The real-valued radial polynomials is defined as,

$$R_{pq}(r) = \sum_{s=0}^{p-|q|}(-1)^s \frac{(2p+1-s)!}{s!(p+|q|+1-s)!(p-|q|-s)!} r^{p-s} \quad , \quad 0 \le |q| \le p,\ p \ge 0 \tag{3}$$

3 Stable Bitstring Generation

Multi-bit Scheme (MbS), proposed by Chang et al. [5], is a stable bitstring generation scheme that transforms moment feature vector $\{\boldsymbol{x}\}$ to a stable bitstring $\{\boldsymbol{y}\}$. The distinguishability of each biometrics feature is examined to generate a stable bitstring [5][10]. Features that fall within a certain standard deviation from the genuine mean will be assigned multiple bits. Let $\boldsymbol{x}$ be a feature vector. Based on the information of $\boldsymbol{x}$, $\boldsymbol{y}$ of each face is generated by the following steps, see Fig. 2:

(1) Compute the left and right boundaries, *LB* and *RB*, respectively, for each feature:

$$LB = \min\left(m_g - k_g\sigma_g, m_a - k_a\sigma_a\right)$$
$$RB = \max\left(m_g + k_g\sigma_g, m_a + k_a\sigma_a\right)$$

m_a and σ_a are the mean and standard deviation of the authentic feature distribution, and m_g and σ_g are those of global feature distribution. A suitable value is set to k_g to cover almost 100% of the global distribution and k_a is to control the range of authentic feature distribution $\left[m_a - k_a\sigma_a, m_a + k_a\sigma_a\right]$ as the authentic region.

(2) Determine the number of segment with the same size as the authentic region in the ranges of (a) from *LB* to the left boundary of the authentic region, LB_a, and (b) from the right boundary of the authentic region, RB_a, to *RB*.

The number of segment from *LB* to LB_a is $LS = \dfrac{m_a - k_a\sigma_a - LB}{2k_a\sigma_a}$ segment(s).

The number of segment from RB_a to *RB* is $RS = \dfrac{RB - m_a - k_a\sigma_a}{2k_a\sigma_a}$ segment(s).

There are $LS + RS + 1$ segments in $\left[LB, RB\right]$. At least $\log_2\left(LS + RS + 1\right)$ bits are sufficient to specify each segment with a unique binary index.

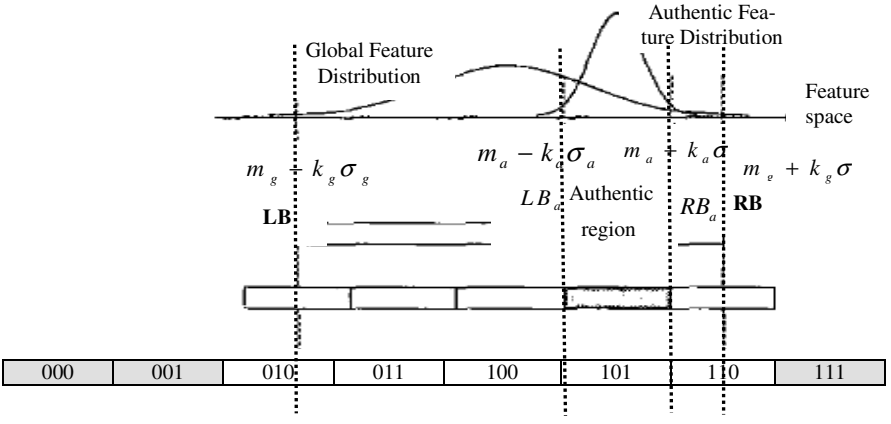


Fig. 2. An example of the bitstream generation for one feature [5]

4 Bio-bit Generation

In this stage, two authentication credentials are combined to generate *bio-Bit*, $\boldsymbol{B}$. The credentials are bitstring, $\boldsymbol{y}$, and a pseudo-random bit sequence, $\boldsymbol{r}$. $\boldsymbol{r}$ is computed based on a seed stored in a token microprocessor through a random bit generator. Same seed is used for both enrollment and verification processes to a same user, but is different among different users and different applications.

Based on the $\boldsymbol{y}$ and $\boldsymbol{r}$, $\boldsymbol{B}$ is generated by the steps: firstly a token is used to generate a set of pseudo-random binary bits with the length of m, which can be represented by $\boldsymbol{r} = \{r_j \mid j = 1, 2, \ldots, m\}$. Then, $\boldsymbol{B}$ is computed by coupling the $\boldsymbol{r}$ and $\boldsymbol{y}$ via XOR-logic operation. This process is represented by $\boldsymbol{B} = \langle \boldsymbol{r} \text{ XOR } \boldsymbol{y} \rangle$.

5 Experimental Results and Discussions

The proposed methodology is evaluated on images from two different types of face databases: Olivetti (OVL) database and Asian Postech Faces'01 Database (PF01). The faces on database for European (OVL database) have definitely different characteristics from those of Asian (PF01). See [11] [12] for more detail. In this paper, the following are the abbreviations used for brevity in our subsequent discussion:

- $PZMl$ denotes pseudo Zernike moments analysis with the first l principal features.
- $mPZMb$ denotes Multi-bit Scheme with a bitstream of b length.
- $mPZM_\text{XOR}b$ denotes our proposed bio-Bit scheme with XOR-logic operation for integrating $\boldsymbol{y}$ and $\boldsymbol{r}$, where b represents the bit code length.

5.1 Performance Measurement: FAR, FRR and EER

Table 1 shows that higher feature length of PZM provides better result because higher order moments contain finer details of the face. But, this is only true to a certain point as the verification rate will level off or even worse if the feature length is extended further. This implies that excessive high order moments contain redundant and unwanted information (e.g. noise).From the table, we can see that the overall error rates of $PZMl$ are relatively high. The reason is the illumination and facial expression variations in face images create a serious non-reproducible issue of biometrics.

By comparing the results of $PZMl$ in Table 1 and $mPZMb$ in Table 2, we observe that $mPZMb$ attains higher recognition rate. This implies that MbS ($mPZMb$) is able to generate a more stable bitstring. From Table 2, we see that $mPZM_\text{XOR}150$ yields FRR=0% when FAR is zero. This is a significant improvement to the contemporary biometric system as the problem of interdependency between FAR and FRR (i.e. specification of low FAR increases FRR and vice versa) is eliminated.

Table 1. Error rates of $PZMl$ with different feature length for OVL and PF01 databases

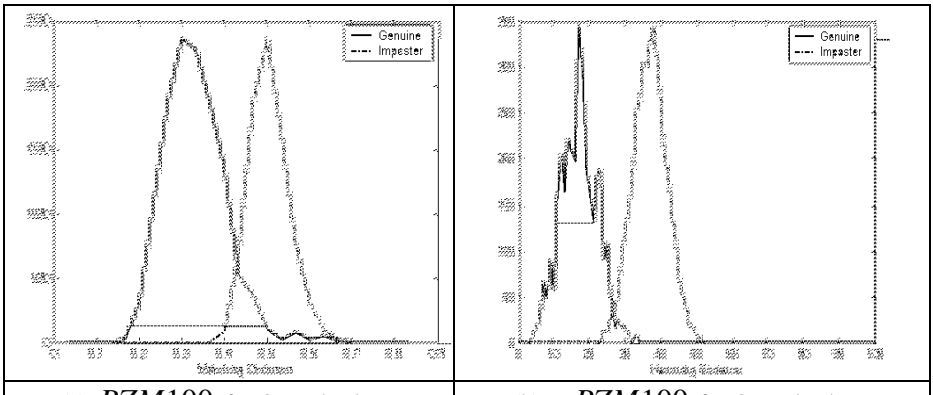
Database	Feature length, l	EER (%)	Database	Feature length, l	EER (%)
OVL	20	20.55	PF01	20	28.04
	40	17.74		40	26.72
	60	17.76		60	25.89
	80	17.23		80	25.46
	100	17.02		100	25.30
	120	16.60		120	26.00
	140	16.33		140	26.39
	160	16.25		160	26.80
	180	16.38		180	27.13
	200	16.40		200	27.29

Table 2. The recognition and error rates of $mPZMb$ and $mPZM_\text{XOR}b$

Database	Bit length, b	$mPZMb$		$mPZM_\text{XOR}b$		
		FAR (%)	FRR (%)	FAR (%)	FRR (%)	Threshold range, $t \in [t_{\min}\ t_{\max}]$
OVL	50	6.64	4.25	0.92	1	-
	100	2.79	2.00	0.15	0.25	-
	150	1.33	1.25	0	0	4 [50 53]
PF01	50	5.67	4.89	0.81	0.95	-
	100	1.55	0.63	0.09	0.10	-
	150	0.32	0.21	0	0	4 [48 51]

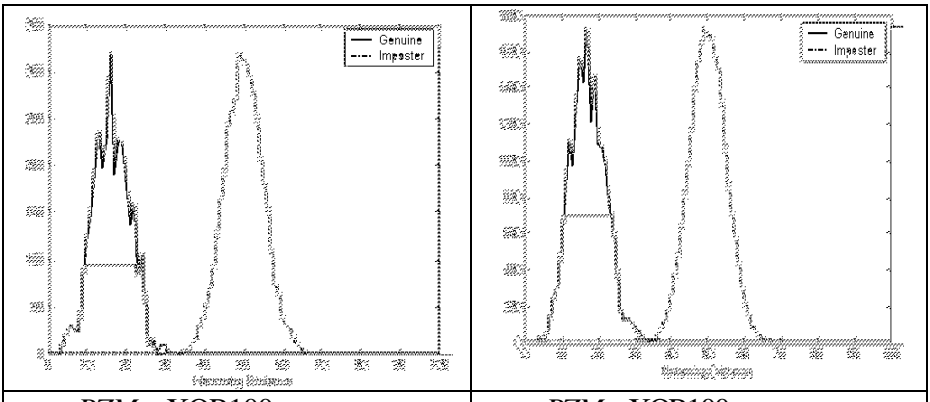
5.2 Genuine and Imposter Population Distributions

The proposed approach is able to provide a clear separation in between the genuine and imposter populations. Fig. 3 shows the genuine and imposter populations of $PZM100$ and $mPZM100$ for OVL database. There is a smaller overlapping in between the populations for $mPZMb$ compared to $PZMl$. This shows that $mPZMb$ is able to produce a distinguishable bitstring. However, $mPZMb$ cannot yield zero EER. This intimates that the FAR-FRR interdependency problem still remains. Fig. 4 illustrates the genuine and imposter population histograms of $mPZM_XOR100$. We observed that $mPZM_XOR100$ is able to separate the populations clearly.



(a) $PZM100$ for OVL database (b) $mPZM100$ for OVL database

Fig. 3. The genuine and imposter populations of $PZM100$ and $mPZM100$



(a) $mPZM_XOR100$ for OVL database (b) $mPZM_XOR100$ for PF01 database

Fig. 4. The genuine and imposter populations of $mPZM_XORb$, b = 100

Let $\boldsymbol{r}_A$ be the random binary data that generated by the user's token and be coupled (via XOR-logic operation) with $\boldsymbol{y}_{EA}$ (enrolled face representation A) with $\boldsymbol{y}_{TA}$ (test face representation A). An experiment is conducted to simulate the below situations.

Case 1: $\boldsymbol{B}_{EA} = \langle \boldsymbol{r}_A \text{ XOR } \boldsymbol{y}_{EA} \rangle \rightleftharpoons \boldsymbol{B}_{TA} = \langle \boldsymbol{r}_A \text{ XOR } \boldsymbol{y}_{TA} \rangle$

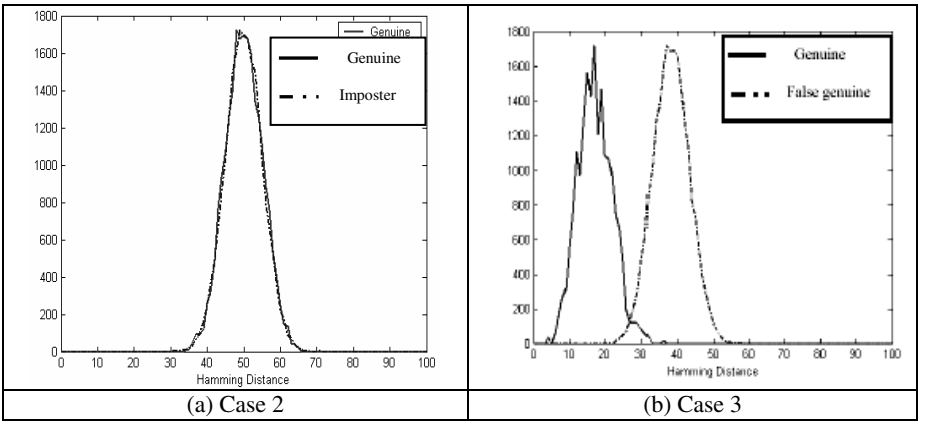
This is the case when A holds his $\boldsymbol{r}_A$ and combines it with his own $\boldsymbol{y}_{EA}$ and $\boldsymbol{y}_{TA}$ during the enrollment and verification processes, respectively. This has been vindicated and discussed in the previous sections.

Case 2: $\boldsymbol{B}_{EA} = \langle \boldsymbol{r}_A \text{ XOR } \boldsymbol{y}_{EA} \rangle \rightleftharpoons \boldsymbol{B}_{TA} = \langle \boldsymbol{r}_o \text{ XOR } \boldsymbol{y}_{TA} \rangle$

This case presumes that A lost his token, i.e. $\boldsymbol{r}_A$ and replace with $\boldsymbol{r}_o$ without updating his bit key in the enrollment session. The result is shown in Fig. 5(a). There is an full overlapping in between the genuine and imposter populations. This reveals that the uniqueness of bio-Bit, $\boldsymbol{B}_{TA}$, for the genuine is vanished when an unauthorized random data, i.e. $\boldsymbol{r}_o$ is used to mix with $\boldsymbol{y}_{TA}$.

Case 3: $\boldsymbol{B}_{EA} = \langle \boldsymbol{r}_A \text{ XOR } \boldsymbol{y}_{EA} \rangle \rightleftharpoons \boldsymbol{B}_{TA} = \langle \boldsymbol{r}_A \text{ XOR } \boldsymbol{y}_{To} \rangle$

This case presumes that adversary o (with feature representation $\boldsymbol{y}_{To}$) uses A's token, $\boldsymbol{r}_A$, to do authentication by claiming himself as A. See Fig. 5(b). The left distribution is the genuine population generated in case 1. The right distribution is the *false genuine* population when person o claims himself as a genuine of A by using $\boldsymbol{y}_{To}$ (unauthorized biometrics) and A's token to do authentication. Since the biometrics is an unauthorized one, for the *false genuine* distribution, the fraction of disagreeing bits is packed around 40%, which is closely to the imposter population in case 1. This shows that person o is treated as an imposter, although an authorized token is used.



(a) Case 2 (b) Case 3

Fig. 5. (a) Case 2 and (b) Case 3

6 Conclusion

In practical, this is almost impossible to get both zero FAR and FRR in biometrics recognition systems because biometrics signals have high uncertainty and the classes are difficult to completely separate in the measurement space. This paper presents a recognition system by using both face data and a unique token deposited with random binary data. The proposed bio-Bit scheme offers 3 main advantages:

(1) The combination of biometrics data and random bit sequence provides a perfect verification performance. The proposed scheme is able to provide a clear separation of the genuine and imposter populations and zero EER level, thereby mitigate the FAR-FRR interdependency problem.
(2) In this proposed approach, two authentication factors are needed in order to derive a bio-Bit for verification. This can strengthen the security of the recognition system as absence of either factor will just handicap the authentication process.
(3) The proposed scheme also addresses the invasion of privacy issue, such as biometrics fabrication. The compromised bio-Bit could be alleviated through the user-specific credential revocation via token replacement.

References

1. Bolle, R. M., Connel, J. H., Ratha, N. K.: Biometric Perils and Patches. Pattern Recognition, 35 (2002) 2727-2738
2. Davida, G., Frankel, Y., Matt, B.J.: On Enabling Secure Applications through Off-line Biometric Identification. Proceeding Symposium on Privacy and Security. (1998) 148-157
3. Andrew, T.B.J., David, N.C.L., Alwyn, G.: An Integrated Dual Factor Verification Based On The Face Data And Tokenised Random Number. In David, Z., Anil, J. (eds.): Lecture Notes in Computer Science, Vol. 3072. Springer-Verlag, (2004) 117-123
4. Savvides, M., Vijaya, B.V.K.,Khosla, P.K.: Cancelable Biometric Filters For Face Recognition. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), (2004) 922-925
5. Chang, Y.J., Zhang, W.D., Chen, T.H.: Biometrics-based Cryptographic Key Generation. Proceedings of the IEEE Int. Conf. on Multimedia and Expo, Vol. 3, (2004) 2203-2206
6. Haddadnia, J., Faez, K., Moallem, P.: Neural Network based Face Recognition with Moment Invariants. Proceedings of the IEEE International Conference on Image Processing, Vol. 1, (Oct. 2001) 1018-1021
7. Teh, C.H., Chin, R.T.: On Image Analysis by the Methods of Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 10, (July 1988) 496-512
8. Khotanzad, A.: Invariant Image Recognition by Zernike Moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, (May 1990) 489-497
9. Bailey, R.R., Srinath, M.D.: Orthogonal Moment Features for use with Parametric and Non-parametric Classifier. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(4), (1996) 389-399
10. Monrose, F., Reiter, M.K., Li, Q., Wetzel, S.: Cryptographic Key Generation from Voice. Proceedings of the 2001 IEEE Symposium on Security and Privacy, (May 2001) 202-213
11. OVL face database: http://www.uk.research.att.com/facedatabase.html
12. PF01 face database: http://nova.postech.ac.kr/archives/imdb.html

Short Critical Area Computational Method Using Mathematical Morphology

Junping Wang[1,3] and Yue Hao[2]

[1] Microelectronics Institute, Xidian University,
Xi' an 710071, Shaanxi, China
[2] Key Lab of Ministry of Education for Wide Band-Gap Semiconductor Materials
and Devices, Xidian University, Xi'an 710071, Shaanxi, China
jpwang@mail.xidian.edu.cn
[3] School of Communication Engineering,
Xidian University, Xi'an 710071, Shaanxi, China
yhao@xidian.edu.cn

Abstract. In current critical area models, it is generally assumed that the defect outlines to be circular and the conductors to be rectangle or merge of rectangles. However, real extra defects and conductors associated with optimal layout design exhibit a great variety of shapes. Based on mathematical morphology, a new critical area computational method is presented, which can be used to estimate critical area of short circuit in semiconductor manufacturing. The results of experiment on the 4*4 shift memory layout show that the new method predicts the critical areas practicably. These results suggest that proposed method could provide a new approach for the yield perdition.

1 Introduction

Yield prediction contributes much to IC manufacturing [1-3]. The yield estimate can be used in production control, material management, timely product delivery and operations planning tool and can be used to highlight designs that have much lower than expected yield, to determine the cost of new devices and their resource implications. It is also the focal point of the industry evolves from production of standard products such as the memories, to the production of ASIC and SOC products. Physical defects have always played an important role in IC yields, and the design sensitivity to these physical elements has continued to increase in today's nanometer technologies [4]. Among typical CMOS process, it is one of the greatest loss factors of the yield that the defect causes the electrical faults. The critical area is defined as the region of the layout where a defect must fall in order to cause an electrical fault. Using the critical area IC yield can be predicted very accurately. The relationship between the yield and critical area can be descried with the famous negative binomial model [5-6].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 796 – 803, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The model of real defect outlines and the features associated with conductor layouts are the two main factors for critical area estimation. The modeling of real defect outlines that exhibit a great variety of defect shapes is usually modeled as circular available, which causes the errors of critical area estimation [7-12]. In addition, the conductor shapes on layouts related critical areas are assumed as the regular geometry form such as rectangles or their merges [13-15]. However, due to the limitation of photolithography, in existing integrated circuit DFY (design-for-yield), the conductor shapes on layouts are to be the non-regular form. For DFY, it is a difficult problem to estimate the critical areas of real defect with the non-regular conductor form on layouts [16-18]. In order to evaluate yield accurately estimation and optimize layout design, it is significant to study the model of the critical areas of real defect with the non-regular conductor form on layouts.

2 Mathematical Morphology and the Core of Real Defect

2.1 The Foundation Operations of Morphology

Mathematical morphology was developed in the 1970's by G. Matheron and J. Serra[19-20]. It has several advantages over other techniques especially when applied to image processing. There are four fundamental operations in morphology, which are dilation, erosion, opening and closing. The foundational operations of binary morphology are defined as follows.

DILATION

$$DILATE(S,E) = \bigcup TRAN(E:i,j) \qquad (i,j) \in Ds \tag{1}$$

where *Ds* represents the *S* domain, which is the set of the known pixel positions of the image *S*. *E* is the structure element, which is converted by *S*, generally. The *TRAN (E: i, j)* is defined as placing the center of the *E* to position *(i,j)* of image *S* and then copying the image *E*.

EROSION

$$[ERODE(S,E)]\ (i,j) = \begin{cases} 1 & if \quad TRAN(E;i,j) \bigcup S = S \\ * & otherwise \end{cases} \tag{2}$$

OPENING

$$OPEN(S,E) = \bigcup \{TRAN(E,i,j) : TRAN(E,i,j) \bigcup S = S\} \tag{3}$$

CLOSING

$$[CLOSE(S,E)]\ (p,q) = \begin{cases} 1 & if \quad for\ all\ [TRAN(E;i,j)](p,q) = 1 \\ & and \quad TRAN(E;i,j) \bigcap S \neq \Phi \\ * & otherwise \end{cases} \tag{4}$$

2.2 The Core of a Real Defect

As critical area is defined as the region where the core of a defect must fall in order to cause an electrical fault, it is necessary to character the core of a defect. The core of a defect can be obtained by using its boundary chain code, which begins with a starting point that has eight neighborhood-points and has one point on the boundary at least. The boundary chain code prescribes the direction that must be used from the current boundary point to a next point.

Given (X_o, Y_o), for chain codes from (X_{j-1}, Y_{j-1}) to (X_j, Y_j), A_{jx} *is* the horizontal displacement , and A_{jy} the vertical displacement. Then, each point coordinates expressed by chain codes is (X_i , Y_i) and the area of a defect is *AREA*:

$$X_i = \sum_{j=1}^{i} A_{jx} + X_o \quad , \qquad Y_i = \sum_{j=1}^{i} A_{jy} + Y_o \tag{5}$$

$$AREA = \sum_{j=1}^{n} A_{jx} * \frac{1}{2}(Y_{j-1} + Y_j) * e^2 \tag{6}$$

where grid resolution is e with 0.3879μm. If *AREA* is not zero, the core of a defect is obtained by finding the center of a defective gravity. With its even density, the center of a defective gravity is equal to the core. For a defect the first-moment matrix Mx of X axis and My of Y axis are:

$$Mx = -\sum_{j=1}^{n} \frac{1}{2} * A_{jx}(Y_{j-1}^{\,2} + A_{jy} * (Y_{j-1} + A_{jy} / 3)) \tag{7}$$

$$My = \sum_{j=1}^{n} \frac{1}{2} * A_{jy}(X_{j-1}^{\,2} + A_{jX} * (X_{j-1} + A_{jX} / 3)) \tag{8}$$

The core of a defect (X_c, Y_c) is:

$$Xc = Mx / AREA \quad , \quad Yc = My / AREA \tag{9}$$

3 Short Critical Area Modeling

There are two main kinds of defects that have remarkable impact on structure of the circuit for IC, the conductor extra defects and the conductor missing defects. Fig.1– Fig. 2 show some real defect shapes in which the extra defects are displayed in Fig. 1 and the missing defects in Fig. 2, with the color images collected by

optical microscopes and the gray images by SEM (scanning electron microscopes). Accordingly, there are two kinds of critical areas that are short circuit critical area and open circuit critical area, respectively. The short critical area formed by the extra defect is researched in the paper, and the open circuit critical area is discussed in later paper.

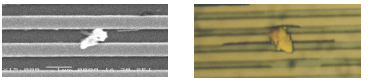


Fig. 1. Extra defect

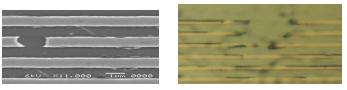


Fig. 2. Missing defect

3.1 Short Circuit Critical Area Modeling

For a conductor $N_1\ (i_1, j_1)$ with row n_1 and column $m1$, where $i_1=1,2..n_1$, $j_1=1,2..m_1$, and another $N_2(i_2,,j_2)$ with row n_2 and column m_2, where $i_2=1,2..\ n_2$, $j_2=1,2..m_2$, the short circuit critical area is the region in which the center of a real defect must fall in order to cause the conductor N_1 to be completely linked to N_2. For example, Fig. 3 gives the points *a, b, c* and *d* on the boundary of the critical area, where the points *a, b, c* and *d* are the center of defects.

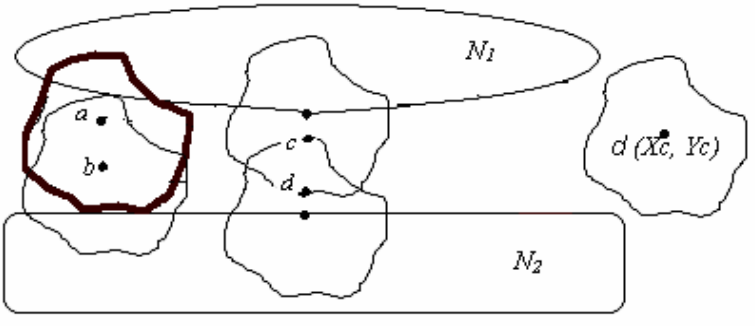


Fig. 3. Short critical area computation for real defect

Based on the definition of mathematical dilation operation and on the concept of short circuit critical area, we put forward the short circuit critical area model. Associated with the real defect $d(X_c\ ,Y_c)$ and with arbitrary conductor shapes N_1, N_2, the short circuit critical area $AS\ (X_c,\ Y_c,\ N_1,\ N_2)$ is composed of the intersection (see Fig. 4), which is caused by dilation conductor N_1 , N_2 with real extra defect $d(X_c\ ,Y_c)$, and is expressed as

$$ASD(\ X_c,Y_c,N_1,N_2\ ) = DILATE(\ N_1,X_c,Y_c\ ) \bigcap DILATE(\ N_2,X_c,Y_c\ ) \tag{10}$$

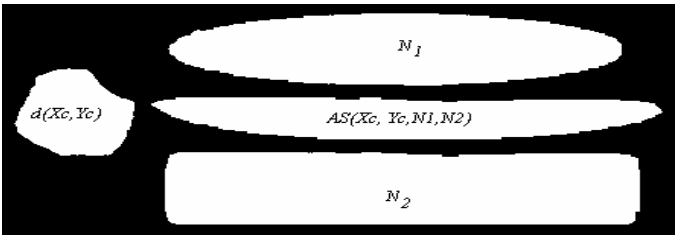


Fig. 4. Short critical area of arbitrary conductive and real defect shapes

$$SN_1 = ASD(X_c, Y_c, N_1, N_2) \bigcap N_1 \tag{11}$$

$$SN_2 = ASD(X_c, Y_c, N_1, N_2) \bigcap N_2 \tag{12}$$

$$AS(X_c, Y_c, N_1, N_2) = \begin{cases} (ASD(X_c, Y_c, N_1, N_2) - SN_1 - SN_2) \bigcup N_1 & \text{if for all } (i_1, j_1) \text{ and } (i_2, j_2) \\ & SN_1 \neq 0 \text{ and } SN_2 \neq 0 \\ ASD(X_c, Y_c, N_1, N_2) & \text{otherwise} \end{cases} \tag{13}$$

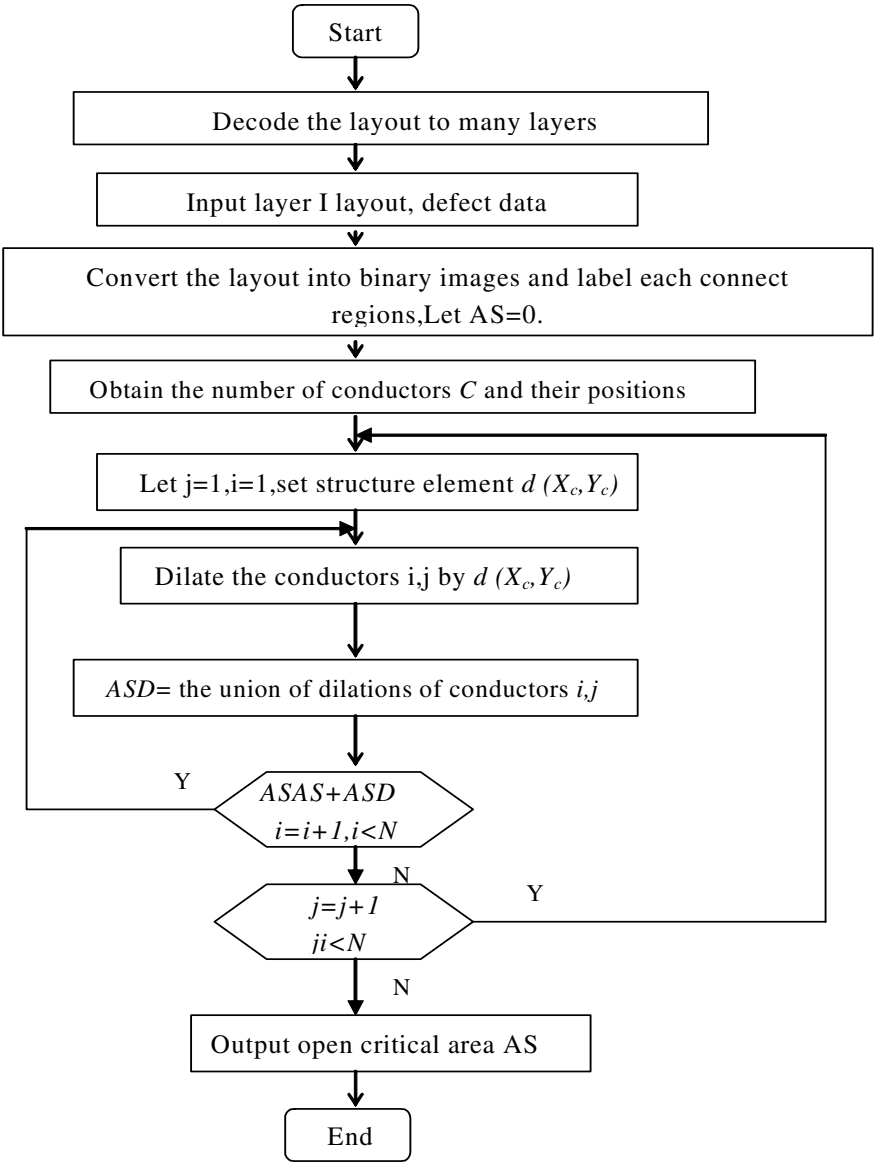
where $d(X_c, Y_c)$ is the core of a defect and the origin position of the structure element, $DILATE(N_1, X_c, Y_c)$ denotes dilating conductor N_1 by defect $d(X_c, Y_c)$, $DILATE(N_2, X_c, Y_c)$ is the dilation of conductor N_2 with defect $d(X_c, Y_c)$, and $ASD(X_c, Y_c, N_1, N_2)$ is the intersection of $DILATE(N_1, X_c, Y_c)$ and $DILATE(N_2, X_c, Y_c)$. If $ASD(X_c, Y_c, N_1, N_2)$ is zero, the critical area is zero, which corresponds to that small size defect will cause short circuit between conductors. If both SN_1 and SN_2 are zero, the critical area of larger size defect is obtained by Eq. (11) and Eq. (12), and shown by the region between N_1 and N_2 in Fig. 4. If neither SN_1 nor SN_2 are zero, the critical area reaches its maximum, that is, the merge of N_1 and the region between N_1 and N_2.Extraction Algorithm for Short Circuit Critical Area.

3.2 Algorithm and Experiment Result of Short Circuit Critical Area

It is assumed that the defect shape be arbitrary form and the layouts be CIF file format, which are changed into the multi-layer plane figure.

For the layer of a conductor layout, we design algorithm A to extract the short circuit critical area. With algorithm A, the counts c of conductors is labeled according to their columns firstly, the critical area of each pair conductors is extracted based on the short circuit critical area model. The critical area at different layers can be obtained by iterating algorithm A.

Algorithm A:



3.3 Experimental Results

The above algorithm is implemented. The experiments were run using an Inter Celeron(TM) AT COMPATIBLE CPU 1.2GHZ (256M memory) on 4*4 shift register metal layout. Some results in experiment are shown in Fig. 5 to Fig. 9. The local experimental layout consists of arbitrary conductor shapes, which is similar to 4*4 shift

register layout (see Fig. 5). By the feature of the real layout, it can be obtained different sizes and different shape defects with 2, which are d1(6,8) and d2(2,3), respectively, (see Fig. 6) ,where (6,8) denotes the center of defect d1 as well as the original position of using d1 as structure element, and (2,3) the center of d2 as well as the original position of structure element d2. Then the critical area associated with Fig. 7 produced by d1 is shown in Fig. 8, and also d2 in Fig. 9. In order to design yield expediently, Fig. 9 shows the union of Fig. 5 and Fig. 8. From Fig.9, it can be seen that the critical area with d2 consists of the elliptic and rectangular regions, as shown in Fig. 9. Altering conductors in those two regions can decrease critical areas and increase IC yield. Since this is the first solution capable of processing such random defect model and layout, comparisons with existing solutions have not been made.

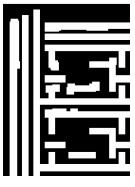


Fig. 5. Layout



Fig. 6. Real defects

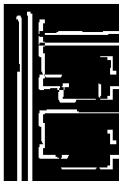


Fig. 7. AS by *d1*



Fig. 8. AS by *d2*

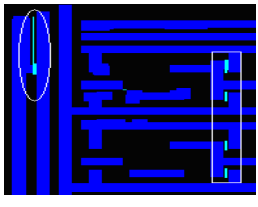


Fig. 9. Union of Fig. 5 and Fig. 8

4 Conclusion

Based on mathematical morphology, the new models and their algorithms for the short circuit critical area have been presented and implemented. Using real defects and arbitrary conductor shapes in our models assure the accuracy and the commonality of the critical area estimations. The experimental results show the efficiency of the new models, which provide the reliable foundations for the optimization layout design and yield accurate estimation.

References

1. Goel, H., Dance, D.: Yield enhancement challenges for 90 nm and beyond**.** IEEEI/SEMI Advanced Semiconductor Manufacturing Conference and Workshop(2003)262-265
2. Asami, T., Nasu, H., Takase, H., Oyamatsu, H., Tsutsui, M.: Advanced yield enhancement methodology for SoCs. IEEEI/SEMI Advanced Semiconductor Manufacturing Conference and Workshop(2004)144 – 147
3. Stephen A. Campbell: The science and engineering of microelectronic fabrication, publishing house of electronics industry, bejing(2003)
4. Mark Rencher: Physical DFY techniques improve yields, EETimes(2004)
5. Linda S.Milor: Yield modeling based on in-line scanner defect sizing and circuit's critical area. IEEE Trans. Semiconductor Manufacturing. Vol. 12(1999) 26-35

6. C. Stapper: The Effects of Wafer to Wafer Defect Density Variations on Integrated Circuit Defect and Fault Distributions. IBM Journal of Research and Development. Vol. 29(1985) 87-97
7. Hess, C.Weiland, L.H.: Issues on the size and outline of killer defects and their influence on yield modeling. IEEE/SEMI ASMC 96 Proceedings(1996)423-428
8. Wang Junping, Hao Yue: Segmentation of IC Real Defect Color Images Based on an LS Space. Chinese Journal of Electronics. Vol. 33(2005)954-956
9. Xiaohong Jiang, Yue Hao: Equivalent Circular Defect Model of Real Defect Outlines in the IC Manufacturing Process. IEEE Trans on Semiconductor Manufacturing. Vol. 11 (1998) 432-440
10. Papadopoulou, E.: Critical area computation for missing material defects in VLSI circuits, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Vol. 20(2001)583-597
11. De Vries, D.K., Simon, P.L.C.: Calibration of open interconnect yield models. 18th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems(2003)26-33
12. Allan, G.A., Walton, A.J.: Efficient extra material critical area algorithms. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. Vol.18(1999)1480-1486
13. W. Maly: Computer-Aided Design for VLSI Circuit Manufacturability. Proceeding of the IEEE. Vol. 78(1990)356-392
14. Milor, L.S.: Yield modeling based on in-line scanner defect sizing and a circuit's critical area. IEEE Transactions on Semiconductor Manufacturing. Vol. 12(1999) 26-35
15. Gerard A. Allan and Anthony J. Walton: Critical Area Extraction for Soft Fault Estimation. IEEE Trans. On Semiconductor Manufacturing. Vol. 11(1998) 146-154
16. Junping Wang, Hao Yue: Yield Modeling Based on Circular Defect Size and a Real Defect Rectangular Degree. ICSICT Proceedings. Vol. 2(2004)1104-1107
17. Allan, G.A.: A comparison of extra material critical area extraction methods**.** IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop(2000)142-151
18. Zachariah, S.T, Chakravarty, S.: Algorithm to extract two-node bridges**,** IEEE Transactions on Very Large Scale Integration Systems. Vol. 11(2003)741-744
19. Allan,G.A.: Targeted Layout Modifications for Semiconductor Yield/Reliability Enhancement. IEEE Transactions on Semiconductor Manufacturing. Vol. 17(2004)573-581
20. Serra J.: Image analysis and Mathematical Morphology. Academic Press New York (1992)
21. Kenneth R. Castleman: Digital Image Processing, Inc., a Simon & Schuster company, Prentice Hall(1996)

A Robust Lane Detection Approach Based on MAP Estimate and Particle Swarm Optimization

Yong Zhou, Xiaofeng Hu, and Qingtai Ye

School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
{zhou, wshxf, yqingtai}@sjtu.edu.cn

Abstract. In this paper, a robust lane detection approach, that is primary and essential for driver assistance systems, is proposed to handle the situations where the lane boundaries in an image have relatively weak local contrast, or where there are strong distracting edges. The proposed lane detection approach makes use of a deformable template model to the expected lane boundaries in the image, a maximum a posteriori (MAP) formulation of the lane detection problem, and a particle swarm optimization algorithm to maximize the posterior density. The model parameters completely determine the position of the vehicle inside the lane, its heading direction, and the local structure of the lane. Experimental results reveal that the proposed method is robust against noise and shadows in the captured road images.

1 Introduction

Driver assistance systems aim to increase the comfort and safety of traffic participants by sensing the environment, analyzing the situation, and generating warning for the driver of the vehicle, or autonomously driving the vehicle [1]. Lane detection, the process of estimating the geometry structure of the lane in front of the vehicle and the position and orientation of the vehicle inside the lane, is primary and essential for driver assistance systems, including lane departure warning, intelligent cruise control, and lateral control etc.

Most lane detection methods are edge-based [2]-[7]. After an edge detection step, the edge-based methods organize the detected edges into meaningful structure (lane marking) or fit a lane model to the detected edges. Most of the edge-based methods use simple lane models to characterize the lane boundaries.

Edge-based methods are highly dependent on the methods used to extract the edges corresponding to the lane boundaries. Often in practice the strongest edges are not the road edges, so that the detected edges do not necessarily fit a straight-line or a smoothly varying model. Edge-based methods often fail to locate the lane boundaries in images with strong distracting edges.

To overcome the demerits of the methods described in the above, a novel lane detection and tracking method is developed in this paper. Firstly, a deformable

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 804–811, 2005.
© Springer-Verlag Berlin Heidelberg 2005

template model of the projective projection of the lane boundaries is introduced assuming that the lane boundaries are parabolas in the ground plane. Then, the lane detection problem is formulated as a maximum a posteriori (MAP) estimate. Due to the non-concavity of the function involved, a particle swarm optimization algorithm (PSO) is used to obtain the global maxima. The model parameters calculated completely determine the position of the vehicle inside the lane, its heading direction, and the local structure of the lane. The proposed lane detection approach can handle the situations where the lane boundaries in an image have relatively weak local contrast, or where there are strong distracting edges.

The rest of this paper is organized as follows: Section 2 introduces the lane model in the image coordinate system. Section 3 describes the proposed lane detection approach including a maximum a posteriori (MAP) formulation and a particle swarm optimization algorithm to maximize the posteriori density. Experiment results are provided in Section 4 and Section 5 gives the conclusions.

2 Lane Model

Lane model plays an importance role in lane detection and tracking algorithms. In order to fully recover 3D information from the 2D lane model, some assumptions about the lane's structure in real world have to be made. In this study, we assume that the two boundaries of the lane are parallel on the ground plane. To derive the model of the lane boundary in the image coordinate system, the model of the lane boundary must be mathematically defined. Let (η, ε) denote a point on the lane boundary in an earth fixed coordinate system with its origin represented by a reference point on the road boundary and η, ε pointing to front (along the tangent to the lane boundary), and right, respectively. According to [7], the coordinates of the point on the lane boundaries can be approximated by the parametric equation (1).

$$\begin{cases} \varepsilon = C_0 l^2/2 \\ \eta = l \end{cases} \tag{1}$$

where, C_0 is the curvature of the lane boundary, and l is the arc length of the lane boundary segment. By a complicated derivation, the lane boundary model in the image coordinate system can be expressed as equation (2).

$$n = K/(m - m_0) + B(m - m_0) + n_0 \tag{2}$$

where, m and n are the row number and the column number of the image, respectively, $K = \alpha_u \alpha_v C_0 H/(2\cos^3\phi)$, $B = \alpha_u d\cos\phi/(\alpha_v H)$, $m_0 = -\alpha_v \tan\phi + u_0$, is the row in the image plane corresponding to the horizon of the ground plane, $n_0 = -\alpha_u\theta/\cos\phi + v_0$, α_u and α_v are the horizontal scale factor and the vertical scale factor of the camera, respectively, ϕ is the camera tilt angle, θ is the angle between the vehicle head direction and the tangent to the lane boundary, H is the height of the camera focal point from the ground plane, d is the distance from the projection point of the focal point of the camera on the

ground plane to the tangent to the lane boundary, (u_0, v_0) is the image center coordinates.

The model introduced above is similar to that introduced by [6] and [8] in form, but the meanings of our model's parameters are explicit. The coefficients K, B and n_0 are related to the curvature of the lane, the position and heading direction of the vehicle in the lane, respectively. The parameters (u_0, v_0), (α_u, α_v), H and ϕ can be determined by an extrinsic camera calibration procedure and an intrinsic camera calibration procedure. Assumed that the left lane boundary is the shifted version of the right lane boundary at a distance along X axis in the ground plane, the left and right lane boundaries have equal curvature and equal tangential orientation where they intersect the X axis, so K and n_0 will be equal for the left and right lane boundaries. As a result, the lane shape in an image can be defined by the four parameters K, B_L, B_R and n_0.

3 Lane Detection Based on Deformable Model and MAP Estimate

3.1 The Likelihood Function

For a given deformable model, how to evaluate its fitness with the road image is an important problem. Here, we define a likelihood function $L(z|x)$, which specifies the probability of seeing the real lanes in the input road images, when given a lane model with specific parameters $x = [K, B_L, B_R, n_0]$. The likelihood function we propose here only uses the edge information in the input image.

In many road scenes, it is difficult, if not impossible, to select a suitable threshold that eliminates clutter edges without eliminating many of the lane edge points of interest. We argue that a very low threshold value can insure the existence of true edges corresponding to the lane markings or road boundaries. So, we calculate the gradient values of the input image using the 3×3 Sobel operator with very low threshold. Then, two images can be obtained: a gray level edge magnitude map $f_m(u, v)$, denoting the gradient magnitude of the input image, and a gray scale edge direction map $f_d(u, v)$, denoting the ratio of vertical and horizontal gradient magnitude of the input image. Based on the edge information, the likelihood function is defined by equation (3).

$$L(z|x) \propto \sum_{u,v} (f_m(u,v) \times W(D_L(u,v)) \times |\cos(\beta_L)| + f_m(u,v) \times W(D_R(u,v)) \times |\cos(\beta_R)|) \tag{3}$$

where, $D_L(u, v)$ and $D_R(u, v)$ are the minimum distances from the point (u, v) to the left lane boundary and to the right lane boundary respectively, β_L and β_R are the angles between the gradient direction of the point (u, v) and the tangential directions of the left lane boundary, and the right lane boundary at the point that has minimum distance from the point (u, v), respectively. $W(D)$ is a weighting function, and can be interpreted as a fuzzy member function with

membership radius R if $W(0) = 1$,$W(D)$ decreases monotonically from 0 to R, and $W(D) = 0$ for $D \geq R$. $W(D)$ is defined use equation (4).

$$W(D) = \begin{cases} e^{-D^2/\sigma^2} & 0 \leq D < R \\ 0 & \text{else} \end{cases} \tag{4}$$

The likelihood function definition requires that the lane model agrees with the image edges not only in position, but also in the tangential direction. The contribution made by a pixel to the likelihood is the gradient magnitude at that pixel, multiplied by a weighting function whose value decreases as the pixel gets further away from the lane boundary defined by the given parameter set x, and a function whose value decreases as the gradient direction at that pixel moves farther away from the tangential direction of the lane model. The higher this likelihood, the better the lane model matches the edges in the input image.

3.2 MAP Estimate

The problem of lane detection in gray level images can be formulated as an equivalent problem of determining the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis. In this study, finding the MAP hypothesis involves the maximization of a function over a compact subspace of R^4.

A prior probability density function (pdf) $P(x)$ describes the priori constraints on the location of lane edges in the images of the road scenes. A likelihood pdf $P(z|x)$ denotes the probability of observing the image in which the lane shape is represented by $x = [K, B_L, B_R, n_0]^T$. Then, the MAP estimate can be given by

$$x^* = \arg\max_{x \in R^4} P(x|z) \tag{5}$$

By using Bayes theorem,

$$x^* = \arg\max_{x \in R^4} P(z|x)P(x) \tag{6}$$

Real world lanes are never too narrow or wide, so a prior probability density function (pdf) $P(x)$ is constructed over the lane model parameters set to reflect a priori knowledge as follow.

$$P(x) \propto \exp\left(-\frac{(B_R - B_L - \mu_R)^2}{\sigma_R^2}\right) \tag{7}$$

So,

$$\begin{aligned} x^* = \arg\max_{x \in R^4} \exp\left(-\frac{(B_R - B_L - \mu)^2}{\sigma_R^2}\right) \times \sum_{u,v} (f_m(u,v) \times W(D_L(u,v)) \\ \times |\cos(\beta_L)| + f_m(u,v) \times W(D_L(u,v)) \times |\cos(\beta_L)|) \end{aligned} \tag{8}$$

The maximization, however, cannot be achieved by a simple hill-climbing approach, due to the non-concavity of the function involved. Simulated annealing algorithm and genetic algorithm can be used to find global extremes. But simulated annealing algorithm exists some problems such as slow convergence and efficiency, and the probability to get the global optimal value is small [9]. To obtain global maxima, we use the particle swarm optimization technique. In next subsection, particle swarm optimization of the function involved in the MAP estimate is present.

3.3 Particle Swarm Optimization Based MAP Estimate

The particle swarm optimization (PSO) technique is a population-based stochastic optimization technique first proposed by Eberhart and Kennedy [10], and the population is called a swarm. The PSO technique can be likened to the behavior of a flock of birds. This technique finds the optimal solution using a population of particles. Each particle represents a alternative solution in the multi-dimensional search space. Some of the attractive features of the PSO include ease of implementation and the fact that no gradient information is required. In this study, the gradient of the function is difficult to determinate. They have been used to solve a range of optimization problems, including neural network training [11] and function minimization [12]. The PSO's operation is explained as follows. Each particle represents a possible solution to the optimization task to be solved. During each iteration each particle accelerates in the direction of its own personal best solution found so far, as well as in the direction of the global best position discovered so far by any of the particles in the swarm. This means that if a particle discovers a promising new solution, all the other particles will move closer to it, exploring the region more thoroughly in the process.

In this study, let $x = [x_1, x_2, x_3, x_4]^T = [K, B_L, B_R, n_0]^T$. The side condition can be formulated as follow:

$$R^4 = \{x | a_i \leq x_i \leq b_i, \quad i = 1, \cdots, 4\} \tag{9}$$

For the parameters are related to the lane structure and the vehicle position and orientation which are within certain range, the side condition can be determined by the prior knowledge about the lane structure and vehicle position and orientation.The optimization problem can be converted into the following forms: $\max f(x) = P(x)P(z|x)$ with the side condition (9). The details about the PSO based MAP optimization are present as follows.

Let s denote the swarm size. Each individual $1 \leq i \leq s$ has the following attributes. A current position x^i in the search space, a current velocity v^i and a personal best solution y^i. The initial positions of the particles is produced stochastic. The global best position is the solution found to be the best solution in the swarm until current iteration. Equation (10) and (11) define how the personal and global best solutions are update at time t, respectively.

$$y^i(t+1) = \begin{cases} y^i(t) & \text{if} \quad f(y^i(t)) \geq f(x^i(t+1)) \\ x^i(t+1) & \text{if} \quad f(y^i(t)) < f(x^i(t+1)) \end{cases} \tag{10}$$

$$\hat{y}(t+1) = \arg\max_{y^i} f(y^i(t+1)) \qquad 1 \le i \le s \tag{11}$$

During each iteration, every particle in the swarm is updated using equation (12) and (13). The velocity update step is

$$\begin{aligned} v_j^i(t+1) \;=\; & w v_j^i(t) + c_1 r_{1,j}(t)[y_j^i(t) - x_j^i(t)] \\ & + c_2 r_{2,j}(t)[\hat{y}_j(t) - x_j^i(t)] \qquad j \in 1, \cdots, 4 \end{aligned} \tag{12}$$

where, x_j^i, y_j^i and v_j^i are the current position, current personal best position, and velocity of the jth dimension of the ith particle. c_1 and c_2 are the acceleration coefficients, which control how far a particle will move in a single iteration. Here, they are both set to a value of 2.0. $r_1 \sim U(0,1)$ and $r_1 \sim U(0,1)$are elements from two uniform random sequences in the rang $(0,1)$. The new velocity is then added to the current position of the particle to obtain its next position.

$$x^i(t+1) = x^i(t) + v^i(t+1) \tag{13}$$

To reduce the likelihood of the particle leaving the search space, the value of each dimension of every velocity v^i is clamped to the range $[-v_{j\max}, v_{j\max}]$. We specify different range for every dimension here, which is different with the other PSOs. However, it has been noted in [13] that the particles may still occasionally fly to a position beyond the defined search space, and hence produce an invalid solution. Here, we adapt the damping boundary condition proposed in [14]. Whenever a particle tries to escape the search space in any one of the dimensions, part of the velocity in that dimension is absorbed by the boundary and the particle is then reflected back to the search space with a damped velocity along with a reversal of sign. We first start by performing the same procedure as for the reflecting boundary where the magnitude and the sign of the velocity for the reflected particle are determined. The velocity is then multiplied by a damping factor, Δd, which is a random variable uniformly distributed between $[0,1]$ to create the damping effect.

The inertial weight w is set according to the following equation:

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{\mathrm{ite}_{\max}} \cdot \mathrm{ite} \tag{14}$$

where $\mathrm{ite}_{\max}$ is the maximum number of iterations, and ite is the current iteration number. w linearly decreases from $w_{\max} = 1.0$ to near $w_{\min} = 0$ over the execution.

4 Experiment Results

This subsection presents the performance of the proposed method in the real road scenes. The left and right lane boundaries in the images can be detected. If the extrinsic camera calibration results and an intrinsic camera calibration results are available, the structure parameter of the lane and the vehicle position and orientation parameters inside the lane can be easily determined. The

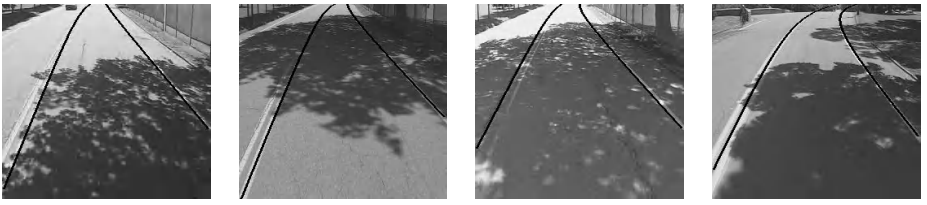


Fig. 1. The detection results of the lanes with strong shadow



Fig. 2. The detection results of the curved lanes

algorithm is tested on some images from video grabbed by an onboard camera in our laboratory and images provided by Robotics Institute, Carnegie Mellon University (CMU), Pittsburgh, PA. All experimental images are gray images of size 240×256. The detected lane boundaries are superimposed onto the original images. These lane images include curve and straight road, with or without shadows, with dashed or solid lines. Four examples of applying the proposed lane detection algorithm to roads with strong shadows are shown in Fig. 1. All results are correctly matching to the lane markings despite the shadows lay over the pained lines. The detection results of four curved roads are shown in Fig. 2. It can be seen that the used lane model can achieve very good results on approaching lane markings on the both sides of the roads. Please notice in the top of the third image the result is not matching very well due to weak lane edge pixels. However, the matching is still maintaining high accuracy in the bottom of the image. The experimental results show that the proposed method is robust against noise, shadows, and illumination variations in the captured road images.

All the results shown in this paper are generated by running the PSO algorithm with the particle number $s = 10$, the maximal 20 iterations.

5 Conclusion

We have addressed the problem of lane detection and tracking in this paper. A deformable template model of the projective projection of the lane boundaries is introduced assuming that the lane boundaries are parabolas in the ground plane. The lane detection problem is formulated as a maximum a posteriori (MAP) estimate. A particle Swarm optimization algorithm is used to obtain the global maxima. The model parameters calculated completely determine the position of

the vehicle inside the lane, its heading direction, and the local structure of the lane. The experimental results obtained by our algorithm are good and accurate, even under the shadowy and noisy conditions.

References

1. Thomas Bcher, Cristobal Curio, Johann Edelbrunner, Christian Igel, and David Kastrup, et al, Image processing and behavior planning for intelligent vehicles *IEEE Transactions on Industrial Electronics*,vol. 50, no.1,pp. 62-75, Jan. 2003.
2. Qing Li, Nanning Zheng, and Hong Cheng, Springrobot: a prototype autonomous vehicle and its algorithms for lane detection, *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp300-308, Dec. 2004.
3. Yue Wang, Eam Khwang Teoh, And Dinggang Shen, Lane detection using B-Snake, *Image and Vision Computing*, Vol. 22, no. 4, pp. 269-280, Apr. 2004.
4. Roland Chapuis, Romuald Aufrere, and Frdric Chausse, Accurate road following and reconstruction by computer vision, *IEEE Transactions on Intelligent transportation systems*, vol. 3, no. 4, pp. 261-270, Oct.2002.
5. Jong Woung Park, Joon Woog Lee, Kyung Young Jhang, A lane-curve detection based on an LCF, *Pattern Recognition Letters*, vol. 24, no.14, pp. 2301-2313, Oct. 2003.
6. Karl Kluge and Sridhar Lakshmanan, A deformable-template approach to lane detection, in *Proc. IEEE Intelligent Vehicles Symposium*, 1995, pp. 54-59.
7. Ernst D. Dickmanns and Birger D. Mysliwetz, Recursive 3-D road and relative ego-state recognition,*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 199-213, Feb. 1992.
8. Antonio Guiducci, Parametric Model of the perspective projective projection of a road with application to lane keeping and 3D road reconstruction, *Computer Vision and Image Understanding*," vol. 73, no. 3, pp. 414-427, Mar.1999.
9. E. -G. Talbi, T. Muntean, Hill-climbing, simulated annealing and genetic algorithm: a comparative study and application to the mapping problem, *in Proc. of the Twenty-Sixth Hawaii International Conference on System Sciences*, vol. 2, 1993, pp.565-573.
10. R. C. Eberhart and J. Kennedy, A new optimizer using particle swarm theory, in *Proc. 6th Int. Symp. Micro Machine and Human Science*, Nagoya, Japan, 1995, pp. 39-43.
11. A. P. Engelbrecht and A. Ismail, Training product unit neural networks, *Stability Control: Theory Appl.*, vol. 2, no. 1C2, pp. 59C74, 1999.
12. Y. Shi and R. C. Eberhart, Empirical study of particle swarm optimization, in *Proc. Congr. Evolutionary Computation,Washington, DC, July* 1999, pp. 1945C1949.
13. J. Robinson and Y. Rahmat-Samii,Particle swarm optimization in electromagnetics,*IEEE Trans. Antennas Propag.*, vol. 52, no. 2, pp. 397C407, Feb. 2004.
14. Tony Huang, and Anada Sanagavarapu, A hybrid boundary condition for robust particle swarm optimization, *IEEE Antennas and Wireless Propagation Letters*, vol. 4, no.1, 2005.

MFCC and SVM Based Recognition of Chinese Vowels*

Fuhai Li[1], Jinwen Ma[1,**], and Dezhi Huang[2]

[1] Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China
lfh@water.pku.edu.cn, jwma@math.pku.edu.cn
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract. The recognition of vowels in Chinese speech is very important for Chinese speech recognition and understanding. However, it is rather difficult and there has been no efficient method to solve it yet. In this paper, we propose a new approach to the recognition of Chinese vowels via the support vector machine (SVM) with the Mel-Frequency Cepstral Coefficients (MFCCs) as the vowel's features. It is shown by the experiments that this method can reach a high recognition accuracy on the given vowels database and outperform the SVM with the Linear Prediction Coding Cepstral (LPCC) coefficients as the vowel's features.

1 Introduction

Recently, automatic speech processing becomes very important and significant since it can contribute to the natural language recognition and understanding as well as the applications to robot and automation. In fact, speech processing technology has been developed very quickly in many fields, especially in the Internet, telecom and security. Speech recognition, understanding and synthesis are the main contents of the speech processing [1, 2, 3, 4], in which the vowel recognition generally plays an important role.

Chinese language (Mandarin) is the largest natural language in the world. So it is very important and significant to study the Chinese speech processing. Actually, Chinese has five major vowels, i.e., /a/, /o/, /e/, /i/, /u/. Clearly, the recognition of the five Chinese vowels is very useful for Chinese speech recognition and understanding. However, there has not been any efficient method to do such a difficult task, to our best knowledge. For the purpose of Chinese vowel recognition, there are two tasks. The one is to extract a set of features from each Chinese vowel such that should be quite different from those of the other vowel and speaker-independent. The other is to find out a reasonable classifier to classify the vowels by their features.

* This work was supported by the Natural Science Foundation of China for Project 60471054.

** The corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 812–819, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In signal processing, cepstral analysis has turned out to be an efficient tool for speech analysis [4]. Actually, the Mel-Frequency Cepstral Coefficients (MFCCs) [2, 4] can be considered as a favorable choice of the features of a vowel. Moreover, since the number of features for a vowel is large, the Supporter Vector Machine (SVM) [5, 6, 7, 8, 9] can be used to classify the vowels with the selected features.

In this paper, we will use the MFCCs as the vowel's features and the SVM as the classifier for Chinese vowel recognition. It is shown by the experiments that the MFCCs and SVM based Chinese vowel recognition system can reach good recognition accuracy on the given data and outperform the SVM with the Linear Prediction Coding Cepstral (LPCC) coefficients as the vowel's features.

2 The Mel-Frequency Cepstral Coefficients (MFCCs) of the Vowels

We begin to introduce the MFCCs as the features of the vowels. Actually, the MFCCs is one of the cepstral analysis technology that has been applied to speech analysis successfully. The Mel frequency describes that the human's apperception of the frequency is nonlinear. The following formula shows the relationship between the Mel frequency and the real frequency:

$$Mel(f) = 2595 \log(1 + f/700). \tag{1}$$

where f is the real frequency and the unit is Hz.

The extraction process of MFCCs is given as follows.

(1). **Pretreatment**

A. Pre-emphasis. The aim of pre-emphasis is to elevate the high frequencies of the speech and make the spectrum smooth. The used first-order digital filter and it's output are given as follows:

$$H(z) = 1 - \mu z^{-1}. \tag{2}$$

$$y(n) = x(n) - \mu x(n-1). \tag{3}$$

where $\mu = 0.9375$ and $x(n)$is the original signal.

B. Windowing. We divide each speech signal into certain frames by Hamming sliding window $h_j(n)$ whose length is 512 and offset is 128.

$$h_j(n) = \begin{cases} 0.54 - 0.46\cos[2n\pi/(N-1)], & \text{if } 128(j-1) \le n \le N-1+128(j-1); \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$s_j(n) = y(n)h_j(n). \tag{5}$$

where $h_j(n)$ is the $j-th$ window, $s_j(n)$ is the $j-th$ frame signal (We neglect the zero signals outside the Hamming window), and $j = 1, 2, ..., 13$.

(2). **The Spectrum Calculation**

We then calculate the amplitude spectrum [4] of each frame $|X_j(k)|$ through the Discrete Fourier Transformation as follows.

$$X_j(k) = \sum_{n=0}^{N-1} s_j(n) e^{-j\frac{2\pi kn}{N}}. \tag{6}$$

(3). $m(l)$ **Calculation**
We further calculate the Mel-scaled triangular filter's output $m(l)$. The triangular filters are equal interval in the mel-frequency scale.

$$m(l) = \sum_{k=o(l)}^{k(l)} W_l(k)|X_j(k)|, l = 1, 2, ..., L. \tag{7}$$

where

$$W_l(q) = \begin{cases} \frac{k-o(l)}{c(l)-o(l)}, o(l) \leq k \leq c(l); \\ \\ \frac{h(l)-k}{h(l)-c(l)}, c(l) \leq k \leq h(l). \end{cases} \tag{8}$$

and $o(l)$,$c(l)$,$h(l)$ are the triangular filter's lower limit, central, and upper limit respectively. L is the number of the triangular filters and we select $L = 16$.

(4). **MFCCs Calculation**
Finally, the MFCCs can be got via the Discrete Cosine Transformation.

$$c_{\text{mfcc}}(i) = \sqrt{2/L} \sum_{l=1}^{L} \log m(l) \cos[(l - 1/2)i\pi/L]. \tag{9}$$

3 The Chinese Vowel Recognition System Via SVM

In this section, we introduce the SVM and then construct a Chinese vowel recognition system via SVM.

3.1 Support Vector Machine (SVM)

The basic ideas of the SVM is to find the optimal hyperplane of two linear separation classes. Figure 1 shows an intuitional example in the two dimension case, where the white and black balls represent two type of samples, respectively. The line H separates the two types of balls without any mixture. H1 and H2 drill through balls which are the nearest to H and are parallel to H. The word of "margin" means the distance between H1 and H2. The so-called optimal hyperplane (Here we means the optimal classification line) is to maximize the margin. In this case, the samples on the lines H1 and H2 are called support vectors which actually support the optimal hyperplane.

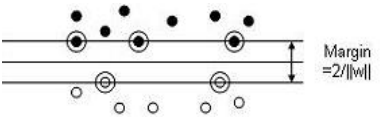


Fig. 1. The optimal classification lines of two-class case

Mathematically, we consider two linearly separated classes C1 and C2 with the samples set $\{(x_i, y_i), i = 1, 2, ..., n, x \in \mathbb{R}^m, y \in \{+1, -1\}\}$. And the linear classification function takes the general form of

$$g(x) = \omega \cdot x + b. \tag{10}$$

Via certain normalization, all the samples in the two classes satisfy $|g(x)| \geq 1$, i.e., the samples which are the nearest to the hyperplane H satisfy $|g(x)| = 1$. So the margin is equal to $2/ \parallel \omega \parallel$. In this setting, the problem of solving the optimal hyperplane can be described as the following optimization problem:

$$\text{minimize} \qquad \phi(\omega) = \frac{1}{2} \parallel w \parallel^2 = \frac{1}{2}(\omega \cdot \omega); \tag{11}$$

$$\text{subject to} \qquad y_i[(w \cdot x_i) + b] - 1 \geq 0. \tag{12}$$

By the method of Lagrange Multipliers, We can get the following dual optimization problem:

$$\text{maximize} \quad Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i x_j); \tag{13}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \ \ and \ \ \alpha_i \geq 0, \ \ i = 1, 2, ..., n. \tag{14}$$

In this way, the optimization problem turns to find a quadratic function's extremum with a linear equation and positive constraints. Actually, it has a unique solution. If the optimal solution is α_i^*, we have

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i. \tag{15}$$

According to the Kühn-Tucker condition, generally there are only small part of α_i^* are nonzero which correspond to the support vectors.

Finally, we have the discriminant function:

$$g(x) = \text{sgn}\{(\omega^* \cdot x) + b^*\} = \text{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\}. \tag{16}$$

where b^* is the threshold value and can be calculated by any support vector via the inequality (12).

From the function (16), we notice that the classification function just has relation with the inner products of the new sample and the support vectors. For the nonlinear case, we can project the original feature space into a higher dimension space in which the samples are linearly separable. But for SVM, we just need to find an appropriate inner product $K(x_i, x_j)$, to replace (x_i, x_j), i.e., we do nothing for the space transformation. In the new feature space, the objective function and the discriminant function are:

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j). \tag{17}$$

$$g(x) = \mathrm{sgn}\{(\omega^* \cdot x) + b^*\} = \mathrm{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^*\}. \tag{18}$$

There are three commonly used kernel functions:

(1). linear: $K(x_i, x_j) = x_i^T x_j$;
(2). polynomial: $K(x_i, x_j) = (x_i^T x_j + 1)^q$;
(3). radial basis function (RBF): $K(x_i, x_j) = \exp\{-\frac{|x_i - x_j|^2}{\sigma^2}\}$.

For the multi-class problem, we can combine some two-class SVMs together. There are two possible ways to do such a task: "one vs. all" and "one vs. one", which are given and compared in detail in [9]. Actually, there are many softwares of SVM available on the web and we will use a typical software of SVM given at a website.

3.2 The Chinese Vowel Recognition System

Based on the MFCCs and SVM, we construct the Chinese vowel recognition system. Figure 2 gives the frame of the Chinese vowel recognition system. The sample data are divided into 2 groups: train data and test data. We firstly use the train data to construct the vowel model via the SVMs in the combination of "one vs. one", and use the test data to check the vowel model. The output of the system is to give the class of an input vowel speech signal.

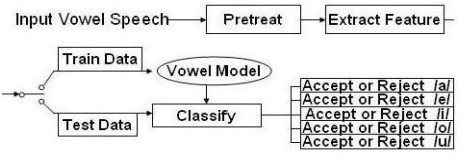


Fig. 2. The vowel recognition training and Testing system

4 Experimental Results and Comparisons

In order to test the Chinese vowel recognition system, we generate one corpora of mandarin speech. The corpora's speakers are required to naturally utter 5000 sentences that have 21 characters in average. Actually, these sentences cover all Chinese tonal syllables. From this corpora we further construct a Chinese vowel database that contains 1567 Chinese vowels which only contain five important Chinese vowels: /a/,/o/,/e/,/i/,/u/. In our experiments, each Chinese vowel speech signal is sampled at the rate of 16000Hz with 16-bit sampling decision. The length of hamming window is 512 and the offset is 128 and so we get 13 frames of each signal. We further set 16 triangular filters on the amplitude spectrum of each frame. Finally, each vowel is represented by its MFCCs which are composed as a 208-dimension vector.

4.1 Experimental Results

For comparison, we use the three kinds of SVMs for the Chinese vowel recognition system: Radial basis function SVM (RBF Kernel), 3-poly SVM (cubic polynomial kernel), and Linear SVM (no kernel). We get the SVM software from website: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/. We use the 4-fold cross-validation test method. We divide the total 1567 vowel samples into 5 classes according to their class labels and divide each class into 4 folds averagely. In each experiment, we choose one fold of each class as the test data set and the remainders as the train data set. Then, we conduct 4 times experiments. At last, we use the average accuracy of the 4 times experiments as the final recognition or classification result. Tables 1-3 give the average accuracies of the vowel recognition system with the different kinds of SVMs.

Table 1. The average recognition accuracies (based on MFCCs) of the Chinese vowel recognition system with the linear SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	96.91%	0.00%	0.00%	3.09%	0.00%
/i/	0.27%	96.98%	1.92%	0.82%	0.00%
/u/	0.00%	0.60%	94.50%	0.89%	4.46%
/e/	3.63%	2.07%	2.07%	85.75%	6.48%
/o/	0.00%	0.00%	12.80%	28.80%	58.40%

Table 2. The average recognition accuracies (based on MFCCs) of the Chinese vowel recognition system with the 3-polynomial SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	99.16%	0.00%	0.00%	0.84%	0.00%
/i/	0.00%	98.35%	0.55%	1.10%	0.00%
/u/	0.00%	0.00%	97.32%	0.60%	2.08%
/e/	1.55%	1.30%	0.52%	94.04%	2.59%
/o/	0.00%	0.00%	16.00%	12.00%	72.00%

Table 3. The average recognition accuracies (based on MFCCs) of the Chinese vowel recognition system with the RBF SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	98.60%	0.00%	0.00%	1.40%	0.00%
/i/	0.00%	98.90%	0.27%	0.82%	0.00%
/u/	0.00%	0.30%	98.51%	0.30%	0.89%
/e/	1.04%	1.04%	0.00%	96.11%	1.81%
/o/	0.00%	0.00%	10.40%	12.80%	76.80%

From tables 1-3, we can find that the MFCCs and SVM based Chinese vowel recognition system is very efficient to recognize the vowels in Chinese speech. Actually, the recognition accuracy on each vowel is very high. Comparing the three kinds of SVMs, we can find that on the average, the Chinese vowel recognition system with the RBF SVMs is slightly better than that with the linear or 3-poly SVMs. However, the system with the 3-poly SVMs get the highest recognition accuracy 99.16% for the vowel /a/.

4.2 Comparisons with the LPCC Based System

We further compare our Chinese vowel recognition system with the same SVM system under the LPCC coefficients as the features of a vowel. Actually, the LPCC coefficients are also widely used in the speech signal processing and have certain favorable properties on the vowel recognition. Here, each vowel also is

Table 4. The average recognition accuracies (based on LPCC coefficients) of the Chinese vowel recognition system with the Linear SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	90.73%	0.00%	0.00%	9.27%	0.00%
/i/	1.10%	96.43%	0.27%	2.20%	0.00%
/u/	0.00%	0.89%	91.37%	2.38%	5.36%
/e/	7.77%	4.15%	1.55%	81.61%	4.92%
/o/	0.80%	0.00%	30.4%	20.80%	48%

Table 5. The average recognition accuracies (based on LPCC coefficients) of the Chinese vowel recognition system with the 3-poly SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	95.51%	0.00%	0.00%	4.49%	0.00%
/i/	0.82%	97.53%	0.00%	1.65%	0.00%
/u/	0.00%	0.60%	94.35%	1.49%	3.57%
/e/	5.70%	4.40%	1.30%	84.72%	3.89%
/o/	0.00%	0.00%	26.40%	25.60%	48.00%

Table 6. The average recognition accuracies (based on LPCC coefficients) of the Chinese vowel recognition system with the RBF SVMs

vowel	/a/	/i/	/u/	/e/	/o/
/a/	96.35%	0.00%	0.00%	3.65%	0.00%
/i/	0.00%	98.90%	0.27%	0.82%	0.00%
/u/	0.00%	0.00%	94.35%	1.49%	4.17%
/e/	3.89%	4.66%	0.26%	88.86%	2.33%
/o/	0.00%	0.00%	23.2%	24.00%	52.80%

Table 7. The differences of the average accuracies of the MFCCs based Chinese vowel recognition system over those of the LPCC based Chinese vowel recognition system

vowel	/a/	/i/	/u/	/e/	/o/
Linear SVM	6.18%	0.55%	2.68%	4.14%	10.40%
3-Poly SVM	3.65%	0.82%	2.97%	9.32%	24.00%
RBF SVM	2.25%	0.00%	4.16%	7.25%	24.00%

represented by a 208 $(13 * 16)$ LPCC coefficients. Tables 4-6 show the average accuracy of Chinese vowel recognition system with three kinds of SVMs. We can see that the Chinese vowel recognition system based on the LPCC coefficients can also lead to a high recognition accuracy, but its recognition accuracy is considerably lower than that of the MFCCs based system, which is given more accurately in Table 7.

5 Conclusion

We have investigated the recognition of Chinese vowels. The MFCCs of a vowel signal are extracted as the features of the vowel and we then use the SVMs to construct the Chinese vowel recognition system. It is shown by the experiments that this MFCCs and SVM based system can reach a high recognition accuracy on the given vowels database and outperform the SVM with the Linear Prediction Coding cepstral (LPCC) coefficients as the features of each vowel.

References

1. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ (1993)
2. Shao, Y., Liu, B., Li, Z.: Speaker Recognition System Based on MFCC and Weighted Vector Quantization. Computer Engineering and Applications (in Chinese), **5** (2002) 127–128
3. Yang, X., Chi, H., et al.: Digital Processing of Speech Signals (in Chinese). Publishing House of Electronics Industry, Beijing (1995)
4. Zhao, L.: Speech Signal Processing (in Chinese). China Machine Press, Beijing (2003)
5. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification . 2rd edn(in Chinese). China Machine Press and CITIC Publishing House, Beijing (2003)
6. Bian, Z., Zhang, X., et al.: Pattern Recognition (in Chinese). Tsinghua University Press, Beijing (2000)
7. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
8. Wang, X., Paliwal, K. K.: Feature Extraction and Dimensionality Reduction Algorithms and Their Applications in Vowel Recognition. Pattern Recognition, **36** (2003) 2429–2439
9. Hsu, C. W., Lin, C. J.: A Comparison of Methods for Multi-class Support Vector Machines. IEEE Trans. on Neural Networks, **13** (2002) 415–425

A Spatial/Frequency Hybrid Vector Quantizer Based on a Classification in the DCT Domain

Zhe-Ming Lu[1,2], Hui Pei[2], and Hans Burkhardt[1]

[1] Institute for Computer Science, University of Freiburg,
D-79110 Freiburg i.Br., Germany
zhemingl@yahoo.com, zheminglu@informatik.uni-freiburg.de
[2] Department of Automatic Test and Control, Harbin Institute of Technology,
Harbin 150001, China
zheminglu@dsp.hit.edu.cn

Abstract. A simple and efficient image vector quantizer (VQ) based on a classification in the DCT domain is presented. Each 8×8 image block is quantized with ordinary VQ (OVQ) or transformed VQ (TVQ), which is determined by the three low-frequency DCT coefficients. Experimental results show that the proposed approach can achieve promising performance.

1 Introduction

Vector quantization (VQ) [1] is an efficient image-coding technique to achieve the low bit-rate, which is usually performed in the spatial domain. Generally, we regard as ordinary VQ (OVQ) the vector quantization process based on full search encoding algorithm with the codebook generated by LBG [2] training algorithm. However, the vector dimension used in OVQ should be low since the storage complexity and the time complexity increase exponentially as the vector dimension increases. Also there are many transforms [3] that have been extensively used in image compression. The principle of transform coding [4] is to quantize and code the transform coefficients. Most compression algorithms are either based on transform coding or by direct VQ of the original image. However, by combining the DCT with the VQ, the DCT-VQ technique has been proposed in [5]. Among the variance DCT-VQ techniques, it was shown that a truncated transformed vector quantization (TVQ) [6] proposed recently which uses the truncated transform coefficients as input vectors with decreased dimension for VQ is simple and can substantially save the complexity. Another DCT-VQ technique called DCT-classified VQ [7-9], provides a satisfactory performance but its classification procedure is too complicated. Based on DCT, a compression algorithm using two VQs (DVQ) [10] also can obtain good performance.

In order to improve the conventional DCT-VQ further, we propose a DCT-based hybrid VQ scheme in this Paper, which employs a simple classification technique to combine both ordinary VQ coding and transformed VQ coding hoping to achieve the best result in terms of bit-rate with acceptable recovery quality. That is to say, this scheme uses both spatial VQ and VQ in transform domain, i.e., this new approach is based on Spatial/Frequency Hybrid Vector Quantization (SFHVQ).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 820 – 825, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Ordinary VQ (OVQ)

VQ is basically a clustering method, grouping similar vectors (blocks) into the same class. Every block through the entire image is grouped. The vectors are obtained from image data by extracting non-overlapping square blocks of size $n \times n$ (such as 8×8). The pixels in each block are arranged in a line-by-line order. VQ process can be defined as a mapping from k-dimensional Euclidean space R^k to a limited subset $C=\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N | \boldsymbol{y}_i \in R^k\}$, where C is the codebook, N is the size of codebook, and $\boldsymbol{y}_i$ is the codeword, i.e., $Q: R^k \rightarrow C$. This mapping must satisfy both $Q(\boldsymbol{x} | \boldsymbol{x} \in R^k) = \boldsymbol{y}_p$ and

$$d(\boldsymbol{x}, \boldsymbol{y}_p) = \min_{1 \le i \le N} d(\boldsymbol{x}, \boldsymbol{y}_i) \tag{1}$$

where $\boldsymbol{x}=(x_1, x_2, \ldots, x_k)$, $\boldsymbol{y}_p=(y_{p1}, y_{p2}, \ldots, y_{pk})$. The distortion between vector $\boldsymbol{x}$ and codeword $\boldsymbol{y}_i$ can be measured by the squared Euclidean distance, i.e.,

$$d(\boldsymbol{x}, \boldsymbol{y}_i) = \sum_{l=1}^{k} (x_l - y_{il})^2 \tag{2}$$

In brief, VQ maps input vectors into a set of code vectors (codewords), and then similar vectors are mapped to the same code vectors in the codebook. VQ consists of two parts: the encoder and the decoder. The encoder outputs the index of code vector of each input vector while the decoder uses the received index to fetch corresponding code vector in the codebook only by simple table lookup operation.

3 Transformed VQ (TVQ)

TVQ is a combined approach of transform coding and VQ techniques. The TVQ starts by mapping the original image data into the transform domain such as DCT domain. Then the produced transformed coefficients are viewed as vector components and can be vector quantized. After vector quantizing, the quantizer outputs are inverse transformed to produce the quantized approximation of the original input image. In our work, a truncated TVQ system is adopted. DCT often results in that the high energy components would be concentrated in the low frequency region. This means that the transformed vector components in the higher frequency region have very little information. These low energy components might be discarded entirely and hence the dimension of transformed vectors and the complexity of VQ are both reduced. In the truncated TVQ system, the low energy components which account for a certain percentage of all transformed components are truncated from each subblock leaving only the high energy components. The following list summarizes steps for the truncated TVQ algorithm:

Step 1: Each image block of dimension k is transformed to frequency domain using DCT.

Step 2: The low energy components are truncated from each image block.

Step 3: VQ is performed on each truncated transformed image block of dimension p ($p<k$).

Step 4: At the receiving end, each decoded coefficient block is inversely transformed using IDCT to compose the reconstructed image.

In a truncated TVQ system, a k-dimensional input vector is mapped into a p-dimensional ($p<k$) vector. Then the p-dimensional VQ is used to vector quantize the

truncated vector. When k is large, the savings in complexity can be very substantial while degradation can be negligible. In other words, TVQ with truncation is a good complexity reducing technique for image coding.

4 Proposed Algorithm (SFHVQ)

As we know, a block of DCT coefficients consists two parts: DC and ACs. DC coefficient associates with the mean value, which is the most important information of the block. But DC coefficient can't reveal the structure in the block. On the contrary, AC coefficients are related to detail structure in the block. There is a useful property that the edge direction is highly correlated with the corresponding DCT coefficients. The number of nonzero AC coefficients denotes the complexity of the block, which means that the block is smooth if there is no nonzero AC coefficient. As reminded in above sections, DCT has the energy compaction property distributing the total signal energy among few coefficients, that is, the high energy components would be concentrated in the low frequency region. Therefore we adopt a simple classification algorithm[11-12] only employing the first three DCT coefficients (see Fig.1) of low frequency region of the input block to classify the input block into two classes: 'relatively complex block' and 'relatively smooth block'. The method is as follows. Given a threshold T, if one of the absolute values of the first three DCT coefficients, that is AC_{01}, AC_{11}, AC_{10}, is larger than T, the block is deemed to 'relatively complex block', otherwise 'relatively smooth block'.

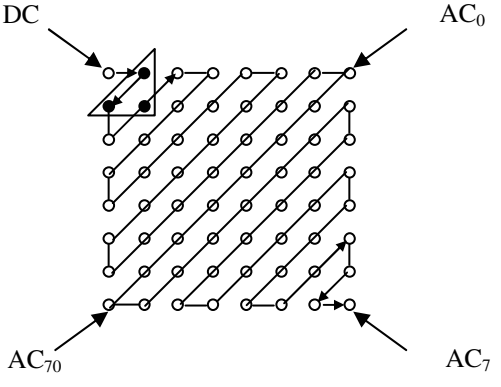


Fig. 1. Zigzag scan of DCT coefficients

Based on the above cases, we propose a hybrid VQ scheme. The proposed algorithm (SFHVQ) is schematized in Fig.2. It is characterized by the novel concept to use both VQ in spatial domain (OVQ) and VQ in transform domain (TVQ) simultaneously. On the other hand, we adopt a simple classification algorithm employing only the first three DCT coefficients of the input block to determine which kind of VQ should be selected. The detail method is as follows. Given a threshold T, if one of the absolute value of the first three DCT coefficients is larger than T, OVQ is selected to encode the original block. Otherwise, truncated TVQ is selected to encode the

original block. Given OVQ codebook and truncated TVQ codebook, the steps for the proposed SFHVQ approach can be outlined below.

Step 1: The input image is divided into blocks of size 8×8.

Step 2: Each block is discrete cosine transformed.

Step 3: Classification algorithm is used to determine which kind of VQ should be selected to encode each block. If OVQ is selected, an extra bit '1' should be added before the index of OVQ codeword. Otherwise, if truncated TVQ is selected, an extra bit '0' should be added before the index of truncated TVQ codeword.

Step 4: At the receiving end, if the index is started with bit '1', the OVQ codebook is looked up for the suitable codeword. Otherwise the truncated TVQ codebook is looked up. All decoded blocks compose the reconstructed image.

5 Experimental Results

We performed experiments on a Pentium-III PC using a 512×512 monochrome image, Lena, with 256 gray scales and the dimension of each block is set to be 8×8. The proposed SFHVQ is compared with ordinary VQ (OVQ), truncated transformed VQ (TVQ) and double VQ (DVQ) in Table 1 and Table 2, where various bit rates are achieved by adjusting the codebook size. In our experiment, threshold value in SFHVQ is selected as 140 and the first 26 DCT coefficients remain to be quantized by TVQ. In Tables 1 and 2, '256+256' in column 'booksize' means the size of OVQ codebook is 256 and the size of truncated TVQ codebook is 256. For all VQ schemes in Tables, LBG algorithm is used to generate codebooks. It is easy to show, the SFHVQ employs a tradeoff between OVQ and truncated TVQ in encoding speed. Although DVQ needs less encoding time, its reconstructed quality is poor. We can

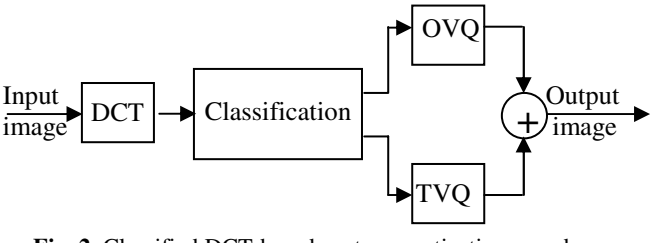


Fig. 2. Classified DCT-based vector quantization encoder

Table 1. Performance Comparison of OVQ, TVQ, DVQ and DCHVQ (Booksize=512)

Algorithm	OVQ	TVQ	DVQ	DCHVQ
Booksize	512	512	256+32	256+256
PSNR(dB)	28.529	28.411	28.055	28.928
Time(s)	2.078	0.906	0.562	1.593
Bit-rate(bpp)	0.141	0.141	0.141	0.141

also see that the proposed scheme exhibits great preponderance over ordinary VQ both in encoding time and reconstructed quality. In addition, we compare the performance with JPEG and JPEG2000 in Table 3. It is easy to see that the performance of the proposed scheme is better than JPEG and JPEG2000 in relatively low bit rate.

Table 2. Performance Comparison of OVQ, TVQ, DVQ and DCHVQ (Booksize=1024)

Algorithm	OVQ	TVQ	DVQ	DCHVQ
Booksize	1024	1024	256+64	512+512
PSNR(dB)	29.821	29.466	28.632	31.133
Time(s)	4.125	1.843	0.688	2.172
Bit-rate(bpp)	0.156	0.156	0.156	0.156

Table 3. Performance Comparisons of JPEG, JPEG2000 and DCHVQ

Algorithm	JPEG	JPEG2000	DCHVQ
PSNR(dB)	26.769	29.674	31.133
Bit-rate(bpp)	0.181	0.156	0.156

6 Conclusion

A simple but efficient image vector quantization (VQ) technique based on a classification in the DCT domain is presented. The algorithm combines both ordinary VQ and transformed VQ technique. Experimental results show that the hybrid VQ scheme outperforms ordinary VQ, transformed VQ, double VQ and JPEG.

Acknowledgement

This work was supported by Alexander von Humboldt Foundation Fellowship (Germany), ID: CHN 1115969 STP.

References

1. Gersho, A., Gray, R. M.: Vector Quantization and Signal Compression, Boston: Kluwer Academic Publishers (1992)
2. Linde, Y., Buzo, A., Gray, R. M.: An Algorithm for Vector Quantizer Design. IEEE Trans. Commun, Vol. 28, No. 1, (1980) 84–95
3. Helsingius, M., Kuosmanen, P., Astola, J.: Image Compression Using Multiple Transforms. Signal Processing. Image communication, Vol. 15, No. 6, (2000) 513–529
4. Netravali, A. N., Limb, J. O.: Picture Coding: A Review. Proceedings of IEEE, Vol. 68, No.3, (1980) 366-406
5. Clarke, R. J.: Transform Coding of Image. New York: Academic Press (1985)

6. Li, R. Y., Kim, J., AL-Shamakhi, N.: Image Compression Using Transformed Vector Quantization. Image and Vision Computing, Vol. 20, No.1, (2002) 37–45
7. Kim D. S., Lee, S. U.: Image Vector Quantizer Based on A Classification in the DCT Domain. IEEE Transactions on Communications, Vol. 39, No.4, (1991) 549–556
8. Kim, J. W., Lee, S. U.: A Transform Domain Classified Vector Quantizer for Image Coding. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 2, No.1, (1992) 3–14
9. Lu, Z. M., Pan, J. S., Sun, S. H.: Image Coding Based on Classified Side-match Vector Quantization. IEICE Transactions on Information and Systems, Vol. E83-D, No. 12, (2000) 2189–2192
10. Zhou, J. P., Yang, Y. X.: Still Picture Compression Algorithm Using Two VQ Based on DCT. Signal Processing, Vol. 14, No. 2, (1998) 177–182
11. Iyoh, Y.: An Adaptive DCT Coding with Edge Based Classification. Fifth International Symposium on Signal processing and Its applications, ISSPA'99, Australia, Vol. 1, (1999) 83-85
12. Busch, C., Funk, W., Wolthusen, S.: Digital Watermarking: from Concepts to Real-time Video Applications. IEEE Computer Graphics and Applications, Vol. 19, No.1, (1999) 25–35

Removing of Metal Highlight Spots Based on Total Variation Inpainting with Multi-sources-flashing

Ji Bai, Lizhuang Ma, Li Yao, Tingting Yao, and Ying Zhang

Digital Media Lab, Computer Science, Shanghai Jiao Tong University,
Shanghai 200240, China
{baiji, ma-lz, yaolily77, yaotingting, zhang-ing}@sjtu.edu.cn

Abstract. The removing of specular highlight spots has been the hotspot in the field of computer vision. A variety of methods have been designed to remove the specular on the surface of the objects. Unfortunately few paper on removal of (specular) highlight spots on the metal surface have been reported. Since the region of highlight spots on metal surface is much larger than that of non-metal, it is harder to remove them. We propose a novel Inpainting with Multi-Sources-Flashing method to resolve the problem. The outcome of the experiments demonstrate that the light-spots on the surface of the metal objects with complex texture are removed effectively, and the physical property of the surface maintains well. The pyramid structure of our algorithm flow makes the innovation function efficiently.

1 Introduction

The reflection of light from the surface in real world is generally classified into two categories; diffuse reflection components and specular reflection components. The diffuse reflection represents the physical property of the surface of the objects, while the specular reflection is associated with the light source. The highlight-spots or specularities which are incident on the surface, provide a true sense of real environment, demonstrate important local curvature information, and may reveal cue for object recognition. However the influence of the spots-lights is not fully considered in the classical algorithm of the computer vision, the removal of the spots is essential step in advance to obtain the accurate information from images, otherwise the error is inevitable.

1.1 Previous Work

A variety of photometric techniques have been proposed for removing specularities using polarization [1, 2, 4], color [3, 5, 6] .However , most of these techniques assume that highlights appear in regions with no variation of material type or surface normal, especially for removing the light spots on the metal surface. Since there is no reflection of diffuse components on the surface of the metal objects, only mirror reflection which is the main source of highlight, the light-spots of metal are much larger than that of dielectric ones. Although several methods exist for the highlight removing [10], these methods have only partially succeeded in it.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 826 – 831, 2005.
© Springer-Verlag Berlin Heidelberg 2005

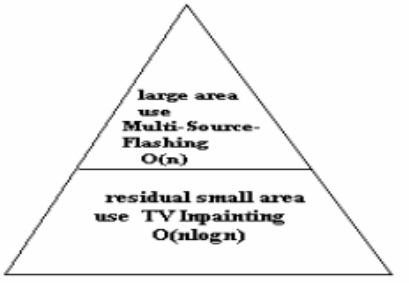


Fig. 1. The Pyramid framework of our algorithm

New method for removing the spots on metal objects is urgently needed. Roderio Feris[7] proposed a method which can deal with the overlapped region so as to reduce the effect of specularities in digital images, however the light-spots can not be removed completely. Previously TV inpainting [8] was formulated to restore edge features in image.

1.2 Contribution

Up to now, little work on the light-spots removing on metal surface has been reported, as a pioneering our method has considerable engineering and research value. To function highly effective, we propose a novel pyramid structured algorithm which is shown in Fig1 to remove the light-spots. We propose the algorithm called TV-Inpainting with MS-Flashing.

The pyramid structure of the algorithm flow is well constructed: the top level is MS-Flashing which is highly effective and can solve the intermediate result in $O(n)$ operations, while the lower level is TV-inpainting which is time consuming and has the time complexity approximately of $O(nlogn)$, however the TV-inpainting will address to move the residual light spots utterly, so as to make the whole process more integral, united, coherent and efficient.

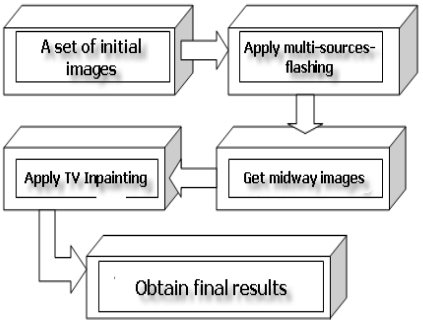


Fig. 2. The basic flow of the Total variation with Multi-sources-Flashing Inpainting

1.3 Overview

The basic flow of our algorithm is demonstrated in Fig2: We apply the set of 4 images with MS-Flashing. Then we obtain a midway image in which the highlight spots are primarily reduced. We address the residual light spots in the midway image with TV-Inpainting. Then we gain image in which the specular light totally removed.

2 Multi-source-flashing

We propose a method of MS-Flashing [7] to reduce the highlight spots on the metal surface. In this paper, we adopt light source from 4 different directions resulting 4 images. The steps taken in the implementation are as follows: (a) For every image with its lighting condition, calculate the intensity gradient (b) Find the median of gradients at the corresponding position in the 4 images (c) Reconstruct the desired image I , that minimizes $|\nabla \hat{I} - G|$. In the paper, we use the standard full multigrid method to reconstruct the image.

Generally speaking, the midway result after applying MS-Flashing algorithm, still has highlight spots on, although shrank sharply. Thus, to deal with the residual highlight spots gracefully, we invoke the TV-inpainting method to work on the residual spots in order to reduce them completely.

3 Total Variation Inpainting

In the paper, the method of TV-inpainting [8] has been adopted to remove the minimized highlight spots. We adopted the mathematic equation in [8] to solve the TV inpainting problem, and we should apply the TV inpainting algorithm to the midway result after the MS-Flashing algorithm in the following steps.

4 Experiment Results

In this paper, the algorithm is implemented on the Matlab6.5 running on a PentiumⅢ PC (128Mb RAM 650GHz) under Windows 2000.To demonstrate the effect of the algorithm in this paper, we use two sets of original images of metal objects. Each set



Fig. 3. The setting of lighting condition

has 4 images taken under 4 different lighting conditions. The original images and the midway image after multi-source-flashing step in the first set come from [7]. The images in the second set are captured with a Olympus C-5060 camera under the lighting condition as showed in Fig3 as follows.

In this paper, we use ordinary lamp as illumination sources. Every target metal objects is imaged 4 times. At every time, turn on one of the 4 lamps as illumination source, while holding position of the camera and the target objects stationary.

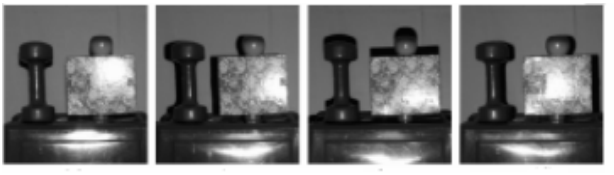


Fig. 4. First image set

The original images and the midway image after MS-flashing step of the first set are illustrated in Fig4 and Fig5(a).It is obviously that although the light-spots in the midway image showed in Fig5(a) are reduced distinctly, some residues of light-spots.

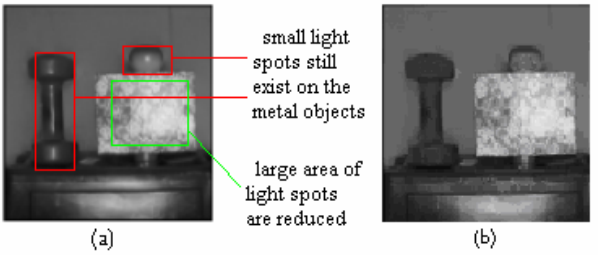


Fig. 5. Midway and final result of the image (a)Midway result after the multi-flashing (b) Final result after the TV inpainting still exist

The final result after the TV inpainting is displayed in Fig5(b), which is almost free from the highlight spots definitely.

The images in the second set are described in Fig 6 as the following.



Fig. 6. The second set of images

The midway result and the final result can be found in the Fig7(a), 7(b) respectively, and there is hardly any highlight spots on the final result.

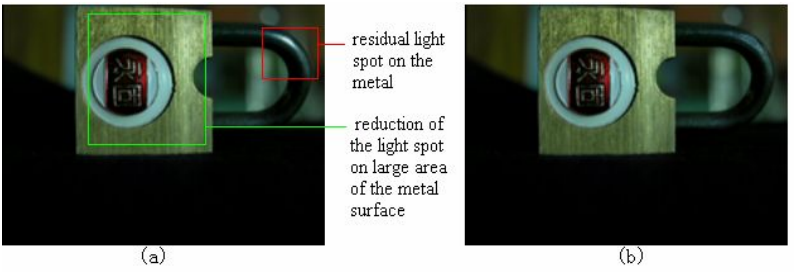


Fig. 7. Midway and final result of the second image set (a)Midway result after the multi-source-flashing (b) Final result after the TV inpainting

Since most of the classical algorithms in the computer vision do not take account of the effect of the high light spots, a comparison experiment is devised here to verify the functionality of the light-spots reducing algorithm in our paper. The graph cuts [9] algorithm is used in the paper, and the result group after being applied with graph cuts is demonstrated in Fig8 as a objective criterion to test the effectiveness of the our algorithm for removing the highlight spots on the surface of the metal objects.



Fig. 8. Comparison of the disparity map after the processing of Graph Cuts (a)Directly apply the Graph Cuts to the original image with highlight (b)Apply the Graph Cuts to the image with highlight reduction by the method in the paper

5 Discussion and Conclusion

There are two main measures for removing the light-spots. The color techniques based on dichromatic model is not meant for metal. The reflection model of metal is different from dielectric so we can not simply use the color one. Polarization techniques can not be used to solve the problem because the polarized angle which is indispensable condition for the polarization measure makes the intensity tend to be too saturated to be suitable for removing the light-spots on the metal surface.

Taking account that the metal objects in the image will have a large region of the saturated pixels owing to the reflection nature of the metal materials, the method using a single image is not desirable. It is required to make use of the method addressing with several images to recover the diffuse factors in the saturated regions.

We have presented a novel method, the total variation inpainting with multi-source-flashing based on the information that highlights often shift according to the position of the light sources.The outcome of algorithm implementation confirms that our method can not only remove the light-spots on the metal surface effectively but also maintain the intrinsic properties of the objects, compared with any other methods published on the metal specular removing.

Acknowledgments

This work was supported by Omron company.

References

1. Wolf, L.B. and Boult, T. :Constraining object features using polarization reflectance model. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 13.(1991)636-657
2. Nayar, S.K., Fang, X.S.,and Boult, T.: Separation of reflection components using color and polarization, International Journal of Computer Vision.(1997)163 - 186
3. Nayar, S. K., Sanderson, A. C., Weiss, L. E. and Simon, D. A.: Specular surface inspection using structured highlight and Gaussian images. IEEE Trans. Rob. Autom, Vol. 6. (1991) 208–218
4. Sato, Y. and Ikeuchi, K.: Temporal-color space analysis of reflection. Journal of Optics Society of America A., Vol. 11.(1992)2990-3002
5. Shafer, S.: Using color to separate reflection components Color Research and applications. Vol. 10.(1985)210-218
6. Klinker, G.J., Shafer, S.A.and Kanade,T.: The measurement of highlights in color images. International Journal of Computer Vision.Vol. 2.(1990)7-32
7. Feris, R., Raskar, R., Tan, K. and Turk, M.: Non-photorealistic Camera: Depth Edge Detection and Stylized Rendering using Multi-Flash Imaging. ACM SIGGRAPH 2004
8. Chan, T. F., Shen, J. H.: Mathematical models local Deterministic inpainting. SIAM J. Appl. Math., Vol. 62.(2001)1019-1043
9. Boykov, Y., Olga, V., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. IEEE PAMI Vol. 23. (2001)1222-1239
10. Lin, S. and Shum, H.Y.: Separation of diffuse and specular reflection in color image. Proceeding of IEEE conference on Computer Vision and Pattern Recognition, Vol. 1. (2001)341-346

Component-Based Online Learning for Face Detection and Verification

Kyoung-Mi Lee

Department of Computer Science, College of Information Engineering,
Duksung Women's University, Seoul 132-714, Korea
kmlee@duksung.ac.kr
http://namhae.duksung.ac.kr/~kmlee

Abstract. Component detectors can accurately locate facial components, and component-based approaches can be used to build detectors that can handle partial occlusions. This paper proposes a face detection and verification method using component-based online learning. The main difference from previously reported component-based approaches is the use of online learning, which is ideal for highly repetitive tasks. This results in faster and more accurate face detection, because system performance improves with continued use. Further, uncertainty is added by calculating the standard deviation of face components and their relations.

1 Introduction

Face detection and face verification is the essential front end of any face recognition system, which locates and segments face regions from video or still images. There are numerous face recognition applications in areas like surveillance and security control systems, content-based image retrieval, video conferencing and intelligent human-computer interfaces. Most face recognition algorithms assume that the face location is known. However, in reality, most images are complicated and may contain extraneous visual information or multiple faces.

Recently, component-based approaches have produced better results than global approaches, since individual components vary little while the variations induced by pose changes are mainly geometric. Component detectors can accurately locate facial components, and component-based approaches can be used to build detectors that can handle partial occlusions [4,6]. In color images, Hsu *et. al* proposed a component-based face detection algorithm that constructs eye, mouth, and boundary maps to verify candidate faces [1]. The facial components were detected using feature maps derived from models based on skin color, and were combined based on the geometries and orientations of the detected components. Since this procedure used simple geometrical relations, the system may be limited in terms of flexibility.

In this paper, we propose a component-based online learning process, and apply this to a face detection and verification system which is able to handle a wide range of variations of facial poses in color images. Section 2 proposes a flexible component-based model and component-based online learning. In Section 3, the proposed model and learning are applied to face detection and verification. Experimental results and conclusions are presented in Sections 4 and 5.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 832 – 837, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Component-Based Online Learning

In this paper, we propose a component-based online learning, which is ideal suited for highly repetitive tasks. Therefore, instead of gathering and training the entire data set, the proposed learning algorithm learns every time new data is collected. It is more desirable to learn such a pattern incrementally, based on examples provided by the user while she is solving a problem.

2.1 Flexible Component-Based Model

Objects can be described well by using a few characteristic components and their geometrical relations. Suppose an object consists of n_c components and n_r relations. The object can be defined as

$$F = (C_i, R_j), \qquad i=1,\ldots, n_c, j=1,\ldots, n_r \tag{1}$$

where C_i is an *i*-th component and R_j a *j*-th relation.

Suppose a set of components are given and are to be partitioned into clusters such that components within the same cluster have a high degree of similarity, while components belonging to different clusters have a high degree of dissimilarity. Each of these clusters is represented by its center, called a *template* (μ), and the center of a cluster is computed by averaging the components that belong to that particular cluster. So do relations.

To provide flexibility, we add uncertainty to the above object model *F* (Eq. (1)). In this paper, uncertainty is measured using standard deviation (σ), the spread of components or relations around an averaged component or an averaged relation [3]. In a flexible component-based model, each component and relation has uncertainty. In other words, the *i*-th component C_i has k_i templates ($\mu_{i_t}^c$, $t=1,\ldots, k_i$) and a standard deviation ($\sigma_{i_t}^c$). So does the *j*-th relation R_j with k_j templates ($\mu_{j_t}^r$, $t=1,\ldots, k_j$) and a standard deviation ($\sigma_{j_t}^r$).

2.2 Component-Based Online Learning

For a component-based object model, the proposed approach incrementally clusters individual components and relations. During training, each new object is compared to all existing component and relation templates. We compute the minimum difference between a component-based object *F* (Eq. (1)) and a template as follows:

$$d_F = \sum_{i=1}^{n_c} \min_{t=1\cdots k_i} \left(\frac{\left\| C_i - \mu_{i_t}^c \right\|_p}{\sigma_{i_t}^c} \right) + \sum_{j=1}^{n_r} \min_{t=1\cdots k_j} \left(\frac{\left\| R_j - \mu_{j_t}^r \right\|_p}{\sigma_{j_t}^r} \right) \tag{2}$$

where $\|\bullet\|_p$ is an Euclidean distance.

If all the minimum distances are less than a predefined threshold, the online learning algorithm adds components and relations to corresponding clusters and updates the clusters by recalculating their centers and uncertainties. Here, similarity thresholds

are set empirically and can be adjusted by a user. If some of the distances are significantly larger than the threshold, a new template is created with the new component or relation and the number of clusters, k_i or k_j, is increased. Thus, after training on a collection of component-based objects, each template represents a cluster of components or relations that are near each other in a component or relation space.

3 Application to Face Detection and Verification

In this section, we applied the proposed component-based algorithm to face detection and verification algorithm. The algorithm first detects skin color pixels using the skin color model and the detected pixels are then grouped into face candidates. After finding candidate eyes and lips as components of a face, the algorithm uses a flexible component-based face model to detect faces based on their geometrical relations and tests for validity by matching nose templates. Each step for detection and verification updates the corresponding template of each component and relation.

It is well known that component-based face detection and verification performs better for pose and illumination variations and occlusions, since components are less prone to these changes and each component is detected independently. In general, faces can be represented by two eyes, a lip and a nose.

3.1 Eye and Lip Detection

To find eyes and a lip as components of a face in a color image (Fig. 1(a)), we used *EyeMap* and *LipMap*, as proposed in [5]. Eye maps are constructed using chrominance and luminance components and then these two maps are combined using an AND operation.

$$EyeMap = \frac{1}{3}\left\{\left(C_b^2\right) + \left(255 - C_r\right)^2 + \left(C_b / C_r\right)\right\} \text{ AND } \frac{Y^+(x,y)}{Y^-(x,y)+1} \tag{3}$$

where Y^+ is a dilation operation and Y^- an erosion operation. Fig. 1(b) shows the resulting image using the ANDed *EyeMap*, Eq. (3), and Fig. 1(c) displays candidate eye regions with *EyeMap* values greater than a predefined threshold value.

For the lip map, the color of the lip region contains a stronger red component and weaker blue component than other facial regions. Hence, the chrominance component C_r is greater than C_b in the lip region. The lips have a relatively low response in the C_r / C_b feature, but a high response in C_r^2.

$$LipMap = C_r^2 \cdot \left(C_r^2 - \eta\, C_r / C_b\right)^2 \tag{4}$$

where η means a ratio of the average C_r^2 to the average C_r / C_b. Fig. 1(d) shows the construction of the lip map and Fig. 1(e) displays candidate lip regions.

3.2 Component-Based Face Detection

The detected eye candidates and lip candidates can be fused by testing their all possible triangular combinations [5]. But, to compute the face score, in other words the

probability of the existence of a face, they used only components satisfying a face triangle. Since they did not consider flexibility of components and relations, it is difficult to detect rotated or skewed faces.

We compute the face score using a flexible component-based face model (Eq. (1)), proposed in Section 3.1. Using Eq. (1) with uncertainty helps to detect faces which do not satisfy some of the conditions for components or relations. Because all possible components and relations are considered, we can locate an optimal face combination from images containing several faces and skewed faces.

$$FaceScore = \prod_{i=1}^{n_c} \alpha_i L(C_i) + \prod_{j=1}^{n_r} \beta_j L(R_j) \tag{5}$$

where *L()* represents the likelihood of the flexible model, and α_i and β_j importance in the model. As discussed in Section 3.1, a component, C_i, can be one *EyeMap* and one *LipMap*, and a relation, R_j, is an angle and the ratios of distances of components. Fig. 1(f) shows a result image obtained using Eq. (5).

3.3 Component-Based Face Verification

To determine whether a candidate is a real face or not, component-based face verification first locates the nose in the detected face in Section 4. The nose is a very useful component since it is positioned between two eyes and a lip. Furthermore, the nose is mostly rigid, so its locus cannot change with facial expression. In most images, the nose is one of the brightest regions of the face. It protrudes from the face and is thus better illuminated than other regions. Simultaneously, the nostrils and the base are significantly darker than the rest of the nose. Even if the nostrils are not present in the image, a dark contour around the bottom of the nose is visible due to the shading under the nose and the steep foreshortening below the bottom of the nose tip. In this paper, we use vertical signatures of the projected gradient map to locate the nose in the detected triangle area between the eyes and lip [2]. From the gradient and phase maps derived by Sobel edge detection, the projection of the gradient magnitude of each edge is computed along the vertical. The edges contained by this triangle are summed through the *x* axis. These summed values form the vertical signature of the nose. The value with the strongest edge content is the one that corresponds to the vertical position of the nose's bottom. The nose's bottom position corresponds to the peak value of the signature.

$$NoseMap = \sum_{x \in \text{face triangle}} Edge \tag{6}$$

The next step in verification is to use Eq. (5) to compute similarity using the nose map and its relations with eyes and a lip. This computation will result in accepting or rejecting the claim by thresholding the dissimilarity measure. Fig. 1(g) shows edged images derived by the Sobel operations in the detected triangle areas from Fig. 1(f); Fig. 1(h) displays the candidate nose regions. Fig. 1(i) show the result images verified with the nose components.

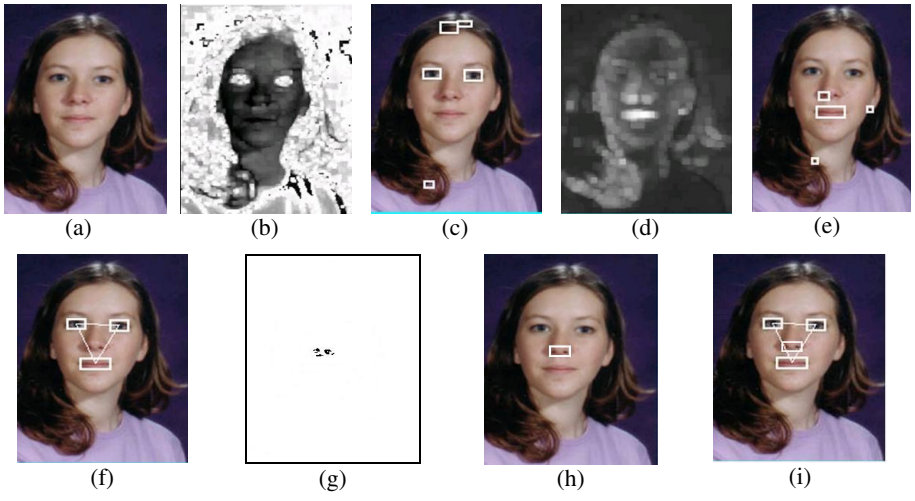


Fig. 1. Component-based face detection and verification: (a) color images, (b) result images applying Eq. (3), (c) eye candidates, (d) result images using Eq. (4), (e) lip candidates, (f) detected faces with components using Eq. (5), (g) edged images by Sobel operators in detected facial triangles, (h) nose candidates using Eq. (6), and (i) verified faces with components

4 Experimental Results and Conclusions

The proposed algorithm was implemented in JAVA, and tested in Windows 2000 with a Pentium-IV 1.8 GHz CPU with a memory of 512 MB. Accuracy was measured on the 713 test images as the training proceeded. Fig. 4 shows that as the proposed system training progresses, its ability to correctly detect the face in the test images increases. After the first few faces, the detection rate stabilized at 74.1% for the image database. Adding verification results in a 93.3% detection rate. We compared our experimental results to those obtained using Hsu's approach [1]. The Hsu algorithm performed at 87.9% accuracy. As expected, detecting several faces in one image is more challenging; our proposed algorithm performs well in these cases.

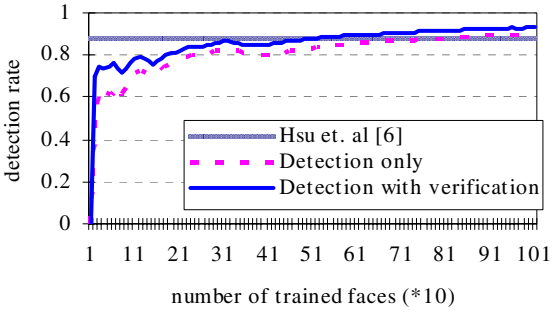


Fig. 2. Face detection and verification performance on the trined 1,980 faces

This paper presents an online learning system for component-based face detection. The proposed algorithm can automatically detect the components and relations of a face. The algorithm showed faster detection times than Hsu's approach based on the Hough transform, which is a time-demanding but accurate method. This paper also introduces an incremental learning scheme for component-based approaches. Our on-line learning algorithm clusters the detected components and relations.

Acknowledgements

This work was supported by Korea Research Foundation Grant No. R04-2003-000-10092-0 (2005).

References

[1] R.L. Hsu, M. Abdel-Mottaleb, and A.K. Jain,: Face detection in color images, *IEEE transactions on pattern analysis and machine intelligence* 24(5) (2002) 696-706.

[2] T. Kanade,: Picture processing system by computer complex and recognition of human faces*, PhD. Thesis*, PhD thesis, Kyoto University, Japan, (1973).

[3] K.-M. Lee and W. N. Street,: Model-based detection, segmentation, and classification using on-line shape learning, *Machine vision and applications* 13(4) (2003) 222-233.

[4] T.K. Leung, M.C. Burl, and P. Perona,:Finding faces in cluttered scenes using random labeled graph matching, *Proceedings of the fifth international conference on computer vision*, (1995).

[5] B. Xie, D. Comaniciu, V. Ramesh, M. Simon, and T. Boult,: Component fusion for face detection in the presence of heteroscedastic noise, *Annual conference of the german society for pattern recognition*, (2003), pp. 434-441.

[6] K.C. Yow and R. Cipolla,: Feature-based human face detection, *Image and vision computing* 15(9) (1997) 713-735.

SPIHT Algorithm Based on Fast Lifting Wavelet Transform in Image Compression

Wenbing Fan, Jing Chen, and Jina Zhen

College of Information Engineering, Zhengzhou University, 450052 China
iewbfan@zzu.edu.cn

Abstract. This paper presents the improved algorithm according to image compression technology to pledge the real time of the image transmission and gain the high compression ratio under the image quality. The improved SPIHT image coding algorithm based on fast lifting wavelet transform presents fast lifting wavelet transform to improve trasform course, because of many consecutive zero appearing in SPIHT quantification coding, adopting the simultaneous encoding of entropy and SPIHT. Entropy coding adopts run-length-changeable coding. Experiment results show that encoding by this method can improve PSNR and efficiency. This method can apply in the image data transmission and storage of remote image surveillance systems.

1 Introduction

The discrete wavelet transform (DWT) is being increasingly used for image coding due to the fact that DWT supports features like progressive image transmission (by quality, by resolution)[1]. The main feature of the lifting based DWT scheme is to break up the highpass and lowpass filters into a sequence of upper and lower triangular matrices and convert the filter implementation into banded matrix multiplications[2,3]. The widely know embedded zerotree encoding (EZW) by Shapiro[4] is an excellent example of such coding scheme, and also a good reference to wavelet based image compression is general. SPIHT (Set Partitioning In Hierarchical Trees) by Said and Pearlman[5,6] is clearly a descendant of EZW using similar zerotree structure and bitplane coding.

This paper is organized as follows. The section 2 describes process of the fast lifting wavelet transform scheme and analysis all the filters. Section 3 gives a brief overview of set partitioning in hierarchical trees (SPIHT) and The proposed SPIHT algorithm based on fast lifting wavelet transform is given in the section 4. Finally, conclusions are presented in Section 5.

2 Fast Lifting Wavelet Transform Algorithm

The basic principle of the lifting scheme is to factorize the polyphase matrix of a wavelet filter into a sequence of alternating upper and lower triangular matrices and a diagonal matrix[1,2]. This leads to the wavelet implementation by means of banded-matrix multiplications.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 838 – 844, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Let $\tilde{h}(z)$ and $\tilde{g}(z)$ be the lowpass and highpass analysis filters, and let $h(z)$ and $g(z)$ be the lowpass and highpass synthesis filters. The corresponding polyphase matrices are defined as

$$\tilde{P}(z)=\begin{bmatrix}\tilde{h}_e(z) & \tilde{h}_o(z)\\ \tilde{g}_e(z) & \tilde{g}_o(z)\end{bmatrix} \text{ and } P(z)=\begin{bmatrix}h_e(z) & h_o(z)\\ g_e(z) & g_o(z)\end{bmatrix} \tag{1}$$

It has been shown in [1] and [2] that if $(\tilde{h},\tilde{g})$ is a complementary filter pair, then $\tilde{P}(z)$ can always be factored into lifting steps as

$$\tilde{P}(z)=\begin{bmatrix}1/K & 0\\ 0 & K\end{bmatrix}\prod_{i=1}^{m}\begin{bmatrix}1 & \tilde{s}_i(z)\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ \tilde{t}_i(z) & 1\end{bmatrix} \tag{2}$$

Where K is a constant. The lifting schemes are shown in Fig.1.

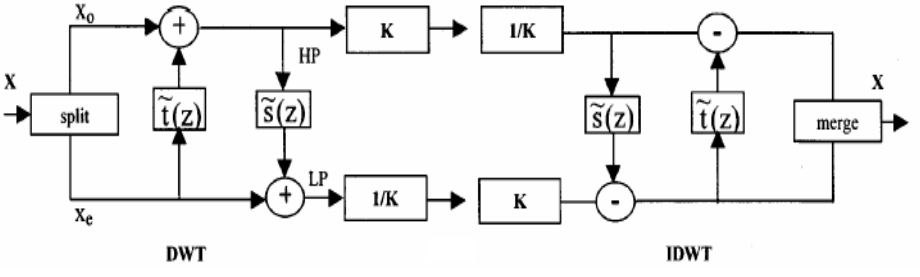


Fig. 1. Fast lifting wavelet transform scheme

Lifting wavelet transform consist of three steps:

$$\begin{aligned}\begin{bmatrix}X^{\mathrm{LP}}(z)\\ X^{\mathrm{HP}}(z)\end{bmatrix}&=\tilde{P}(z)X(z)=\begin{bmatrix}\tilde{h}_e(z) & \tilde{h}_o(z)\\ \tilde{g}_e(z) & \tilde{g}_o(z)\end{bmatrix}\begin{bmatrix}X_e(z)\\ X_o(z)\end{bmatrix}\\ &=\begin{bmatrix}1/K\cdot\{X_e(z)+\tilde{s}(z)[\tilde{t}(z)X_e(z)+X_o(z)]\}\\ K[\tilde{t}(z)X_e(z)+X_o(z)]\end{bmatrix}\end{aligned} \tag{3}$$

1) Split step, where the input samples are split into two even and odd samples by the time domain;

$$X_e(n)=X(2n) \quad \text{and} \quad X_o(n)=X(2n+1) \tag{4}$$

2) Predict step, where the even samples are multiplied by $\tilde{t}(z)$ and are added to the odd samples and the results are multiplied by scaling factor K.

$$X^{\mathrm{HP}}(z)=K[\tilde{t}(z)X_e(z)+X_o(z)] \tag{5}$$

3)Update step, where updated odd samples are multiplied by $\tilde{t}(z)$ and are added to the even samples and the results are multiplied by scaling factor $1/K$.

$$X^{\mathrm{LP}}(z)=1/K\cdot\{X_e(z)+\tilde{s}(z)[\tilde{t}(z)X_e(z)+X_o(z)]\} \tag{6}$$

The inverse DWT is obtained by traversing in the reverse direction, changing the factor K to $1/K$, factor $1/K$ to K, and reversing the signs of coefficients in $\tilde{t}(z)$ and $\tilde{s}(z)$.

Due to the linearity of the lifting scheme, if the input data is in integer format, it is possible to maintain data to be in integer format throughout the transform by introducing a rounding function in the filtering operation. Due to this property, the transform is reversible (i.e., lossless) and is called the integer wavelet transform (IWT). It should be noted that filter coefficients need not be integers for IWT. However, if a scaling step is present in the factorization, IWT cannot be achieved. It has been proposed in[2] to split the scaling step into additional lifting steps to achieve IWT.

3 Set Partitioning in Hierarchical Trees (SPIHT)

The SPIHT algorithm is one of the most efficient algorithms for still image compression. This algorithm was developed to transmit images progressively with SNR scalability. A tree structure, called spatial orientation tree (SOT), naturally defines the spatial relationship on the hierarchical pyramid. Fig.2 shows how the SOT is defined in a pyramid constructed with recursive four-subband splitting.

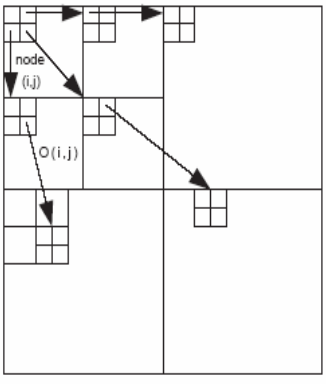


Fig. 2. Spatial-orientation tree in a wavelet pyramid

Each node of the tree corresponds to a pixel and is identified by the pixel coordinate. Its direct descendants (offspring) correspond to the pixels of the same spatial orientation in the next finer level of the pyramid. The tree is defined in such a way that each node has either no offspring (the leaves) of four offspring, which always form a group of 2x2 adjacent pixels. In Fig.2, the arrows are oriented from the parent node to its four offspring. The pixels in the highest level of the pyramid are the tree roots and are also grouped in 2 x 2 adjacent pixels. However, their offspring branching rule is different, and in each group, one of them (indicated by the star in Fig.2) has no descendants. When a threshold value is given, it partitions the coefficients or the sets of coefficients into significant or insignificant coefficients. Significant coefficients are added to the list of significant pixels (LSP), and insignificant coefficients to

the list of insignificant pixels (LIP) or the list of insignificant sets (LIS). The following sets of coordinates are used to present the new coding method:

$O(i, j)$: set of coordinates of all offspring of node (i, j);
$D(i, j)$: set of coordinates of all descendants of the node (i, j);
H : set of coordinates of all spatial orientation tree roots (nodes in the highest pyramid level); $L(i, j) = D(i, j) - O(i, j)$.

The set partitioning rules are simply the following:

1) The initial partition is formed with the sets $\{(i, j)\}$ and $D(i, j)$, for all $(i, j) \in H$.
2) If $D(i, j)$ is significant, then it is partitioned into $L(i, j)$ plus the four single-element sets with $(k, l) \in O(i, j)$.
3) If $L(i, j)$ is significant, then it is partitioned into the four sets $D(i, j)$, with $(k, l) \in O(i, j)$.

4 Modified SPIHT Encoding Algorithm Based on Fast Lifting Wavelet Transform

4.1 SPIHT Encoding Algorithm[5]

Since the order in which the subsets are tested for significance is important, in a practical implementation the significance information is stored in three ordered lists, called list of insignificant sets (LIS), list of insignificant pixels (LIP), and list of significant pixels (LSP). The encoding algorithm is described as follow.

1) Initialization:
Set the LSP as an empty list, and add the coordinates (i, j)εH to the LIP,and only those with descendants also to the LIS, as type A entries.
2) Sorting Pass:
2.1) for each entry (i, j) in the LIP do:
2.1.1) output Sn(i, j);
2.1.2) if Sn(i, j) = 1 then move (i, j) to the LSP and output the sign of ci,j;
2.2) for each entry (i, j) in the LIS do:
2.2.1) if the entry is of type A then
- output Sn(D(i, j));
- if Sn(D(i, j)) = 1 then
  - for each (k, l) ∈ O(i, j) do:
    - output Sn(k, l);
    - if Sn(k, l) = 1 then add (k, l) to the LSP and output the sign of ck,l;
    - if Sn(k, l) = 0 then add (k, l) to the end of the LIP;
  - if L(i, j) ≠ 0 then move (i, j) to the end of the LIS, as an entry of type B, and go to Step 2.2.2); otherwise, remove entry (i, j) from the LIS;

2.2.2) if the entry is of type B then
- output Sn(L(i, j));
- if Sn(L(i, j)) = 1 then
  - add each (k, l) ∈ O(i, j) to the end of the LIS as an entry of type A;
  - remove (i, j) from the LIS.

3) Refinement Pass: for each entry (i, j) in the LSP, except those included in the last sorting pass (i.e., with same n), output the nth most significant bit of |ci,j|;

4) Quantization-Step Update: decrement n by 1 and go to Step 2).

This SPIHT algorithm uses the principles of partial ordering by magnitude, set partitioning by significance of magnitudes with respect to a sequence of decreasing thresholds, ordered bit plane transmission, and self-similarity across scale in an image wavelet transform. The algorithm that operates through set partitioning in hierarchical trees (SPIHT) accomplishes completely embedded coding.

4.2 SPIHT and Run-Length Coding Algorithm

After the above-mentioned SPIHT quantification coding, the coding stream appeared many consecutive zero which affect the high compression ratio under the image quality. To pledge the real time of the image transmission and gain the high compression ratio, this paper adopts an improved SPIHT image coding algorithm based on fast lifting wavelet transform and run-length-changeable coding. We can use run-length-changeable coding with the three unused states to gain the high compression ratio. Run-length mark code and coefficient type code are shown in Table 1. The coding scheme is given as follow.

1)Decompose image using fast lifting wavelet algorithm.

2)Find out max absolute value of coefficients and calculate binary bits.

3)Output the threshold T=2B

4)Scanning data according to SPIHT encoding scheme, and then go to 10). If important coefficients, these pixels need not scanned and coded.

5)If the data sets are insignificant pixels or zero data sets, we statistics the count C of consecutive zero, go to 4)

6)If the count C of insignificant pixels or zero data sets is not zero, output the type code and count code corresponding to unimportant coefficient.

7)If the data sets are positive or negative important coefficient, output the type code and magnitude code and got to 9)

8)If insignificant pixels and data, output corresponding type code and clear the value of count C, and then go to 4)

9)Checking to reach transmission rate or not, if not B=B-1 go to 3), otherwise end the SPIHT coding scheme.

Table 1. Type of SPIHT quantification coefficient

Type of coefficient	code
positive important coefficient	101
negative important coefficient	110
significant pixels	100
insignificant pixels and data	000
	001
	010
	011
End code	111

5 Simulation

The following simulation results are obtained from a 512×512 pixel monochrome Lena image having 8 bits per pixel for fixed rate coding, i.e., with run-length coding algorithm. We use 7-level fast lifting-wavelet transform to decompose and reconstruct the image. And the simulation results comparing with SPIHT algorithm are listed in table 2. The wavelet function use 9/7-tap biorthogonal wavelet filters to decompose and reconstruct the image. In the table 2, the PSNR and encoding-decoding time are listed at threshold T=8 with two algorithms, yet the transmission rate hold the line. From the simulation result, the algorithms of this paper increase PSNR and decrease encoding-decoding time.



(a) (b) (c)

Fig. 3. Coding results: (a) the original 512×512 pixel Lena image; (b) the reconstructed Lena image using the algorithm of this paper; (c) the reconstructed Lena image using the SPIHT algorithm

Table 2. Simulation results

Compared index	Algorithm of this paper	SPIHT
PSNR/dB	31.8	31.6
Encoding time/s	2.25	2.76
Decoding time/s	2.12	3.11

6 Conclusion

A new method to improve the performance of the original SPIHT coder is proposed in this paper. This method is based on fast lifting wavelet transform, and combines SPIHT with run-length-changeable coding algorithm. Simulation results show that the proposed method consistently outperforms the original SPIHT coder for the test images. In particular, the proposed method can achieve good performance in real time of the image transmission and the compression ratio.

References

1. Sweldens, W.: The lifting scheme: A construction of second wavelets.SIAM, J.Math.Anal, Vol:29(1996)511–546
2. Kim,K.L., Ra,S.W.:Performance improvement of the SPIHT coder.Signal Processing: Image Communication.19(2004)29–36
3. Taubman,D.: High Performance Scalable Image Compression with EBCOT. IEEE Trans. Image Processing.9(2000)1158–1170.
4. Shapiro,J.M.:Embedded image coding using zerotrees of wavelet coefficients.IEEE Trans. Signal Process.12(1993)31–39
5. Said,A., Pearlman,W.A.: A new, fast, and efficient image codec based on set partitioning in hierarchical trees.IEEE Trans. Circuit Syst. Video Technol.6(1996)243–250.
6. Park,K.H., Lee,C.S.:A multiresolutional coding method based on SPIHT.Signal Processing: Image Communication.17(2002)467–476
7. Kishore,A., Chaitali,C., Tinku,A.: A VLSI architecture for lifting-based forward and inverse wavelet transform.IEEE Trans Actions on Signal Processing.50(2002)966–977

Modified EZW Coding for Stereo Residual*

Han-Suh Koo and Chang-Sung Jeong**

Department of Electronics and Computer Engineering, Korea University,
5-1, Anam-dong, Sungbuk-gu, Seoul 136-701, Korea
{hskoo, csjeong}@korea.ac.kr

Abstract. The objective of this study is to make stereo image coding more efficient by estimating the characteristics of stereo residual. The proposed method is based on the embedded zerotree wavelet algorithm, which is modified to improve its performance for stereo residual images by determining the thresholds during the dominant pass and the repetition times of the subordinate passes dynamically by considering the edge tendency of the input image. The experimental results show that the proposed method provides better performance than the conventional coder for the processing of stereo residuals.

1 Introduction

Most stereo image coding algorithms take advantage of disparity estimation (DE) used for the compression of the monocular image sequences, which includes block matching to predict one image from the other one. The reference image which is one of the images in the stereo pair is coded independently, while the other image which is known as the target image is predicted from the reference image. The residual which is the error between the actual and predicted target images is coded successively using a suitable transform coder. Although the characteristics of disparity compensated stereo residuals are different from those of general images, less research has been devoted to residual image coding with the result that JPEG-like methods including Discrete Cosine Transforms (DCTs) which are used to compress common monocular images are mainly employed.

As for stereo residual image coding, JPEG-like DCT-based schemes are frequently used, because they can be easily combined with existing monocular image coding methods. However, more recently, DWT-based compression schemes have attracted much attention in the image coding research field, because they provide many useful features, such as the progressive property and better performance than DCT-based schemes [1]. Boulgouris and Strintzis attempted to incorporate an embedded coding algorithm into a stereo image coding scheme [2]. Their algorithm employs a disparity compensated prediction process, followed by transform coding on both the reference and the residual (error) images. The Embedded Zerotree Wavelet (EZW) coder presented by Shapiro [3] is one of

* This work was partially supported by the Brain Korea 21 project in 2005 and KIPA-Information Technology Research Center.

** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 845–850, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the standard progressive coder for monocular images. As its term suggests, it is based on progressive (embedded) encoding and designed to use wavelet transforms.

In this paper, we present a new stereo residual image coding algorithm which is based on the EZW algorithm. By taking into consideration the features of stereo residual images, we propose a modified method which determines the thresholds during the dominant pass and the repetition times of the subordinate passes dynamically according to the edge tendencies of the input image.

2 Stereo Residual Coding

2.1 DWT-Based Stereo Residual Coding

The residual, which is the error between the actual and predicted images, is transformed by DWT first, and it then coded using the proposed coder which is referred to as the Modified EZW (MEZW). The characteristics of disparity compensated stereo residuals are different from those of normal images. Moellenhoff and Maier investigated some features of disparity compensated stereo residuals, which are caused by the strictly horizontal apparent motion between the two views [4]. By making use of these properties, they modified the DCT quantization matrix based on the anticipated DCT energy distribution, in order to emphasize the horizontal structure in the image. However, residual images, as well as typical images, have various edge tendencies. Moreover, their distribution is dependent on both the contents of the image and the performance of the disparity compensated prediction module. Nevertheless, the optimal quantization matrix, which is best adapted to the variation in the conditions of the input images caused by both the baseline of the two cameras and the characteristics of the subjects, cannot be determined easily, because the DCT quantization matrix of the input images is fixed and confined within local 8×8 areas. Determining a common matrix applicable to the entire image would be difficult, and finding an individual one for each local area would pose a performance problem within the traditional JPEG framework. In the ordinary block-based coding scheme, DCT coefficients are grouped within a common spatial location and each of them indicates different subbands. However, the DWT coefficients indicate different spatial locations in a common subband. Because the overall distribution of intensity in a residual is easily evaluated using the DWT approach, adaptive schemes which adjust to the image conditions are devised using the DWT rather than the DCT.

2.2 Modified EZW Coding

To satisfy our research goal, which is to develop a scheme which adapts to the image conditions as well as being applicable over the entire image, the EZW coder was chosen and modified to compress the stereo residual effectively by stressing some wavelet subbands containing more important information. After DWT, the first step of the proposed method is to calculate the edge tendencies

of those wavelet subbands containing horizontal and vertical edge information. The edge tendency is defined as $\alpha_l^{(k)} = E_l^{(k)}/(E_l^{(1)} + E_l^{(2)})$, where l and k are the level and orientation of the wavelet subband, respectively; and $E_l^{(k)}$ refers to the intensity energy of the corresponding subband, that is, orientations 1, 2, and 3 denote the Low/High (LH), High/Low (HL), and High/High (HH) subbands, respectively. In these notations, subband LH represents the horizontal edges and subband HL represents the vertical edges. Subband HH, which is located in the lower right-hand corner of the wavelet decomposed image, contains the information pertaining to the diagonal edges. The thresholds $t_l^{(k)}$ for each level and orientation are calculated as follows:

$$\begin{aligned} t_l^{(k)} &\Leftarrow (0.5 - A|\alpha_l^{(k)} - 0.5|) \times t_l^{(k)}, \text{for } (\alpha_l^{(k)} > 0.5) \\ t_l^{(k)} &\Leftarrow 0.5 \times t_l^{(k)}, \text{for } (\alpha_l^{(k)} \leq 0.5) \end{aligned} \tag{1}$$

where A is a scaling factor used to control the threshold and is set to 1 in this scheme. By means of (1), the thresholds are updated in each loop of the MEZW, according to the edge tendency. If a subband LH (or HL) has a higher edge tendency than the counterpart subband HL (or LH), the threshold of subband LH (or HL) is lowered in comparison with that of the conventional EZW. Otherwise, the threshold is set to one half of its value in the previous loop.

In the first threshold determination, the largest coefficient magnitude in the residual image is employed as the seed value, which is used as $t_l^{(k)}$ on the right-hand side of (1). If a coefficient is higher than the threshold, its information is added to the subordinate list to increase the encoding accuracy. Through this process, the thresholds are estimated adaptively in proportion to the edge strength, and the coefficients in those subbands having important information can be coded preferentially. In the dominant pass, the POS, NEG, IZ, and ZTR symbols are generated in a similar way to that in the conventional EZW. However, the important coefficients are encoded preferentially in the bit stream, by increasing the frequency of the significant symbols in those subbands which have a strong edge tendency. After the individual calculation of the threshold for each subband during the dominant pass, the uncertainty intervals of the important subbands are enlarged and some significant values in the subordinate list are degraded in consequence, as compared with conventional EZW coding. To compensate for this quality degradation, the subordinate pass is repeated only for symbols from the subband having a strong edge tendency until the uncertainty interval becomes shorter than that of the opposite subband in the wavelet tree.

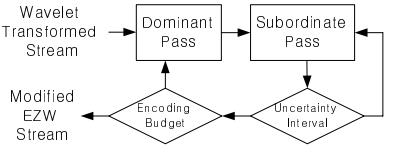


Fig. 1. Main loop of proposed stereo residual image encoding

This combination of a dominant pass and subordinate passes constitutes the main loop of the proposed method, which is shown in Fig. 1, and this process continues until the predefined encoding budget matches. Before entering a new loop, the thresholds are modified using (1). The thresholds for each subband become different, because the edge tendencies differ from each other. After the encoding process, the disparity vector and residual data including the edge tendencies of one orientation are transmitted to the decoder along with the information of the first view.

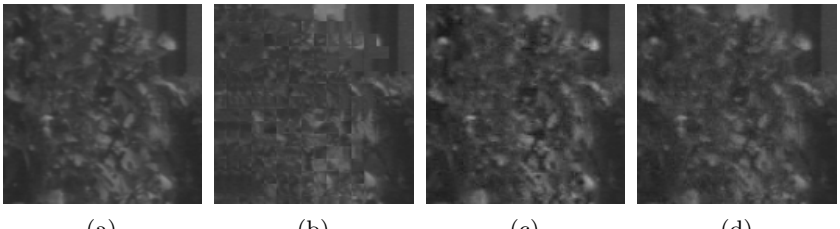
3 Experimental Results

Our algorithm was tested extensively with various stereo test sets shown in Fig. 2. Since the focus of this paper is to develop an algorithm for stereo residual coding based on EZW, we compared the performance of the proposed method with that of the EZW coder. In our experiments, the conventional EZW algorithm was used to code the stereo residual after DE, and its performance was compared with that of the proposed method. Experiments were carried with some well-known DWT transforms, such as the Haar and Adelson's transforms [5]. The 9/7-F transform [6] by Antonini *et al.* was also tested, because it was reported to yield good performance in the case of lossy compression [7]. The wavelet decomposition



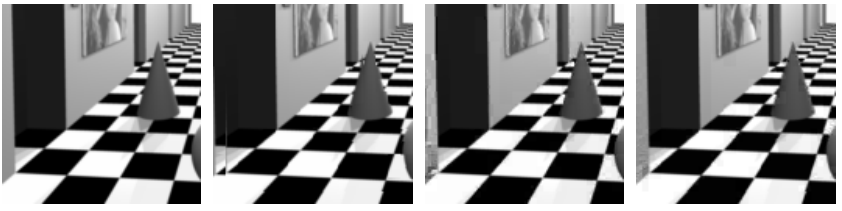
(a) (b) (c) (d)

Fig. 2. Stereo test sets. (a) Pentagon (courtesy of CMU) (b) Parking meter (courtesy of CMU) (c) Corridor (courtesy of Computer Vision Group at the University of Bonn) (d) Sawtooth (downloaded from `http://www.middlebury.edu/stereo`).



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

Fig. 3. Experimental results with Parking meter stereo pair at 0.35 bpp. (a) A part of the original target image (b) Decoded image using DE (c) Decoded image using EZW-9/7-F coder (d) Decoded image using MEZW-9/7-F coder.



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

Fig. 4. Experimental results with Corridor stereo pair at 0.35 bpp. (a) A part of the original target image (b) Decoded image using DE (c) Decoded image using EZW-Haar coder (d) Decoded image using MEZW-Haar coder.

Table 1. Performance comparison of the proposed method and EZW (in dB) with Haar (Hr), Adelson's (Ad), and 9/7-F (An) transforms

Image		Pentagon							Parking meter						
DE		26.45							33.67						
Bit rate		0.1	0.15	0.2	0.25	0.3	0.35	Av.	0.1	0.15	0.2	0.25	0.3	0.35	Av.
EZW	Hr	0.14	1.16	1.56	2.10	2.48	2.67	1.75	0.36	2.11	3.33	3.34	3.72	4.15	2.93
	Ad	0.15	1.15	1.55	2.08	2.45	2.56	1.73	0.38	2.37	3.57	3.55	3.95	4.38	3.14
	An	0.15	1.15	1.55	2.13	2.45	2.66	1.75	0.38	2.37	3.22	3.53	3.96	4.42	3.09
MEZW	Hr	0.15	1.41	1.80	2.16	2.82	3.42	1.99	0.50	2.91	3.74	4.63	5.56	5.97	3.97
	Ad	0.15	1.37	1.87	2.21	2.56	3.14	1.92	0.50	3.18	3.96	3.94	5.79	6.21	4.20
	An	0.15	1.37	1.83	2.17	2.56	3.25	1.91	0.54	3.10	4.01	4.99	5.83	6.23	4.23
Image		Corridor							Sawtooth						
DE		21.79							27.32						
Bit rate		0.1	0.15	0.2	0.25	0.3	0.35	Av.	0.1	0.15	0.2	0.25	0.3	0.35	Av.
EZW	Hr	6.18	7.60	8.64	8.83	9.12	9.65	8.40	0.98	2.23	3.30	3.50	3.78	4.16	3.04
	Ad	3.92	4.64	4.91	5.31	5.42	5.56	5.03	0.93	2.23	3.28	3.56	3.89	4.27	3.09
	An	3.97	4.57	5.13	5.41	5.51	5.67	5.09	0.93	2.23	3.38	3.61	3.34	4.37	3.13
MEZW	Hr	7.17	9.29	10.46	11.02	11.48	12.04	10.33	1.17	2.84	3.56	4.88	5.30	5.61	3.97
	Ad	7.21	8.61	9.27	10.21	10.51	11.05	9.52	1.08	2.72	3.93	4.90	5.36	5.71	4.05
	An	7.28	8.71	9.64	10.22	10.58	11.14	9.61	1.11	2.85	3.62	5.05	5.53	5.92	4.10

levels used were 4 for the *Pentagon*, *Parking meter*, and *Sawtooth* pairs and 3 for the *Corridor* pair. The quality of the reconstructed image is evaluated by the PSNR improvement after DE under various bit rates for each divided 8×8 block in the target image. The search range is confined to an area of 31×3 pixels.

For the subjective performance comparison, a part of the *Parking meter* stereo image which is the left part of the hedge, and its results under various conditions are presented in Fig. 3. As shown in these results, the detail of the leaves was depicted more softly as compared with the conventional method. Figure 4 is presented for the purpose of evaluating the performance of the occlusion description. The test set that is used is the lower left patch of the *Corridor* image. The gray wall which appears on the left side is completely occluded in the reference image.

As shown in Fig. 4(b), the left wall is hardly reconstructed at all by DE, due to the lack of proper textures in the search range of the reference image. However, the occluded regions are decoded well with the proposed method. Objective evaluations are given in Table 1. The table presents the PSNR values for DE and the PSNR improvements afforded by residual coding for each test set. The values are evaluated at short intervals with various DWT transformed residual images. The average (Av.) PSNR values clearly show that the stereo residual coding performance of the proposed method is better than that of the EZW coder. Another observation can be made concerning the relation between the type of DWT transforms and the quality of the reconstruction. The 9/7-F transform is known to provide good performance in lossy compression, because of its regularity and smoothing properties. Therefore, it is popular for general image coding. However, it does not guarantee the best performance for stereo residuals whose characteristics are quite different from those of general images. Through these experimental results, we can see that shorter filters such as the Haar transform are better for cases in which the disparity is poorly estimated such as the *Corridor* pair, because there are likely to be many apparent narrow edges in the residual. On the other hand, longer filters such as the 9/7-F transform perform better in the case of well estimated residual images such as the *Parking meter* pair, because the intensity variation is plainer, as in the case of general images.

4 Conclusion

In this paper, we proposed a modified EZW algorithm for coding stereo residual images. To emphasize the more significant information contained in the stereo residual, the proposed method modifies the coding sequence according to the edge tendency of the residual image.

References

1. Xiong, Z., Ramchandram, K., Orchard, M. T., and Zhang, Y. Q.: A Comparative Study of DCT- and Wavelet-Based Image Coding. IEEE T. Circ. Syst. Vid. **9** (1999) 692–695
2. Boulgouris, N. V., Strintzis, M. G.: Embedded Coding of Stereo Images. Proc. IEEE Int. Conf. Image Process. **3** (2000) 640–643
3. Shapiro, J. M.: Embedded Image Coding Using Zerotrees of Wavelet Coefficients. IEEE T. Signal Process. **41** (1993) 3445–3462
4. Moellenhoff, M. S., Maier, M. W.: Transform Coding of Stereo Image Residuals. IEEE T. Image Process. **7** (1998) 804–812
5. Adelson, E. A., Simoncelli, E., and Hingorani, R.: Orthogonal pyramid transforms for image coding. Proc. SPIE, **845** (1987) 50—58
6. Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I.: Image coding using wavelet transform. IEEE T. Image Process. **1** (1992) 205–220
7. Adams, M. D., Kossentini, F.: Reversible interger-to-integer wavelet transform for image compression: performance evaluation and analysis. IEEE T. Image Process. **9** (2000) 1010–1024

Optimal Prototype Filters for Near-Perfect-Reconstruction Cosine-Modulated Filter Banks[†]

Xuemei Xie, Guangming Shi, and Xuyang Chen

School of Electronic Engineering, Xidian University, Xi'an, P.R. China
xmxie@see.xidian.edu.cn

Abstract. In this paper, we propose a simple method for designing *M*-channel near-prefect-reconstruction (NPR) cosine-modulated filter banks (FBs). By employing the Parks-McClellan algorithm in constructing the prototype filter, an ideal magnitude response of the filter is achieved. Furthermore, the transition band of the prototype filter is constrained in such a way that it follows a cosine function. As a result, the FBs are approximately power complementary and therefore possess the NPR property. There are two main advantages in our proposed method. The first one is that no objective function in the optimization is explicitly required and the second one is that the resulting prototype filter is a global optimal solution. Compared with the traditional design method using general optimization methodology, the proposed method here is very simple and efficient. In addition, the stopband attenuation of the resulting FB is significantly higher than those offered by the traditional methods.

1 Introduction

Cosine-modulated filter bank (CMFB) is attractive due to its low complexities of design and implementation, and high filter quality. In many practical applications, such as subband lossy coding and subband quantization, some degree of distortions and certain amount of aliasing error are in general allowed as long as their effects are smaller than that introduced by the coding itself. Thus, the near-prefect-reconstruction (NPR) in these cases is sufficient and even desirable.

There are many methods available for the designs of NPR CMFBs. Nonlinear constrained optimization schemes are quite popular in the design of NPR CMFBs. It is evident that all the methods in [1]-[6] rely on a kind of nonlinear optimization. However, optimization procedures are computationally intensive, and worst of all, the cost functions have generally many local minima. Thus some efforts have been made aiming at overcoming the above difficulties, for instance, by reducing the number of the optimized parameters. A design method is presented in [5] in which equiripple FBs are obtained by the Parks-McClellan algorithm with a single parameter on a convex error surface being optimized. A technique on the NPR FBs design is frequency sampling approach [6] where the only optimized parameter is a set of magnitude response values of samples in the transition band of the prototype filter.

[†] Work supported by National Natural Science Foundation of China. (Grand No. 60372047).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 851 – 856, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we propose a simple method for designing *M*-channel NPR CMFBs. We employ the Parks-McClellan algorithm to design the prototype filter. The desired magnitude response of the filter is specified as an optimization objective. Rather than resulting in a local optimal by nonlinear optimization methods, a global optimal solution is obtained. In addition, the attenuation at the stopband of the filter can be significantly high. By following a cosine roll-off characteristic of the transition band, the FB is approximately power complementary and hence has the NPR property. The example has shown that the performance of the NPR FBs by the proposed approach is much better than that due to conventional method with an assumption of the same distortion level. Most importantly, the given example demonstrates that the proposed approach is very simple, effective, and efficient.

2 Cosine-Modulated Filter Banks with NPR Property

In the general structure of *M*-channel maximally decimated FBs, the input-output relationship of this system can be expressed as

$$\hat{X}(z) = \sum_{l=0}^{M-1} A_l(z) X(zW_M^l), \quad 0 \le l \le M-1 \ . \tag{1}$$

where

$$A_l(z) = \frac{1}{M} \sum_{k=0}^{M-1} H_k(zW^l) F_k(z), \tag{2}$$

and $W_M = e^{-j2\pi/M}$. It can be seen that the term $X(zW^l)$ is a shifted version of $X(z)$ for all l except for $l=0$. We call the terms, $X(zW^l)$ and $A_l(z)$, $l>0$, aliasing components and the alias transfer functions, respectively.

Without considering the aliasing distortion, the input-output relationship of the system can be described as

$$\hat{X}(z) = T(z)X(z), \tag{3}$$

where

$$T(z) = A_0(z) = \frac{1}{M} \sum_{k=0}^{M-1} H_k(z) F_k(z). \tag{4}$$

$T(z)$ is referred to the distortion transfer function. It determines the distortion level of the system with respect to the non-aliased component $X(z)$. In one word, without imposing any constraint on the analysis and synthesis filters the output signal $\hat{X}(z)$ would be corrupted by three distortions: aliasing, phase and magnitude distortions.

Two quantities are used to characterize the reconstruction degree of the FB precisely. The notation E_{pp} is used to measure the worst possible amplitude distortion, and the notation E_a measures the worst possible peak aliasing distortion.

In a CMFB, $H_k(z)$ and $F_k(z)$, are generated by cosine modulating a prototype filter $P_0(z)$ with an order of *N*. Their corresponding impulse responses are given by

$$h_k(n) = 2p_0(n)\cos\left(\frac{\pi}{M}(k+0.5)\cdot\left(n-\frac{N}{2}\right)+(-1)^k\frac{\pi}{4}\right),$$

$$f_k(n) = 2p_0(n)\cos(\frac{\pi}{M}(k+0.5)\cdot(n-\frac{N}{2}) - (-1)^k\frac{\pi}{4}), \quad (5)$$

$$0 \le n \le N,\ 0 \le k \le M-1,$$

where N is the length of each filter. If the prototype filter is linear-phase, the distortion transfer function $T(z)$ is also linear-phase. This can be easily seen from the following analysis. In the orthogonal CMFB system, $f_k(n)$ is the time reversed version of $h_k(n)$ with N samples delay, that is,

$$f_k(n) = h_k(N-n). \quad (6)$$

Or equivalently,

$$F_k(z) = z^{-N}H_k(z^{-1}) \quad (7)$$

Substituting Eq. (7) into (4), $T(z)$ becomes

$$T(z) = \frac{1}{M}\sum_{k=0}^{M-1} z^{-N}H_k(z)H_k(z^{-1}). \quad (8)$$

Or in the frequency domain

$$MT(e^{j\omega}) = e^{-j\omega N}\cdot\sum_{k=0}^{M-1}|H_k(e^{j\omega})|^2, \quad (9)$$

Which indicates that $T(z)$ is linear phase. In the CMFB, due to the cosine modulation the alias between adjacent channels can be effectively canceled.

Now, phase distortion of the reconstructed signal is free and the aliasing error is reduced by nature. Only amplitude distortion needs to be considered. Next, we will discuss how to design the prototype filter so as to minimize the remaining distortion.

3 Design of Prototype Filter

To begin with, let us consider Eq. (9) carefully. Firstly we notice that if the stopband attenuation of $H_k(z)$ is sufficiently high, $|T(e^{j\omega})|$ will be as nearly flat as the magnitude response of the individual $H_k(z)$ in the region with ω belonging to the passband of all $H_k(z)$. Secondly, if we are able to make sure that the sum of the square amplitude of two adjacent $H_k(z)$ and $H_{k+1}(z)$ in their transition band is closed to unity, then there would be no 'bump' or 'dip' around the transition frequency $(k+1)\frac{\pi}{M}$ and $-(k+1)\frac{\pi}{M}$, implying that $|T(e^{j\omega})|$ in those regions is approximately flat. Recall that, in the CMFB, the analysis and synthesis filters are the shifted versions of the prototype filter. Therefore, to meet the requirements described above, it is sufficient to impose the following constraint on the prototype filter $P_0(z)$,

$$|P_0(e^{j\omega})|^2 + |P_0(e^{j(\omega-\frac{\pi}{M})})|^2 = 1, \text{ for } 0 < \omega < \pi/M. \quad (10)$$

We notice that the above requirement in the transition band would be met if we force the response of $P_0(z)$ in the right (left) transition interval to follow a cosine (sine) function of $\theta(\omega)$, such that $\cos^2\theta + \sin^2\theta = 1$, where $0 \le \theta(\omega) \le \pi/2$. Here $\theta(\omega)$ is a linear function of ω.

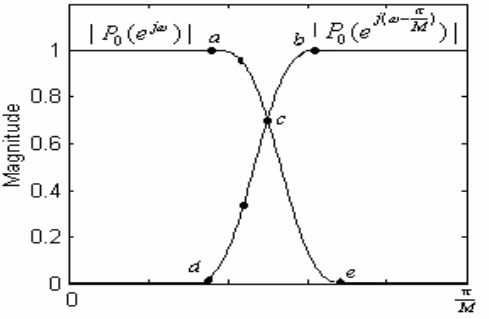


Fig. 1. The magnitude response of prototype filter $P_0(z)$ and its shifted version

Fig. 1 illustrates the desired responses of $|P_0(e^{j\omega})|$ and $|P_0(e^{j(\omega-\pi/M)})|$, where it is assumed that the response in the passband takes a constant value of unity and the response in the stoband a constant value of zero. Points *a* and *e* indicate the passband and stopband edges of $|P_0(e^{j\omega})|$, respectively, denoted by ω_p and ω_s. Evidently, points *b* and *d* associated with $|P_0(e^{j(\omega-\pi/M)})|$ are respectively the counterparts of points *a* and *e,* denoted by ω_p' and ω_s'. Point *c* is the intersection point of the two filters. It can be easily verified that (i) $|P_0(e^{j\omega})|$ and $|P_0(e^{j(\omega-\pi/M)})|$ are symmetric with respect to point *c*; and (ii) point *c* corresponds to $\omega_c = \pi/2M$ which is the cutoff frequency of $P_0(z)$. Thus ω_p should be equal to ω_s'. Similarly, ω_p' should be equal to ω_s. Denote $A[\theta(\omega)]$ and $B[\theta(\omega)]$ as the magnitude responses of $|P_0(e^{j\omega})|$ and $|P_0(e^{j(\omega-\pi/M)})|$ in the transition band interval $\omega_p \le \omega \le \omega_s$. $A[\theta(\omega)]$ and $B[\theta(\omega)]$ are chosen to meet all the constraints and requirements described above. To be more specific, we have,

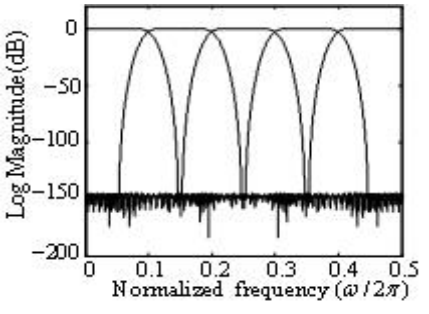
$$A[\theta(\omega)] = \cos[\theta(\omega)],\ B[\theta(\omega)] = \sin[\theta(\omega)], \tag{11}$$

where $\theta(\omega) = \dfrac{\omega - \omega_p}{\omega_s - \omega_p} \cdot \dfrac{\pi}{2}$, $\omega_p \le \omega \le \omega_s$. As $\theta(\omega)$ varies from 0 to $\pi/2$ in the transition band, $A[\theta(\omega)]^2 + B[\theta(\omega)]^2 \equiv 1$. Also giving $A[\theta(\omega_c)] = B[\theta(\omega_c)] = 1/\sqrt{2}$ as expected, $\theta(\omega_c) = \theta((\omega_p + \omega_s)/2) = \pi/4$ at ω_c. Finally, we notice that the transition band of $|P_0(e^{j(\omega-\pi/M)})|$ discussed above is exactly similar with that of $|P_0(e^{j\omega})|$ in the negative frequency side. If the transition band of the prototype filter $P_0(z)$ satisfies Eq. (11), the analysis and synthesis filters also satisfy this relation in their transition band and possess the desired property.

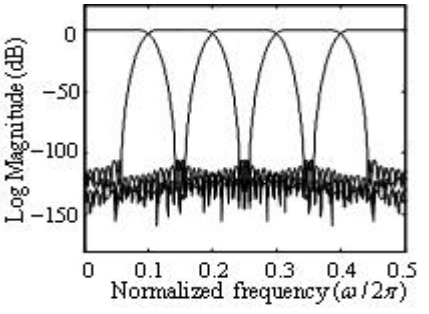
In our design, we employ the Parks-McClellan algorithm to obtain the prototype filter. In general, the filter quality by using this method is of optimum, leading to the filters with low order. More importantly, $|P_0(e^{j\omega})|$ are completely specified and are ready to be used as the design target. The Parks-McClellan algorithm for the filter design is also very useful in constructing nonuniform FBs and *M*-uniform linear phase FBs. Interested readers please refer to [7].

4 Design Example

In this section, we present an example of a 5-channel CMFB. The length of the prototype filter is 110 and the stopband cutoff frequency is $\omega_s = 0.195\pi(rad)$. By using the proposed method and applying the above design procedures, we get the desired FB. Fig. 2(a) shows the magnitude responses of the analysis filters. The stopband attenuation level is about 147.1 dB, E_a is 2.224×10^{-8}, and E_{pp} is 1.734×10^{-3} (Due to page limitation, the figures of amplitude distortion and aliasing error are omitted).



(a) Magnitude response of the analysis filter with our proposed method



(b) Magnitude responses of the analysis filters with nonlinear optimization

Fig. 2. 5-channel NPR CMFB

For comparison, we designed a 5-channel NPR CMFB by the conventional nonlinear optimization method. The frequency responses of this FB are shown in Fig. 2(b). In this example, we got $E_a = 3.111\times10^{-7}$ and $E_{pp} = 1.978\times10^{-4}$. The stopband attenuation level is about 106 dB.

The results have indicated that our proposed method is capable of achieving significant larger attenuation in the stopband under the similar distortion level.

5 Conclusions

In this paper, we propose a method for designing *M*-channel NPR CMFBs. We employ the Parks-McClellan algorithm in the design of the prototype filter. By forcing the transition band to follow a cosine function, we can obtain the NPR FBs.

Compared with the traditional nonlinear optimization method, our proposed method is much simpler, more effective and efficient, and especially can obtain larger attenuation in the stopband.

References

1. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs, New Jersey (1992)
2. Truong, T.Q.: Near-prefect-reconstruction pseudo-QMF banks. IEEE Trans. Signal Processing, Vol. 42. No. 1. January (1994) 65-75
3. Rossi, M., Zhang, J.Y., and Steenaart, W.: Iterative constrained least squares design of near perfect reconstruction pseudo QMF banks. Lecture Notes in Proc. IEEE CCECE, Vol. 2. November (1996) 766-769
4. Cruz-Roldan, F., Amo-Lopez, P., Maldonado, S., and Lawson, S.S.: An efficient and simple method for designing prototype filters for cosine-modulated pseudo-QMF banks. IEEE Signal Processing Letter, Vol. 9. No. 1. January (2002) 29-31
5. Creusere, C.D., and Mitra, S.K.: A simple method for designing high quality prototype filters for M-band pseudo-QMF banks. IEEE Trans. Signal Processing, Vol. 43. No. 4. April (1995) 1005-1007
6. Cruz-Roldan, F., Santamaria, I., and Bravo, A.M.: Frequency sampling design of prototype filters for nearly perfect reconstruction cosine-modulated filter banks. IEEE Signal Processing Letter, Vol. 11. No. 3. March (2004) 397-399
7. Xie, X.M., and Chan, S.C.: A new method for designing linear-phase recombination nonuniform filter-banks. Lecture Notes in The 47th Midwest Symposium on Circuits and Systems, Vol. 2. July (2004) 101-104

Fast Motion Estimation Scheme for Real Time Multimedia Streaming with H.264

Chan Lim[1], Hyun-Soo Kang[2], and Tae-Yong Kim[3]

[1] KTF Technologies Co. Ltd., Seong-nam, Korea
mj23cb@ktftech.com
[2] School of ECE, ChungBuk National University, Chungju, Korea
hskang@chungbuk.ac.kr
[3] Graduate School of AIM, Chung-Ang University, Seoul, Korea
kimty@cau.ac.kr

Abstract. In this paper, we propose a novel fast motion estimation algorithm based on successive elimination algorithm (SEA), which can dramatically reduce complexity of the variable block size motion estimation by removing the unnecessary computation of SAD in H.264 encoder. The proposed method, which accumulates current sum norms and pre-computed SAD for the bigger block sizes than 4×4 drives tighter bound in the inequality than an ordinary SEA depending on the availability of SAD. Experimental results explain that our method reduces computation complexity. In addition, the proposed method is an extended version of the rate constrained block matching for variable block-sized applications. It surely works on variable block-based motion estimation with just a little degradation.

1 Introduction

The purpose of block matching algorithm (BMA) is to search the optimal reference block in the previous image for a block in a current image. To do this, SAD is widely used as a criterion for finding the most correlated block. The variable block-based motion estimation (VBME) of H.264 supports block matching of the variable sizes, i.e. from 16×16 to 4×4. However, the use of hierarchically structured block modes causes complexity problem in spite of high compression efficiency. Furthermore, applying full search algorithm (FSA) to VBME makes the computational cost heavier. Therefore fast full search algorithm (FFSA), in which SADs of other six modes are computed by reusing pre-computed 4×4 SADs was used to eliminate the redundant SAD computation in JM 7.3 of H.264. Nevertheless, variable block approach requires more cost than fixed size approach in computation for block matching [1]. Successive elimination algorithm (SEA) was proposed as a solution to reduce computational cost by removing unnecessary computation [2][3][4].

In this paper, we propose a fast search algorithm using SEA, which may be useful to decide the optimal mode among seven variable blocks. And a rate-distortion (R-D) optimized BMA was introduced in [5] where the cost function considers both of SAD and the number of bits used in motion vectors.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 857–862, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Rate-Constrained VBME with SEA

SEA among many fast algorithms is quite an efficient method to reduce computational complexity by removing unnecessary blocks for SAD computation with a necessary condition for a reference block to be the optimal block. The necessary condition is well known as Minkowski's inequality, which identifies the relationship between SAD and sum norm. The inequality in [3] for the optimal motion vector to have to satisfy the necessary condition is as follows

$$R^i - SAD^i_{min}(m,n) \leq M^i(x,y) \leq R^i + SAD^i_{min}(m,n), \tag{1}$$

where i is each mode, R and $M(x,y)$ are sum norms of a block in the current frame and in the previous frame, and $SAD_{min}(m,n)$ denotes the minimal SAD among SADs of the candidate blocks before the current search point. (x,y) and (m,n) are the motion vectors of each position.

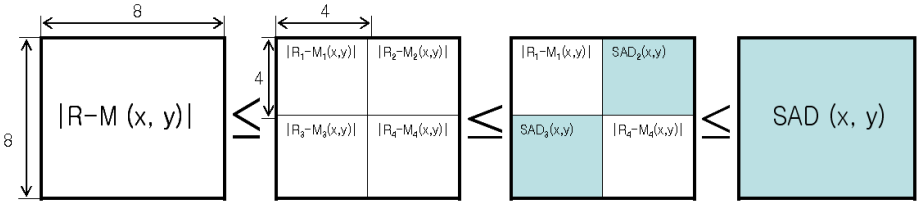


Fig. 1. The comparison of the bound of 8×8 block by the accumulation of SAD and sum norm for 4×4 blocks

In this paper, we propose a method to exclude the blocks for which SAD computation is not needed by the inequality, which compares the minimal SAD with SAD or sum norm of a block. In case that the inequality in [4] is applied to 4×4 blocks, an 8×8 sub-block satisfies the first inequality of Eq. (2).

$$\begin{aligned} |\sum_{k=1}^{4} R_k - M_k(x,y)| \leq \sum_{k=1}^{4} |R_k - M_k(x,y)| \leq |R_1 - M_1(x,y)| + SAD_2(x,y) \\ + SAD_3(x,y) + |R_4 - M_4(x,y)| \leq \sum_{k=1}^{4} SAD_k(x,y) = SAD(x,y) \end{aligned} \tag{2}$$

where k is the index of each 4×4 block. The first inequality of Fig. 1 and Eq. (2) means that the sum of the sum norms of four 4×4 sub-blocks to which an 8×8 block is partitioned is larger than the sum norm of the 8×8 block. That is to say, the sum of sum norms of the split blocks makes bound of SEA tighter than the sum norm of the original block before splitting. Undoubtedly, being split smaller, the bound is getting tighter by the sum of the partial sum norms because the sum of sum norms of a pixel by a pixel is SAD of the block in case of being split by pixels. Therefore, as presented in the third term of Eq. (2), if pre-computed SADs are partially reused, we can get almost SAD for the tight

bound of SEA without SAD computation for the mode (larger blocks than 4×4) and it is feasible that unnecessary SAD computation is reduced by the inequality of SEA. Fig. 1 shows the relationship of SADs according to the ways of block partitioning and the selections of SAD or sum norm. Furthermore, suppose that the reuse of SADs of the second and the third 4×4 blocks, denoted by $SAD_2(x, y)$ and $SAD_3(x, y)$, are available by the mean to make the bound of 8×8 tighter.

So far we described the fast algorithm considering only distortion regardless of the rate. However, we can minimize the distortion for a given rate constraint. On consideration of the rate, a cost function with Lagrange multiplier is as follows.

$$J^i(x, y, \lambda) = SAD^i(x, y) + \lambda r^i(x - p, y - q), \tag{3}$$

where λ is Lagrange multiplier, (x, y) is a motion vector, and $r(x-p, y-q)$ is the number of bits associated to the difference between the motion vector and the predicted vector (p, q). By means of Eq. (3), we can achieve the rate-constrained block matching for variable sized blocks.

The following procedure presents the method to apply bit rate with Lagrange multiplier instead of only SAD for cost functions of 4×4. Assume that search point index is given by a look-up table determined by searching strategy like the spiral search. The optimal motion vectors for each mode require repeating the procedure.

- Step1. When search point index equals 0, compute the cost function $J(0, 0, \lambda)_{4\times4}$ for each index of 4×4 block. (where, $J(0, 0, \lambda)_{4\times4}$ is $J(m, n, \lambda)_{min4\times4}$.
- Step2. When search point index is larger than 0 (that is, position vector is (x, y)), if SAD for the block index was computed, compare $SAD(x, y)_{4\times4}$ with $J(m, n, \lambda)_{min4\times4}$ according to each index.
  - 2.1. If $SAD(x, y)_{4\times4} \geq J(m, n, \lambda)_{min4\times4}$, go to Step 5.
  - 2.2. If $SAD(x, y)_{4\times4} < J(m, n, \lambda)_{min4\times4}$, go to Step 3.
- Step3. Compute $J(x, y, \lambda)_{4\times4}$ and compare $J(x, y, \lambda)_{4\times4}$ with $J(m, n, \lambda)_{min4\times4}$.
  - 3.1. If $J(x, y, \lambda)_{4\times4} \geq J(m, n, \lambda)_{min4\times4}$, go to Step 5.
  - 3.2. If $J(x, y, \lambda)_{4\times4} < J(m, n, \lambda)_{min4\times4}$, go to Step 4.
- Step4. Replace $J(m, n, \lambda)_{min4\times4}$ with $J(x, y, \lambda)_{4\times4}$ as the minimum cost function from the initial position to the current position (x, y) for 4×4 blocks.
- Step5. Maintain the minimum cost function of previous position as that of current position for 4×4 blocks.

3 Experimental Results

For the performance evaluation, we compare the proposed method with FFSA in JM. We assume ADD (addition), SUB (subtraction), and COMP (comparison) as one cycle operation and ABS (absolute value) as two cycle operations considering 2's complement operations. The experiment was performed

for four QCIF (176×144) standard image sequences. The input image sequences of 100 frames were used and only one previous frame was referenced as a reference frame. The total number of search points for motion estimation was $(2N+1)\times(2N+1) = 1089$ for $N = 16$. Moreover, we calculate bit rate (kbit/s) and PSNR (dB) of the test images according to each quantization parameter 28, 32, 36, and 40 with the input image sequences of 150 frames. The frame structure is IPPP and the frame rate is 15. The simulation considers only H.264 baseline profile that does not contain B slice nor CABAC, and the option of R-D optimization and Hadamard transform were turned on.

Table 1 shows the ratio of the number of 4×4 blocks with SAD computation to the total number of 16×16 MBs, without and with pre-computed $SAD_{4\times4}$, according to each mode and test images. As presented in Table 1, the difference of the computational costs between NOT and REUSED for pre-computed $SAD_{4\times4}$ is remarkable except 4×4 mode. That supports that the proposed algorithm reduces unnecessary SAD computation by the tight bound of the larger modes than 4×4 with reusing pre-computed $SAD_{4\times4}$ in Minkowsi's inequality as shown in Fig. 1.

Table 2 shows the computational complexity of SAD and sum norm for a 4×4 block. For the computation of M in Table 2, we refer to [3] where it is

Table 1. Without and with reusing pre-computed $SAD_{4\times4}$, the comparison for the average ratio of the number of each mode with SAD computation to the total number of each block-size per 16×16 MB

mode	Foreman		Coastguard		Stefan		Table tennis	
$SAD_{4\times4}$	not	reused	not	reused	not	reused	not	reused
4×4	7.828	7.828	17.205	17.205	13.427	13.427	19.183	19.183
4×8	3.662	0.457	10.734	0.364	7.701	0.587	14.327	0.665
8×4	3.936	0.421	13.666	0.369	9.333	0.614	14.077	0.606
8×8	2.195	0.296	8.625	0.239	5.379	0.435	10.906	0.446
8×16	1.271	0.284	5.461	0.234	3.291	0.38	8.646	0.459
16×8	1.361	0.281	6.595	0.228	3.894	0.384	8.331	0.407
16×16	0.918	0.24	4.22	0.222	2.614	0.331	6.878	0.439
$4\times4_{total}$	10.694	8.364	24.372	17.483	19.273	14.067	29.23	20.109

Table 2. The comparison of the computational cost of SAD and that of sum norm per 4×4 block: the computational cost for each step was represented in this table

		$SAD_{4\times4}$	$\lvert R-M\rvert_{4\times4}$			
Search range		0 ∼ 1088	0	1 ∼ 1088	1	2 ∼ 1088
ADD(+)	SAD	15				
	R		15			
	M				15	4
SUB(-)		16		1		4
ABS($\lvert\cdot\rvert$)		16		1		4

Table 3. The comparison of the computational cost of fast full search algorithm of JM 7.3 and that of the proposed algorithm for all test image sequences (1) $|R-M|$ (2) COMP ($|R-M|$ vs. SAD_{min}) (3) $SAD_{4\times4}$ (4) $SAD+|R-M|$ (5) $COMP(SAD+|R-M|$ vs. $SAD_{min})$

	Common		Individual (Proposed + SEA)								
Image	All		Foreman		Coastguard		Stefan		Table tennis		FFSA
Operation	(1)	(2)	(3)	(4) +(5)	(3)	(4) +(5)	(3)	(4) +(5)	(3)	(4) +(5)	(JM7.3)
$SAD_{4\times4}$		1008	91854		191709		154287		220563		1097712
$\|R-M\|_{4\times4}$	155872										
$4\times4(+)$		17408		1353		2957		2305		3347	
$4\times8(+)$	8712	8704		80		64		102		116	8712
$8\times4(+)$	8712	8704		74		64		108		106	8712
$8\times8(+)$	4356	4352		26		20		38		38	4356
$8\times16(+)$	2178	2176		12		10		16		20	2178
$16\times8(+)$	2178	2176		12		10		16		18	2178
$16\times16(+)$	1089	1088		6		4		8		10	1089
Sum	228713		93417		194818		156878		224218		1124973
Common+Individual			322130		423531		385591		452931		
Proposedx100/FFSA(%)			28.6		37.6		34.3		40.3		100

Table 4. The comparison of JM 7.3 and the proposed algorithm for bit rate and PSNR

	QP	Foreman		Coastguard		Stefan		Table tennis	
		JM	Prop.	JM	Prop.	JM	Prop.	JM	Prop.
Bit Rate (kbps)	28	174.75	175.42	280.66	280.53	582.13	582.08	161.88	162.04
	32	104.20	104.76	129.48	129.92	335.12	332.33	95.73	95.59
	36	63.91	64.28	58.82	58.62	185.13	185.19	56.43	56.57
	40	39.73	40.12	27.99	28.40	101.62	103.71	35.20	35.18
PSNR (dB)	28	35.79	35.78	34.11	34.11	34.26	34.26	35.89	35.86
	32	32.99	32.99	31.07	31.06	30.66	30.66	33.25	33.21
	36	30.43	30.42	28.48	28.46	27.46	27.47	30.55	30.57
	40	27.94	27.91	26.17	26.17	24.63	24.63	28.17	28.16

shown that the sum norms of reference blocks can be computed with just a little computation since the previously obtained sum norm of a neighbor block is used to compute a current block's sum norm. Adopting [3], complexity of M is 8 cycles on average. Provided that there are two 4×4 blocks, which are horizontally overlapped by 3 pixels, the block on the right requires four addition operations and four subtraction operations. Table 3 shows the computational complexity of FFSA in JM7.3 and the complexity of the five operation processes in the proposed algorithm for all test image sequences, respectively. The computational complexity of the proposed algorithm comes from the operation frequency for five main processes as presented in Table 3. We can get the outstanding result from the table. The ratio for FFSA denotes the complexity of our method when we consider FFSA as 1. As presented in Table 3, the proposed algorithm has gain of 60∼70% in complexity with the same optimal motion vector as FFSA has.

Table 4 shows bit rate and PSNR of JM and the proposed algorithm for the test images. As presented in Table 4, generally, the bit rate of the proposed algorithm is more increased and its PSNR is smaller than those of JM. That result from the fact that the number of the candidate blocks for the cost function to decide the optimal vector in the proposed algorithm is less than that of JM under control of rate. However, there is no remarkable difference except for just a little degradation in the list of Table. It means that our method is still effective for not only the computational complexity but also its performance.

4 Conclusion

In this paper, we proposed a fast search method based on SEA for the variable block size motion estimation in H.264. It was realized that accumulating not only sum norm but also pre-computed SAD reduced SAD computation for 4×4 blocks. In experimental results, we showed that our method had the consistent gain of 60 $\sim$ 70% without degradation of performance with the various test images whose motion characteristics are quite different from one another. Moreover, in case that rate is constrained in variable block-based block matching, our algorithm still takes effect for the performance in spite of a little degradation. In conclusion, the proposed algorithm is quite effective for the reduction of computation and its performance under restricted bandwidth.

Acknowledgement. This work was supported by the IT Research Center (ITRC), Ministry of Information, Korea.

References

1. Z. Zhou, M. T. Sun, and Y. F. Hsu, "Fast variable block-size motion estimation algorithm based on merge and slit procedures for H.264 / MPEG-4 AVC," International Symposium on Circuits and Systems, Vol.3, pp. 725-728, May 2004.
2. T. G. Ahn, Y. H. Moon, J. H. Kim, "Fast Full-Search Motion Estimation Based on Multilevel Successive Elimination Algorithm," IEEE Trans. on Circuits and Systems for Video Technology, Vol.14, No.11, pp. 1265-1269, Nov. 2004.
3. W. Li, E. Salari, "Successive elimination algorithm for motion estimation," IEEE Trans. on Image Processing, Vol.4, No.1, pp. 105-107, Jan. 1995.
4. Y. Noguchi, J. Furukawa, H. Kiya, "A fast full search block matching algorithm for MPEG-4 video," International Conference on Image Processing, Vol.1, pp. 61-65, 1999.
5. M. Z. Coban and R. M. Mersereau, "A fast exhaustive search algorithm for rate-constrained motion estimation," IEEE Trans. on Image Processing, vol.7, no.5, May 1998.

Motion-Compensated 3D Wavelet Video Coding Based on Adaptive Temporal Lifting Filter Implementation

Guiguang Ding, Qionghai Dai, and Wenli Xu

Tsinghua University, Beijing 100084, China
{dinggg, qhdai, xuwl}@tsinghua.edu.cn

Abstract. In wavelet-based video coding with motion-compensated lifting, efficient compression is achieved by exploiting motion-compensated temporal filtering (MCTF). The performance of a 3D wavelet video codec is greatly dependent on the efficient of MCTF. In this paper, an adaptive motion-compensated temporal filter scheme is proposed for wavelet video coding. Our method focused to control the level number of the temporal decomposition by detecting the number of the un-connection pixels in low-pass frame to avoid the inefficiency of MCTF while scene changing quickly. Moreover, this method has most useful features of predictive wavelet codecs, such as multiple reference frame and bi-directional prediction. Our experimental results show that, compared with the conventional MCTF, the proposed scheme has better coding performance for most sequences.

1 Introduction

With the development of Internet and wireless communication, more and more applications start to use the streaming video techniques. In response to the increasing demand on streaming video applications, the coding objective for streaming video is changing to optimize the video quality for a wide range of bit rates and channel conditions. Scalable video coding (SVC) has a great advantage due to its very easy adaptation to unpredictable bandwidth variations and network conditions.

Wavelet based SVC has received considerable attention for streaming video applications in terms of both coding algorithms and standardization activities. Broadly speaking, two families of wavelet-based video codecs have been described in the literature: t+2D [1,2] schemes and 2D+t [3,4] schemes. MCTF is a key technology for both reducing temporal redundancy and providing flexible temporally scalable architecture in the two families of codecs. MCTF was first introduced by Ohm [5] and later improved by Choi and Woods [6]. Recently, many extensions have been proposed to increase the coding efficiency and improve the visual quality of MCTF schemes. Among these are the lifting-based MCTF schemes introduced by Pesquet-Popescu and Bottreau [7], as well as the Unconstrained Motion Compensated Temporal Filtering (UMCTF) proposed by Turaga and van der Schaar [8]. In the past researches, many experimental results have shown that the UMCTF without update-step performs better than the original MCTF [9].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 863 – 868, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In these schemes, all don't take the condition of scene changing into account. Actually, MCTF is inefficiency while scene changing quickly. At the time, the intra-coding mode is more efficient than the MCTF mode. Therefore, it is necessary to detect the changing of scene to stop the temporal decomposition in order to improve the coding efficiency. In this paper, an adaptive motion- compensated temporal lifting filtering (AMCTF) scheme is proposed, which is based on MCTF scheme to control the level number of temporal decomposition by designing a detector to estimate the changing of scene. The lifting implementation of wavelet filter is also used for MC temporal decomposition to provide sub-pixel accuracy, multiple reference frames' prediction and flexible temporal scalability.

2 MCTF Based on Lifting Filter Implementation

MCTF has proved to be an effective coding tool in the design of scalable video codec. Both DCT and wavelet based video coding architectures recently considered by the MPEG Ad Hoc group on SVC adopt MCTF in order to reduce temporal redundancy. For wavelet transforms, the so-called lifting scheme [7] can be used to construct the kernels. A two-channel decomposition can be achieved with a sequence of prediction and update steps that form a ladder structure. The advantage is that this lifting structure is able to map integers to integers without requiring invertible lifting steps. Further, motion compensation can be incorporated into the prediction and update steps. The fact that the lifting structure is invertible without requiring invertible lifting steps makes this approach feasible. The transform also allows a fully in-place calculation, which means that no auxiliary memory is needed for the computations.

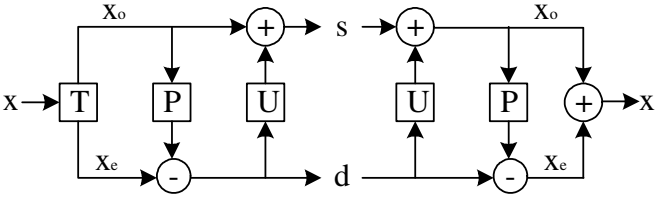


Fig. 1. The idea of lifting scheme

The idea of lifting is based on splitting a given set of data into tow subsets. Fig.1 depicts the idea of lifting scheme. Input samples are first split into odd and even samples. In a prediction step, the odd samples are predicted from the even samples, resulting in a high-pass subband signal. In an update step, the high-pass subband signal is filtered and added back to the even sample sequence, resulting in a low-pass subband signal. Because the lifting decomposition is easily invertible, any type of operation, linear or non-linear, can be incorporated into the prediction and update steps. The lifting implementation of the wavelet transform allows for a motion-compensated temporal transform, based on any wavelet kernel and any motion model, without sacrificing the perfect reconstruction property.

3 Proposed Scheme

In this paper, we utilize 5/3 DWT as the temporal filter. To avoid the ghosting artifacts in low-pass temporal frames provide flexibility temporal scalability, the update step is omitted in temporal lifting decomposition. This reduces the temporal transform to a 1/3 DWT, in which the low-pass temporal subband filter has only one tap. Equivalently, the low-pass subband frames are simply generated by choosing even indexed frames from the original sequence.

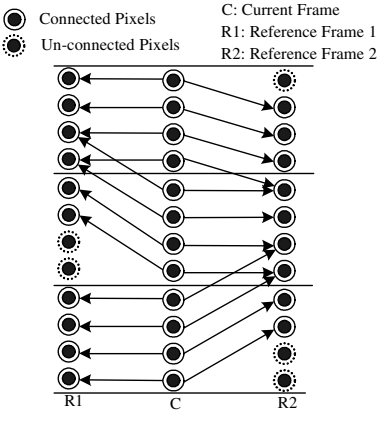


Fig. 2. The definition of connected and un-connected pixels

In MCTF, the motion compensation assumes all pixels in the same block are the same motion. However, the motion compensation would be failed when the scene is changing. Under this condition, the intra coding mode is more efficient than the MCTF mode. Therefore, MCTF should be adaptively finished according to the performance of the motion compensation.

In this paper, the adaptive temporal decomposition is implemented by detecting the number of the un-connection pixels in low-pass frame. As shown in Fig. 2, the pixels in the low-pass frame (reference frame) can be defined connected pixels and un-connected pixels. The connected pixels are the ones are used while generating high-pass frame. Then, the unused pixels are the un-connected pixels. If there are more un-connected pixels in low-pass frame, the motion compensation must be inefficient. So, the number of the un-connected pixels N is first detected during MCTF. If N is smaller than a threshold T, MCTF is gone on, otherwise MCTF is finished.

Fig. 3 shows the construct of AMCTF decomposition using bi-directional prediction. In Fig.3, “A frame” means an unfiltered frame, which is an original frame, and “H frame” means a high-pass filtered frame. In this paper, the first frame of a GOP is set as high-pass frame and the last frame is set as intra-frame. The reason is that the first frame of the next GOP must be first decoded to decode the last high-pass frame of the current GOP if the first frame of GOP was selected as low-pass frame. If the

size of GOP is eight, frame index 7 is coded as intra-frame at the highest temporal level, and frame index 3 is coded as high-pass frame (H frame) at the next temporal level by using frame index 7 of the current GOP and frame index 7 of the previous GOP. Like literature [9], for each temporal level, only necessary frames are located, for example, at the highest temporal level, only the last frame of a GOP is needed. At the next temporal level, temporal resolution is successively refined, and missing frames are predicted by already processed frame index.

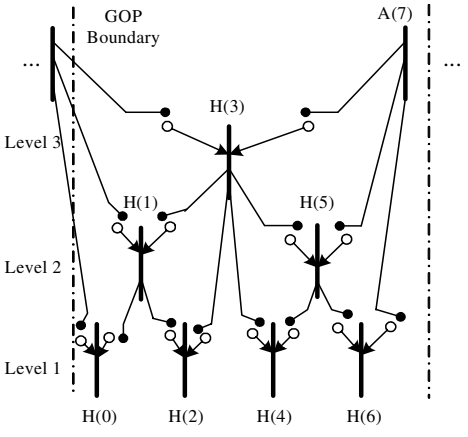


Fig. 3. Diagram of temporal processing flow of AMCTF

As shown in the figure, the number of un-connected pixels in the low-pass frame is first detected before every temporal decomposition level. If N is smaller than a pre-set threshold T, the switch is open and MCTF is not performed at this temporal level. Otherwise, the switch is close and MCTF is performed at this temporal level. This estimation is very useful to improve the coding efficiency, especially for sequence including large motion object and scene changing. The experimental results have also proven the method is efficient.

4 Experimental Results

The experiments in this section compare the proposed adaptive MCTF (AMCTF) given in Section 3 with the conventional motion-compensated 3D wavelet coder MC-EZBC. To test the performance of adaptive MCTF, we use the CIF video sequence "Table Tennis" as the source data. We extract 120 frames from the sequence for performance measure. The frames are partitioned into 15 GOPs containing 16 frames per GOP. The bit-rate is 768 kbps, the frame rate is set as 30fps. T is set as 30%. Video codec uses AMCTF and no AMCTF, respectively. Because there is obvious scene changing between the 97th frame and the 98th frame, AMCTF can obtain the better image quality than MCTF. The experimental result shows PSNR is improved by up to 2.5dB.

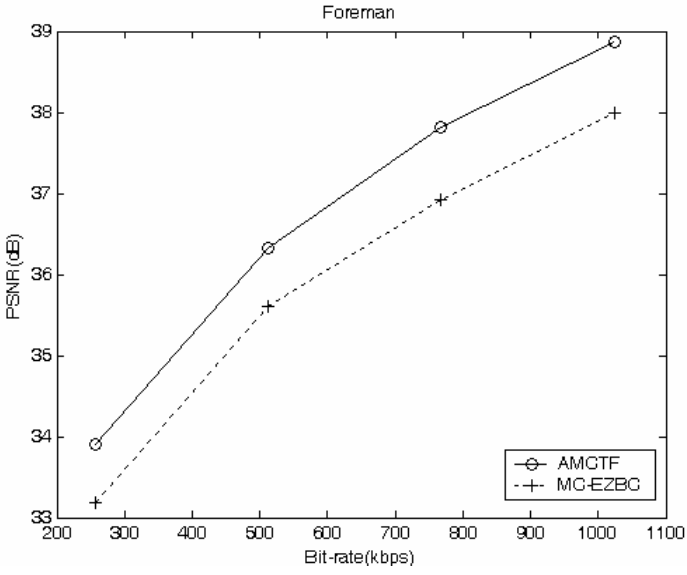


Fig. 4. Rate-distortion curves for Foreman sequence

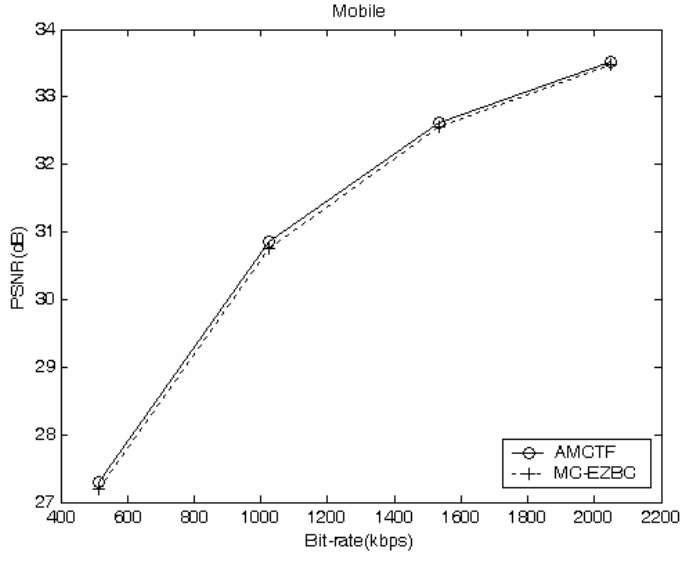


Fig. 5. Rate-distortion curves for Mobile sequence

To verify the whole compression performance of AMCTF, two standard video sequences Foreman and Mobile with CIF format (352×288), are used in our test to compare our proposed AMCTF with the MC-EZBC scheme. The frame rates of these three sequences are set as 30fps. In this test, 80 frames (from 121th frame to 200th frame) were processed with GOP size of 16.

For the two specified test sequences, some experimental results are given in Fig.4 and Fig.5. As shown in the figures, for sequence with scene changing (Foreman), AMCTF improves the encoding performance about 1dB, compared with MC-EZBC. For mobile sequence (without scene changing), two schemes have almost the same encoding performance.

5 Conclusions

In this paper, we propose a novel temporal decomposition method for scalable video coding. The method improves the coding performance by controlling the level number of the temporal decomposition. Moreover, it has more useful features of conventional coders, such as multiple reference fame and bi-directional prediction. The overall coding efficiency is improved by about 1dB for video sequence with scene changing when compared to MC-EZBC. The proposed scheme is very attractive for wavelet based scalable video coding.

References

1. Chen, P, Woods, J. W.: Bidirectional mc-ezbc with lifting implementation. IEEE Trans. Circuit and System: Video Technology, 10 (2004) 1183-1194.
2. Secker, A., Taubman, D.: Lifting-based invertible motion adaptive transform (limat) framework for highly scalable video compression. IEEE Trans. Image Procesing, 12 (2003) 1530-1542
3. Andreopoulos, Y. (ed.): In-band motion compensated temporal filtering. Signal Processing: Image Communication, 8 (2004) 653-673.
4. Wang, Y., Cui, S., Fowler, J. E.: 3D video coding using redundant-wavelet multihypothesis and motion-compensated temporal filtering. in Processing of the Int. Con. On Image Processing, 9 (2003) 755-758.
5. Ohm, J.R.: Three-dimensional subband coding with motion compensation. IEEE Trans. Image Processing, 9 (1994) 559-571.
6. Choi, S. –J., Woods, J. W.: Motion compensated 3-D subband coding of video. IEEE Trans. Image Processing, 2 (1999) 155-167.
7. Pesquet-Popescu, B., Bottreau, V.: Three-dimensional lifting schemes for motion compensated video compression. in Processing IEEE Int. Con. Acoustics, Speech, and Signal Processing, 5 (2001) 1793-1796.
8. Ruraga, D. S., Schaar, M. van der: Unconstrained motion compensated temporal filtering. MPEG Contrib. M8388, 5(2002).
9. Han, Woo-Jin: Low-delay unconstrained motion compensated temporal filtering technique for wavelet-based fully scalable video coding. IEEE 6th Workshop on Multimedia Signal Processing, 9 (2004) 478-481.

Accurate Contouring Technique for Object Boundary Extraction in Stereoscopic Imageries

Shin Hyoung Kim[1], Jong Whan Jang[1], Seung Phil Lee[2], and Jae Ho Choi[2]

[1] PaiChai University, Daejeon, Korea 302-435
{zeros, jangjw}@mail.pcu.ac.kr
[2] Chonbuk National University, RCIT of Engineering Research Institute,
Chonju, Korea 561-756
dalgitong@daum.net, wave@chonbuk.ac.kr

Abstract. A new snake-based algorithm is presented to segment objects from a pair of stereo images. The proposed algorithm is built upon a new energy function defined in the disparity space in such a way to successfully locate the boundary of an object found in a stereo image pair. The distinction of our algorithm comes from its superb segmentation capability even when the objects in the image are occluded and the background behind them is cluttered. The computer simulation shows out-performing results over the well-known conventional snake algorithm in terms of segmentation accuracy.

1 Introduction

Object segmentation is a very important issue for many applications such as machine vision, video games, 3D interactive TV, interactive multimedia system, and image and video coding [1-3].

Over the past two decades, various object segmentation schemes have been developed for extracting an object from an ordinary image. The most recent of these algorithms is the active contour (snake) algorithm [4-8]. The snake algorithm was first introduced by Kass et al. [4]. It is an energy-minimizing spline guided by internal and external forces that pull the spline toward features like lines and edges. However, as any other methods that rely on energy minimization, the snake algorithm may fall into a local minimum. Also, the performance of this algorithm is sensitive to the initial placement of the snake points [5]. Amini et al. [6] proposed a dynamic programming method to overcome the limitations of Kass's algorithm. His approach is numerically more stable; however, Amini's algorithm has the disadvantage of relatively long processing time. Williams et al. [7] proposed an algorithm to reduce the processing time by using a greedy algorithm. Influenced by Williams' algorithm, Lam et al. [8] proposed a fast greedy snake algorithm that significantly reduced processing time.

Nevertheless, all the aforementioned algorithms are suitable only for an image with a single object against a uniform background. In other words, the performance of these algorithms degrades rapidly when the image contains more than one object, especially when the objects overlap one another and the image background is cluttered.

In this paper, we propose a new snake-based algorithm that can exploit the disparity information obtained from a pair of stereo images. A new energy function has

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 869 – 874, 2005.
© Springer-Verlag Berlin Heidelberg 2005

been defined in the disparity space in order to successfully locate the boundary of an intended object (IO). The proposed algorithm works well even when the objects in the image are occlude and the background behind them is cluttered.

2 A Novel Snake Algorithm Using New Energy Functions

Consider the behavior of the snake algorithm. The objective is to lock the snake points around IO by repetitive energy minimization computation. Snake points are indexed by Cartesian coordinates and iteration number. Thus, $v_{i,j} = (x_{i,j}, y_{i,j})$ for $i = 0,..., N-1$ and $j = 1,..., M$ is the i^{th} snake point at the j^{th} iteration, where N is the total number of snake points, and M is the total number of iterations. Note that j is 0 for initial snake points, and it is 1 on the first iteration. The snake points $v_{0,0}, v_{1,0},..., v_{N-1,0}$ are initial snake points, and the closed contour linking these initial snake points defines an initial snake contour C_0. In a successful execution of the snake algorithm, the renewed contour C_j at each iteration moves toward the boundary of IO in an energy minimization sense.

Now, consider the energy function $E_{2D_snake}(v_j)$ involved in the energy minimization process of the snake algorithm that utilizes a single 2D image. The typical energy function for a snake point is defined as follows [4]:

$$E_{2D_snake}(v_j) = \sum_{i=0}^{N-1} \left(E_{int}(v_{i,j}) + E_{ext}(v_{i,j}) \right) \tag{1}$$

It contains two energy components, i.e., an internal energy and external energy. The performance of a snake algorithm is highly determined by the energy function definition. While the energy function as above is based on a single image, however, our energy function is based on a pair of stereo images. In other words, in our case a snake point $v^d_{i,j} = (x_{i,j}, y_{i,j}, d_{i,j})$ is defined in the 3-D disparity space [9-13]. The proposed energy function contains three energy terms; and the first term is the normalized continuity energy defined as follows:

$$E_{con}(v^d_{i,j}) = \frac{\left| \overline{d^d_{j-1}} - \left\| v^d_{i,j} - v^d_{i-1,j} \right\| \right|}{\overline{d^d_{j-1}}} \tag{2}$$

The main role of the continuity energy term is to make even spacing between the snake points by minimizing the distance between $\overline{d^d_{j-1}}$ and $\left\| v^d_{i,j} - v^d_{i-1,j} \right\|$, where $\overline{d^d_{j-1}}$ is the average distance between neighboring snake points at the $(j-1)^{th}$ itera-

tion; $v_{i,j}^{d}$ is the current snake point at the j^{th} iteration; and $v_{i-1,j}^{d}$ is the previous snake point at the j^{th} iteration.

The second term in our new energy function is the normalized curvature energy defined as follows:

$$E_{cur}(v_{i,j}^{d}) = \left\| \vec{u}_{i,j-1} / \left\| \vec{u}_{i,j-1} \right\| - \vec{u}_{i-1,j} / \left\| \vec{u}_{i-1,j} \right\| \right\| \tag{3}$$

where $\vec{u}_{i,j-1}$ and $\vec{u}_{i-1,j}$ represent $(x_{i+1,j-1} - x_{i,j}, y_{i+1,j-1} - y_{i,j}, d_{i+1,j-1} - d_{i,j})$ and $(x_{i,j-1}x_{i-1,j}, y_{i,j} - y_{i-1,j}, d_{i,j} - d_{i-1,j})$, respectively, where x, y and d represents row, column, and disparity. The main role of the curvature energy is to form a smooth contour between neighboring snake points.

Finally, the third term is the external energy defined as follows:

$$E_{ext}(v_{i,j}^{d}) = -\frac{\left| f_{MLDS}(x_{i,j}, y_{i,j}, d_{OOI}) \right|}{e_{\max}} \tag{4}$$

where $e_{\max}$ is the maximum of possible edge value in neighborhood, and $f_{MLDS}(x_{i,j}, y_{i,j}, d_{OOI})$ is a modifiable layer image, which contains IO objects separated for each corresponding disparity information d_{OOI}. When IO get influenced by other objects or background edges, the behavior of external energy changes; and therefore, obtaining edge information about IO with minimal influence of other edges is important. Here, watershed transformation and region merging techniques are applied in the course of obtaining $f_{MLDS}(x_{i,j}, y_{i,j}, d_{OOI})$.

Now, based on the new energy functions defined in Eqs. [2-4], a novel object segmentation algorithm is designed. The flowchart in Fig. 1 illustrates the whole processes taken in our snake algorithm for segmenting a boundary of IO from a pair of stereo images. The first step is to select one IO and place a rectangular window (or an arbitrary shaped window) around IO, called the region of interest (ROI). Next, the disparity map is created for ROI, and the other stages are followed, serially and iteratively. Particularly, in the stage where the energy computation is involved, the three terms in the new energy function is computed as follow:

$$E_{new-snake}(x_j, y_j, d_{OOI}) = \sum_{i=0}^{N-1} \left[\alpha E_{con}(x_{i,j}, y_{i,j}, d_{OOI}) + \beta E_{cur}(x_{i,j}, y_{i,j}, d_{OOI}) + \gamma E_{ext}(x_{i,j}, y_{i,j}, d_{OOI}) \right] \tag{5}$$

where α, β and γ are selected to balance the relative influence of the three terms in the energy function. In our experiments, $\alpha = 1$ and $\beta = 1$ if $v_{i,j}^{d}$ is not a corner point; otherwise, $\alpha = 0$, $\beta = 0$ and $\gamma = 1.2$. These controlling parameter values are selected in order to accentuate the external energy than two others.

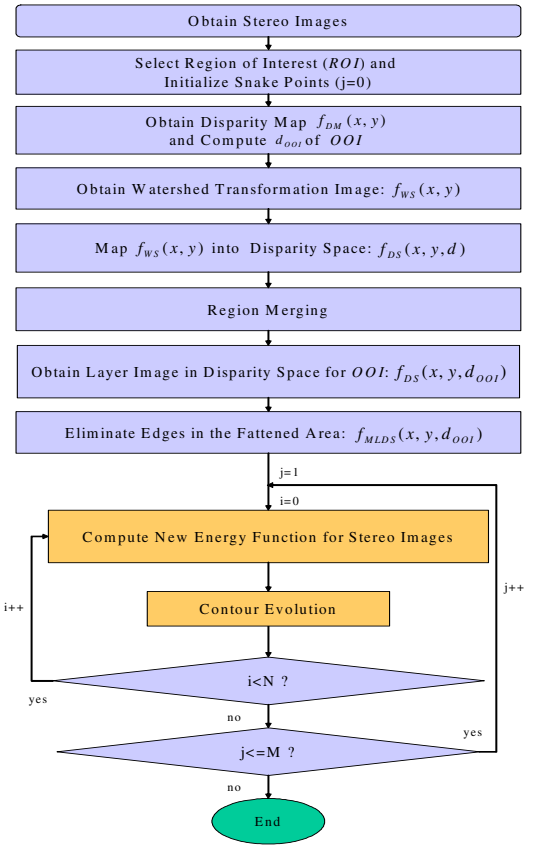


Fig. 1. Flowchart of our snake algorithm for object segmentation

3 Computer Simulations

Computer simulations are performed to verify the performance of our algorithm. To obtain comprehensive performance verification, a set of nine stereo image pairs are used. Fig. 2 demonstrates typical results of segmentation on stereo image pairs, which have varying degrees of complexity. The first column shows the initialization of *ROI* in each image. The results of the conventional algorithm for each image are illustrated in the second column. The results of our proposed algorithm are shown in the third column. While the proposed algorithm is successful, the conventional algorithm could not detect the boundary of IO. Generally, for the conventional snake algorithm, the segmentation error increases in accordance to scene complexity, and it has a relatively high dependency on the placement of the initial snake points. When the initial snake points are placed closer to the boundary of IO, the segmentation performance of the conventional algorithm improves. However, the manual manner of *ROI* placement is not desirable, and the segmentation performance for the conventional snake algorithm degrades rapidly as the image complexity increases. On the other hand, for our snake algorithm, the snake point migration toward the true boundary of the *IO* was successful in each set of the experiments.

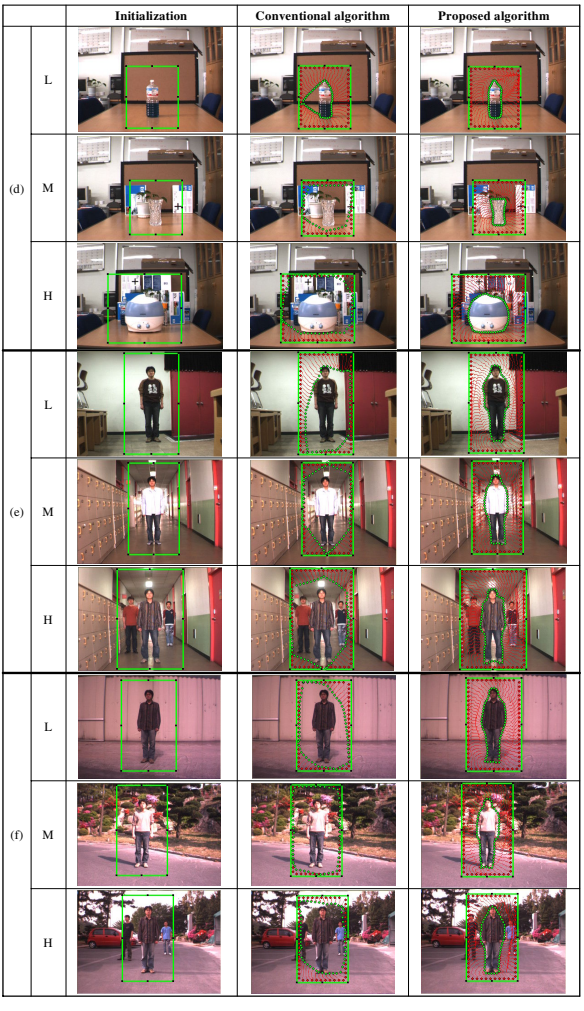


Fig. 2. Results on the various stereo image pairs

4 Conclusions

A novel snake algorithm for object segmentation using a pair of stereo images has been presented in this paper. The proposed algorithm improves and extends the conventional snake algorithm. Our snake algorithm is built upon newly defined energy functions that exploit the disparity information found in the disparity space; and it allows a successful segmentation of IO in the image even when IO objects are overlapping one other and the background is cluttered. Of course, the true boundary extraction was successful at the cost of increased computation for processing a set of stereo images, however, the computer simulation results have yielded us a superior capability of new method over the conventional one.

References

1. Sikora, T.: The MPEG-4 Video Standard Verification Model. IEEE Transaction Circuits System and Video Technology, Vol. 7. No. 1. (1997) 19-31
2. Bais, M., Cosmas, J., Dosch, C., Engelsberg, A., Erk, A., Hansen, P. S., Healey, P., Klungsoeyr, G.K., Mies, R., Ohm, J.R., Paker, Y., Pearmain, A., Pedersen, L., Sandvancd, A., Schafer, R., Schoonjans, P., Stammnitz, P.: Customized Television Standards Compliant Advanced Digital Television. IEEE Transaction Broadcasting, Vol. 48. No. 2. (2002) 151-158
3. ISO/IEC, JTC/SC29/WG11/W4350.: Information Technology. Coding of Audio-Visual Objects Part2-Visual. ISO and IEC (2001) 14496
4. Kass, M., Witkin, A., Terzopoulos, D.: Snake Active Contour Models Int. J. Computer Vision, Vol. 1. No. 4. (1987) 321-331
5. Amini, A. A., Weymouth, T. E., Jain, R. C.: Using Dynamic Programming for Solving Variational Problems in Vision. IEEE Transaction Pattern Analysis and Machine Intelligence, Vol. 12. No. 9. (1990) 855-867
6. Williams, D. J., Shah, M.: A Fast Algorithm for Active Contours and Curvature Estimation. Computer Vision Graphics and Image Processing, Vol. 55. No. 1. (1992) 14-26
7. Lam, K.M., Yan, H.: Fast Greedy Algorithm for Active Contours. Electron Letter, Vol. 30. No. 1. (1994) 21-23
8. Kanade, T., Okutomi, M.: A Stereo Matching Algorithm with an Adaptive Window. Theory and Experiment. IEEE Transaction Pattern Analysis and Machine Intelligence, Vol. 16. No. 9. (1994) 920-932
9. Birchfield, S., Tomasi, C.: Depth Discontinuities by Pixel to Pixel Stereo. Int. J. Computer Vision, Vol. 35. No. 3. (1999) 269-293
10. Redert, A., Hendriks, E., Biemond, J.: Correspondence Estimation in Image Pairs. IEEE Signal Processing Magazine, Vol. 16. No. 3. (1999) 29-46
11. Tasi, C. J., Katsaggelos, A. K.: Dense Disparity Estimation with a Divide and Conquer Disparity Space Image Technique. IEEE Transaction Multimedia, Vol. 1. No. 1. (1999) 18-29
12. Zitnick, C. L., Kanade, T.: A Cooperative Algorithm for Stereo Matching and Occlusion Detection. IEEE Transaction Pattern Analysis and Machine Intelligence, Vol. 22. No. 7. (2000) 675-684
13. Christopoulos, V.A., Muunck, P.D., Cornelis, J.: Contour Simplification for Segmented Still Image and Video Coding-Algorithms and Results. Image Communication in Signal Processin, Vol. 41. No. 2. (1999) 335-357

Robust Object Tracking Based on Uncertainty Factorization Subspace Constraints Optical Flow

Yunshu Hou, Yanning Zhang, and Rongchun Zhao

College of Computer, Northwestern Polytechnical University, Xi'an 710072, China
yunshuhou@263.net

Abstract. The traditional methods of optical flow estimation have some problems, such as huge computation cost for the inverse of time-varying Hessian matrix, aperture phenomena for the points with 1D or little texture and drift phenomena with long sequences. A novel nonrigid object tracking algorithm based on inverse component uncertainty factorization subspace constraints optical flow is proposed in this paper, which resolves the above problems and achieves fast, robust and precise tracking. The idea of inverse Component is implemented in each recursive estimation procedure to make the algorithm fast. Uncertainty factorization is used to transform the optimization problem from a hyper-ellipse space to a hyper-sphere space. SVD is correspondingly performed to involve the subspace constraints. The proposed algorithm has been evaluated by both the standard test sequence and the consumer USB camera recorded sequence. The potential applications vary from articulated automation to structure from motion.

1 Introduction

In an image sequence, moving objects are represented by their feature points detected prior to tracking or during tracking. Consequently how to automatically and precisely track these feature points becomes a key step to the future process and analysis [1].

There have been presented a number of feature points tracking algorithms for rigid objects in the computer vision literature. The most famous may be the Lucas-Kanade tracker [2]. However the original L-K tracker has many problems such as the aperture phenomena for the points with line texture, hard to estimate the points with depth discontinuity, huge computation cost for the inverse of time-varying Hessian matrix and so on. Recently some papers have discussed the above problems [5] ~ [7], which facilitate and inspire the research of this paper.

Recently nonrigid objects tracking has gained more and more interests by computer vision researchers. The traditional rigid trackers can not achieve the feasible results any more. In computer vision and computer graph literatures there has been proven that the structured flexible objects can be modeled as a linear combination of basic structures [3], [4], [8]. As a result there is a booming for the AAM-like tracker; however they need large pre-labeled samples to make off-line training. In this paper the popular subspace modeling technique is introduced into the traditional L-K tracker. However the subspace is not learned from samples, instead it can be inferred from the tracking procedure.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 875 – 880, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The rest of the paper is organized as follows. In section two a detailed theoretical analysis of the proposed methodology and formulation is deduced. In section three the system implementation is proposed and the experimental results show the superiority of the proposed algorithms. The conclusion is made in section four.

2 Methodology and Formulation

The original L-K tracker [2] is to minimize the sum of squared difference between the template image T and the object image I warped back onto the template:

$$\min \sum_{x} [I(W(x;p)) - T(x)]^2 \tag{1}$$

Three refinements are introduced below, which are inverse component to reduce the computing cost, uncertainty factorization to deal with aperture phenomena and subspace constraints to make the tracker fitted to nonrigid tracking. These techniques form the foundation of the proposed nonrigid objects tracking algorithm.

2.1 Inverse Component

Two refinements are presented to reduce the computation cost of Hessian matrix and to make the optimization more robust. Firstly the role of the template image and object image is exchanged to make Hessian matrix time-independent. Secondly the compositional parameter incremental method is involved to iteratively solve for an incremental warp rather than an additive update, which has been proven to be more efficient in image mosaicking. The new form of the L-K tracker and its solution is obtained respectively:

$$\min \sum_{x} [T(W(x;\Delta p)) - I(W(x;p))]^2 \tag{2}$$

$$\begin{cases} W(x;p) \leftarrow W(x;p) \circ W(x;\Delta p)^{-1} \\ \Delta p = \left[\sum_{x} \left[\nabla T \frac{\partial W}{\partial p} \right]^T \left[\nabla T \frac{\partial W}{\partial p} \right] \right]^{-1} \cdot \sum_{x} \left[\nabla T \frac{\partial W}{\partial p} \right]^T [I(W(x;p)) - T(x)] \end{cases} \tag{3}$$

From (3) we can observe that Hessian matrix is time independent and the parameter refinement is the incremental warp instead of the directly additive update.

2.2 Uncertainty Factorization

Suppose the considered warp is 2D displacement, L-K tracker can be rewritten as:

$$g_{2\times 1} = \begin{bmatrix} g_{ij} \\ h_{ij} \end{bmatrix} = \begin{bmatrix} \sum g_{xi} g_{tj} \\ \sum g_{yi} g_{tj} \end{bmatrix} = \begin{bmatrix} \sum g_{xi}^2 & \sum g_{xi} g_{yi} \\ \sum g_{xi} g_{yi} & \sum g_{yi}^2 \end{bmatrix} \cdot \begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} = \begin{bmatrix} a_i & b_i \\ b_i & c_i \end{bmatrix} \cdot \begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} = X_{2\times 2} \cdot w_{2\times 1} \tag{4}$$

Suppose we are tracking multiple points in a video, (4) can be stacked to form the more compact L-K tracker, so-called multi-frame multi-point formulation:

$$G_{2N\times F} = X_{2N\times 2N} \cdot W_{2N\times F} \tag{5}$$

where $W_{2N\times F}$ denotes the optical flow matrix to be estimated, $X_{2N\times 2N}$ encodes the distribution of the uncertainty and $G_{2N\times F}$ denotes the spatial and temporal gradient.

If the rank of X is not full it means the local image has degenerate structure [1]. Uncertainty factorization is introduced to deal with the famous aperture problem. Suppose we need to minimize the error norm for M or S given W:

$$\min\left\|E\right\|^2 = \left\|W - M * S\right\|^2 \tag{6}$$

with the constrains W = M * S. The traditional method is performing SVD on W. It is reasonable when no prior knowledge is known about the noise. However if the noise is anisotropic it is better to use an elliptical error norm which can be formulated as:

$$\min\left\|E\right\|^2 = \left\|M * S - W\right\|^2 = \left((vecE)^T \cdot \sum^{-1} \cdot vec(E)\right) \tag{7}$$

In equation (5), under gauss-noise assumption the matrix X can be seen as the inverse covariance of W. Consequently we can eigen-decompose X to obtain Q and use Q to warp the problem of a hyper-ellipse space into a hyper-sphere space and use traditional SVD to solve the MSE problem. Geometrically, Q can rotate the problem's directions of the largest and least uncertainty into axis-alignment which is with the same uncertainties, and then scale each axis proportional to its certainty.

2.3 Subspace Constraints

Uncertainty factorization is based on local feature representation. Here we will introduce the idea of subspace analysis into the proposed algorithm to involve the global constraints of the tracked object. It has been proven that a nonrigid object S lies in a linear space and can be represented as a linear combination of K basic shapes.

Under the week-perspective projection model we can obtain:

$$W_f = R_f\left(C_{f,1}S_1 + \cdots + C_{f,k}S_k\right) + T_f \tag{8}$$

in which $(S_1, S_2, \cdots, S_k)$ is the basic shape of the modeled nonrigid object, R_f and T_f is the camera rotation and translation of frame f.

For videos (8) can be stacked into a big matrix to gain a more compact form:

$$W_{2F\times N} = M_{2F\times 3K} S_{3K\times N} \oplus T_{2F\times 1} \tag{9}$$

where M denotes the information of the camera, S denotes the linear coefficients of the basis shapes and W is the stack of W_f, which is to be estimated in (5).

It can be inferred from (9) that the multi-frame multi-point optical flow matrix W lies in a rank 3K subspace. It is straightforward to perform SVD to involve the global constraints. The geometrical explanation is that the multiple frames multiple points flow matrix W is the effect of the projection of the real 3D nonrigid object through the different views. The constraints are not violated at the depth discontinuities or when the camera motion changes abruptly. Through SVD there is no need to perform pre-training; instead the basic shapes can be inferred from SVD.

3 System Implementation and Experimental Results

3.1 The Proposed ICUFSCOP Tracking System

The three techniques discussed above can be fused together to implement the proposed system. The idea of inverse Component is implemented in the recursive estimation procedure to make the algorithm fast. The idea of uncertainty factorization and subspace constraints is alternately applied in each iteration circle to deal with the aperture problem and nonrigid object tracking, which is the key of the proposed system. Firstly uncertainty factorization is used to warp W from space with hyper-ellipse distribution into space with hyper-sphere distribution. Secondly SVD is performed in the warped hyper-sphere space to make the subspace constraints. Below is the flowchart of the proposed system.

1. Set up gauss pyramid;
2. Computer X in the reference frame and set the initial estimation of W;
3. Compute $\sum_G = X = V\Lambda V^T$, $R = V^T$, $Q = \Lambda^{1/2} V^T$, $IH = QX^+$;
4. Compute G and Warp the current estimation as: $GG = -IH \cdot G + R \cdot W$;
5. Perform SVD on GG to make the 3K subspace constraints, then get UV;
6. Unwrap UV back and update the new estimation of optical flow;
7. If convergence stop, otherwise return to step 4.

3.2 Experimental Results

A number of experiments have been implemented to evaluate the performance of the proposed nonrigid objects tracking algorithm. Here two typical face sequences will be used to show the superiority of the system. There are many nonrigid variations in the testing sequences and many of the tracked points have only 1D or little texture.



Fig. 1. The MISSA Sequence

The first sequence (Fig. 1) is the standard face testing sequence named as American Lady MISSA sequence. It has the notable head translations. The eyes and mouth are often closed and opened. The proposed algorithm can track all the points in all the frames. Even when the eyes and mouth are closed the tracker doesn't lose them and when they are opened the tracker can continue tracking properly.



Fig. 2. The USB Camera Recorded Sequence

The second sequence (Fig. 2) is much more difficult to be tracked. Firstly it is recorded with the consumer USB camera with low imaging quality and low frame rate. Secondly it has also a little out-of-plan rotation. Finally it has a lot of in-plane rotation, scaling and many expression variations, even with the blur for the focusing error of the USB camera. However it is quite exciting that we can track all the points in all the frames, which shows the superiority of the proposed algorithm.

4 Conclusions

A robust nonrigid object tracking algorithm is proposed in this paper. Three refinements is fused to develop the traditional L-K tracker, which are inverse component to accelerate the computation, uncertainty factorization to deal with aperture phenomena and subspace constraints to make the tracker fitted to nonrigid tracking and make the tracking more robust and stable.

The experimental results of the standard testing sequence and the consumer USB camera recorded sequence demonstrate that the proposed algorithm is fast, precise and robust. The most attractive feature is that it can track the points with little texture on the nonrigid object. The potential applications vary from face expression analysis to obtaining the semi-dense correspondence to feed to SFM.

Acknowledgements

This Research is supported by Chinese Defense Fundamental Research Grant.

References

1. Shi J., Tomasi C.: Good Features to Track. IEEE Int. Conference Computer Vision and Pattern Recognition (1994) 593-600
2. Lucas B.D., Kanade T.: An Iterative Image Registration Technique with an Application to Stereo Vision. DARPA Image Understanding Workshop (1981) 121-130
3. Blanz V., Vetter T.: A Morphable Model for the Synthesis of 3D Faces. ACM SIGGRAPH (1999) 187-194
4. Cootes T.F., Edwards G.J, Taylor C.T.: Active Appearance Models. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 23(6), (2001) 681-685
5. Anandan P., Irani M.: Factorization with Uncertainty. Int. J. Computer Vision Vol. 49(2-3) (2002) 101-116
6. Irani M.: Multi-Frame Optical Flow Estimation Using Subspace Constraints. IEEE Int. Conference Computer Vision (1999) 626-633
7. Brand M.E., Bhotika R.: Flexible Flow for 3D Nonrigid Tracking and Shape Recovery. IEEE Int. Conference Computer vision and Pattern Recognition (2001) 315-322
8. Hager G.D., Belhumeur P.N.: Efficient region tracking with parametric models of geometry and illumination. IEEE Trans. Pattern Analysis and Machine Intelligence (1998) 1025–1039

Bearings-Only Target Tracking Using Node Selection Based on an Accelerated Ant Colony Optimization

Benlian Xu and Zhiquan Wang

Department of Automation, Nanjing University of Science & Technology,
Nanjing 210094, China
{Xu_benlian, Wangzqwhz}@yahoo.com.cn

Abstract. In order to improve target localization accuracy and conserve node power simultaneously, a multi-objective accelerated ant colony optimization algorithm is presented to deal with the problem of node selection in the sensors network. On the base of it, the interacting multiple model algorithm (IMM) is introduced to track bearings-only maneuvering target, and simulation results show that the proposed algorithm not only has better tracking performance, but also meets the needs of conserving battery power of node.

1 Introduction

It is noted that the target location accuracy depends on the geometry relations between nodes and target in the bearings-only target tracking (BOT) field [3]. Kaplan proposed four kinds of node selection algorithms [1, 2], i.e., the Closest Nodes, the Virtual Closest Nodes, the Simplex and the Autonomous Selection. Based upon the Kaplan's idea, the ant colony optimization (ACO) [4, 5] is investigated and used to find the very nodes satisfying two requirements, i.e. location accuracy and energy requirement. First, we use the predict value obtained from the IMMEKF as the predicted target location, then the proposed multi-objective accelerated ACO algorithm determines which nodes will be selected to measure target bearing angles.

The paper is arranged as follows. Section 2 introduces the multi-objective optimization problem for the multi-static BOT system, and the accelerated ACO algorithm is described in Section 3. The multi-static bearings-only maneuvering target tracking (BOMTT) algorithm using the IMMEKF is given in Section 4. Finally, numerical result and conclusion are given in Section 5 and 6.

2 The Multi-objective Optimization for the Multi-static BOT

The purpose of node selection is to obtain smaller geometric dilution of precision (GDOP) and conserve energy. We consider 2D localization problem. The root of mean square position error can be computed from the covariance of the position error

$$GDOP = \sqrt{trace\left(\left(\sum_{i\in\Omega_a}\frac{1}{\sigma_i^2 r_i^2}\begin{pmatrix}\cos^2\phi_i & -\sin\phi_i\cos\phi_i\\ -\sin\phi_i\cos\phi_i & \sin^2\phi_i\end{pmatrix}\right)^{-1}\right)}, \tag{1}$$

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 881 – 886, 2005.
© Springer-Verlag Berlin Heidelberg 2005

where Ω_a is the set of nodes in a given searching region. σ_i and ϕ_i are the measurement deviation and bearing angle corresponding to node i, respectively.

When part of nodes in the set of Ω_a is used to locate the target and centralized tracking structure is implemented, the total energy consuming is

$$E = \sum_{i \in \Omega_a} \left(l_i \varsigma_{ampl} d_i^4 \right), \tag{2}$$

where ς_{ampl} is the energy per bit to run the amplifier, l_i is the data to be transmitted for node i, and d_i is the distance between the transmitting node i and the centralized processor position. Therefore, multi-objective optimization problem can be summarized to minimize both Eq. (1) and (2).

3 The Multi-objective Ant Colony Optimization Algorithm

The multi-objective optimization problems are characterized by finding Pareto optimal solutions. In terms of the slow convergence speed of the ACO, an accelerated ACO algorithm is proposed in the Pareto solution sense. When an ant has to select the next node to be visited, it will use the following law

$$j = \begin{cases} \arg\max\limits_{j \in \tilde{\Omega}_a} \left(\left[\sum\limits_{k=1}^{K} w_k \bullet \tau_{ij}^k \right]^{\alpha} \bullet \eta_{ij}^{\beta} \right) & \text{if } q \le q_0 \\ \tilde{j} & \text{otherwise} \end{cases}, \tag{3}$$

where K is the number of considered objectives, τ_{ij}^k is the pheromone amount on edge (i, j), w_k is the weight, η_{ij} is an aggregated value of attractiveness of edge (i, j), α and β are the pheromone updating parameters, q is a random number between 0 and 1, $q_0 \in [0,1]$ is a parameter allowing to regulate the elitism of the algorithm, and $\tilde{\Omega}_a$ is the set of nodes that will be visited. Node $\tilde{j}$ is to be visited according to the probability distribution given by:

$$P(j) = \begin{cases} \left[\sum\limits_{k=1}^{K} w_k \bullet \tau_{ij}^k \right]^{\alpha} \bullet \eta_{ij}^{\beta} \Big/ \sum\limits_{m \in \tilde{\Omega}_a} \left[\sum\limits_{k=1}^{K} w_k \bullet \tau_{im}^k \right]^{\alpha} \bullet \eta_{im}^{\beta}, & \text{if } j \in \tilde{\Omega}_a \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

For the objective k, the local update is executed according to $\tau_{ij}^k = (1-\rho)\tau_{ij}^k + \rho\tau_0$, with ρ being the pheromone evaporation rate, and τ_0 being the initial pheromone amount.

The global pheromone updating is performed by using two different ants: the best and the second-best solutions generated in the current iteration for each objective k. Once each ant has completed its tour, the global pheromone updating is executed by

$$\tau_{ij}^{k} = (1-\rho)\tau_{ij}^{k} + \rho\Delta\tau_{ij}^{k}. \tag{5}$$

When both the best and the second best solutions run across edge (i,j), we only consider the best one. Therefore we have

$$\Delta\tau_{ij}^{k} = \begin{cases} 15, & \text{if edge(i, j)} \in \text{best solution} \\ 7, & \text{if edge(i, j)} \in \text{second best solution} \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

In terms of the real-time requirement, an accelerated ACO algorithm is presented. An efficient way is to preprocess the searching region: we assume that only m nodes are randomly selected from the N nodes in the searching region to measure the target, then we obtain C_N^m node combinations constituting new searching region $\hat{\Omega}_a$, i.e., the searching region $\hat{\Omega}_a$ is composed of C_N^m new "nodes". Each ant starts from the same point, and reach its destination (new "node") after one time node selection according to Eq. (5),(6). From the discussed above, the searching region is regulated efficiently, and the search time is saved. The node selection procedure is described as follows:

Step1: parameter initialization. We assume that the target prediction position is known at the current time instant, then searching region centered on the target position with the radius of r_c is determined, and nodes in the region can be used to constitute $\hat{\Omega}_a$. The maximum number of iteration is set to be N_0, and also we generate $\tilde{N}$ ants totally. We have $NumofIter = 0$ and $NumofAnts = 0$.

Step 2: the number of iteration $NumofIter \leftarrow NumofIter + 1$.

Step 3: the number of ants $NumofAnts \leftarrow NumofAnts + 1$. The objective weight w_k is generated according to a uniform distribution. Each ant starts from the target position, and visit next node according to Eq. (5), (6), meanwhile, the pheromone value on the trail is updated locally according to Eq. (7). η_{ij} is the inversion of the distance between target position and geometrical center of the m nodes.

Step 4: if $NumofAnts \le \tilde{N}$, go to step 3, otherwise go to step 5.

Step 5: for each objective, we execute single objective optimization, and obtain the best and the second best solution. Update the global pheromone information according to Eq. (8), (9).

Step 6: when the number of iteration $NumbofIter < N_0$, go to step 2, otherwise stop computation print the obtain solutions in Pareto sense.

4 Multi-static BOMTT Algorithm Using IMMEKF

The target state is denoted by $\boldsymbol{X}_T(k) = [x_T(k), y_T(k), \dot{x}_T(k), \dot{y}_T(k)]^T$ at time index k, and the observer state is $\boldsymbol{X}_{ol}(k) = [x_{ol}, y_{ol}, 0, 0]^T$ for node l ($l = 1, 2, ...m$). The measurement equation is given by

$$\boldsymbol{\beta}(k)=\boldsymbol{h}(\boldsymbol{X}_T(k))+\boldsymbol{\eta}(k)=\left(\tan^{-1}\left(\frac{x_T(k)-x_{o1}(k)}{y_T(k)-y_{o1}(k)}\right),\cdots,\tan^{-1}\left(\frac{x_T(k)-x_{om}(k)}{y_T(k)-y_{om}(k)}\right)\right)^T+\boldsymbol{\eta}(k)\,, \tag{7}$$

where $\boldsymbol{\beta}(k)$ denotes measurement vector of the target bearing angles, and $\boldsymbol{\eta}(k)$ is the measurement noise, assumed to be zero mean with covariance $\boldsymbol{R}(k)$ .The IMM algorithm cycle consists of the following four steps:

1) Starting with the mode probability weights $\hat{\mu}_t(k-1)$, the means $\hat{\boldsymbol{X}}_T^t(k-1)$ and the associated covariance $\hat{\boldsymbol{P}}_t(k-1)$. The initial condition for the t th mode is

$$\begin{aligned}&\hat{\boldsymbol{X}}_T^{t,0}(k-1)=\sum_r H_{tr}\hat{\mu}_r(k-1)\hat{\boldsymbol{X}}_T^r(k-1)/\bar{\mu}_t(k),\quad \bar{\mu}_t(k)=\sum_r H_{tr}\hat{\mu}_r(k-1)\\&\hat{\boldsymbol{P}}_t^0(k-1)=\sum_r H_{tr}\hat{\mu}_r(k-1)[\hat{\boldsymbol{P}}_r(k-1)+(\hat{\boldsymbol{X}}_T^r(k-1)-\hat{\boldsymbol{X}}_T^{t,0}(k-1))(...)^T]/\bar{\mu}_t(k)\end{aligned}, \tag{8}$$

where H_{tr} is the state transition probability from state r to t .

2) According to the N pairs $\hat{\boldsymbol{X}}_T^{t,0}(k-1)$, $\hat{\boldsymbol{P}}_t^0(k-1)$, one step prediction $\bar{\boldsymbol{X}}_T^t(k)$ and $\bar{\boldsymbol{P}}_t(k)$ are obtained, respectively. Then measure updating yields $\hat{\boldsymbol{X}}_T^t(k)$ and $\hat{\boldsymbol{P}}_t(k)$.

3) The mode probability is updated according to the following law:

$$\begin{aligned}&\hat{\mu}_t(k)=c\bullet\bar{\mu}_t(k)\bullet N(\boldsymbol{\beta}(k);\boldsymbol{h}(\bar{\boldsymbol{X}}_T^t(k)),\boldsymbol{S}_t(k))\\&\boldsymbol{S}_t(k)=\tilde{\boldsymbol{H}}_t(k)\bar{\boldsymbol{P}}_t(k)\tilde{\boldsymbol{H}}_t(k)+\boldsymbol{R}(k),\ \tilde{\boldsymbol{H}}_t(k)=\partial\boldsymbol{h}(\boldsymbol{X}_T)/\partial\boldsymbol{X}_T\big|_{\bar{\boldsymbol{X}}_T^t(k)}\end{aligned} \tag{9}$$

with c denoting a normalizing constant.

4) Once the mode probability is updated, the total estimation is obtained:

$$\hat{\boldsymbol{X}}_T(k)=\sum_r\hat{\mu}_r(k)\hat{\boldsymbol{X}}_T^r(k),\hat{\boldsymbol{P}}(k)=\sum_r\left\{(\boldsymbol{X}_T(k)-\hat{\boldsymbol{X}}_T^r(k))(...)^T+\hat{\boldsymbol{P}}_r(k)\right\}\hat{\mu}_r(k)\,. \tag{10}$$

The predicted state, which will be used to select nodes based on the accelerated ACO algorithm discussed above, is

$$\bar{\boldsymbol{X}}_T(k)=\sum_r\hat{\mu}_r(k-1)\bar{\boldsymbol{X}}_T^r(k)\,. \tag{11}$$

5 Numerical Results

The target is flying with a constant velocity in the first 100 samples. Then in the next 30 samples, it is making a constant rate turn with acceleration of $a_n=9.8m/s^2$. In the last segment of its trajectory the target switches back to uniform linear motion lasting for 70 samples. The target initial state is $X_T(0)=[0km,20km,9m/s,-5m/s]^T$, and we assume that any node measurement deviation is identical and is set to be $\sigma=\sigma_l=0.5^O$. The interval of sampling is $T=2s$. We use three models: the first mode is uniform rectilinear motion with process noise covariance

$\boldsymbol{Q}_1 = diag(0.2^2, 0.2^2)$ and the other two are constant rate turn models with $a_n = 9.8m/s^2, -9.8m/s^2$ and process noise covariance $\boldsymbol{Q}_2 = \boldsymbol{Q}_3 = diag(0.1^2, 0.1^2)$, respectively. The initial mode probability is set to be $\hat{\mu}(0) = [1/3, 1/3, 1/3]$, and the state transition matrix is $\boldsymbol{H} = [0.6, 0.2, 0.2; 0.2, 0.6, 0.2; 0.2, 0.2, 0.6]$.

We set $\alpha = 1$, $\beta = 2$, $q_0 = 0.98$, $\rho = 0.2$, $\tau_0 = 5$, $NumofIter = 10$, $\tilde{N} = 30$, $r_c = 1km$, $\tau_{ij}^{k} = 5$, and $\Delta\tau_{ij}^{k} = 0$. The network consists of 100 nodes with uniform distribution and placed in a $20km \times 12km$ region centered on $(10km, 14km)$ .50 Monte-Carlo runs are conducted.

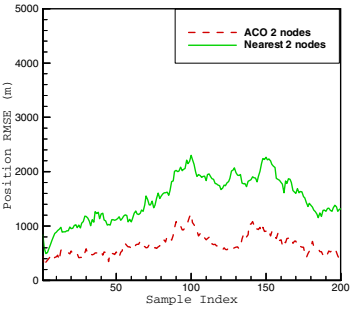


Fig. 1. Position RMSE

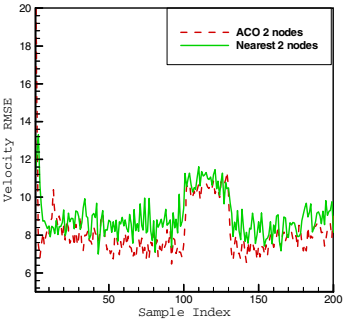


Fig. 2. Velocity RMSE

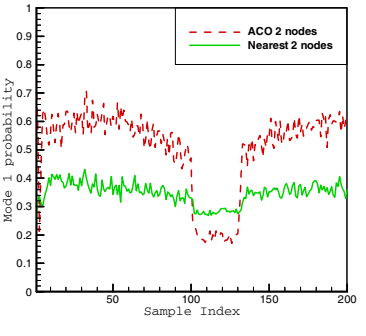


Fig. 3. Mode 1 probability

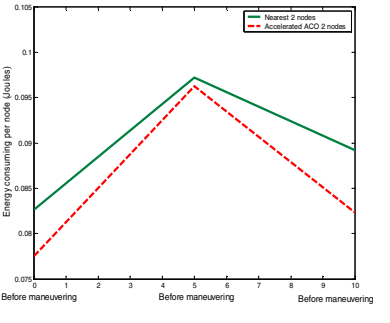


Fig. 4. Energy consuming per node

Fig.1-2 show the results of RMSE using different two methods, i.e., the ant colony optimization algorithm (2 nodes used) proposed in this paper, and the nearest nodes algorithm presented by [1]. It can be observed from figures that the performance of the ACO algorithm is superior to the one of the nearest 2 nodes. The search time is limited to the range of $1.06-1.25$ s less than sampling interval. Fig.3 shows the probability of mode 1, it can be seen that the ACO algorithm may improve the estimation probability of mode and thus obtain satisfactory tracking performance. Fig. 4 shows

the energy consuming per node for different tracking segments. Under the same nodes used, the accelerated ACO conserves power efficiently.

6 Conclusion

This paper deals with two problems: node selection and its application to the multi-static BOMTT with IMMEKF. An accelerated ACO algorithm is presented, which considers not only localization accuracy, but also energy requirement per node. Simulation results show that the accelerated ACO algorithm, together with IMMEKF, tackles these two requirements successfully.

References

1. Kaplan, L.M.: Node Selection for Target Tracking Using Bearing Measurements from Unattended Ground Sensors. In: IEEE Aerospace Conference Proceedings, Vol. 5, Big Sky, Montana, (2003) 2137-2152
2. Kaplan, L.M.: Transmission Range Control During Autonomous Node Selection for Wireless Sensor Networks. In: IEEE Aerospace Conference Proceedings, Vol. 3, Big Sky, Montana, (2004) 2072-2087
3. Ivan Kadar: Optimum Geometry Selection for Sensors Fusion. In: Proceedings of SPIE on Signal Processing, Sensors Fusion and Target Recognition Ⅶ, Vol.3347. Orlando, U.S.A, (1998)96-107
4. Dorigo M, Maniezzo V, Colorni A: Ant System: Optimization by a Colony of Cooperating Agents. IEEE Trans. on Systems, Man and Cybernetics- Part B, Vol. 26, No. 1, (1996) 29-41
5. Verbeeck K, Nowe A: Colonies of Learning Automata. IEEE Trans. on Systems, Man and Cybernetics- Part B, Vol. 32, No. 6, (2002) 772-780
6. H.A.P Blom, Y.Bar-Shalom: The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients. IEEE Trans. on Automatic Control, Vol. 33, No. 8, (1988) 780-783

Image Classification and Delineation of Fragments

Weixing Wang

Department of Computer Science & Technology,
Chongqing University of Posts & Telecommunications 400065, China
wangwx@cqupt.edu.cn, znn525d@yahoo.com

Abstract. This paper shows that an algorithm technique involving image classification and valley-edge based image segmentation is a highly efficient way of delineating densely packed rock fragments. No earlier work on segmentation of rock fragments has exploited these two building blocks for making robust segmentation. Our method has been tested experimentally for different kinds of closely packed fragment images which are difficult to detect by ordinary edge detections. The reason for the powerfulness of the technique is that image classification (knowledge of scale) and image segmentation are highly cooperative processes. Moreover, valley-edge detection is a nonlinear filter picking up evidence of valley-edge by only considering the strongest response for a number of directions. As tested, the algorithm can be applied into other applications too.

1 Introduction

The earliest image analysis system for rock fragments was developed by Gallagher [1] in 1976. In his PhD study, he set up a system aimed to measure fragment size parameters on a conveyor belt. The camera was mounted above the particle stream with its optical axis aligned normal to the moving bed of particles. The size distribution of the fragments was then computed by finding the spacing of edges with a chord sizing method. In most applications, the quality of rock fragment images varies too much, which make image segmentation hard. Therefore, this research subject becomes a hot topic in the world during last thirty years. Today, a number of image systems have been developed for measuring fragments in different application environments such as fragments on/in gravitational flows, conveyor belts, rockpiles, and laboratories. The research and development has been and is being carried out in many countries, the detailed information can be found in [2-5].

The common problem for the research and development is to delineate every rock fragment, but in most industrial cases, rock fragment images are difficult to segment due to rough surface, overlapping, and size variation etc. Hence, it is crucial to extract qualitative information about a rock fragment image to characterize images before starting segmentation. In this paper, characterization of rock fragment images have been thoroughly investigated by extensive tests on hundreds of images, using several packages of commercial software for image segmentation, and some previous image segmentation algorithms[6-11] coded by the author.

Image classification was recognized by author to be essential in segmentation of rock fragments. Therefore, it was started by developing procedures for crude determi-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 887 – 892, 2005.
© Springer-Verlag Berlin Heidelberg 2005

nation of number of rock fragments in an image, the basic idea being that "edge density" is a rough measure of average size in images of densely packed fragments. This work is an essential component in the current segmentation process. The studied segmentation algorithm is based on grey value valleys which is a grey-value structure occurring more frequently than traditional step edges. However, without knowledge of scale (approximate size of rock fragments in the image) such an approach would be hard to realize. In fact, this goes for any segmentation technique which normally "handles" the problem by adjustment of various smoothing parameters, thresholds etc. Since it needs an automatic image segmentation process to perform "image classification" first, and to avoid making "smoothing parameters" crucial for good results.

2 Outline of the Whole Segmentation Procedure

The whole segmentation procedure consists of the two parts in this study: image classification, and fragment delineation. Since the delineation algorithm needs thin edges of fragments, the classification algorithm first classify image into different classes, then based on the classification results, the procedure shrink the image into a certain scale size, provided for fragmentation delineation. After the delineation, the delineated image is converted to the original image size, re-mapping the contours of fragments. The rock fragment image classification algorithm was developed for general-purpose of rock fragment image segmentation. The algorithm evaluates image quality and produces image class labels, useful in subsequent image segmentation. Because of the large variation of rock fragment patterns and quality, the image classification algorithm produces five different labels for the classes: (1) images in which most of the fragments are of small size; (2) images in which most of the fragments are of medium size; (3) images in which most of the fragments are of relative large size; (4) images with mixed fragments of different sizes; and (5) images with many void spaces. If most fragments in an image are very small, the fine-detail information in the image is very important for image segmentation, and the segmentation algorithm must avoid destroying the information. On the contrary, if fragments are large, it is necessary to remove the detailed information on the rock fragment surface, because it may cause image over-segmentation. If most fragments are of relative large size (e.g. medium size), the segmentation algorithm should include a special image enhancement routine that can eliminate noise of rock fragment surface, while keeping real edges from being destroyed. Since the delineation algorithm was developed for densely packed fragments, the void spaces have to be removed before the delineation.

Consider the case of an image containing closely packed rock fragments, which can be approximated by ellipses in the image plane. The approximation is not done for the purpose of describing individual fragment shape, but for setting up a model for relating edge density to average size. With known average shape factor= $\bar{s}$ for fragments, average size $\bar{L}$ of fragments in a single frame can be solved from $\bar{L} + \frac{\sigma^2{}_L}{\bar{L}} = \frac{\bar{s}}{\sum P / \sum A} = \frac{\bar{s}}{\hat{\delta}_*}$, there is a relation between average length $\bar{L}$ and a kind of edge density $\hat{\delta}^*$.

After image classification, the image scale is reduced. Let $x = 1,...,n$, $y = 1,...,m$, and $f(x, y)$ is the original image. Then

$$f(x_2, y_2),\ x_2 = 1,...,n/2,\ y_2 = 1,...,m/2 \text{ (CLASS 2)} \quad (1)$$

$$f(x_3, y_3),\ x_3 = 1,...,n/4,\ y_3 = 1,...,m/4 \text{ (CLASS 3)} \quad (2)$$

are by sub-sampling (or weighted sub-sampling).

A recursive segmentation process is used for Class 4 images, separated into two or three steps: In the first step, the image is treated in the same way as an image with large-size (or medium-size) fragments, but thereafter the strategy is to examine each delineated fragment to see if the region can be processed further, based on edge density estimates, in the non-shrunk image.

In order to detect fragment edges clearly and disregard the edges of the texture on a fragment, the algorithm just detects valley-edges between fragments in the first step; it detects each image pixel to see if it is the lowest valley point in a certain direction. If it is, the pixel is a valley-edge candidate.

There is also a special class of images, Class 5. This class refers to any of Classes 1 to 4 on a clear background, hence only partially dense. In this special case, Canny edge detection [11] is a good tool for delineating background boundaries for clusters of rock fragment.

3 Rock Fragment Image Segmentation Algorithm

A simple edge detector uses differences in two directions: $\Delta_x = g(x+1, y) - g(x, y)$, $\Delta_y = g(x, y+1) - g(x, y)$. In the valley detector, four directions are used. Obviously, in many situations, the horizontal and vertical grey value differences, do not characterize a point.

Assume that P is surrounded by strong negative and positive differences in the diagonal directions: $\nabla_{45} < 0$, and $\Delta_{45} > 0$, $\nabla_{135} < 0$, and $\Delta_{135} > 0$, whereas, $\nabla_0 \approx 0$, and $\Delta_0 \geq 0$, $\nabla_{90} \approx 0$, and $\Delta_{90} \approx 0$. Where Δ are forward differences: $\Delta_{45} = f(i+1, j+1) - f(i, j)$, and ∇ are backward differences: $\nabla_{45} = f(i, j) - f(i-1, j-1)$, etc. for other directions. It uses $\max(\Delta_\alpha - \nabla_\alpha)$ as a measure of the strength of a valley point candidate. It should be noted that sampled grid coordinates are used, which are much more sparse than the pixel grid $0 \leq x \leq n$, $0 \leq y \leq m$. f is the original grey value image after weak smoothing. One of examples is shown in Fig. 1, the newly developed algorithm is much better than other existing algorithms.

What should be stressed about the valley edge detector are: (a) It uses four instead of two directions; (b) It studies value differences of well separated points: the sparse $i \pm 1$ corresponds to $x \pm L$ and $j \pm 1$ corresponds to $y \pm L$, where $L >> 1$, in our case, $3 \leq L \leq 7$. In applications, if there are closely packed particles of area > 400 pixels, images should be shrunk to be suitable for this choice of L. Section 3 deals

with average size estimation, which can guide choice of L; (c) It is nonlinear: only the most valley-like directional response $(\Delta_\alpha - \nabla_\alpha)$ is used. By valley-like, it means $(\Delta_\alpha - \nabla_\alpha)$ value. To manage valley detection in cases of broader valleys, there is a slight modification whereby weighted averages of $(\Delta_\alpha - \nabla_\alpha)$ - expressions are used. $w_1\Delta_\alpha(P_B) + w_2\Delta_\alpha(P_A) - w_2\nabla_\alpha(P_B) - w_1\nabla_\alpha(P_A)$, where, P_A and P_B are two end points in a section. For example, $w_1 = 2$ and $w_2 = 3$ are in our experiments; (d) It is one-pass edge detection algorithm; the detected image is a binary image, no need for further thresholding; (e) Since each edge point is detected through four different directions, hence in the local part, edge width is one pixel wide (if average particle area is greater than 200 pixels, a thinning operation follows boundary detection operation); and (f) It is not sensitive to illumination variations.

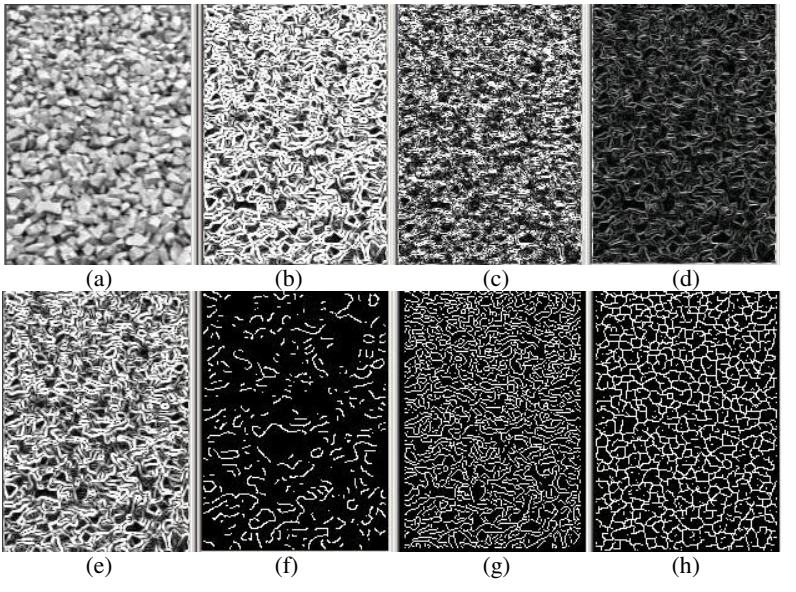


Fig. 1. Testing of edge detection algorithms: (a) Original image; (b) Sobel detection; (c) Robert detection; (d) Laplacian detection; (e) Prewitt detection; (f) Canny detection with a high threshold; (g) Canny detection with a low threshold; and (h) Boundary detection result by the new algorithm

Without image classification, there is a substantial difficulty choosing an appropriate L, the spacing between sampled points. Let L refer to spacing in an image of given resolution, where the given resolution may be a down sampling of the original image resolution. Since the image classification described earlier leads to an automatic down-sampling, the choice of L is not critical.

For Class 4 images – mixed size images – multiple values of L are used. Within each detected coarse-scale fragment, edge density variations in the un-shrunk image (may) trigger a new valley-detection search. The L-value has not changed, numerically, but it is operating in higher resolution variations of the image (Class 2 and Class 1), hence L in terms of the original image decreases.

After valley edge point detection, there are pieces of valley edges, and a valley edge tracing subroutine, filling gaps is needed (Some thinning is also needed). As a background process, there is a simple grey value thresholding subroutine which before classification creates a binary image with quite dark regions as the bellow-threshold class. If this dark space covers more than a certain percentage of the image, and has few holes, background is separated from fragments by a Canny edge detector [11] along the between-class boundaries.

In that case, the image is then classified into Class 1 to 4, only after separation of background. This special case is not unusual in quarry rock fragment data on dark conveyor belt. This is reasonable cooperative process. If background is easily separable from brighter rock fragments this is done, and dense sub-clusters are handled by the image classification and valley-edge segmentation. This part of the segmentation process is specific for rock fragment images where part of a homogeneous (dark) background is discernible.

4 Experimental Testing Results

To test the segmentation algorithm, a number of different fragment images from a laboratory, a rockpile, and a moving conveyor belt, have been taken (Fig. 2).

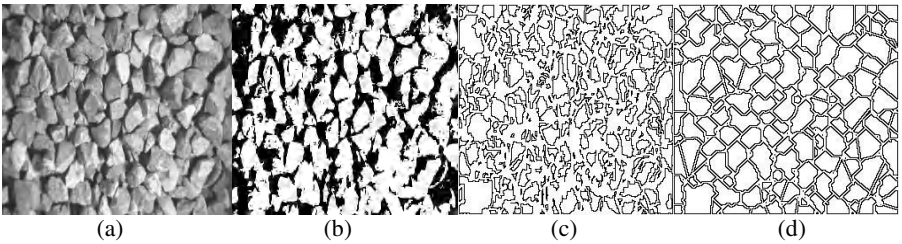


Fig. 2. A densely packed fragment image and segmentation results: (a) Original image (512x512 resolution, in a lab, Sweden), (b) Auto-thresholding, (c) Segmentation on similarity, and (d) New algorithm result

As an illustration of typical difficulties encountered in rock fragment images, Fig. 2 shows an example. Where, thresholding [9] (Fig.2b) and similarity-based segmentation algorithms [10] (Fig. 2c) are applied. The segmentation results are not satisfactory. Auto-thresholding gives an under-segmentation result, and similarity-based segmentation gives an over-segmentation result. By using the new algorithm, the image is classified into class 2, in which most fragments are of medium size, according to the classification algorithm. After image classification, fragments are delineated by the new segmentation. The segmentation result is reasonable.

5 Conclusion

In this paper, a new type of segmentation algorithms has been studied and tested; the combination of image classification algorithm and fragment delineation algorithm has

been described. The rock fragment image classification algorithm was developed based on valley edge detection for the general-purpose of image segmentation of densely packed rock fragments. The classification algorithm produces image class labels, useful in subsequent image segmentation. The fragment delineation algorithm studied is actually based on both valley-edge detection and valley-edge tracing. The presented rock fragment delineation algorithm seems robust for densely packed complicated objects; it is also suitable for other similar applications in the areas of biology, medicine and metal surface etc.

References

1. Gallagher, E: Optoelectronic coarse particle size analysis for industrial measurement and control. Ph.D. thesis, University of Queensland, Dept. of Mining and Metallurgical Engineering, (1976).
2. Franklin, J.A., Kemeny, J.M., Girdner, K.K: Evolution of measuring systems: A review. In: Franklin JA, Katsabanis T. (eds): Measurement of Blast Fragmentation, Rotterdam: Balkema (1996), 47-52.
3. Schleifer J, Tessier B: Fragmentation Assessment using the FragScan System: Quality of a Blast. Int. J. Fragblast, Volume 6, Numbers 3-4 (2002) 321 – 331.
4. Kemeny J, Mofya E, Kaunda R, Lever P: Improvements in Blast Fragmentation Models Using Digital Image Processing. Int. J. Fragblast, Volume 6, Numbers 3-4 (2002) 311 – 320.
5. Norbert H Maerz, Tom W, Palangio: Post-Muckpile, Pre-Primary Crusher, Automated Optical Blast Fragmentation Sizing. Int. J. Fragblast, Volume 8, Number 2 (2004) 119 – 136.
6. Pal, N.R, Pal, S.K: A review of image segmentation techniques. Int. J. Pattern Recognition, Vol. 26, No. 9 (1993) 1277-1294.
7. Wang, W.X., 1999, Image analysis of aggregates, J Computers & Geosciences, No. 25, 71-81.
8. Wang, W.X:, Binary image segmentation of aggregates based on polygonal approximation and classification of concavities. Int. J. Pattern Recognition, 31(10) (1998) 1503-1524.
9. Otsu, N: A threshold selection method from gray-level histogram. *IE*EE Trans. Systems Man Cybernet, SMC-9(1979) 62-66.
10. Suk, M., Chung, SM: A new image segmentation technique based on partition mode test. Int. J. Pattern Recognition Vol. 16, No. 5 (1983).469-480.
11. Canny, J.F: A computational approach to edge detection. Int. J. PAMI-8, No.6 (1986).

A Novel Wavelet Image Coding Based on Non-uniform Scalar Quantization

Guoyuo Wang[1] and Wentao Wang[1,2]

[1] Institute for Pattern Recognition and Artificial Intelligence,
Huazhong University of Science and Technology,
1037 Luoyu Road, Wuhan 430074, China
wangguoyou@hust.edu.cn
[2] College of Computer Science, South-Central University for Nationalities,
Minyuan Road, Wuhan 430074, China
wangwt_long@yahoo.com.cn

Abstract. In this paper, we investigate the problem of how to quantize the wavelet coefficients in the lowest frequency subband with non-uniform scalar method. A novel wavelet image coding algorithm based on non-uniform scalar quantization is proposed. This algorithm adopts longer step to quantize the wavelet coefficients in the lowest frequency subband and uses shorter step for other ones. According as the results of the experiment we design a coding approach by using two labels 0 or 1 to code a coefficient bit of decimal plane. Experiment results have shown the proposed scheme improves the performance of wavelet image coders. In particular, it will get better coding gain in the low bit rate image coding.

1 Introduction

Image compression encoding based on wavelet is one of hot research problem. Several very competitive algorithms, e.g. embedded zero-tree wavelets (EZW) of Shapiro, set partitioning in hierarchical trees (SPIHT) of Said and Pearlman, and embedded block coding with optimized truncation (EBCOT) of Taubman have been developed [1-4]. EZW is effective and computationally simple algorithm. With an embedded bit stream, the reception of code bits can be stopped at any point, and the image can then be reconstructed immediately. SPIHT is an improved version of EZW. It improves the coding performance by exploiting the self-similarity of the coefficients across subbands more efficiently than EZW. Although it is less efficient in coding performance than EBCOT, which forms the basis of the JEPG2000 standard in image coding, it has much lower computational complexity than EBCOT. So there are many researchers have great interest in improving the performance of SPIHT.

In [5], a pre-processing method, which applies the discrete sine transform (DST) or the discrete cosine transform (DCT) to the wavelet coefficients in the highest frequency subbands and in the next highest frequency subbands before the SPIHT encoding, is proposed. First, it gets the correlation coefficients of each of the highest frequency subbands. Second, it applies the DST or DCT according to the correlation

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 893 – 898, 2005.
© Springer-Verlag Berlin Heidelberg 2005

coefficients of the wavelet coefficients in the subbands. This method can helpfully pack the energy of the correlated coefficients into a relatively few components of a DST or DCT block, but it increases the computational complexity greatly. Reference [6] thinks that the refined and re-refined bit streams of SPIHT in the preceding phase may not transmit and replace them with the bit streams derived from the important coefficients in the next phase. The problem of this method is that it should add some synchronization bits in the bit streams. This will reduce the compression ratio. The rate-distortion (R-D) optimized significance map pruning method [7] uses a rate-distortion criterion to decision the significance coefficients. They all add complexity of the coder algorithm. In [8], DPCM (differential pulse code modulation) and SPIHT are adopted. But it is not fit for the low bit rate encoding.

In this paper, we check the characteristic of the wavelet coefficients of the multi-level wavelet decomposition, especially 9/7 wavelet filter. The Daubechies 9/7 wavelet [9] is one of the most popular ones, since it combines good performance and rational filter length. It is stressed here that the 9/7 wavelet is a default filter of the JPEG2000 standard [10] and included in the MPEG4 standard [11]. From the results of our experiment we find that it is possible to quantize the wavelet coefficients using non-uniform strategy. It is useful to improve the coding ratio. Meanwhile, a new simple and efficient pre-processing coding method: like-DPCM is proposed.

2 Quantization and Coder of Wavelet Coefficient

2.1 SPIHT Coder

The SPIHT coder [2] organizes the wavelet coefficients of an image into the spatial orientation trees by using wavelet pyramid. It uses three types of sets: $D(i, j)$ denoting the set of all descendants of a node (i, j), $O(i, j)$ representing the set of all offspring of the node (i, j), and $L(i, j)$ representing the set of all descendants excluding the immediate four offspring of the node. The SPIHT coder starts with the most significant bit plane. At every bit plane, it tests the tree lists in order, starting with LIP (list insignificant pixels) are coded. Those that become significant are moved to the end of the LSP (list of significant pixels) and their signs are coded. Similarly, set $D(i, j)$ and $L(i, j)$ are sequentially coded following the LIS (list insignificant sets) order, and those that become significant are partitioned into subsets. Finally, each coefficient in LSP except the ones added in the last sorting pass is refined in each refinement pass. The algorithm then repeats the above procedure for next resolution, and then stops at desired bit rates.

2.2 Quantization Plane

Probabilistic quantization model of wavelet coefficients will be influence the performance of wavelet coders. We know that a wavelet coefficient can be represented by a binary polynomial (binary uniform quantization) or a decimal polynomial (decimal uniform quantization). It is show as equation (1).

$$\begin{aligned} w_{i,j} &= b_n \times 2^n + b_{n-1} \times 2^{n-1} + \ldots + b_0 \times 2^0 + \cdots \\ &= d_k \times 10^k + d_{k-1} \times 2^{k-1} + \ldots + d_0 \times 10^0 + \cdots \end{aligned} \tag{1}$$

Here, $w_{i,j}$ is a wavelet coefficient; b_n is the coefficient of n-level bit plane; d_k is the coefficient of k-level "10 plane". The relation between k and n is show in equal (2). Figure 1 shows the quantization step of "10 plane" is longer than "bit (2) plane".

$$k \le n, \quad k, n \in N \tag{2}$$

From equation (2) we know that the number of plane of a wavelet coefficient denoted by using decimal system is fewer than by using binary system. So we can assume that if we could use two labels, for example 0 and 1, to represent the quantization coefficient (d_k) of each decimal plane we would get more coding efficiency.

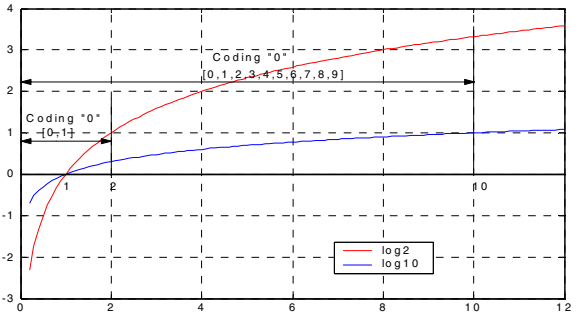


Fig. 1. The example of log2 and log10 for quantization

3 The Proposed Method

How to quantize, represent and code efficiently the wavelet transform coefficients of the image is one of the main problems for a wavelet image coder. The SPIHT coder uses the binary uniform quantization. Its coding structure and exploiting strategy are based on zero-tree root. Yet, these simple methods have two main obvious shortcomings. First, it treats each quantization plane equally. Since there may be very few significant coefficients in some or other plane but the code for locating the position of insignificant coefficients in the same subband should be still needed. The position bits are output by SPIHT to represent the proper positioning of the value bits in the bit stream so that the decoder can reproduce the same value information correctly. Second, the wavelet coefficients in high frequency subbands are always scanned and coded with other wavelet coefficients in the low frequency subband at the same time. But the energy of each wavelet coefficients in the low frequency subband is always higher than the high frequency subbands. So, some labels 0, which denote the zero-tree root but they are not always necessaries, will be coded in the bit streams. In this way the coding efficiency for significant coefficients is not high.

We analyze the wavelet coefficients by using the 9/7 filter to multi-decompose some test images which size are all 512×512. The results of the 7-level lowest frequency subband of wavelet coefficient show as Table 1. From the results of Table 1 we can find that the differences of the highest bit between the wavelet coefficients are almost 0 or 1. This result will give us a chance to quantize the these wavelet coefficients by using terminal plane and code the leftmost bit of them only using two labels of 0 and 1. This method of non-uniformity quantization accords with the characteristic of the human vision system (HVS).

The novel coding algorithm is based on four main concepts: 1) do non-uniform quantization for wavelet coefficients; 2) adjust the value of the leftmost bit of the wavelet coefficients in the high frequency subband to fit to code by using the two labels 0 and 1; 3) like DPCM; 4) use SPIHT to code the remain wavelet coefficients.

The steps of the proposed coding algorithm are summarized below.Equal (3) shows the decimal system quantization:

$$w_p^{(i,j)} = d_k^{(i,j)} \times 10^k + d_{k-1}^{(i,j)} \times 2^{k-1} + \cdots, w_p^{(i,j)} \in LL_p \tag{3}$$

a) do { 9/7 DWT for original image } until (

$$abs(\hat{d}_k^{(i,j)} - \max(d_k^{(m,n)} \mid w_p^{(m,n)} \in LL_p)) \in \{0,1\} \quad i,j,m,n = 1,2,...,size(LL_p)$$

or $size(LL_p) = 1$)

here LL_p denotes the p-level wavelet coefficients in the lowest frequency subband. $\hat{d}_k^{(i,j)} = d_k^{(i,j)}$ or its value is adjusted. The value of k is as equation (4).

$$k = \lfloor \log_{10}(w_p^{(i,j)}) \rfloor \tag{4}$$

If $size(LL_p) = 1$ then goto d).

b) Coding:

for each $\hat{d}_k^{(i,j)}; i, j = 1,2,...,size(LL_p)$ do

{ If ($\hat{d}_k^{(i,j)} = \max(d_k^{(m,n)} \mid d_k^{(m,n)} \in LL_p)$)

{ output 1 ; }

else

{ output 0; } }

The values of k and $\max(d_k^{(m,n)} \mid d_k^{(m,n)} \in LL_p)$ are coded into the head of bit streams.

c) $\tilde{w}_p^{(i,j)} = \hat{w}_p^{(i,j)} - \hat{d}_k^{(i,j)} \times 10^k$,then get new $\widetilde{LL}_p$ ($\tilde{w}_p^{(i,j)} \in \widetilde{LL}_p$, $i, j = 1,2,...,size(LL_p)$).

d) adapt binary system quantization and use SPIHT to code the remain coefficients.

4 Experimental Results and Discuss

In the implementation of the proposed algorithm, the input image is transformed using a seven-level wavelet decomposition based on the 9/7 DWT. Experiments are performed on the 512×512 gray scale images such as Barbara, Lena, Airplane and Woman etc. The distortion is computed from actually decoded images and measured by the peak signal to noise ration (PSNR).

Table 1 gives the obtained results of the wavelet coefficients in the lowest frequency subband after an image has been transformed seven levels and all data are represented by decimal system. From Table 1, we can find that the proposed method can be almost used for all testing images directly. But some values such as Worman2 should be adjusted before encoding. For example, the number 0.7235 can be written $(1.0+(-0.2765))$, the number 0.4788 can be written $(1.0+(-0.5212))$ etc. After finished the adjusting for the wavelet coefficients we can use the proposed method (like-DPCM) to code.

Table 2 summarizes the performance comparison of the proposed method in PSNR (dB) at various bit rates simulated on test images such as Barbara, Lena, Airplane, and Woman2. The filter is the biorthogonal 9/7 wavelet and the level of wavelet transform is seven.

We can show in Table 2 that the improvement in PSNR obtained with new method. In particular, the new method proposed in the paper offers good performance for the images in low bit rate. This is because the proposed method mainly increases the coding ratio in the lowest frequency subband.

Table 1. The wavelet coefficients of the lowest frequency subband of the 7 levels of the 9/7 wavelet decomposition

Barbara: (1.0e+004)	**Lena**: (1.0e+004)	**Airplane**: (1.0e+004)
1.1386 1.0519 1.4902 1.5121	1.4361 1.2758 1.7088 1.8182	2.1554 2.4142 2.3613 2.0645
1.5045 1.9389 2.3921 1.4577	1.5196 1.5563 2.0975 1.7652	2.3487 2.1971 2.3730 2.4529
2.0495 1.6029 1.7665 1.1654	1.8559 1.2462 1.6478 1.4625	2.0918 1.7638 1.9605 2.2399
1.2393 1.0840 1.4278 1.2205	1.7504 1.0536 1.5136 1.6981	2.4738 2.1006 1.7280 2.0674
Woman1: (1.0e+004)	**Woman2**: (1.0e+004)	**Pepper**: (1.0e+004)
1.4850 1.4704 1.8191 1.5661	1.6302 1.5646 1.6209 1.9144	1.8133 1.7793 1.2785 1.5694
1.4971 1.8641 1.8489 2.1093	1.6575 0.7235 0.8219 1.1443	1.8074 1.4383 1.3796 1.5814
1.5420 1.5245 1.6870 2.0472	2.0708 1.2471 1.6575 1.0003	1.1354 1.8345 1.6609 1.6942
1.5067 1.9543 1.8550 2.0906	1.7749 1.2985 1.6851 0.4788	0.9280 1.6124 1.1935 1.4124

Table 2. Performance comparison of the proposed method in PSNR (dB) at various bit rates

	Barbara		**Lena**		**Airplane**		**Woman2**	
Bit rates	SPIHT	Our method	SPIHT	Our method	SPIHT	Our method	SPIHT	Our method
0.01dpp	19.500	19.540	21.936	22.063	22.068	22.180	25.321	25.447
0.05dpp	22.633	22.647	26.641	26.669	26.184	26.198	32.372	32.423
0.10dpp	24.145	24.161	29.278	29.291	28.455	28.479	35.578	35.601
0.50dpp	31.387	31.388	36.582	36.586	37.053	37.059	42.022	42.025

5 Conclusions

A new image coding algorithm based non-uniform scalar quantization for the wavelet coefficients has been proposed in this paper. Experiments show that the proposed method consistently outperforms the original SPIHT coder for all the popular test images. In particular, the proposed method increases the performance of low bit rate image coding. The computation complexity is not increased. By contrast, our approach is applicable to 9/7 wavelet filter and typical test images.

Acknowledgements

This work supported by the National Natural Science Foundation of China (NSFC General Projects, Grant No. 60372066). The authors would like to thank the reviewers.

References

1. J.M. Shapiro: Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans. Signal Processing, Vol. 41, (1993)3445-3463
2. A. Said and W.A. Pearlman: A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. Circuits Syst. Video Technol., Vol. 6, (1996) 243-250
3. Servetto, S.D., Ramchandran, K., Orchard, M.T.: Image coding based on a morhplogical representation of wavelet data. IEEE Trans. IP 8, (1999)1161-1174
4. D. Taubman, "High Performance Scalable Image Compression with EBCOT," IEEE Trans. Image Processing, Vol. 9, (2000)1158-1170
5. Ki-Lyug Kim, Sung-Woong Ra: Performance improvement of the SPIHT coder. Signal Processing: Image Communication, vol. 19, (2004)29-36
6. Chun-lianf Tung, Tung-Shou Chen, Wei-Hua Andrew Wang and Shiow-Tyng Yeh: A New Improvement of SPIHT Progressive Image Transmission. IEEE Internat. Conf. on Multimedia Software Engineering,(2003)
7. Ulug Bayazit: Significance map pruning and other enhancements to SPIHT image coding algorithm. Signal Processing: Image Communication, Vol. 18, (2003)769-785
8. Min Shi, Shengli Xie: A Lossless Image Compression Algorithm by Combining DPCM with Integer Wavelet Transform. IEEE Internat. Conf. on Mobile and Wireless Comm. (2004)293-296
9. A. Cohen, I. Daubechies and J.C. Feauvcau: Biorthogonal bases of compactly supported wavelets. Communications on Pure and Appl. Math., Vol. 5, (1992)485-560
10. ISO/IEC FCD15444-1:2000 V1.0: JPEG 2000 Image Coding System. offical release expected Mar.(2001)
11. ISO/IEC JTC1/SC29/WG11, FDC 14496-1: Coding of Moving Pictures and Audio. (1998)

A General Image Based Nematode Identification System Design

Bai-Tao Zhou, Won Nah, Kang-Woong Lee, and Joong-Hwan Baek

School of Electronics and Communication Engineering, Hankuk Aviation University, Korea
{zhou, nahwon, kwlee, jhbaek}@hau.ac.kr

Abstract. Nematodes are primitive organisms which nonetheless devour many of the essential resources that are critical for human beings. For effective and quick inspection and quarantine, we propose a general image based system for quantitatively characterizing and identifying nematodes. We also describe the key methods ranging from gray level image acquisition and processing to information extraction for automated detection and identification. The main contributions of this paper are not only presenting a framework of the system architecture, but also giving detail analysis and implementation of each system component with instance of *Caenorhabditis elegans*. Therefore with a little modification, this system can be applied to other nematode species discrimination and analysis.

1 Introduction

Nematodes are the most numerous multicellar animals on the earth. A handful of soil contains thousands of the microscopic worms, many of them parasites of insects, plants or animals. It's well known that some species are threatening the development of farming and forestry all over the world. From the NCFAP report [6], U.S commercial pineapple production centered in Hawaii have reduced as high as 47% in 2002 due to uncontrolled nematode populations. So detection and identification of nematodes are the prerequisite for controlling them and checking their spread.

Traditional methods for nematode identification request repetitive work and also expertise from nematologists. But sometimes naked eye observations are subjective and imprecise. To enhance the reliability in great extent, some automated recording and analysis systems have been described. Geng et al. [3] solved behavior phenotype classification of some *C. elegans*. Their work focused on a single type of worm, not concerning about different species identification problems. Silva et al. [7] proposed an intelligent nematode detection system. But the system only detected whether worms exist or not. Based on our previous work [1], we designed an advanced system that can automatically identify different nematodes simultaneously supported by nematode feature database.

This paper is organized as following way. In section 2, we present the system architecture. In section 3, section 4 and section 5, we describe the system from the aspects of image acquisition and processing, feature extraction, and problem domain modeling and identification respectively. The *C. elegans* worms as examples are employed to illustrate theses procedures. Section 6 concludes this article.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 899 – 904, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 System Architecture

The General system is composed of four functional blocks, which are Image Acquisition Block (IAB), Image Processing Block (IPB), Feature Extraction Block (FEB), and Problem Domain Modeling Block (PDMB). The PDMB is used to build different classification models according to their domain requirements. Fig.1 illustrates system architecture of the proposed system.

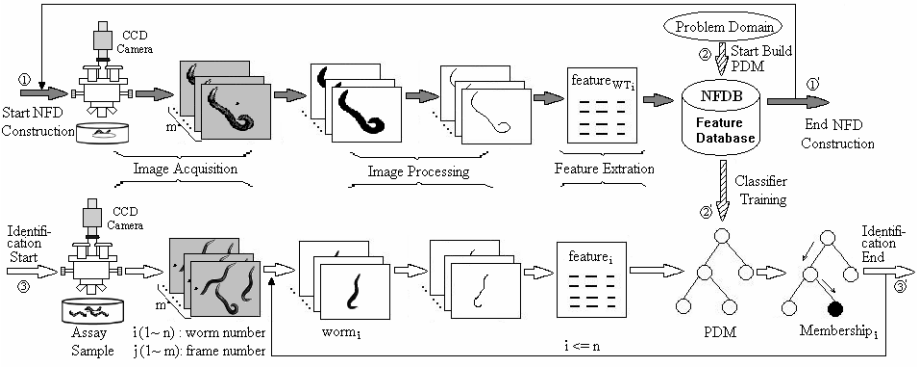


Fig. 1. The system architecture and working procedures

The system starts with constructing a general Nematode Feature Database (NFDB). By feeding IAB each time with a new type-known assay sample, NFDB can record as much species information as possible so as to let it cover most of the problem domains. In the second phase, the system builds the PDM by querying NFDB. The querying terms come from the domain requirement, and the queried result is the domain-scope nematode features. These features, as training data, are used to build the classifier model for that domain nematode identification. In the third phase, the classifier decides the assay nematode memberships by feeding the assay sample features to the PDM.

3 Image Acquisition and Processing

We designed an automatic track system motorized microscope that could record an individual animal's behaviors at high magnification. In detail, the *C. elegans* locomotion was tracked with a stereomicroscope mounted with a high performance CCD video camera. And a computer-controlled tracker was employed to put the worms in the center of the optical field of the stereomicroscope [1]. The animal was snapped every 0.5 second for at least five minutes to record the locomotion. The collected data is a series of time-index image frames, which are saved as eight-bit gray scale data.

To facilitate feature extraction, first, the gray scale image is converted into binary image based on the distribution of the gray value. Secondly, the spots inside the worm body are removed with a morphological closing operator [4]. Thirdly, the multiple objects are segmented with the sequential algorithm for component labeling [5]. The areas containing worms are processed respectively. The areas containing isolated

objects, such as eggs, are removed by setting off pixel. At last, the morphological skeleton was obtained by applying a skeleton algorithm [8]. By iterating dilation operation along the binary image, a clean skeleton can be obtained. The image process procedure and results are depicted in Fig. 2 and Fig.3.

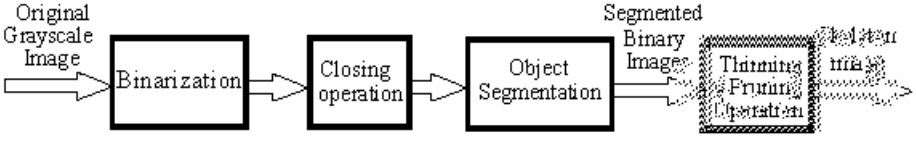


Fig. 2. The image processing procedure

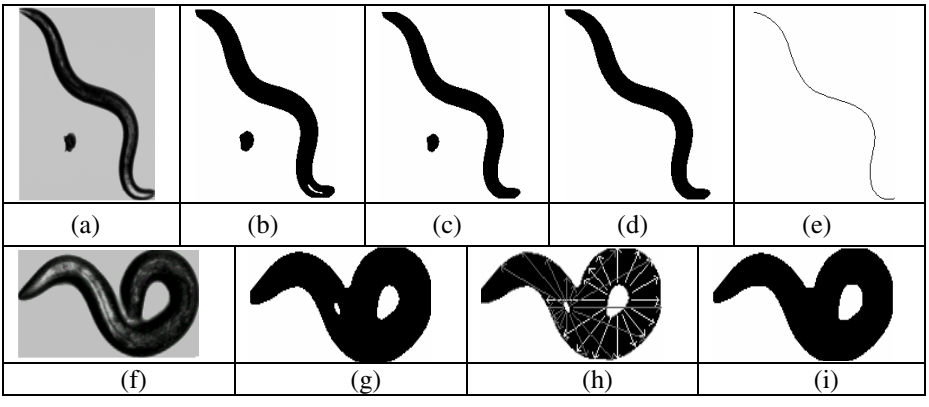


Fig. 3. (a) Original gray level image, (b) Binary image after thresholding operation, (c) Binary image following closing operation, (d) Final clean binary image after removal of isolated object, (e) Skeleton obtained through thinning and pruning, (f) *C. elegans* gray level image, (g) The result of binarization with median filtering, (h) Test each labeled area except worm body using 16 vectors, (i) The result of hole detection and noise removal.

Another aspect need to mention here is to distinguish the looped holes and noise spots inside the worm body, illustrated in Fig. 3 (f)-(i). If we use the operation mentioned in the second step, some holes may be filled as the spots inside the worm body. To remove the spots inside the worm body and to remain the holes simultaneously, we measure thickness of these two regions with 16 vectors. The number of pixels, or thickness, from the centroid to the background in each vector direction is counted for every region. If the total thickness is less than a threshold, in our experiment 25, the region is considered as noise. Otherwise, we preserve the region as a hole.

4 Feature Extraction

Based on the segmented binary images and skeleton images, the system can extract the features to describe the body shapes and locomotion patterns. The measured features in our system include the minimum, maximum, and average values of the following: distance moved in 0.5, 5, 10, 15, 20, 25, and 30s; and 5 min, number of reversals in 40, 80, and 120s; and 5 min, worm area, worm length, width at center and

head/tail, ratio of thickness to length, fatness, eccentricity and lengths of major/minor axes of best-fit ellipse, height and width of minimum enclosing rectangle (MER), ratio of MER width to height, ratio of worm area to MER area, angle change rate etc. Some measurements of these features are described as following:

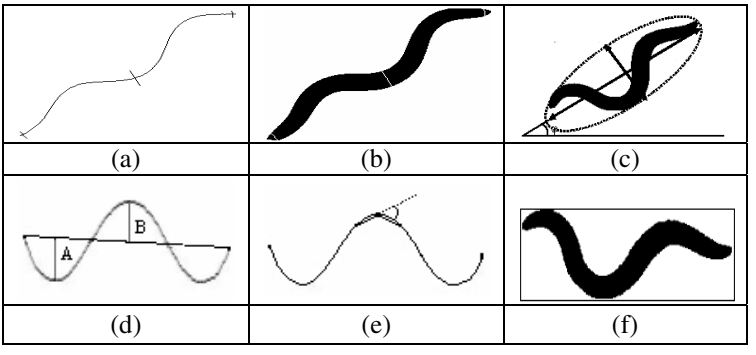


Fig. 4. Some body feature measurements. (a) Center and head/tail position to measure thickness, (b) The center and head/tail thickness, (c) The best-fit ellipse of the shape, (d) The largest and second largest peaks of the skeleton. The sum of the two peak distance is the animal's amplitude, (e) Measurement of the angle change rate, (f) The MER of the shape.

4.1 Features to Describe the Worm Shape

1) The worm thickness/width: measured at the center, head and tail position of the skeleton. The center width can be measured by finding the best-fit line with 9-pixel-long segment near the center position, rotating the line by 90° to get a perpendicular line to cross the contour of the worm, and the distance of the two cross point is the width. The head and tail width are measured in the same way (Fig. 4b).
2) The eccentricity: defined as the ratio of the distance between the foci of the best-fit ellipse of the worm shape (Fig.4c.) to its major axis length.
3) The minimum enclosing rectangle (MER): By rotating the image according to the orientation of the best-fit ellipse's major axis, the MER of the shape can be obtained (Fig.4f). The height and width of the MER indicate the elongation tend.
4) The approximate amplitude: the sum of the two maximal distances (A and B) among the perpendicular distance from every skeleton point to the line connecting the two end points of the skeleton (Fig.4d). It is used to indicate the worm skeleton wave. Then amplitude ratio is defined as the ratio of min (A, B) to max (A, B).
5) The angle change rate (R): defined as the ratio of the average angle difference between every two consecutive segments along the skeleton to the worm length (Fig.4e). The skeleton points are spaced apart by 10 pixels. It can be represented by the following equations:

$$R = \left\{ \frac{1}{n-1} \sum_{i=1}^{n-1} \theta_i \right\} / L, where, \theta_i = \arctan \frac{y_{i+2} - y_{i+1}}{x_{i+2} - x_{i+1}} - \arctan \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \quad (1)$$

n is the number of segments and L is the worm length.

4.2 Features to Describe Locomotion Pattern

1) Speed: measured by sampling the centroid position over a constant time interval.
2) The reversals: measured by sampling the trajectory of the centroid at intervals of constant distance. The turning angle at every vertex is computed; if the angle is greater than 120, the position is considered to be a reversal.

Other features such as the amount of time to keep coiled and the how often it coiled can be calculated by looped hole detection method described in section 3. In order to eliminate the extreme values introduced by noise and angle of view, the data were summarized with such quantities as the 90^{th} and 10^{th} value out of the total data.

5 Problem Domain Modeling

To model a problem domain, such classification tools as SVM, Neural Nets and CART can be employed after querying NFDB with given worm types. We use the classification and regression tree (CART) algorithm [2]. In our experiment, the training set contained 8 worm types: wild-type, *goa-1*, *unc-36*, *unc-38*, *egl-19*, *nic-1*, *unc-29* and *unc-2*. In NFDB, each strain has 100 5-min recordings, with images captured every 0.5s. So the training samples consisted of 800 data points, which are fed to CART as training data. The original measurement vector is consisted of 94 features.

Table 1. Classification results from 10-fold cross validation

Actual	Predicted							
	Wild	Goa-1	Nic-1	Unc-36	Unc-38	Egl-19	Unc-2	Unc-29
Wild	0.950	0.000	0.000	0.000	0.000	0.000	0.020	0.030
Goa-1	0.010	0.930	0.000	0.020	0.010	0.000	0.000	0.020
Nic-1	0.000	0.000	0.880	0.000	0.060	0.000	0.000	0.000
Unc-36	0.000	0.000	0.000	0.760	0.050	0.020	0.120	0.050
Unc-38	0.000	0.000	0.000	0.040	0.730	0.000	0.050	0.170
Egl-19	0.000	0.000	0.000	0.020	0.000	0.930	0.000	0.020
Unc-2	0.033	0.067	0.017	0.150	0.100	0.000	0.717	0.000
Unc-29	0.000	0.000	0.000	0.020	0.080	0.010	0.010	0.760

According to the report from CART, we build an optimal classification tree using only 11 parameters (average and max distance moved in 0.5s, max angle change rate, number of reversals in 5m and in 40s, average and min center thickness, number of looped, worm area, average and max ratio of worm length to MER fill). And the 10-fold cross validation classification probability for each type is given in Table 1. The success rates are listed along the diagonal and the misclassification error rates are listed in the off-diagonal entries.

From the results of CART analysis, we can see an optimal classification tree with only 11 parameters can provide better classification with approximately 90% or higher success rates except for the Unc types. We checked the feature data, and found that the centers for *unc-29, unc-38, unc36* and *unc-2* were closer to each other in the feature space. Considering the similarities between Unc types, we are considering to

divide 2 clusters first: Unc type and other type, and then divide the Unc types respectively or looking for new measurements to identify the Unc type worms better.

6 Conclusion

This image based automatic identification of nematode system offers several advantages for the characterization of nematode phenotypes. First, the quantitative features based on image computation avoid the subjectivity from real-time observation. Second, the computerized imaging system has the potential to be much more reliable at detecting abnormalities that are subtle or manifested over long time scales. Finally, those images from computerized system can record multiple aspects of behavior that can be used to other relevant research work.

Acknowledgement

This research was supported by the Internet Information Retrieval Research Center (IRC) in Hankuk Aviation University. IRC is a Regional Research Center of Gyeonggi Province, designated by ITEP and Ministry of Commerce, Industry and Energy.

References

1. Baek, J., Cosman, P., Feng, Z., Silver, J., Schafer, W.R.: Using machine vision to analyze and classify *C.elegans* behavioral phenotypes quantitatively. J. Neurosci. Methods, Vol. 118 (2002) 9-21
2. Breiman, L., Fried J.H., Olshen R.A., Stone C.J.: Classification and regression trees. Belmont, CA, Wadsworth (1984)
3. Geng,W., Cosman, P., Berry, C.C., Feng, Z., Schafer, W.R.: Automatic Tracking, Feature Extraction and Classification of *C. elegans* Phenotypes. IEEE Trans. Biomedical Engineering. Vol. 51 (2004) 1181-1820
4. Gonzalez, R. C., Woods R. E.: Digital image processing (2nd Edition). PrenticeHall. (2002)
5. Jain, R., Rangachar, R., Schunck, B.: Machine Vision. McGraw-Hill New York (1995)
6. Leonard P. Gianessi, Cressida S. Silvers, Sujatha Sankula: Current and Potential Impact For Improving Pest Management In U.S. Agriculture An Analysis of 40 Case Study. National Center for Food & Agricultural Policy (2002)
7. Silva, C.A., Magalhaes, K.M.C., Doria Neto, A.D.: An intelligent system for detection of nematodes in digital images. Proceedings of the International Joint Conference on Neural Networks, Vol. 1 (2003) 20-24
8. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Comm. ACM, Vol. 27, No.3. (1984) 236-239

A Novel SVD-Based RLS Blind Adaptive Multiuser Detector for CDMA Systems

Ling Zhang and Xian-Da Zhang*

Department of Automation, Tsinghua University,
Tsinghua National Laboratory for Information Science and Technology,
Beijing 100084, China
lingzhang03@mails.tsinghua.edu.cn
zxd-dau@tsinghua.edu.cn

Abstract. In this paper, we propose a novel blind adaptive multiuser detector using the recursive least squares (RLS) algorithm based on singular value decomposition (SVD) for code division multiple access systems. The new presented algorithm can overcome the disadvantages of numerical instability and divergence of the conventional RLS algorithm. Simulation results show that the novel SVD-based RLS algorithm is superior to the conventional RLS algorithm in convergence rate, numerical stability and robustness.

1 Introduction

Recently, the blind adaptive multiuser detection (MUD) schemes for code division multiple access (CDMA) systems have been widely studied. It is well-known that the recursive least squares (RLS) algorithm can implement blind adaptive multiuser detector [1], [2]. Since the RLS algorithm has the advantages of conceptual simplicity and ease of implementation, it is widely used in MUD [1]-[4].

However, the blind adaptive multiuser detector based on the conventional RLS algorithm performs poorly or may be invalid in certain circumstances due to its disadvantages of numerical instability and divergence. The reason is that in the conventional RLS algorithm [1], the inverse of the covariance matrix $\mathbf{R}(n)$ in the update formulation is very sensitive to the truncation errors and there is no guarantee that $\mathbf{R}^{-1}(n)$ will always be positive and symmetric in the recursive procedure. To overcome these disadvantages, a singular value decomposition (SVD) based RLS algorithm is presented by [5]. However, the algorithm has the disadvantages of increased computation complexity and complicated implementation. Hence, it is unsuitable for blind adaptive MUD. At present, no SVD-based RLS blind adaptive multiuser detector has been presented in the literature. So this paper presents a novel blind multiuser detector using an SVD-based RLS algorithm.

This paper is organized as follows. In Section 2, we formalize the problem of blind adaptive MUD and give a brief introduction of the conventional RLS algorithm. In Section 3, a novel SVD-based RLS algorithm for blind adaptive MUD is proposed. This is followed by a comparison of the new algorithm with other related algorithms in Section 4. Simulation results are presented in Section 5 to demonstrate the performance of the new detector. Finally, the paper is concluded in Section 6.

---

* Senior Member (IEEE).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 905 – 910, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Problem Formulation

Consider an antipodal K-user synchronous direct-sequence CDMA (DS-CDMA) system signaling through an additive white Gaussian noise (AWGN) channel. By passing through a chip-matched filter followed by a chip-rate sampler, the discrete-time output of the receiver during one symbol interval can be modeled as [7]

$$y(n) = \sum_{k=1}^{K} A_k b_k s_k(n) + \sigma v(n), \quad n = 0,1,...,T_s - 1 \tag{1}$$

where, $T_s = NT_c$ is the symbol interval; T_c is the chip interval; N is processing gain; $v(n)$ is the additive channel noise; σ^2 is the variance of the noise $v(n)$; K is the number of users; A_k is the received amplitude of the k th user; b_k is the information symbol sequence of the k th user; $s_k(n)$ is the k th signature waveform that is assumed to have unit energy and with supported interval $[0, T_s - 1]$.

Defining $\mathbf{r}(n) \triangleq [y(0), y(1), ..., y(N-1)]^T$; $\mathbf{v}(n) \triangleq [v(0), v(1), ..., v(N-1)]^T$; $\mathbf{s}_k(n) \triangleq [s_k(0), s_k(1), ..., s_k(N-1)]^T$, we can express (1) in vector form

$$\mathbf{r}(n) = \sum_{k=1}^{K} A_k b_k \mathbf{s}_k(n) + \sigma \mathbf{v}(n). \tag{2}$$

For the above system, the blind adaptive multiuser detector based on the conventional RLS algorithm is proposed in [1]. However, as discussed in Section 1, this algorithm has some disadvantages due to the unstable recursion of the matrix $\mathbf{R}^{-1}(n)$, where, $\mathbf{R}(n)$ is the covariance matrix satisfying the following equation

$$\mathbf{R}(n) = \lambda \mathbf{R}(n-1) + \mathbf{r}(n)\mathbf{r}^T(n). \tag{3}$$

Hence, to overcome the numerical instability of the conventional RLS algorithm, we propose a novel SVD-based RLS algorithm for blind adaptive MUD in Section 3.

3 A Novel SVD-Based RLS Algorithm

In this section, we propose a novel SVD-based RLS algorithm for MUD. Using the knowledge of SVD [6], since $\mathbf{R}^{-1}(n)$ is symmetric positive definite matrix, applying the SVD to $\mathbf{R}^{-1}(n)$, we get

$$\mathbf{R}^{-1}(n) = \mathbf{U}(n)\Sigma^2(n)\mathbf{U}^T(n) \tag{4}$$

where $\mathbf{U}(n)$ is orthogonal matrix and $\Sigma(n)$ is a diagonal matrix with positive diagonal elements.

The objective of the SVD-based RLS algorithm is to get the update formulation of $\mathbf{U}(n)$ and $\Sigma^2(n)$. This method can ensure the matrix $\mathbf{R}^{-1}(n)$ be symmetric during iteration.

Using (4), we have

$$\mathbf{R}(n) = \mathbf{U}(n)\Sigma^{-2}(n)\mathbf{U}^T(n) . \tag{5}$$

Substituting (5) into (3) yields

$$\mathbf{U}(n)\Sigma^{-2}(n)\mathbf{U}^T(n) = \lambda\mathbf{U}(n-1)\Sigma^{-2}(n-1)\mathbf{U}^T(n-1) + \mathbf{r}(n)\mathbf{r}^T(n) . \tag{6}$$

Considering the $(N+1)\times N$ matrix

$$\begin{bmatrix} \sqrt{\lambda}\Sigma^{-1}(n-1)\mathbf{U}^T(n-1) \\ \mathbf{r}^T(n) \end{bmatrix}$$

of rank N and computing its SVD, we have

$$\begin{bmatrix} \sqrt{\lambda}\Sigma^{-1}(n-1)\mathbf{U}^T(n-1) \\ \mathbf{r}^T(n) \end{bmatrix} = \mathbf{U}_1(n-1)\begin{bmatrix} \Sigma_1(n-1) \\ 0 \end{bmatrix}\mathbf{V}^T(n-1) . \tag{7}$$

Multiplying each side on the left by its transpose yields

$$\begin{bmatrix} \sqrt{\lambda}\Sigma^{-1}(n-1)\mathbf{U}^T(n-1) \\ \mathbf{r}^T(n) \end{bmatrix}^T \begin{bmatrix} \sqrt{\lambda}\Sigma^{-1}(n-1)\mathbf{U}^T(n-1) \\ \mathbf{r}^T(n) \end{bmatrix} = \mathbf{V}(n-1)\Sigma_1^2(n-1)\mathbf{V}^T(n-1) .$$

From the above equation and (6), we have

$$\mathbf{U}(n)\Sigma^{-2}(n)\mathbf{U}^T(n) = \mathbf{V}(n-1)\Sigma_1^{\ 2}(n-1)\mathbf{V}^T(n-1) . \tag{8}$$

Comparing the two sides of (8), we get

$$\mathbf{U}(n) = \mathbf{V}(n-1) , \tag{9}$$

$$\Sigma^2(n) = \Sigma_1^{-2}(n-1) . \tag{10}$$

Thus, (7), (9) and (10) provide a new SVD scheme for the update of $\mathbf{R}^{-1}(n)$. Therefore, the novel SVD-based RLS blind adaptive multiuser detector is developed. In practice, we only need to save the right singular vector matrix $\mathbf{V}(n-1)$ and the diagonal matrix $\Sigma_1(n-1)$ in order to reduce the computational requirement.

4 Comparison with Other Related Algorithms

The SVD-based RLS algorithm increases computation complexity over the RLS algorithm. The latter has complexity $O(N^2)$, whereas the former has complexity $O(N^3)$. However, the SVD-based RLS algorithm performs better than the conventional RLS algorithm in convergence rate, numerical stability and robustness [5]. Simulation results will be shown in Section 5 to demonstrate the performance of the new algorithm for blind adaptive MUD.

As compared with the previous SVD-based RLS algorithm [5], our new algorithm is simpler and more efficient. Especially, our algorithm does not need the matrix in-

version lemma (with $O(N^3)$ complexity) as in Eq.8 and Eq.9 in [5]. Furthermore, the implementation of the new algorithm is very simple. Using (7), the update formulas of $\Sigma^{-1}(n)$ and $\mathbf{U}(n)$ can be obtained directly from (9) and (10). Different from Eq.15 and Eq.16 in [5], we do not need to compute the matrix-matrix multiplication (with $O(N^3)$ complexity) in the update formulation. Hence, our novel SVD-based RLS algorithm is more efficient and practical.

5 Simulation Results

In this section, we present several simulation results to compare the novel SVD-based RLS algorithm with the conventional RLS algorithm [1] for blind adaptive MUD.

Assume that user 1 is the desired user. To compare the convergence rate of the two algorithms, we use the excess output energy (EOE) as criterion. The time-averaged measured EOE at the n th iteration is expressed as

$$EOE(n) = \frac{1}{M}\sum_{l=1}^{M}[\mathbf{c}_{1l}^{T}(n)(\mathbf{y}_l(n) - b_{1,l}(n)\mathbf{s}_1)]^2 \tag{11}$$

where, M is the number of independent runs, and the subscript l indicates that the associated variable depends on the particular run. From (11), we can see that the "ideal" EOE value is zero.

The typical simulation example takes from [7]. In this example, synchronous DS-CDMA systems in an AWGN channel are simulated. All signal energies are given in decibels relative to the additive noise variance σ^2, i.e. $SNR = 10\log(E_k/\sigma^2)$, where E_k is the bit energy of user k. Assume SNR of user 1 is 20 dB. There are nine multiple-access interfering users among which five users have SNR of 30 dB each, three users have SNR of 40 dB each, and another user has an SNR of 50 dB. Here, M =500; iteration number n =2000; the forgetting factor $\lambda = 0.997$; processing gain N=31.

1) *Convergence Rate Comparison:* Fig.1 shows the time-averaged measured EOE versus time for the conventional RLS and the new SVD-based RLS algorithms applied in a synchronous CDMA system.

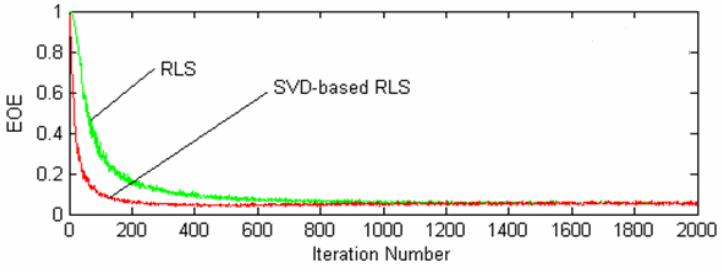


Fig. 1. Time-averaged measured EOE versus time for 500 runs when using the conventional RLS algorithm and the new SVD-based RLS algorithm to a CDMA system

2) *Numerical Stability Comparison:* Fig. 2 depicts the time-averaged measured EOE versus time for the two algorithms in a CDMA system when the matrix $\mathbf{R}^{-1}(n)$ is not symmetric. Fig.3 shows the time-averaged measured EOE versus time for the novel SVD-based RLS algorithm at the same condition as in Fig. 2.

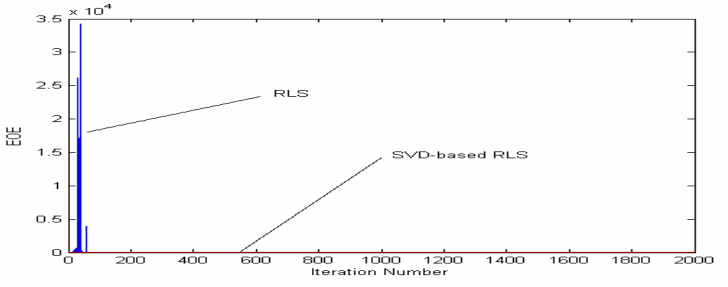


Fig. 2. Time-averaged measured EOE versus time for 500 runs when using the conventional RLS algorithm and the new SVD-based RLS algorithm to a CDMA system when the matrix $\mathbf{R}^{-1}(n)$ is not symmetric

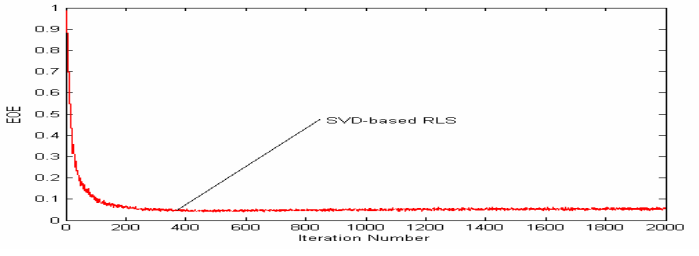


Fig. 3. Time-averaged measured EOE versus time for 500 runs when using the novel SVD-based RLS algorithm to a CDMA system when the matrix $\mathbf{R}^{-1}(n)$ is not symmetric

From the simulation results, we can see the following facts.

a) The new SVD-based RLS algorithm has faster convergence rate than the conventional RLS algorithm. This can be validated from Fig. 1. The SVD-based RLS algorithm achieves convergence at $n \approx 200$, while the conventional RLS algorithm at $n \approx 600$.

b) The new SVD-based RLS algorithm performs better than the conventional RLS algorithm in numerical stability and robustness. This can be validated from Fig. 2 and Fig. 3. From the simulation results, we can see that when the matrix $\mathbf{R}^{-1}(n)$ cannot keep symmetric, the conventional RLS algorithm performs poorly, whereas the new SVD-based RLS algorithm still performs well in this situation as shown in Fig. 3.

6 Conclusions

In this paper, we propose a novel blind adaptive multiuser detector using the SVD-based RLS algorithm for CDMA systems. The principle of the new algorithm is to

develop a stable update of the matrix $\mathbf{R}^{-1}(n)$ via computing its SVD. So it can overcome the numerical instability and divergence of the conventional RLS algorithm. Furthermore, our new algorithm is simpler and more efficient than the previous SVD-based RLS algorithm. Simulation results show that the novel SVD-based RLS algorithm has performance advantages over the conventional RLS algorithm in faster convergence rate, numerical stability and robustness.

Acknowledgement

This work was mainly supported by the Major Program of the National Natural Sciences Foundation of China under Grant 60496311.

References

1. Poor, H.V., Wang, X.D.: Code-Aided Interference Suppression for DS/CDMA Communications—Part II: Parallel Blind Adaptive Implementations. IEEE Transactions on Communications. 45 (1997) 1112-1122
2. Thanh, B.N., Krishnamurthy, V., Evans, R.J.: Detection-aided Recursive Least Squares Adaptive Multiuser Detection in DS/CDMA Systems. IEEE Signal Processing Letters 9 (2002) 229 – 232
3. Wan, P., Du, Z.M., Wu, W.L.: Novel Concatenated Space-time Multiuser Detector Based on Recursive Constrained CMA with Modulus Correction. Proceedings of the 8th IEEE International Symposium on Spread Spectrum Techniques and Applications (2004) 32-36
4. Nam, J.Y.: A Fast RLS Algorithm for Multiuser Detection. Proceedings of the 56th IEEE Vehicular Technology Conference 4 (2002) 2503-2507
5. Zhang, Y.M., Li, Q.G., Dai, G.Z., Zhang, H.C.: A New Recursive Least-Squares Identification Algorithm Based on Singular Value Decomposition. Proceedings of the 33rd IEEE Conference on Decision and Control 2 (1994) 1733-1734
6. Heath, M.T.: Scientific Computing: An Introductory Survey. 2nd edn. McGraw-Hill Companies, Boston (2002)
7. Zhang, X.D., Wei, W.: Blind adaptive multiuser detection based on Kalman filtering. IEEE Transactions on Signal Processing 50 (2002) 87-95

New Electronic Digital Image Stabilization Algorithm in Wavelet Transform Domain

Jung-Youp Suk, Gun-Woo Lee, and Kuhn-Il Lee

School of Electrical Engineering and Computer Science,
Kyungpook National University,
1370 Sankyuk-dong, Puk-gu, Daegu 702-701, South Korea
youbi75@naver.com, teiler75@lycos.co.kr, kilee@knu.ac.kr

Abstract. In this paper, we presented a new wavelet-based digital image stabilization (DIS) algorithm on the motion estimation for the stabilization of the 2-axes rotation sight system. In proposed algorithm, we first estimate the local motion defined in terms of translational motion vector by using fine-to-coarse (FtC) multi-resolution motion estimation (MRME). Next, we estimate the global motion defined as the rotational motion parameters such as the rotational center and angular frequency by using the vertical and horizontal component in wavelet domain. The rotational center and angular frequency are estimated from the local motion field and the special subset of the motion vector respectively. The experimental results show the improved stabilization performance compared with the conventional digital image stabilization algorithm.

1 Introduction

Generally, the sight system mounted on the vehicle requires high stabilization function, which removes all of the translational and rotational motion disturbances under stationary or non-stationary conditions. For eliminating the disturbances, which come from vehicle engine, cooling pan, and irregular terrains, the conventional system adopts just 2-axes mechanical stabilization using accelerometers, gyros, or inertial sensors because the 3-axes mechanical stabilization is too bulky and expensive. However, in case of the 2-axes mechanical stabilization system, the uncompensated roll component of the unwanted motion is presence, which causes the deteriorated performance of the object detection and recognition. This shortcoming has led to only the use of DIS for the roll motion, which was not mechanically compensated in the 2-axes stabilization system. The DIS is the process of generating the compensated video sequences where any and all unwanted motion is removed from the original input.

Recently, several studies on the DIS have been presented such as global motion estimation using local motion vectors [1], motion estimation based on edge pattern matching [2], fast motion estimation based on bit-plane and gray-coded bit-plane matching [3, 4], and phase correlation-based global motion estimation [5]. These approaches can only correct the translational movement. But they produce poor performance when the image fluctuation contains rotational motion dominantly. Moreover, the multi-resolution feature based motion estimation

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 911–916, 2005.
© Springer-Verlag Berlin Heidelberg 2005

scheme [6] has been presented for the rotational motion estimation. However, it needs to predetermine prominent features such as the horizon. Chang [7] proposed digital image translational and rotational motion stabilization using optical flow technique. The algorithm estimates the global rotation and translation from the estimation of angular frequency and rotational center. However, it contains time-consuming process since it finds the rotational center by searching basis, which is not appropriate to the real time application.

Accordingly, we proposed a new wavelet-based DIS algorithm on the rotational motion estimation for the stabilization system. First, for finding the translational motion vector that is local motion vector, proposed algorithm is used FtC MRME. Second, we estimate rotational motion vector that represents center and angular frequency by using the local motion component that are vertical and horizontal motion vectors of decomposition level 2 in wavelet domain. Meanwhile, the global motion field is defined as the rotational center and the angular frequency. The estimation of rotational center is achieved from the zero-crossing points of directions of vertical and horizontal motion vectors obtained by FtC MRME with block matching (BM). Then, rotational angle is computed from the special subset of motion vector. Finally, the motion compensation process is achieved by the bilinear interpolation method. The experimental results show the improved stabilization performance compared with the conventional DIS algorithm.

2 The Proposed Algorithm

The proposed efficient digital image stabilization algorithm based on wavelet transform is basically composed of two main modules: motion estimation by using FtC MRME and motion compensation.

2.1 Motion Estimation

In order to get the true translational local motion information, BM, as one of the several motion estimation techniques, is easy to be implemented and is used in the conventional image compression standard. However, this simple method is prone to fall into local minima, instead of desired global minima. Overcome to this problem can be done by using a FtC approach in wavelet domain [8]. This approach exploits the multi-resolution property of the wavelet transform. Here, the initial motion estimation is executed in the pixel domain. In the other word, the motion vectors at the finest level of the wavelet transform are first estimated using the conventional motion estimation algorithm based on BM. Then, scale and refine that at coarser resolutions. Therefore, we achieved accuracy motion estimation without local minima. In this technique, because accurate motion estimation are formed at the finest resolution and then scaled to coarser resolutions in the encoded process, these motion estimates better track the true motion and exhibit low entropy, providing high quality, both visually and quantitatively. We show the structure of wavelet transform and FtC MRME method in Fig. 1. After FtC MRME, for selection rotational center point

estimation, we used the local motion vectors of horizontal and vertical component in decomposition level 2. Therefore, h and v are respectively formulated as

$$h = V_2^H(x, y) \quad and \quad v = V_2^V(x, y) \ . \tag{1}$$

In case where motion of images composed of only rotational components, the amplitude of the motion vector at rotational center (x_0, y_0) is equal to zero. In addition, when horizontal axis with y_0 is regarded as a fiducial line the horizontal motion vectors $h(x, y)$ in the upper side and the lower side at the same column coordinates are heading for different direction. In a similar way, the vertical motion vectors $v(x, y)$ in the left side and the right side at the same row coordinates are also heading for different direction.

As shown in Fig. 2, the proposed method uses the local motion field obtained by FtC MRME into the horizontal motion components $h(x, y)$ and the vertical motion components $v(x, y)$. In order to find a zero-crossing point of motion vectors, they are scanned in the direction of column and row, respectively. The set of possible rotational center candidates from the zero-crossing points is given by as follows:

$$X_0 = \{x_k| \ \ |v_1(x_k) - v_2(x_k)| > max\{|v_1(x_k)|, |v_2(x_k)|\}\} \tag{2}$$

and

$$Y_0 = \{y_k| \ \ |h_1(y_k) - h_2(y_k)| > max\{|h_1(y_k)|, |h_2(y_k)|\}\} \tag{3}$$

where $v_1(x_k) = v(x_k + 1, y)$, $v_2(x_k) = v(x_k - 1, y)$, $h_1(y_k) = h(x, y_k + 1)$, and $h_2(y_k) = h(x, y_k - 1)$. The coordinates of rotational center (x_0, y_0) can be obtained by Eqs. (2) and (3) as follows:

$$x_0 = \frac{1}{N_x} \sum_{x_i \in X_0} x_i \quad and \quad y_0 = \frac{1}{N_y} \sum_{y_i \in Y_0} y_i \tag{4}$$

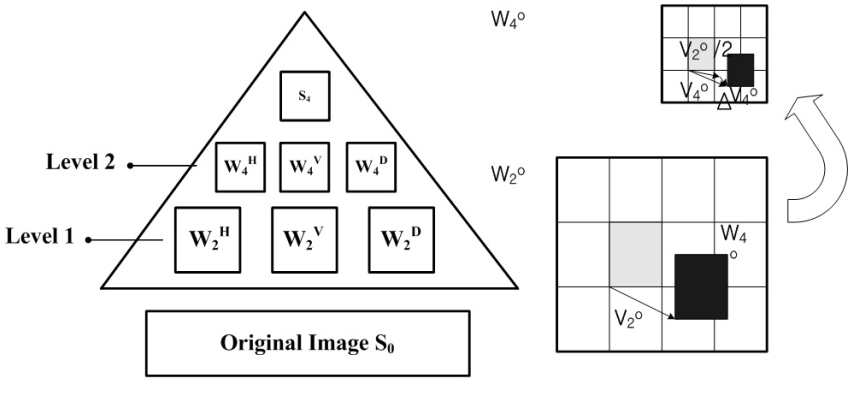


Fig. 1. The structure of wavelet transform and FtC method

where N_x and N_y is the number of x_i and y_i detected by Eqs. (2) and (3), respectively. Next, after the rotational center is estimated, the angular frequency ω can be computed from the special subset of motion vectors by least square method. The angular frequency can be computed on those motion vectors as follows:

$$\omega = \frac{dh}{dy} \quad and \quad \omega = \frac{dv}{dx} \tag{5}$$

where dx and dy are the distance from rotational center (x_0, y_0) and dh and dv are the amplitude of motion vectors at each pixel. A least squares solution is usually adopted to minimize the discrepancy from Eq. (5) for the wavelet coefficient pairs in the image.

2.2 Motion Compensation

When unwanted motion estimation of the frame is detected, the motion compensation is performed. The objective of motion compensation is to keep some kind of history of the motion estimation in order to create a stabilized sequence.

The way of motion estimation and compensation is generally divided into two algorithms: one that always use two consecutive frames from the input image sequence to estimate the motion parameters, which is referred to as the frame-to-frame algorithm (FFA), and a second one that keeps a reference image and uses it to estimate the motion between the reference and current input images, which is referred to as the frame-to-reference algorithm (FRA).

The large displacement between a pair of frames causes increment in the computational cost or makes the estimated motion unreliable. To this end, the proposed algorithm achieves the motion estimation and compensation using FFA.

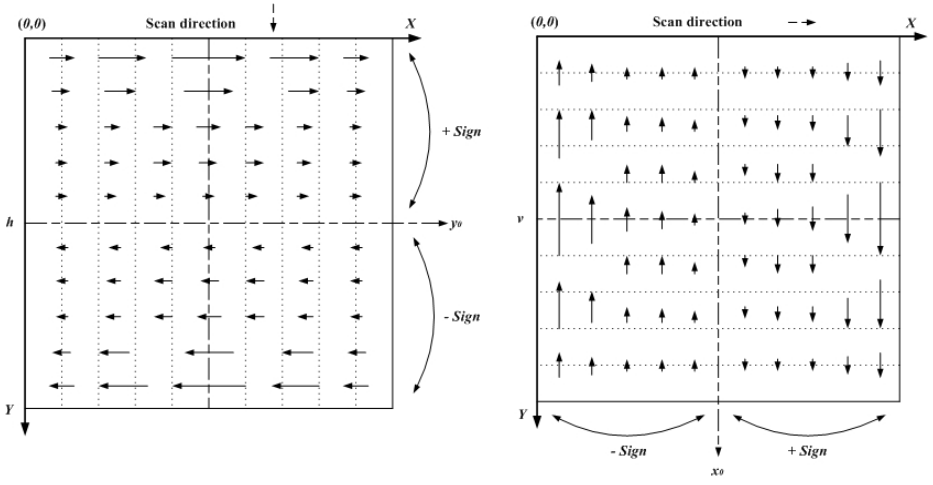


Fig. 2. Rotational center estimation in decomposition level 2 of wavelet domain

For the rotational motion compensation, the relationship between coordinates of the original point and the rotated point takes the form as follows:

$$\begin{pmatrix} x_{org} \\ y_{org} \end{pmatrix} = \begin{pmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{pmatrix} \begin{pmatrix} x_{new} - x_0 \\ y_{new} - y_0 \end{pmatrix} + \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{6}$$

where (x_{new}, y_{new}) and (x_{org}, y_{org}) are the coordinates of rotated point and original point, respectively. In case that computed coordinates do not have integer, motion compensation is achieved by the bilinear interpolation method to reduce noise of grid-to-grid matches.

3 Experimental Results

In order to evaluate the performance of the proposed algorithm, we use the peak signal-to-noise ratio (PSNR). Rotating the reference image frame with the

Table 1. The experimental result on the synthetic images

Rotational frequency estimation			Rotation center estimation		
Exact value	Chang's algorithm	Proposed algorithm	Exact value	Chang's algorithm	Proposed algorithm
-3 degree	-3.0927	-3.0014	(200,256)	(219,200)	(199,254)
-2 degree	-1.9533	-1.9978	(200,256)	(217,253)	(199,257)
-1.1 degree	-1.3027	-1.0935	(200,256)	(243,258)	(200,258)
1.1 degree	1.3027	1.0730	(200,256)	(240,261)	(199,256)
2 degree	1.9529	1.9971	(200,256)	(207,245)	(200,256)
3 degree	3.0844	2.9972	(200,256)	(197,200)	(201,257)

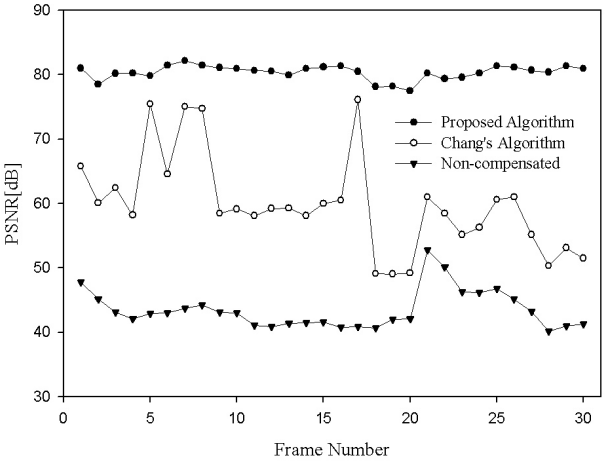


Fig. 3. Experimental results of PSNR

predefined rotational center and angular frequency generates the synthetic image sequence. The proposed DIS system shows good estimation compared with conventional method as shown in Table 1. The test real image sequence is obtained from the 2-axes stabilizer mounted on test bed. The size of image frame is 400×400 pixels. For the 30 frames, the experimental results are shown in Fig. 3 and show the improved performance of the proposed algorithm.

4 Conclusions

In this paper, we proposed a new wavelet based stabilization algorithm for roll motion, which has been uncompensated in 2-axes stabilization system. From the estimated local motion field in FtC MRME, the proposed algorithm estimates the rotational center and angular frequency to define the unwanted motion. The input image sequence containing the unwanted motion is stabilized with the estimated rotational motion parameters. The experimental results show the improved stabilization performance compared with the conventional algorithm.

References

1. K. Uomori, A. Morimura, and H. Ishii.: Electronic Image Stabilization System for Video Cameras and VCRs. SMPTE Journal. Vol. 101. (1992) 66-75
2. J. K. Paik, Y. C. Park, and D. W. Kim.: An Adaptive Motion Decision System for Digital Image Stabilizer Based on Ddge Pattern Matching. IEEE Trans. on Consumer Electronics. Vol. 38. (1992) 607-615
3. S. J. Ko, S. H. Lee, and K. H. Lee.: Digital Image Stabilizing Algorithms Based on Bit-Plane Matching. IEEE Trans. on Consumer Electronics. Vol. 44. (1998) 617-622
4. S. J. Ko, S. H. Lee, S. W. Jeon, and E. S. Kang.: Fast Digital Image Stabilizer Based on Gray-Coded Bit-Plane Matching. IEEE Trans. on Consumer Electronics. Vol. 45. (1999) 598-603
5. S. Erturk and T. J. Dennis.: Image Sequence Stabilization Based on DFT Filtering. IEE Proceedings on Image Vision and Signal Processing. Vol. 127. (2000) 95-102
6. C. Morimoto, and R. Chellappa.: Fast Electronic Digital Image Stabilization for Off-Road Navigation. Real-Time Image. Vol. 2. (1996) 285-296
7. J. Y. Chang, W. F. Hu, M. H. Cheng, and G. S. Chang.: Digital Image Translational and Rotational Motion Stabilization Using Optical Flow Technique. IEEE Trans. on Consumer Electronics. Vol. 48, No. 1. (2002) 108-155
8. G. J. Conklin, S. S. Hemami.: Multiresolution Motion Estimation. Proc. of ICASSP. (1997) 2873-2876

Line Segments and Dominate Points Detection Based on Hough Transform

Z.W. Liao[1], S.X. Hu[2], and T.Z. Huang[1]

[1] School of Applied Mathematics, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2] School of Physical Electronics, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
{liaozhiwu, hushaox}@163.com

Abstract. Hough Transform (HT) is a powerful tool to detect straight lines in noisy images since it is a voting method. However, there is no effective way to detect line segments and dominate points, which are more important in pattern recognition and image analysis. In this paper, we propose a simple way to detect lines segments and dominate points simultaneously in binary images based on HT using generalized labelling. The new framework firstly detects straight lines using HT and then labels each black point of the image by considering the discrete errors of HT. Finally, the connectivity among the points having the same labels is checked in order to reduce the effect of noises and detect line segments properly. The experimental results show that our new framework is an powerful and effective way to detect line segments and dominate points in noisy binary images.

1 Introduction

Line segment and dominate point detection is an important problem in pattern recognition and image analysis. However, there is no effective way to detect line segments and dominate points in noisy images. Some segment detection and point detection algorithms proposed that these features can be detected by HT.

However, it is well known that there are two faults in feature detection by HT: one is the endpoints and dominate points can not be detected correctly in noise, the other is the discrete errors leads to HT can not correctly detect some lines in images. Some works were proposed to overcomes these difficulties and reported good detection results. But they are very complex.

Generally, the system of line segment detection based on HT is a two-step process: extracting the straight lines using HT; and then detecting line segments based on line tracking [2]-[7].

Some algorithms focus on reducing the discrete errors in HT to obtain the precise parameters of straight lines; and then searching line segments using line tracking. Therefore, they analyze the voting patterns around peaks in the accumulator space to improve the precision of the HT [1]–[3] and [5].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 917–922, 2005.
© Springer-Verlag Berlin Heidelberg 2005

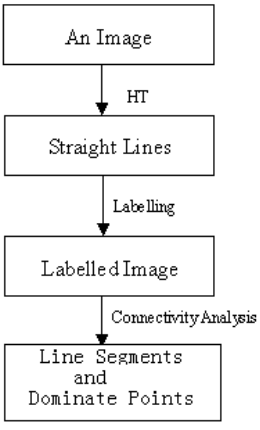


Fig. 1. The flow chart of our new framework

Other frameworks detect line segments by considering the connectivity of points on a line segment. Two parameters need to be decided in advance, they are the minimum acceptable line length and maximum acceptable gap length [6]. The connectivity should be represented by the runs of consecutive non-zero values based on two parameters.

Although the connectivity analysis is more straight than line tracking and it reduces discrete errors of HT by considering the discrete error in line recover, the basic procedures of two methods are almost the same. The algorithms based on this two methods are very time-consuming in complex images.

In this paper, we proposed a straight and novel algorithm based on generalized labelling. In this new framework, the line segments are detected by three steps: firstly, the straight lines are checked using HT and each of straight line is coded as a unique number, named the code of the line; then every points in the image is labelled by the codes of straight lines or zero by considering discrete errors and the equations of the lines; finally, the connectivity analysis is carried on the points having same labels (see Fig. 1). Since the connectivity analysis is only carried on small number points, the three-step framework simplifies the complexity of the traditional connectivity analysis.

The following parts of the paper starts at introducing the HT; and then the basic theory about labelling is discussed; after that, we give the connectivity analysis of our framework; finally, the experimental results are presented and future work will be discussed.

2 Hough Transform

The Hough transform is the most popular technique for digital straight line detection in a digital image. The form proposed by Duda and Hart [8] is

$$\rho = x \cos\theta + y \sin\theta \tag{1}$$

where ρ is the distance of the point of origin to the line and θ is the angle of the normal to the straight line with the x-axis.

Each point in $(\rho,\ \theta)$ space corresponds to a straight line in $(x,\ y)$ space. The number of votes of a certain straight line $(\rho_j,\ \theta_j)$ is kept in the member of the accumulation matrix $C[i][j]$.

3 Labelling

Many problems in image analysis can be posed as labelling problems in which the solution to a problem is a set of labels assigned to image pixels or feature.

3.1 Sites and Labels

A *labelling* problem is specified in terms of a set of *sites* and a set of *labels*. Let **S** index a discrete set of m sites

$$\mathbf{S} = \{1, \cdots, m\} \tag{2}$$

in which $1, \cdots, m$ are indices. In our framework, a site represents a point in the Euclidean space such as an image pixel. A label set may be categorized as being continuous or discrete.

In the discrete case, a label assumes a discrete value in a set of M labels

$$\mathbf{L}_d = \{\ell_1, \cdots, \ell_M\} \tag{3}$$

or simply

$$\mathbf{L}_d = \{1, \cdots, M\} \tag{4}$$

In line detection, the label set is defined as a integer set from 1 to the number of straight lines detected using HT.

3.2 The Labelling Problem

The labelling problem is to assign a label from the label set **L** to each of the site in **S**. Line detection in an image, for example, is to assign a label f_i from the set $\mathbf{L} = \{1, \cdots, M\}$ to site $i \in \mathbf{S}$ where elements in **S** index the image pixels. M is the number of straight lines detected using HT.

The set

$$f = \{f_1 \cdots f_m\} \tag{5}$$

is called a *labelling* of the sites in **S** in terms of the labels in **L**. When each site is assigned a unique label, $f_i = f(i)$ can be regarded as a function with domain **S** and **L**. Because the support of the function is the whole domain **S**, it is a *mapping* from **S** to **L**, that is

$$f : \mathbf{S} \longrightarrow \mathbf{L} \tag{6}$$

3.3 Generalized Labelling

For detecting the line segments and dominate points, an integer label is not enough. Therefore, the labels are generalized to a vector with different length; that is, the label set can be generalized to

$$\mathbf{L} = \{0, 1, \cdots k, (1,2), (1,3) \cdots (k-1,k), \cdots, (1,2,\cdots k)\} \tag{7}$$

where k is the number of straight lines detected by HT. In our paper, for example, the feature points are black points on the straight lines detected by HT in a binary image. $f_{i,j}$ is a vector with different length in generalized label set $\mathbf{L}$. The label of a pixel (i,j) is defined as

$$f_{i,j} = \begin{cases} 0 & : \text{ if } (i,j) \text{ does not satisfy Eq.1, for all } \rho, \theta \text{ detected by HT} \\ s & : \text{ if } (i,j) \text{ satisfies Eq.1 for } \rho_s, \theta_s \text{ only} \\ \cdots & : \cdots \\ (1,2,\cdots k) & : \text{ if } (i,j) \text{ is the intersection point of straight lines 1 to k} \end{cases}$$

According to the definition, each point in a binary image can be assigned to a unique label in generalized label set $\mathbf{L}$. The noise points whose labels are zero can be found easily and the points of intersection can be obtained from the labels whose dimensions are larger than 1. The line segments can also be detected properly by considering the connectivity.

Considering the discrete errors, the conditions of $f_{i,j}$ can be relax to satisfying Eq. 1 or limiting the errors between feature points and relative value computing by Eq. 1 are smaller than a positive number.

4 Connectivity Analysis

Considering connectivity of points on a line segment, two parameters need to be decided in advance. They are the minimum acceptable line length (MALL) and maximum acceptable gap length (MAGL) [6]. The connectivity should be represented by the runs of consecutive non-zero values based on two parameters; that is, the algorithm detects all segments that are longer than MALL and do not contain gaps longer than MAGL.

The general steps of connectivity analysis in our paper are as follows:

Decide two Parameters and Intersects: MALL and MAGL should be decided according to the natures of images, for example, the MALL of a circuit diagram should be smaller than an engineering drawing, while the MAGL in a heavy noisy image should be bigger than it in a clear image. In order to reduce the effects of the noises, especially, on the corners, all intersects of straight lines are computed and labelled according to the rule defined on Section 3.3. Therefore, if the corners are blurred by the noise, the intersects can help us recover the corners from the noise using *connectivity analysis and postprocessing.*

Connectivity Analysis: the relative connectivity should be analysis among the points having same nonzero labels or a branch of its label vector having

same nonzero values; that is, if the length between two closing points is smaller than the MAGL, the segment can be considered as running between two points; while the length is bigger, it can be considered as these two points are end points of two different segments. In this step, we can modify our original generalized labels defined by straight lines to a new generalized labels defined by segments.

Postprocessing: after deciding the segments labels of pixels of an image, we can check the length of each segment. If the length of a segment is smaller than the MALL, the segment is deleted; while it is bigger than the MALL, the segment is kept. Therefore, we can properly detect all line segments in the images and the dominate points in the images also can be detected according to the generalized labels.

5 Experimental Results

In this section, we design some experiments to support our novel algorithm.

Experiment 1: the generalized labelling can reduce the discrete errors of HT by labelling the points directly. From Fig 2, the HT leads to discrete errors, which is shown on by the difference between the thin line and a thick line, especially at the left top of the figure. While, only the right line is detected in our new framework.

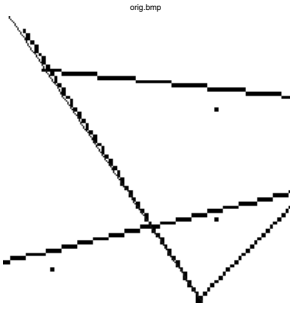


Fig. 2. The discrete errors of the HT: The thin line represents the result of HT, while thick lines represent the original line segments and detected segments by generalized labelling

Experimental 2: detect feature points by the new framework. In this paper, two rectangles: one is with circle corners, the other with straight corners are used. One experiment is designed to demonstrate the accuracy of our algorithm, that is, the framework can detect different feature points and segments in two images. The other experiment is designed to demonstrate the resistance of the new framework. One straight corner is damaged by the noise, the framework can recovery the corner properly; while if we carefully design some parameters to validate the connectivity of line segments, the rectangle with circle corners will not be detected as straight corners.

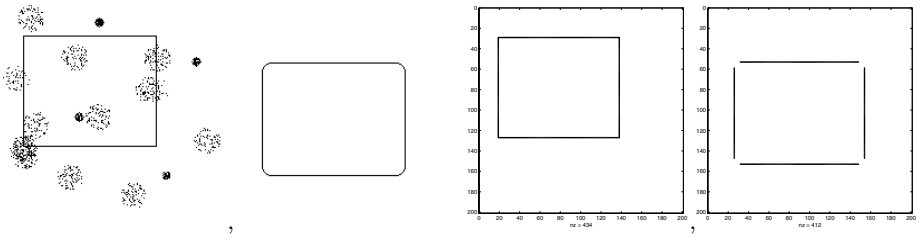


Fig. 3. Two rectangles (left) and their relative detection results (right)

6 Summary and Future Work

In this paper, we propose a simple line segment and feature detection algorithm based on HT and generalized labelling. The algorithm can efficiently detect the line segments and feature points meanwhile. It also preserves the advantage of HT, which can correctly detect the straight lines in noises. However, there are also some interesting works can be done in this field. They includes: find better detection algorithms in line detection to avoid the discrete errors of HT and reduce the computation burden of HT etc. Some important applications should be found as well in future.

We wish the new framework can providing a simple and efficient way in line segments and feature points detection.

References

1. Yasutaka Furukawa, Yoshihisa Shinagawa: Accurate and robust line segment extraction by analyzing distribution around peaks in Hough space. COMPUTER VISION AND IMAGE UNDERSTANDING, Vol. 92 Issue: 1, October (2003)
2. Yuan Jie, Hu Zhenyi, Wang Yanping: Finding Corners by using Hough Transform. J. WUHAN UNIV., VOL. 44, NO. 1, Feb. (1998) 85–88
3. E. Magli and G. Olmo: Determination of a segment endpoints by means of the Radon transform. in Proc. ICECS'99, Sept. (1999)
4. E. Magli, G. Olmo and Letizia Lo Presti: On-Board Selection of Relevant Images: An Application to Linear Feature Recognition. IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 10, NO. 4, APRIL (2001) 543–553
5. D. Ioannou: Using the Hough transform for determining the length of a digital straight line segment. IEE 1995 ElECTRONICS LETTERS, March (1995) 782-784
6. Jiqiang Song, Min Cai, M.R. Lyu, Shijie Cai: A new approach for line recognition in large-size images using Hough transform. in Proceedings of 16th International Conference on Pattern Recognition, Vol. 1, 11-15 Aug. (2002) 33 - 36
7. N. Nagata, T. Maruyama: Real-time detection of line segments using the line Hough transform. in Proceedings of 2004 IEEE International Conference on Field-Programmable Technology (2004) 89 - 96
8. R O OUDA, and P E HART: Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM 15, 1 (1972) 11-15

The Study of the Auto Color Image Segmentation

Jian Zhuang, Haifeng Du, Jinhua Zhang, and Sun'an Wang

School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
Jian_zhj@163.com

Abstract. Auto image segmentation can segment the image without operators interfering and is an important technique in the image processing. The Boltzmann-Color-Image-Segmentation (BCIS), which could control the degree of segmentation by adjusting the temperature parameter, is designed based on the Boltzmann-theory and the Metropolis-rule in the paper. Then the criterion function of image segmentation, which could balance between the number of segmented region and the affinity of the segmented image with the original image, is presented. Based the BCIS and Criterion function, the auto color image segmentation is schemed out by using the artificial immune algorithm. Experiments showed that the color image segmentation algorithm, which we designed in the paper, had the good capabilities.

1 Introduction

Image segmentation is one of the most important tasks in image processing, having a wide range of applications in computer vision, such as pattern recognition, image compression and so on. To date, most of researchers have focused on the image segmentation and many segmentation algorithms have been discussed. However, little work has been done on the auto-image-segmentation that image segmentation algorithm can been performed without interfered by operators.

Image segmentation algorithms can be categorized into four classes: edge-based, clustering-based, region-based and splitting/merge approaches. The edge-based approaches are good at image processing speed, but hard to getting the single pixel and closed contours, such as Canny. The clustering-based approaches are easy to obtain the single pixel and closed contours with high speed, but generally request to preassign the appropriate cluster number or thresholds. If the different cluster number or thresholds is selected, image segmentation results may be different, such as Region-growing, Watersheds and Pyramidal segmentation [1]. Law [2] proposed the RPCCL (Rival Penalize Controlled Competitive Learning) image segmentation algorithm that doesn't need the cluster number, but is sensitive to the proper initial seeds and the selection of de-learning rate. The splitting/merge approaches tessellate an input image into a set of homogeneous regions. Then, the homogenous regions are merged according to a certain rules. Merge algorithm is a key to the splitting/merge approaches, such as region adjacent graph (RAG), nearest neighbor graph (NNG) and so on. Shortcomings of the splitting/merge approaches are segmentation results effected by merge rules and the slow processing speed. Wang [3] designed the criterion of merge algorithm, Cut-Ratio, which is defined as the ratio of the mean affinity of segmenta-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 923 – 928, 2005.
© Springer-Verlag Berlin Heidelberg 2005

tion regions per the length of segmentation boundary, and the minimum Cut-Ratio is tried getting in the image merging. Because of basing iterative bipartition and undirected graph merging, this algorithm computation speed is very slow, but Cut-Ratio is good reference for establishing criterion of image segmentation algorithm. Christopher M [4] developed an EDISON (Edge Detection and Image Segmentation) algorithm that filters image using statistical ways, and then segments the image by the mean-shift. Clear targets are advantages of EDISON, but some characteristics of image may be lost.

In a word, an effective image segmentation algorithm should overcome following four questions. First, results of segmentation should be the single pixel continuous contours. Second, the results should keep main bodies of image and wipe off small details. Third, the speed of algorithm should be fast. Finally, the segmentation should obtain results automatically without operators interfering. The purpose of the present paper is to obtain the auto image segmentation algorithm. We have automatically segmented about 200 images of the Berkeley University image database (www.cs.berkeley.edu\projects\vision\grouping\segbench\default.htm). The results show that auto image segmentation algorithm designed in the paper may be help to solve some kinds of problems of the image segmentation.

The organization of the rest of the paper is as follows. In Section 2, we describe an image splitting algorithm, Boltzmann-Color-Image-Segmentation algorithm (BCIS), based the Boltzmann theory and Metropolis rule. In Section 3, we present the auto image segmentation algorithm. Experiment results are provided in Section 4 and the paper is concluded following discussion in Section 5.

2 Boltzmann-Color-Image-Segmentation Algorithm

From viewpoints of thermodynamics, if a close system has no material exchange but energy exchange with environments, the stable condition of system is defined by the minimal Helmholtz-Free-Energy instead of the maximal entropy. In the close system, there are many moving particles, which are distributed in different Helmholtz-Free-Energy strips. The probability of a particle jumping from the Helmholtz-Free-Energy level of F_i to the level of F_j can be computed as

$$f(F_i,F_j,t)=\exp\frac{-\left|F_i-F_j\right|}{kt} \quad . \tag{1}$$

Where k is a constant and t is the system temperature.

Based the above theory, we assume that the image is a close system, pixels are same as the particles in the close system, and the normalization of attribute values of pixels represents its Helmholtz-Free-Energy. If the pixel $\mathbf{P}^{c}_{ij}$ belongs to the set $\mathbf{C}_k$ and temperature is t, the probability of the $\mathbf{P}^{c}_{ij}$ and its bordering pixel $\mathbf{P}_{i'j'}$ belonging to the same class can be written as

$$f(\mathrm{P}_{i'j'},\mathrm{P}^{c}_{ij},\mathrm{C}_k,t)=\exp\frac{-(a_1\left\|\mathrm{P}_{i'j'}-\mathrm{P}^{c}_{ij}\right\|+a_2\left\|\mathrm{P}_{i'j'}-\mathrm{C}_k\right\|)}{kt} \quad . \tag{2}$$

Where a1 and a2 are distributing coefficients of Helmholtz-Free-Energy.

From above equation, the probability relates to two parts, the distance between two adjacent pixels and the distance between pixel and the pixel set $\mathbf{C}_k$. If f is a big probable case, we consider that the pixel $\mathbf{P}_{i'j'}$ belongs to the set $\mathbf{C}_k$.

In the following, we will present the BCIS algorithm.

1. Set the value of t, the number classified k=0, the temporary edge-points union $\mathbf{E}=\mathbf{\Phi}$, the target-classified union $\mathbf{C}=\mathbf{\Phi}$ and sub-target-classified union $\mathbf{C}_k=\mathbf{\Phi}$.
2. Judge whether all pixels are in the $\mathbf{C}$. If all pixels are in the $\mathbf{C}$, then the process jumps into the step 7.
3. Select the pixel $\mathbf{P}_{ij}$ that does not belong to the $\mathbf{C}$. Then, add the $\mathbf{P}_{ij}$ to the union $\mathbf{E}$ and the union $\mathbf{C}$.
4. Judge whether the $\mathbf{E}$ is empty. If $\mathbf{E}$ isn't empty, then get $\mathbf{P}_{ij}$ form $\mathbf{E}$ and delete it, else jumps into the step 6.
5. Judge the four neighbors ($\mathbf{P}_{(i+1)j}$, $\mathbf{P}_{(i-1)j}$, $\mathbf{P}_{i(j+1)}$, $\mathbf{P}_{i(j-1)}$) of the $\mathbf{P}_{ij}$ whether they are in $\mathbf{C}$. If which pixel is not in the $\mathbf{C}$, the probability of it and $\mathbf{P}_{ij}$ belonging to the same sub union is calculated by the formula 2. And if the probability f is a big probable case, the pixel is added into the union $\mathbf{E}$ and the sub union $\mathbf{C}_k$. Then, jumps into the step 4.
6. Set k=k+1. Jump into the step 2.
7. The target-classified union $\mathbf{C}$ is the image segmentation result.

3 Auto Image Segmentation

From the above algorithm, the segmentation degree can be regulated by the temperature coefficient in the ISABT. If we can find out the optimal temperature coefficient, the auto image segmentation is solved. However, building up the criterion function of image segmentation is a precondition of the optimizing temperature coefficient. Before building up the criterion function of image segmentation, we present a hypothesis that targets, which we want to extract from an input image, have big area. Then, we define some variables.

The dissimilarity between the segmented image and the original image can be expressed as

$$\varsigma=\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\|p_{ij}^{c}-p_{ij}\right\| . \tag{3}$$

The edge set of $\mathbf{C}_k$, $\mathbf{E}_C$, can be written as

$$E^{c}=\left\{\begin{array}{c} P_{ij}\,|\,f(P_{i-1j},P',C)<f_c \ \text{ or } f(P_{i+1j},P',C)<f_c \ \text{ or} \\ f(P_{ij-1},P',C)<f_c \ \text{ or } f(P_{ij+1},P',C)<f_c \end{array}\right\} , \tag{4}$$

Where fc is the initialized probability, Pij and P' are adjacent pixels and belong to the set C.

The difference of Helmholtz-Free-Energy between pixel and its class is defined as

$$d_p^c=\left\|P^c-C_k\right\| . \tag{5}$$

We can now derive the criterion function of image segmentation to Equation (6).

$$\text{Estimate(MAP)}=\tau\max(\text{Length}(E^c))(1.0-\max(d_p^c))(1.0-\varsigma) \quad . \tag{6}$$

Where Length (Ec) is the boundary length function, τ is the total number of the segmentation sets.

The criterion function of image segmentation wants to get a balance between the number of segmented region and the affinity of segmented image with original image. If the affinity is very low and the number of segmented region is small, it means that the difference between the original image and the segmented image is very big, lots of necessary details of image targets are lost, and short-segmentation may be happened. If the affinity is very high and the number of segmented region is big, it means that the segmented image is similar to the original image, image targets may be segmented lots of patches, and over-segmentation may be happened. We have tested the 200 images in the database by the above criterion function. 192 images have the peak values of the function, and others don't accord with the hypothesis. Then, we use the artificial immune algorithm to optimize the temperature parameter according to the criterion function. Thereby we realize the auto segmentation image that is called Auto-Boltzmann-Color-Image-Segmentation (ABCIS). For more details about artificial immune algorithm, interested reader can refer to paper [5].

4 Experiments and Analysis

In order to comparing the segmentation algorithm capabilities, we define two criterions: the ratio of segmentation (RS) and the ratio of error (RE). The ratio of segmentation is described as the degree of the image segmentation and is written as

$$\gamma=(1.0-\frac{\sum E}{mn})\times 100\% \quad . \tag{7}$$

The ration of error is the difference between the segmented image and the original image and can be described as

$$\eta=\frac{1}{3mn}\sum_{i=0}^{m}\sum_{j=0}^{n}\left(\frac{\left|r_{ij}^c-r_{ij}\right|}{r_{ij}}+\frac{\left|b_{ij}^c-b_{ij}\right|}{b_{ij}}+\frac{\left|g_{ij}^c-g_{ij}\right|}{g_{ij}}\right)\times 100\% \quad . \tag{8}$$

If γ is big, the number of the segmented region is small. If η is big, the difference between the original image and segmented image is big.

To show the segmentation performance of ABCIS in comparison with the conventional K-Means, RPCC, Iterative-Self-Organizing (ISO) and EDISON, we conducted two group experiments. In first group experiments, we used the Desert-Snake image and obtained the boundary segmentations of each algorithm.

Fig 1 and Table 1 show that ABCIS and EDISON are better than other three algorithms, and avoid the short-segmented or over-segmented. Although RE of the K-Means is smallest, RS is also smallest and result is short segmented. The result of RPCCL is short segmented too. RE of RPCCL is biggest and targets of image almost can't be distinguished (See Fig. 1 Bottom left). RS of EDISON is biggest, but RE of

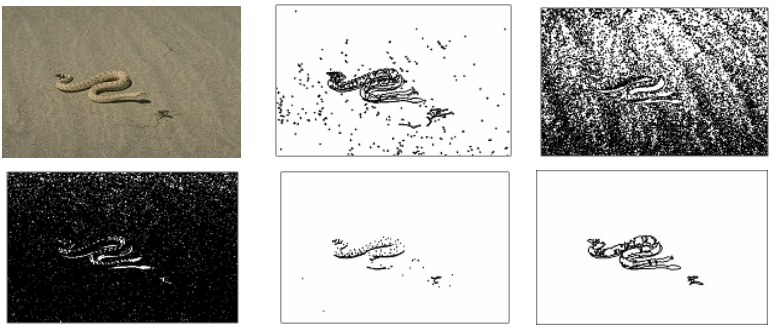


Fig. 1. The Desert-Snake image segmentation results of five algorithms. Top left: The original Desert-Snake image; Top center: The boundary segmentation of ABCIS; Top right: The boundary segmentation of K-Means; Bottom left: The boundary segmentation of RPCCL; Bottom center: The boundary segmentation of ISO; Bottom right: The boundary segmentation of EDISON.

Table 1. The capabilities of five segmentation algorithms

	ABCIS	K-Means	RPCCL	ISO	EDISON
RS (%)	94.43	1.82	97.41	15.97	98.84
RE (%)	8.22	2.5	12.84	4.44	10.79
Short segmented	No	No	Yes	No	No
Over segmented	No	Yes	No	Yes	No

EDISON is bigger 2.57 percents than RE of ABCIS, so details of EDISON is more worse than details of ABCIS (See Fig. 1 Top center and Fig. 1 Bottom right).

In order to further comparing ABCIS with EDSION, the other group experiments had been done. In the experiments, we used the aero-image and obtained the color their segmentations.



Fig. 2. The Airbase image segmentation results of EDSION and ABSCI. Left: The Airbase original image; Center: The segmentation result of EDISON; Right: The segmentation result of ABSCI.

Comparing Fig 2 Center with Fig 2 Right, we find that some planes in the red rectangles disappear in the result of EDISON, but the ABSCI clearly marks out all airplanes in the red rectangles. RE of ABSCI is the 7.02 percents and is smaller 4.67 percents than RE of EDISON. ABSCI has best the balance between the outline and the detail in five segmentation algorithms.

5 Conclusion

ABSCI could segment the image form rough to subtly by adjusting the temperature parameter and has clear physical meaning. The color image segmentation criterion, which we designed in the paper, realized the balance between the affinity and the number of segmented region. The auto image segmentation had been realized by optimizing temperature parameter of ABSCI. The experiments showed that ABSCI is best in five segmentation algorithms.

The program code and the some results, in the paper, can be down load from http://202.117.58.58/files/ImageSoftResult.rar.

References

1. Rezaee, M.R., van der Zwet, P.M.J., Lelieveldt, B.P.E. (ed.): A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering. Vol. 9. IEEE Transactions on Image Processing (2000) 1238–1248
2. Law, L.T., Cheung, Y.M.: Color image segmentation using rival penalized controlled competitive learning. (eds.): Proceedings of the International Joint Conference on Neural Networks, IEEE Neural Networks, Hong Kong (2003) 108–112
3. Wang, S., Jeffrey, M.S.: Image Segmentation with Ratio Cut. Vol. 25. IEEE Transactions On Pattern Analysis And Machine Intelligence (2003) 675–690
4. Meer, P., Weiss, I.: Smoothed differentiation filters for images. Vol. 3. Journal of Visual Communication and Image Representation (1992) 58–72
5. De Castro, L.N., Von Zuben, F.J.: The Clonal Selection Algorithm with Engineering applications. In: Whitley, D., Goldberg, D.E., Cantú-Paz, E., Spector, L., Parmee, I.C., Beyer,H. (eds.): Proceedings of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann, Las Vegas (2000) 36–37

Regularized Image Restoration by Means of Fusion for Digital Auto Focusing*

Vivek Maik, Jeongho Shin, and Joonki Paik

Image Processing and Intelligent Systems Laboratory,
Department of Image Engineering,
Graduate School of Advanced Imaging Science, Multimedia and Film
Chung-Ang University, Seoul 156-756, Korea
paikj@cau.ac.kr

Abstract. This paper proposes a novel digital auto-focusing algorithm using image fusion, which restores an image with out-of-focus objects. Instead of designing an image restoration filter for auto-focusing, we propose an image fusion-based auto-focusing algorithm by fusing multiple, restored images based on regularized iterative restoration. The proposed auto-focusing algorithm consists of (i) *sum-modified-Laplacian (SML)* for obtaining salient focus measure, (ii) iterative image restoration, (iii) *auto focusing error metric (AFEM)* for optimal restoration(iv) *soft decision fusion and blending (SDFB)* which enables smooth transition across region boundaries. By utilizing restored images at consecutive levels of iteration, the soft decision fusion and blending algorithm can restore images with multiple, out-of-focus objects. An *auto-focusing error metric* is used to provide an appropriate termination point for iterative restoration.

1 Introduction

In this paper, a novel auto-focusing algorithm using image fusion is proposed, which can restore multiple objects by iteratively restoring the image and fusing several partially restored images. A sum modified Laplacian operator is used to obtain a numerical measure indicating the degree of focus for an out-of-focus image. For multi-object focusing, the salient focus measures are chosen at each level of iterative restoration, followed by soft decision fusion and blending to produce the in-focus image with maximal focus information. An auto-focusing error metric which uses result from intermediate fusion level and restored image provides an appropriate termination point for the restoration algorithm.

Fusion- based approaches have been extensively studied for auto-focusing applications [1, 2, and 3]. Prior work on the image fusion process has focused on operating on multiple intensity images to decompose the source image in to different spatial scales and orientations [4]. However the fusion-based approaches require a minimum

* This work was supported by Korean Ministry of Science and Technology under the National Research Laboratory Project and by Korean Ministry of Information and Communication under the Chung-Ang University HNRC-ITRC program.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 929 – 934, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of two input source images for proper auto-focusing. Restoration-based techniques have been carried out in the past to overcome the auto-focusing problem [5, 6, and 7]. However restoration of images with different depth of fields tend to cause re-blurring and ringing artifacts in the region with low depth of field or region already in focus [8]. Even with equal depth of field the ill-posed nature of restoration pose a serious limitation to the visual quality of the restored images. Another drawback is the convergence process in iterative restoration [9]. In this paper the proposed auto-focusing algorithm overcomes the drawback of fusion and restoration based approaches, providing an integrated framework for auto-focusing applications.

The rest of the paper is organized as follows. The proposed algorithm is described in section 2. Simulation results and comparisons are shown in section 3. Finally, concluding remarks are outlined in section 4.

2 Proposed Auto-focusing Algorithm

The proposed auto-focusing algorithm uses the following three procedures to obtain a good restored image: (i) salient feature computation, (ii) iterative image restoration, (iii) soft decision fusion and blending (SDB).

2.1 Salient Focus Measure

The saliency function captures the importance of what is to be fused. When combining images having different focus, for instance, a desirable saliency measure would provide a quantitative measure that increases when features are in better focus. The saliency function only selects the frequencies in the focused image that will be attenuated due to defocusing. Since defocusing is a low pass filtering process, its effects on the image are more pronounced and detectable if the image has strong high frequency content. One way to high pass filter an image is to determine its Laplacian or second derivative in our case.

$$\nabla^2 L_K = \frac{\partial^2 L_k}{\partial x^2} + \frac{\partial^2 L_k}{\partial y^2}, \tag{1}$$

Also we know that in the case of Laplacian the second derivatives in the x and y directions can have opposite signs and tend to cancel each other. In the case of textured images, this phenomenon may occur frequently and the Laplacian at times may behave in an unstable manner. We overcome this problem by defining absolute Laplacian as

$$\nabla^2 L_K = \left|\frac{\partial^2 L_k}{\partial x^2}\right| + \left|\frac{\partial^2 L_k}{\partial y^2}\right|, \tag{2}$$

Note that the modified Laplacian is always greater or equal in magnitude to the Laplacian. Finally, the focus measure at a point (i, j) is computed as the sum of modified

Laplacian values, in a small window around (i, j), that are greater than a threshold value.

$$F(i, j) = \sum_{x=i-N}^{i+N} \sum_{y=j-N}^{j+N} M_k(x, y) \, for M_k(x, y) \geq T_1. \tag{3}$$

The parameter determines the window size used to compute the focus measure. In contrast to auto focusing methods, we typically use a small window of size, i.e. $N = 1$. The above equation can be referred to as *sum modified Laplacian (SML).*

2.2 Regularized Iterative Image Restoration

Iterative image restoration is the most suitable for auto-focusing because: (i) There is no need to determine or implement the inverse of an operator, (ii) knowledge about the solution can be incorporated into the restoration process, (iii) the solution process can be monitored as it progresses [8].

A. Regularization Method

The image degradation model is given as

$$y = Hx, \tag{4}$$

where y, H, and x respectively represent the observed image, the degradation operator, and the original image. A general image restoration process based on the constrained optimization approach is to find $\hat{x}$ which minimizes

$$\left\| C\hat{x} \right\|^2, \tag{5}$$

subject to

$$\left\| y - H\hat{x} \right\|^2 \leq \varepsilon^2, \tag{6}$$

where $\hat{x}$, C, and ε^2 respectively represent the restored image, a high-pass filter for incorporating a priori smoothness constraint, and an upper bound of error residual of the restored image. The constrained optimization problem, described in (7) and (8), can be solved by minimizing the following functional

$$\hat{x} = \arg\min_x f(x), \tag{7}$$

where,

$$f(x) = \left\| y - Hx \right\|^2 + \lambda \left\| Cx \right\|^2. \tag{8}$$

B. Convergence

The convergence of the regularized iteration can be established with the use of fused image obtained at previous iteration. We make use of *auto focusing error metric (AFEM)* to determine a good stopping point for restoration. A measure of the differ-

ence between the restored image $C_k^{\,r}$ and fused image $C_{k-1}^{\,f}$ at iteration level k and $k-1$ is used

$$Err(C_k^{\,r}, C_{k-1}^{\,f}) = \frac{\sum_{i,j} \mathrm{E}(C_k^{\,r}(i,j), C_{k-1}^{\,f}(i,j)}{M} \tag{9}$$

Where M is the number of pixels in C^r or C^f, and $E(C^r_{\,k}, C^f_{\,k-1}) = 0$ if $C^r_{\,k} = C^f_{\,k-1}$ and $E(C^r_{\,k}, C^f_{\,k-1}) = 1$ otherwise. If this difference decreases dramatically from one iteration to the next, then the algorithm is terminated and the restored, fused output is taken to be C^f_{k-1} where k is the current iteration.

2.3 Soft Decision Fusion and Blending

The fusion, operates on each iteration level of the restoration process in conjunction with sum modified Laplacian to generate the composite image C. The reconstruction process iteratively integrates information from the lowest to the highest level of the iteration as follows:

$$L_{ck} = F_k . L_{Ak} + (1 - F_K) L_{Bk}, \tag{10}$$

A typical problem that can occur with any type of image fusion is the appearance of unnatural borders between the decisions regions due to overlapping blur at focus boundaries. To combat this, soft decision blending can be employed using smoothing or low pass filtering of the saliency parameter F_k. In this paper Gaussian smoothing has been used for obtaining the desired effect of blending. Then we have,

$$L_{ck} = \begin{cases} L_{Ak}, \tilde{F}_k < l, \\ L_{Bk}, \tilde{F}_k > h, \\ \tilde{F}_k . L_{Ak} + (1 - \tilde{F}_k) L_{Bk}, otherwise. \end{cases} \tag{11}$$

where $\tilde{F}_k$ is now a smoothed version of its former self.

3 Experimental Results

In order to demonstrate the performance of the proposed algorithm, we used real images with one or more differently out-of-focused objects from background. Experiments were performed on a 256-level image of size 640x480. Here, each image contains multiple objects at different distances from the camera. Thus one or more objects naturally become out of focus when the image is taken.



Fig. 1. Experimental results: (a) The source image, (b)-(e) restored image at 1st, 3rd, 5th, 6th levels of iteration, (f)-(i) corresponding salient focus measure, and (j) the restored image obtained by fusion of (b)-(e)

Fig 1(a) represents image with low depth of field, where focus is on the objects far from the camera lens. The final restored image is obtained by fusion of the multiple restored images. Fig. 1. shows the result of fusion based restoration applied to image having different depth of focus. The resulting composite merges the portions that are in focus from respective images. Fig 1(b), (c), (d), (e) represents the images restored at iteration level 1, 3, 5, 6. Fig 1(e), (f), (g), (h) show the optimal focus information extracted from their corresponding restored images. Fig 1(i) is the final restored image formed by fusion of 1(c), (d), (e) and (f).

4 Conclusions

In this paper, we propose an auto-focusing algorithm which restores an out-of-focus image with multiple, differently out-of-focus objects. By utilizing restored images at consecutive levels of iteration, the soft decision fusion and blending algorithm can restore images with multiple, out-of-focus objects. Although the proposed algorithm has been limited to two objects, it can be easily extended to incorporate multiple objects.

References

1. G. Ligthart and F. Groen.: A comparison of different autofocus algorithms, IEEE Int. Conf. Pattern Recognition, (1992)597-600.
2. Z. Zhang and R. S. Blum: "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," Proc. IEEE vol. 87(8), (1999)1315-1326.
3. A. Akerman.: Pyramid techiniques for multisensor fusion, Proc. SPIE, vol.2828, (1992) 124-131.
4. A. Kubota, K. Kodama, K. Aizawa.: Registration and blur estimation method for multiple differently focused images, *IEEE Proc. Int. Conf. Image Processing*, vol.2, (1999)515 -519.
5. M. Subbarao, T. C. Wei, G. Surya,: Focused image recovery from two defocused images recorded with different camera settings, *IEEE Transactions on Image Processing*, vol. 4, No.12, Dec. (1995)1613 –1628.
6. H. C. Andrews and B. R. Hunt,: *Digital Image Restoration*, Prentice-Hall, New Jersey, (1977).
7. S. K. Kim, S. R. Park, and J. K. Paik, Simultaneous Out-of-Focus Blur Estimation and Restoration for Digital AF System, *IEEE Trans. Consumer Electronics*, vol. 44, no. 3, August (1998)1071-1075.
8. A. M. Tekalp, H. Kaufman, and J. W. Woods.: Identification of Image and Blur Parameters for the Restoration of Noncausal Blurs, *IEEE Trans. Acoustics, Speech, Signal Proc*., vol. ASSP-34, no. 4, August(1986)963-972.
9. A. M. Tekalp and H. Kaufman.: On Statistical Identification of a Class of Linear Space-Invariant Image Blurs Using Nonminimum-Phase ARMA Models, *IEEE Trans. Acoustics, Speech, Signal Proc*., vol. 36, no. 8, August (1988)1360-1363.

Fast Ray-Space Interpolation Based on Occlusion Analysis and Feature Points Detection

Gangyi Jiang[1,2], Liangzhong Fan[1,2], Mei Yu[1,3], Rangding Wang[1], Xien Ye[1], and Yong-Deak Kim[4]

[1] Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China
[2] Institute of Computer Technology, Graduate School of Chinese Academy of Sciences, Beijing 100080, China
[3] National Key Laboratory of Machine Perception, Peking University, Beijing 100871, China
[4] Division of Electronics Engineering, Ajou University, Suwon 442-749, Korea

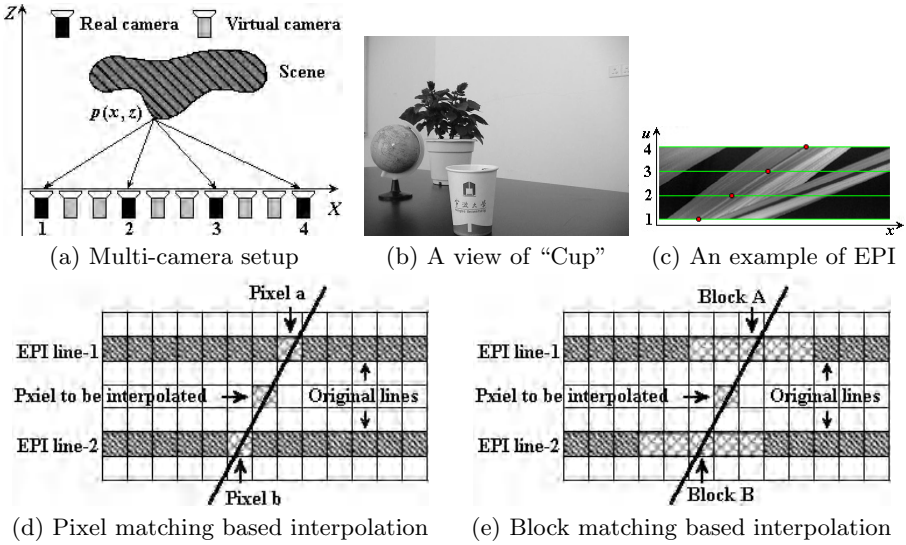
Abstract. A fast algorithm based on depth discontinuity and occlusion analysis is presented for ray-space interpolation. Feature points at depth discontinuity regions in the top and bottom line of sparse epipolar plane image (EPI) is first extracted, and their directions are determined by multiple epipolar matching method. Each of the feature points grows in the nearby epipolar lines until the direction of the grown feature point departures from the direction of the initial point, which means that occlusion occurs or the feature point is in fact a noise. Finally, directional interpolation is implemented within the regions segmented by the feature points. Experimental results show that the new method runs much faster than the pixel matching based and block matching based interpolation methods, additionally, the quality of interpolated EPIs and the rendered arbitrary viewpoint images obtained by the new method is also much better than that of the other two methods.

1 Introduction

Interactivity is a key feature of new and emerging audio-visual media, in the sense that the user shall have the possibility to choose his own viewpoint within a visual scene[1]. Ray-space representation[2] has been approved to be very suited to realize a full real-time Free Viewpoint Video (FVV) system without any restriction on scene[3].

Fig.1(a) shows a camera setup to capture multi-view images for ray-space based FVV system, where the four real cameras are horizontally arranged with the same interval. Virtual viewpoint images corresponding to virtual cameras are interpolated from the real captured images. Fig.1(b) is a view of "Cup" scene, and Fig.1(c) is an epipolar plane image (EPI) of a real scene, in which the four horizontal lines correspond to the same scanlines in the four real captured images. Point $p(x, z)$ in the scene projected to EPI is represented by the four points in Fig.1(c), and these points in the EPI form a locus of straight line, the slope

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 935–940, 2005.
© Springer-Verlag Berlin Heidelberg 2005



(a) Multi-camera setup (b) A view of "Cup" (c) An example of EPI

(d) Pixel matching based interpolation (e) Block matching based interpolation

Fig. 1. Camera setup for multi-view imaging and virtual viewpoint image generation

of which is related to the inverse depth. Ray-space interpolation is associated with arbitrary viewpoint image rendering, and its key problem is how to determine correspondence among points in different epipolar lines. Two interpolation methods had been proposed for this purpose, that is, the pixel matching based interpolation (PMI) method and the block matching based interpolation (BMI) method, as illustrated in Fig.1(d) and Fig.1(e)[4]. The two methods determine correspondence from neighboring epipolar lines with minimum mean square error criterion. The PMI method performs well when disparity is small, but fails in textureless regions. The BMI method achieves better correspondence than the PMI method except for depth discontinuous regions.

As in stereoscope matching, there are many factors that influence the accuracy of correspondence greatly, such as noise, textureless or periodic textured regions, occlusions and depth discontinuities. In this paper, a new ray-space interpolation method is proposed, in which the above factors are taken into account, so that it outperforms the PMI and BMI methods greatly. Firstly, feature points at regions with discontinuous depth are extracted. Then, the correspondences of feature points are determined by utilizing the correlations between multiple epipolar lines. Finally, directional interpolation is implemented within the regions segmented by the feature points.

2 Feature Points Extraction and Multiple Epipolar Matching

Ambiguity in textureless regions is an important factor resulting in inaccurate correspondence for many matching algorithms. In fact, textureless, nearly fronto-

parallel surfaces can be handled quite nicely as long as assumption that intensity variation accompanies depth discontinuities. Given an EPI $I(x,u)$ with the size of $W \times H$, a gradient image $G(x,u)$ is calculated as $G(x,u) = |I(x+1,u) - I(x,u)|$. Since an EPI is composed of the same scanlines of multi-view images, ideally, all lines in an EPI correspond to the same scanline of a scene. Considering computational cost, adaptive threshold is only computed from the first EPI-line for depth discontinuity feature extraction

$$Threshold = \mu + \sigma, \quad \mu = \frac{1}{W}\sum_{i=0}^{W-1} G(i,0), \quad \sigma = \sqrt{\frac{1}{W}\sum_{i=0}^{W-1}(G(i,0)-\mu)^2}. \quad (1)$$

Thus the depth discontinuity feature map $\{f(x,u)\}$ is acquired by

$$f(x,u) = \begin{cases} 1, & \text{if } G(x,u) \geq Threshold \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

As mentioned above, the PMI and BMI methods just utilize nearby EPI line information to determine the correspondence, and often fail at textureless regions or depth discontinuous regions. In view of an EPI is converted from multi-view images, multiple epipolar lines are used to increase the accuracy of correspondence determination. Fig.2 illustrates the multiple epipolar matching method. The pixel at position (x,u) in EPI line u denotes a discontinuous feature point. Multiple epipolar information is incorporated for matching as follows

$$SMSE(d) = \frac{1}{2M}\sum_{k=-M}^{k=M}\left[\frac{1}{2W+1}\sum_{w=-W}^{w=W}|I(x+w+d_{u+k},u+k) - I(x+w,u)|\right] \quad (3)$$

where M is the number of EPI lines nearby EPI line u, W is the size of matching window, d_{u+k} denotes the disparity between EPI line $u+k$ and EPI line u. The range of d_{u+k} is constrained from zero to maximum disparity. Then the direction of feature point (x,u) is determined by

$$d_{best} = \arg\min_{d}\{SMSE(d)\}. \quad (4)$$

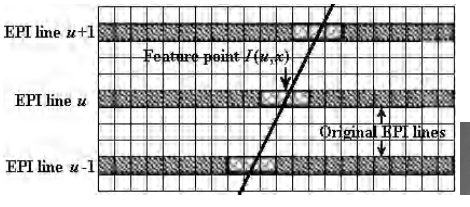


Fig. 2. Multiple epipolar matching method

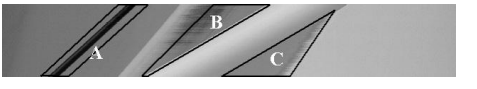
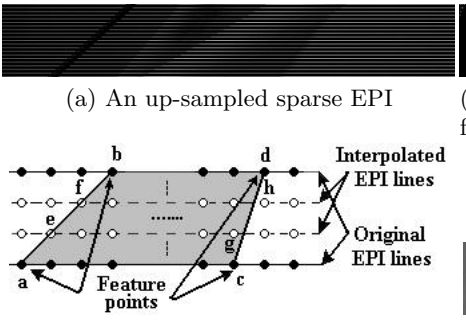


Fig. 3. An EPI with occluded regions

3 Occlusion Analysis and Region-Based Interpolation

In stereoscope or multi-view image coding, objects with different depths in a scene may be occluded with each other when they are projected to the image plane. This makes the correspondence between different views to be more difficult. Fig.3 shows an EPI, the bottom line of which is with respect to the right-most view, while the top line is associated to the left-most view. As we can see, region A has no occlusion, the black object can always be seen from the left-most view to the right-most view. By contrast, the objects with respect to region B and region C are not always visible. We define region B as right-occluded region, for this case, the visible part of object is decreased as camera moves from left to right. Region C, on the other hand, is called as left-occluded region, which is the opposite of the right-occluded region. It is clear that if an object is completely invisible in a view, the correspondence between the view and its neighbors will be failed at the occluded regions.

To handle the occlusion problem, a feature points growing algorithm described as follows is used, during which the correspondence of feature points are also determined. 1) Extract the feature points in the top and bottom EPI-lines as growing seeds, and determine their directions d_{best} by the multiple epipolar matching method. 2) Compute current direction $d_{current}$ of the feature point by the block matching method as shown in Fig.1(e) in which only the neighboring EPI line is considered. 3) If $|d_{current} - d_{best}| \leq 2$, the feature point grows in the neighboring EPI line, otherwise, it is regarded as occluded point or noise point so that the growth of the feature point is stopped. 4) Repeat Step 2 and 3 until the growth of all feature points stops or all EPI lines have been flooded.



(a) An up-sampled sparse EPI



(b) Depth discontinuity feature map after feature point growth

(c) Region-based interpolation



(d) Interpolated EPI ($PSNR$=36.03dB)

Fig. 4. Feature point growth and region-based interpolation

Fig.4(a) gives an up-sampled sparse EPI, while Fig.4(b) is the corresponding depth discontinuity feature map after feature point growth. As we can see, occluded region B and C are detected precisely, and the sparse EPI is split into several regions by the feature points, so a region-based interpolation method is implemented to the up-sampled sparse EPI. Fig.4(c) illustrates the region-based interpolation method. Here a, b and c, d represent four neighboring feature points

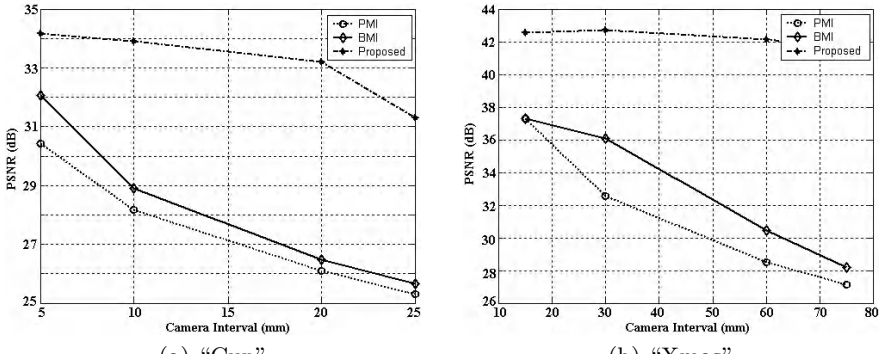
in the original EPI lines, positions of e, f and g, h can be computed from line equation easily, their intensities are linear interpolated from intensities of these feature points, then, the pixels between e and g and the pixels between f and h can be linearly interpolated. Fig.4(d) is the EPI interpolated from Fig.4(a), the quality of which is acceptable.

4 Experimental Results

Experiments are performed on multi-view image sets "Xmas" and "Cup" on a PC with CPU of PIII 850MHz and RAM of 256M. "Xmas" is provided by Tanimoto Lab, and has totally 101 viewpoint images with the size of 640×480, its camera interval is 3mm. For the "Cup" set, the number and the size of images are the same as "Xmas", except that the camera interval is 1mm instead of 3mm, but the maximum disparity is much larger than that of "Xmas". Firstly, the multi-view images are converted into 480 EPIs with different camera interval, the resolutions of EPI are 480×21, 480×11, 480×6, 480×5, respectively. Then, different interpolation algorithms are applied to these sparse EPIs to generate

Table 1. Average *PSNR/Interpolating Time* of 480 EPIs unit: dB/ms

Test sets	Camera interval	PMI	BMI	Proposed method
Xmas	15 mm	40.05 / 22	39.50 / 64	44.29 / 40
	30 mm	35.93 / 41	37.50 / 113	43.76 / 44
	60 mm	31.98 / 87	33.22 / 221	42.87 / 47
	75 mm	30.60 / 94	31.48 / 285	42.46 / 48
Cup	5 mm	33.88 / 41	34.66 / 113	35.85 / 42
	10 mm	32.00 / 78	32.33 / 202	35.18 / 48
	20 mm	30.42 / 149	30.69 / 379	34.30 / 53
	25 mm	29.67 / 190	29.93 / 463	32.59 / 55



(a) "Cup" (b) "Xmas"

Fig. 5. *PSNR*s of rendered virtual viewpoint images under different camera intervals

the dense ones, and the interpolated dense EPIs are compared with actual dense EPIs converted from real captured images with *PSNR* criterion.

Table 1 gives the average *PSNR*s of the 480 interpolated EPIs as well as the corresponding average interpolating times. It is clear that the proposed method achieves much higher *PSNR*s than the PMI and BMI methods. The interpolating time of the PMI and BMI methods increases as camera interval increases, but the proposed method consumes approximately constant time under different camera intervals, and the interpolating time is much saved compared with the PMI or BMI methods especially under large camera interval. Fig.5 gives average *PSNR*s of virtual viewpoint images rendered from interpolated EPIs, which shows that *PSNR*s of rendered images have been improved greatly by the proposed method compared with the PMI or BMI methods.

5 Conclusions

Ray-space representation has superiority in rendering arbitrary viewpoint images of complicated scene in real-time, and ray-space interpolation is one of the key techniques to make the ray-space based FVV system feasible. In this paper, a fast ray-space interpolation method based on depth discontinuity and occlusion analysis is proposed, which improves visual quality as well as *PSNR*s of rendered virtual viewpoint image greatly, and requires lower computational cost, compared with the PMI and BMI methods.

Acknowledgements. This work was supported by the Natural Science Foundation of China (grant 60472100), the Natural Science Foundation of Zhejiang Province (grant RC01057), the Ningbo Science and Technology Project of China (grant 2003A61001, 2004A610001, 2004A630002), and the Zhejiang Science and Technology Project of China (Grant 2004C31105).

References

1. Smolic, A., Kimata, H.: Report on 3DAV Exploration. ISO/IEC JTC1/SC29/WG11 **N5878** (2003)
2. Fujii, T., Tanimoto, M.: Free-Viewpoint TV System Based on Ray-Space Representation. In: Proceedings of SPIE, **4864** (2002) 175–189
3. Smolic,A., McCutchen, D.: 3DAV Exploration of Video-Based Rendering Technology in MPEG. IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications. **14(3)** (2004) 348–356
4. Tehrani, M.P., Fujii, T., Tanimoto, M.: Offset Block Matching of Multi-View Images for Ray-Space Interpolation. The journal of the Institute of Image information and Television Engineers. **58** (2004) 540–548

Non-parametric ICA Algorithm for Hybrid Sources Based on GKNN Estimation

Fasong Wang[1,2], Hongwei Li[1], Rui Li[3], and Shaoquan Yu[1]

[1] School of Mathematics and Physics, China University of Geosciences, Wuhan 430074, China
fasongwang@126.com, hwli@cug.edu.cn
[2] China Electronics Technology Group Corporation, the 27th Research Institute Zhengzhou 450005, China
[3] School of Mathematics and Physics, Henan University of Technology, Zhengzhou 450052, China

Abstract. Novel independent component analysis(ICA) algorithm based on non-parametric density estimation—generalized k-nearest neighbor(GKNN) estimation is proposed using a linear ICA neural network. The proposed GKNN density estimation is directly evaluated from the original data samples, so it solves the important problem in ICA: how to choose nonlinear functions as the probability density function(PDF) estimation of the sources. Moreover the GKNN-ICA algorithm is able to separate the hybrid mixtures of source signals using only a flexible model and it is completely blind to the sources. It provides the way to wider applications of ICA methods to real world signal processing. Simulations confirm the effectiveness of the proposed algorithm.

1 Introduction

Since Comon[1] gave a good insight to ICA problem from the statistical view there has been emerged a set of efficient ICA algorithms, such as the Infomax algorithm[2], the Natural Gradient algorithm[3], the JADE algorithm[4], the Grading Learning algorithm[5], the Flexible Score Function algorithm[6], the FastICA algorithm[7] and so on. But these algorithms still can not capable of universally modeling arbitrarily distributed sources.

Then the nonparametric ICA algorithm[8] was given where the PDF can be estimated from the sample data directly which made the algorithm has a more applied range. In this paper, unlike the nonparametric ICA algorithm given in [8], we propose a novel ICA algorithm that can separate mixtures of symmetric and asymmetric source signals with an efficient score function which is obtained by GKNN non-parametric PDF estimation. As a result, the algorithm proposed in [8] can be seen as a particular case of the one proposed here. Simulations confirm the effectiveness and performance of the proposed algorithm.

This paper is organized as follows: Section 2 introduces briefly the ICA model and maximum likelihood(ML) estimation; Then in section 3, we use a GKNN estimation based non-parametric method to estimate the PDF of the sources and

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 941–946, 2005.
© Springer-Verlag Berlin Heidelberg 2005

get the score function, at the end of this part the new GKNN based algorithm is proposed; Simulations illustrating the good performance of the proposed method are given in section 4; Finally, section 5 concludes the paper.

2 ICA Model and ML Estimation

We consider the ICA model[1,2]:$X(t) = AS(t)$, where $S(t)$ is a unknown stationary source vector. Matrix $A \in R^{N\times N}$ is an unknown real-valued and non-singular mixing matrix and $X(t)$ are sometimes called as sensor outputs. The following assumptions for the model to be identified are needed: a) The sources are statistically mutually independent; b) At most one of the sources has Gaussian distribution; c) Mixing matrix A is invertible. The task of ICA is to recover the original signals from the observations $X(t)$ without the knowledge of A nor $S(t)$. Let us consider a linear feed forward memoryless neural network[9] which maps the observation $X(t)$ to $Y(t)$ by the following linear transform

$$Y(t) = WX(t) = WAS(t), \tag{1}$$

where $W \in R^{N\times N}$ is a separating matrix, $Y(t)$ is an estimate of the possibly scaled and permutated vector of $S(t)$. Inherently there are two indeterminacies in ICA[1]: a)scaling ambiguity; b)permutation ambiguity. That is, the recovered signals are $Y(t) = Q\Lambda S(t)$, where Q is a permutation matrix and Λ is some nonsingular diagonal matrix. So, we get the performance matrix P and $P = WA$.

The Kullback-Leibler(K-L) divergence is an ICA contrast function[1] which measures the mutual statistical independence of the output signals $y_i(t)$. The K-L divergence $D(Y|W) = 0$ if and only if the independence condition $p_y(Y) = \prod_{i=1}^{N} p_i(y_i)$ is satisfied. The likelihood function of this formula is $L(Y, W) = \log p(X) = \log|\det(W)| + \sum_{i=1}^{N} \log p_i(y_i)$. We apply a natural gradient algorithm[3] to obtain the updating rule of W. The general learning rule for updating can be developed as

$$W^{k+1} = W^k + \eta[I - \varphi(Y)Y^T]W^k. \tag{2}$$

In (2), $\eta > 0$ is the learning rate which can be chosen by the method in [5], and $\varphi(\cdot)$ is the vector of score functions whose optimal components are

$$\varphi_i(y_i) = -\frac{d}{dy_i}\log p_i(y_i) = -\frac{p_i'(y_i)}{p_i(y_i)}. \tag{3}$$

As discussed above, the original source distributions must be known. But in most real world applications this knowledge does not exist. Therefore, conventional practical solution is to employ a hypothetical probability density model.

3 The GKNN Estimation of PDF and Separation Algorithm

In this paper we propose a non-parametric method, where the probability density functions are directly estimated from the sample data using a GKNN estimation technique[10].

The basic idea of the KNN method is to control the degree of smoothing in the density estimate based on the size of a box required to contain a given number of observations. The size of this box is controlled using an integer k which is considerably smaller than the sample size. Suppose we have the ordered data sample $x_{(1)}, \cdots, x_{(n)}$. For any point on the line we define the distance by $d_i(x) = |x_i - x|$, so that $d_1(x) \leq \cdots \leq d_n(x)$. Then we define the KNN density estimate by [10]: $\hat{f}(x) = \frac{k-1}{2nd_k(x)}$. By definition we expect $k-1$ observations in $[x - d_k(x), x + d_k(x)]$. Near the center of the distribution d_k will be smaller than in the tails. In application we often use the GKNN estimate $\hat{f}(x) = \frac{1}{nd_k(x)} \sum_{j=1}^{n} K\left(\frac{x - X_j}{d_k(x)}\right)$, where $K(\cdot)$ is a window or kernel function. The most common choice of kernel is: The Gaussian kernel, $K(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$; The Epanechnikov kernel and so on.

The window width $d_k(x)$ is an important characteristic of the GKNN density estimate but there is no firm rule. One chooses $d_k(x)$ keeping in mind the trade-off between too much variability and roughness on the one hand and the lack of fidelity and increase bias on the other hand.

Suppose given a batch of sample data of size M, the marginal distribution of an arbitrary reconstructed signal is approximated as follows[10]: $\hat{p}_i(y_i) = \frac{1}{Md_k(y_i)} \sum_{j=1}^{M} K\left(\frac{y - Y_{ij}}{d_k(y_i)}\right), i = 1, 2, \cdots, N$, where N is the number of the source signals, $K(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$, $Y_{ij} = W_i(X^T)_j = \sum_{n=1}^{N} w_{in} x_{nj}$, W_i and $(X^T)_j$ are the ith row and jth column of the separating matrix W and mixture X^T. We know that this estimator $\hat{p}_i(y_i)$ is asymptotically unbiased estimation and mean square consistent estimation to $p_i(y_i)$. So under the measure of mean square convergence, this estimator converges to the true PDF. Moreover, it is a continuous and differentiable function of the unmixing matrix W

$$\hat{p}_i(y_i) = \hat{p}_i(W_i(X^T)_n) = \frac{1}{Md_k(W_i(X^T)_n)} \sum_{j=1}^{M} K\left(\frac{W_i((X^T)_n - (X^T)_j)}{d_k(W_i(X^T)_n)}\right), \quad (4)$$

From (3), if we want to get the score function, we must get the derivative of (4), $\left(\hat{p}_i(y_i)\right)' = \left(\frac{1}{Md_k(y_i)} \sum_{j=1}^{M} K\left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right)\right)' = \frac{1}{M(d_k(y_i))^2} \sum_{j=1}^{M} K\left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right)$. Then $\left(\hat{p}_i(y_i)\right)' = -\frac{1}{M(d_k(y_i))^2} \sum_{j=1}^{M} \left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right) K\left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right)$. We use the derivative of the estimated PDF as the estimation of the PDF's derivative, that is $\left(\hat{p}_i(y_i)\right)' = \hat{p}_i' y_i)$. So from (3),(4) we can get the score function easily

$$\varphi(y_i) = -\frac{\frac{\partial p(y_i)}{\partial y}}{p(y_i)} = -\frac{\left(\hat{p}_i(y_i)\right)'}{\hat{p}_i(y_i)} = \frac{\sum_{j=1}^{M} \frac{y_i - Y_{ij}}{d_k(y_i)} K\left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right)}{d_k(y_i) \sum_{j=1}^{M} K\left(\frac{y_i - Y_{ij}}{d_k(y_i)}\right)}. \quad (5)$$

Substitute (5) to (2), we can get the GKNN estimation based ICA separation algorithm. In the process of the iteration, in order to confirm the effectiveness of the algorithm, we must update the score function online in every step.

4 Simulation Results

In order to confirm the validity of the proposed GKNN ICA algorithm, several simulations are given below. We simulate it on computer using eight source signals which have different waveforms: four video signals and four imagines which include super-Gaussian, sub-Gaussian sources and asymmetric source signals. The waveforms of them are shown in Fig.1.(a).

In simulations, the mixing matrix A is fixed in each run, but the elements of which are randomly generated. The mixed signals $X(t)$ are plotted in Fig.1.(b). The initial condition of separating matrix W is randomly generated too, the learning rate $\eta = 0.01$ at the first stage of the separation process, then it reduces to $\eta = 0.001$ at the last stage of the optimization, because this procedure can make the algorithm convergence fast and robust[5]. In our simulations, a typical choice of the integer will be used: $k = \sqrt{n}$. The recovered signals $Y(t)$ are plotted in Fig.1.(c)(The sequence of the separated signals have been adjusted in order to contrast to original source signals). It can be seen from the result that the sources were well separated by using the proposed algorithm.

Next, for comparison we execute the mixed signals with different ICA algorithms: JADE Algorithm[4], FastICA algorithm[7], Nonparametric ICA algorithm[8] and Kernel algorithm[11]. At the same convergent conditions, the proposed algorithm was compared along two sets of criteria, statistical and computational. The computational load was measured as CPU time needed for convergence(using Matlab 6.5 and run in the CPU 2.4GHz Pentium4 computer). The statistical performance, or accuracy, was measured using a performance index called cross-talking error index[12](E) defined as $E = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{|p_{ij}|}{max_k |p_{ik}|} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{|p_{ij}|}{max_k |p_{kj}|} - 1 \right)$, where P is the performance matrix. So, if we want to get the good separation result, we must make the E as small as possible. The results of the eight different sources are shown in Table 1 for various ICA algorithms(averaged over 100 Monte Carlo simulation). Then we get that the proposed algorithm can make an ideal separation result.

Table 1. The results of the separation are shown for various ICA algorithms (averaged over 100 Monte Carlo simulations)

	JADE	FastICA	Kernel	NonpICA	GKNN
E	0.1119	0.1092	0.0317	0.0102	0.0096
time(s)	0.5550	1.7862	3649.5	65.501	66.392

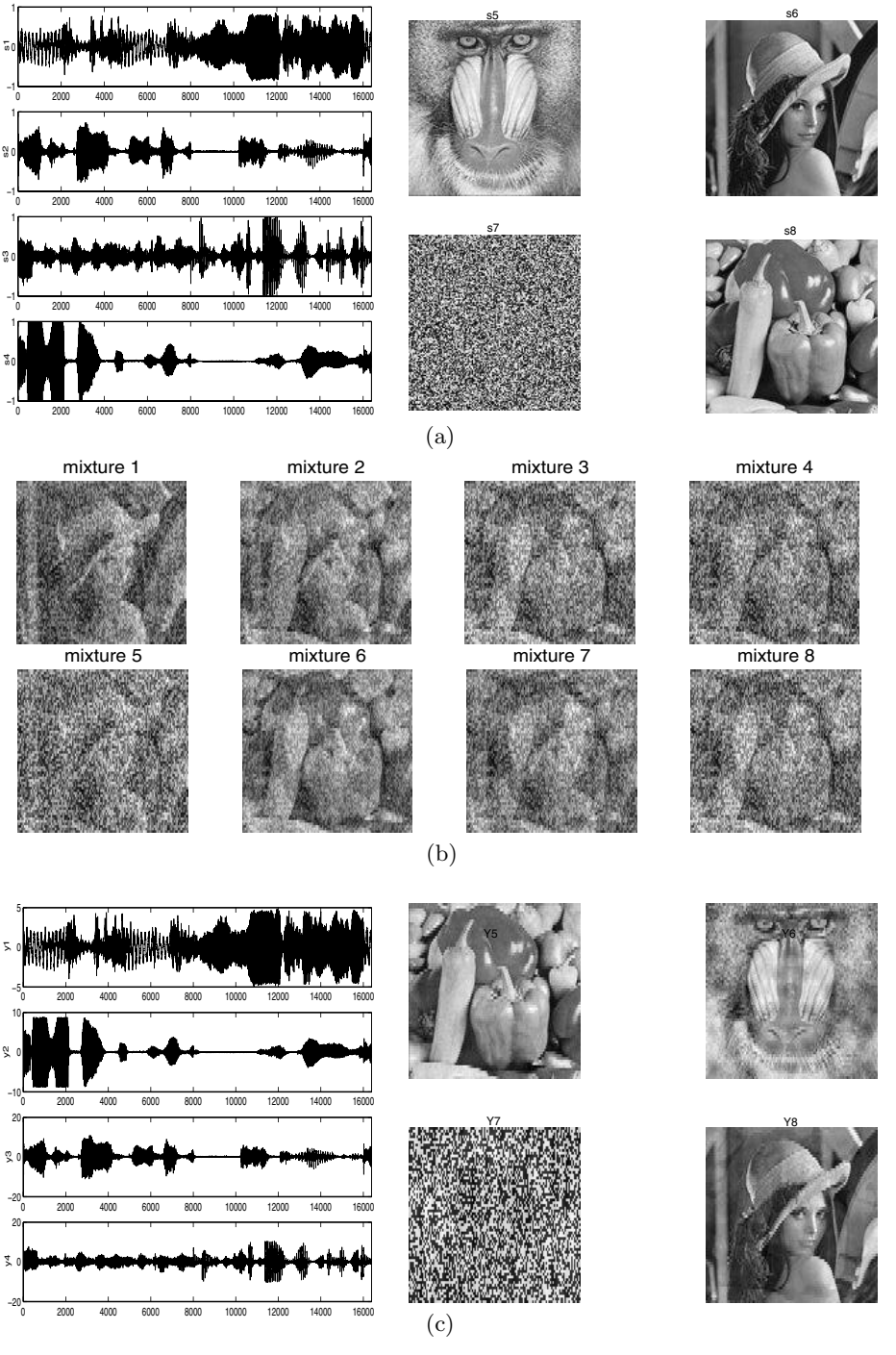


Fig. 1. Experiment Results Showing the Separation of Symmetric and Asymmetric Sources Using the Proposed GNNK Algorithm

5 Conclusion

In this paper, a novel GKNN based non-parametric ICA algorithm has been introduced which is achieved under a natural gradient framework. With its non-parametric property, the approach offers an ICA algorithm which can separate a wide range of source signals, including symmetric(sub-Gaussian and super-Gaussian) and asymmetric sources using only a flexible score function model. The GKNN density estimation method can estimate the PDFs of the sources without any prior knowledge. Finally, the simulation results indicate that the proposed ICA algorithm can make a good performance even if some parametric ICA methods give poor results.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China(Grant No. 60472062) and Natural Science Foundation of Hubei Province, China(Grant No.2004ABA038).

References

1. Comon P., Independent component analysis: A new concept? Signal Processing, 36(1994) 287-314
2. Bell A., Sejnowski T., An information maximization approach to blind separation and blind deconvolution. Neural Computation, 7(1995) 1129-1159
3. Amari S., Natural gradient works efficiently in learning. Neural Computation, 10(1998) 251-276
4. Cardoso J.F., High-order contrasts for independent component analysis. Neural Computation 11(1999) 157-192
5. Zhang X.D., Zhu X.L. and Bao Z., Grading learning for blind source separation. Science in China(Series F), 46(2003) 31-44
6. Wang F.S., Li H.W., Li R. and Shen Y.T., Novel Algorithm for Independent Component Analysis with Flexible Score Functions. 7th Int. Conf. Signal Processing Proceedings, Beijing, Volume 1. IEEE Press, (2004) 132-135
7. Hyvarinen A., Oja E., A fast fixed-point algorithm for independent component analysis. Neural Computation, 9(1998) 1483-1492
8. Boscolo R., Pan H. and Roychowdhury V.P., Independent Component Analysis Based on Nonparametric Density Estimation. IEEE Trans. on Neural Networks, 15(2004) 55-65
9. Zeng,Z.G, Wang,J. and Liao,X.X., Global asymptotic stability and global exponential stability of neural networks with unbounded time-varying delays. IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing, 52(2005) 168-173
10. Silverman B.W., Density Estimation for Statistics and Data Analysis, New York: Chapman and Hall, 1985.
11. Bach F.R., Jordan M.I., Kernel independent component analysis. Journel of Machine Learning Research, 3(2002) 1-48
12. Amari S., Cichocki A. and Yang H.H., A New learning algorithm for blind source separation. Advances in Neural Information Procesing System, Volume 8. MIT Press, Cambridge, MA (1996) 757-763.

SUSAN Window Based Cost Calculation for Fast Stereo Matching*

Kyu-Yeol Chae, Won-Pyo Dong, and Chang-Sung Jeong**

Department of Electronics Engineering, Korea University,
1-5Ka, Anam-dong, Sungbuk-ku, Seoul 136-701, Korea
{wondan, dwp78}@snoopy.korea.ac.kr, csjeong@charlie.korea.ac.kr

Abstract. This paper presents a fast stereo matching algorithm using SUSAN window. The response of SUSAN window is used to calculate the dissimilarity cost. From this dissimilarity cost, an initial match can be found. Then, with this initial match, a dynamic programming algorithm searches for the best path of two scan lines. Since the proposed dissimilarity cost calculation method is very simple, and does not make use of any complicated mathematic formula, its running time is almost as same as SAD in the fixed window. In addition, the proposed matching algorithm only has two control parameters, bright threshold and occlusion penalty, which make it to be easily optimized.

1 Introduction

The stereo vision, which uses the image captured by two different cameras, has been strongly influenced by the rapid performance improvements of computer technology and of embedded systems, thus resulting in the invention of various real-time stereo vision systems[5][6][8] . Many of these systems employ a window-based correlation method. The matching cost calculation is one of the major factors that affect quality of resulting disparity map. Most of real time stereo vision systems use similarity or dissimilarity measurement method in fixed length window[3]. the use of a fixed-length window is due to its simplicity and speed. But in case that disparity jumps are placed in fixed window, the reliability of matching cost becomes poor. To overcome this drawback, many researchers proposed adaptive window method and multiple window method to improve matching cost. Recently, two proposals of window-based stereo cost calculation method (shown in 'variable window'[4], 'Adaptive weight'[7]) have reported successful results.

These two methods employ 'local support information' to find more exact matching cost. However, a decrease in speed from a complex formula and the use of many control parameters make them difficult to be directly adapted to a real-time system. Moreover, the use of optimization process 'winner take all'

* This work was partially supported by the Brain Korea 21 Project and KIPA-Information Technology Research Center.

** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 947–952, 2005.
© Springer-Verlag Berlin Heidelberg 2005

method makes it impossible to find an exact matching in any flat area that is greater than the maximum size of matching window. In order to solve this problem, many global optimization algorithms[2] such as 'dynamic programming', 'maximum flow', and 'graph cut' have been proposed. However, 'maximum flow' and 'graph cut' are inadequate for a fast system due to their immense calculation tasks. consequently, among global optimization algorithms, only 'dynamic programming' is used for the fast stereo vision system.

As a result, in this paper, a fast stereo matching window using SUSAN window[1] is proposed. The size of USAN within the SUSAN window allows the edge detection and feature detection. Similarly, we use the SUSAN window to find the Local Support Area within the window, then, with this result used as a weight, the matching cost of left and right windows can be found. Then, with the method of "Winner takes all," the initial match is found. After performing the reliability check on the initial match, global optimization is done by using a simple dynamic programming. The proposed method only uses two control parameters, which drives the optimization of the proposed algorithm easily. The following sections, from 2 to 5, demonstrate more detailed explanation about the proposed algorithm.

2 SUSAN Operator

SMITH[1] proposed an algorithm that can find the Edge and Feature from the size and weight of Univalue Segment Assimilating Nucleus (USAN) Area within the window. The pixels in the window, which have same or similar brightness, are defined as USAN. The area of USAN delivers the important information about the structure of the window. In the paper[1], USAN is defined by Eq. 1, 3 and the response of SUSAN window is calculated from Eq. 2

$$c(\boldsymbol{r}, \boldsymbol{r}_0) = \begin{cases} 1 & \text{if } |I(\boldsymbol{r}) - I(\boldsymbol{r}_0)| \leq t \\ 0 & \text{if } |I(\boldsymbol{r}) - I(\boldsymbol{r}_0)| > t \end{cases} \tag{1}$$

where $\boldsymbol{r}_0$ is the position of the nucleus in the 2D image, $\boldsymbol{r}$ is the position of any other point within the window,I($\boldsymbol{r}$) is the brightness of pixel, t is the brightness threshold. The output of SUSAN operator n is made ;

$$n(\boldsymbol{r}_0) = \sum_{\boldsymbol{r}} c(\boldsymbol{r}, \boldsymbol{r}_0) \tag{2}$$

$n(\boldsymbol{r}_0)$ is total size of USAN.

$$c(\boldsymbol{r}, \boldsymbol{r}_0) = e^{-(\frac{|I(\boldsymbol{r}) - I(\boldsymbol{r}_0)|}{t})^6} \tag{3}$$

If the similarity function in Eq. 3 is used instead of Eq. 1, a more stable and better result can be obtained. Moreover, if the output of n is less than the pre-computed geometric threshold, it is first defined as an edge or a corner. The geometric threshold of Edge is $\frac{3}{4}n_{max}$ while the corner is defined to be $\frac{1}{2}n_{max}$. The n_{max} is the maximum n value. Since the SUSAN operator does not possess

any calculation of complicated formulas, it is very fast, and the use of only one threshold makes it easy to be controlled. Meanwhile, to obtain an isotropic response from SUSAN operator, Smith proposed a circular window composed with 37 points. The SUSAN window is used to weight of dissimilarity cost. Next section describe it.

3 Matching Cost Calculation

The Local Support Area is the group of pixels which is same or similar to center pixel in windows and follows the smoothness constraint to obtain a similar disparity. In case there is a disparity jump, a more exact cost calculation can be possible if the points inside the local support area are more considered than the ones outside of it in the window cost calculation. To find the local support area, our algorithm uses the SUSAN window operator, and the USAN area is used as the local support area. In order to compare the left and right sides USAN area, it is important to find the common USAN area. Our algorithm uses the intersection of two USAN areas as a weight for the dissimilarity measure of two windows. Intersection is calculated from selecting the minimum value of USAN that is found from each of Left and Right windows. In case of small USAN area like corner, the matching cost value is no longer reliable if the common local support area gets smaller. Thus, in this case, our algorithm uses the union of two USAN areas, which is created by selecting the maximum USAN value of two windows. The Matching cost calculation algorithm is as the following:

The "USAN Area calculation" algorithm

```
for all i,j in Window
{   each USAN_value in position(i,j) is calculated from Eg 1 or 2
   USAN(i,j)=USAN_value;
   totalsum=totalsum+USAN_value; }
```

The "Matching cost calculation" algorithm

```
Find USAN Area calculation in Left SUSAN window
Find USAN Area calculation in Right SUSAN window
for all i,j in Window
{   If left or right totalsum is below geometric corner threshold (1/2) n_max
    then
     weight = max(left USAN(i,j), right USAN(i,j))
    else
     weight = min(left USAN(i,j), right USAN(i,j))
   result = result + weight × ( abs(left pixel(i,j) - right pixel(i,j))
   weight_sum=weight_sum+weight }
result=result/weight_ sum
```

4 Stereo Matching Algorithm

To find the disparity map from already computed cost, a simple dynamic programming is used. First, initial matching is found from the winner-take-all (WTA). Then, the reliability of the initial match is inspected by Left-Right consistency check and the local minima check of best disparity (under WTA optimization, if the cost of the Best disparity is not the local minima then the matching is no loner reliable). Finally, the dynamic programming is used to interpolate the gaps of initial matches. Our algorithm utilize edge information to control occlusion cost. In case there is a right occlusion, there is an edge in the left image, and there is an edge in the right image for a presence of a left occlusion. Then the results obtained from applying the SUSAN operator on the left and right images is compared with the threshold. If it is less than $\frac{3}{4}n_{max}$, then it is decided that there is an edge, and the occlusion cost is consequently decreased. However, if there is an opposite case, the occlusion cost is increased to have a more exact matching as the area does not have an edge.

Penalty computation

$$occ = max(initial\ matches)$$

$$locc = \begin{cases} \text{occ} & right\ n(\boldsymbol{r}_0) < \frac{3}{4}n_{max} \\ & \text{(if a pixel is edge in right image)} \\ \text{occ+occlusion_penalty} & \text{otherwise} \end{cases}$$

$$rocc = \begin{cases} \text{occ} & left\ n(\boldsymbol{r}_0) < \frac{3}{4}n_{max} \\ & \text{(if a pixel is edge in left image)} \\ \text{occ+occlusion_penalty} & \text{otherwise} \end{cases}$$

Here *occ* is the maximum cost of initial matches. As a result, a separate setting is not necessary. So, we define following energy function and finds a path of disparities that minimize the following equation,

$$E = \sum_{match} S(x,d) + \sum_{left\ occlusion} locc + \sum_{right\ occlusion} rocc \tag{4}$$

where $S(x,d)$ is SUSAN cost according to disparity. *locc* is left occlusion cost and *rocc* is right occlusion cost.

5 Experiment Results

In this section, the experiment results are presented. Middlebury test images are used to evaluate our algorithm. All test images are grey Images, the used SUSAN threshold is 15 and Eq. 3 is taken to be used as the similarity function. The circluar SUSAN window with 37 pixels are used. The Occlusion penalty is also set to 10. The "Middlebury web page" provides the test stereo image pairs and ground truth, and the comparison in performance of this program with other algorithms. In the experiment, the percentage of bad matching pixels is

$$B = \frac{1}{N}\sum_{(x,y)} (|d_C(x,y) - d_T(x,y)| > 1) \tag{5}$$

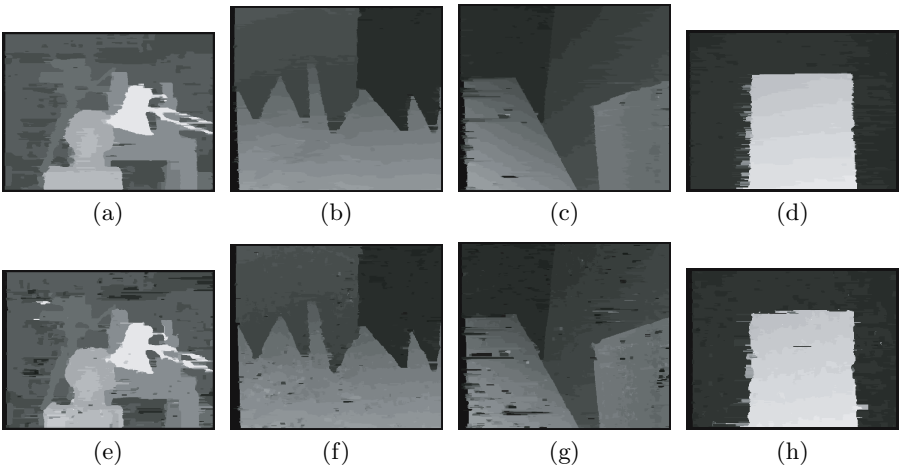


Fig. 1. Final results using DP. (a) Tsukuba(proposed), (b) Sawtooth(proposed), (c) Venus(proposed), (d) Map(proposed), (e) Tsukuba(SAD), (f) Sawtooth(SAD), (g) Venus(SAD), (h) Map(SAD).

Table 1. Comparison between the proposed SUSAN based matching cost and SAD(7x7)

	Tuskuba	Sawtooth	Venus	Map
SUSAN+WTA	11.72	4.36	12.59	4.07
SAD+WTA	18.79	11.27	21.96	4.40
SUSAN+DP	2.21	1.98	3.28	0.74
SAD+DP	6.67	5.77	8.26	2.16

Table 2. Comparison to Middlebury stereo results

Algorithm	Tuskuba			Sawtooth			Venus			Map	
	all	untex	disc	all	untex	disc	all	untex	disc	all	disc
Adapt. weights [7]	1.51	0.65	7.22	1.15	0.29	5.47	1.19	0.72	4.49	1.42	13.4
Reliability-DP [6]	1.36	0.81	7.35	1.09	0.44	4.13	2.35	2.37	13.50	0.55	6.14
Var. win [4]	2.35	1.65	12.17	1.28	0.23	7.09	1.23	1.16	13.35	0.24	2.98
Cooperative [9]	3.49	3.65	14.77	2.03	2.29	13.41	2.57	3.52	26.38	0.22	2.37
Realtime DP [8]	2.85	1.33	15.62	6.25	3.98	25.19	6.42	8.14	25.30	6.45	25.16
Proposed Method	2.21	1.20	11.05	1.98	0.65	10.35	3.28	4.65	16.40	0.74	8.86

where $d_T(x, y)$ is the true disparity at pixel (x, y) and $d_C(x, y)$ is the disparity estimated by our method. This measure B is calculated at various regions of the input image, which have been classified as the entire image (all), untextured (untex) and discontinuity (disc).

Overall test result for SAD and our cost is shown in Fig 1 and table 1. The WTA and the Proposed dynamic programming method are used for performance evaluation. The proposed SUSAN cost improve the quality of disparity map dra-

matically. The new cost takes additionally approximately 10% of execution time than 7x7 SAD window.Although no correction is made to preserve the Scanline Consistency, it is ranked approximately in the middle of Middlebury result rank.(table 2) From the fact that it is not using any complicated calculation, but using a simple algorithm, it is very easy to implement it on any embeded system or GPU based system. Moreover, if the color images were used, the result would nonetheless have been better.

6 Conclusion

In this paper,a new SUSAN cost and a new stereo matching algorithm are proposed. From comparing the proposed cost with other complex matching cost, it might appear to be very simple, but the performances are very similar. Moreover, the simplicity in algorithm and the lack of any complex calculation allow it to be easily implemented into various systems. Only 2 control parameters can easily make the proposed algorithm optimal. Finally, the experimental results successfully show that the proposed algorithm possessed a satisfactory performance. The implementation on a GPU-based real time system with this simpler algorithm is being researched currently.

References

1. S. M. Smith, J. M. Brady: SUSAN–A New Approach to Low Level Image Processing. INT J COMPUT VISION **23** (1997) 45–78
2. D. Scharstein, R. Szeliski: A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. INT J COMPUT VISION **47** (2002) 7–42
3. M. Brown, D. Burschka, and G. Hager : Advances in Computational Stereo. IEEE T PATTERN ANAL (2003) 993–1008
4. O. Veksler: Fast variable window for stereo correspondence using integral images. **1** CVPR. (2003) 556–561
5. R. Yang, M. Pollefeys, S. Li: Improved Real-Time Stereo on Commodity Graphics Hardware. CVPR. (2004) 36–44
6. M. Gong, Y. H. Yang: Near Real-time Reliable Stereo Matching Using Programmable Graphics Hardware. **1** CVPR. (2005) 924–931
7. K. J. Yoon, I. S. Kweon: Locally Adaptive Support-Weight Approach for Visual Correspondence Search. CVPR. **2** (2005) 924–931
8. S. Forstmann, J. Ohya, Y. Kanou, A. Schmitt, and S. Thuering: Real-time stereo by using dynamic programming. CVPRW. (2004) 29–37
9. L. Zitnick, T. Kanade: A cooperative algorithm for stereo matching and occlusion detection. IEEE T PATTERN ANAL **22** (2000) 675–684

An Efficient Adaptive De-blocking Algorithm*

Zhiliang Xu[1,2], Shengli Xie[1], and Youjun Xiang[1]

[1] College of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, Guangdong, China
xzlljl@163.com
[2] Department of Electronic Communication Engineering, Jiang Xi Normal University, Nanchang 330027, Jiangxi, China

Abstract. In this paper, an efficient adaptive de-blocking algorithm is proposed to reduce blocking artifacts. Blocking artifacts are modeled as step functions and then the image blocks are divided into three categories: smooth blocks, texture blocks and edge blocks. For smooth blocks, the expression of amplitude of blocking artifacts is educed firstly in our algorithm, and then the adaptive smooth filter according to the amplitude of blocking artifacts and the smooth degree function is proposed to reduce blocking artifacts. For the texture blocks and edge blocks, the Sigma filter is used to smooth the block boundaries. The experiment results show that the proposed algorithm reduces the blocking artifacts effectively and preserves the original edges faithfully.

1 Introduction

For video coded by BDCT, since each N×N block is transformed and coded independently, the reconstructed image at the side of decoder can generate discontinuities along block boundaries, commonly referred to as blocking artifacts which significantly degrade the visual quality of the reconstructed image.

Many algorithms have been proposed to reduce the blocking artifacts. Blocking artifacts can be modeled as two-dimensional step functions which were firstly proposed by Zeng[1]. In Zeng's algorithm, some of DCT coefficients of some shifted image blocks were set to zero. However, the loss of edge information caused by the zero-masking scheme and the new blocking artifacts are visible. A DCT-domain algorithm based on human vision system for the blind measurement and reduction of blocking artifacts was proposed by Liu [3]. But, the amplitude of step function defined by Liu was computed imprecisely when the shifted block includes texture information. An approach in DCT and spatial domains was proposed to reduce blocking artifacts by Luo[2]. The smoothness constraint set in DCT-domain and quantization constraint set were defined by Paek[4]. This algorithm based on POCS can obtain excellent subjective quality and high PSNR. But this technique is less practical for real-time applications, since it has high computational complexity.

Blocking artifacts are modeled as step functions in our paper, then the image blocks are divided into three categories according to some judging criteria. For smooth blocks, the expression of amplitude of blocking artifacts is educed firstly in our

* The work is supported by the National Natural Science Foundation of China (Grant 60274006), the Natural Science Key Fund of Guang Dong Province, China (Grant 020826).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 953 – 958, 2005.
© Springer-Verlag Berlin Heidelberg 2005

algorithm, and the function of smooth degree is defined. Then, the adaptive smooth filter according to the amplitude of blocking artifacts and the smooth degree is applied to reduce blocking artifacts. For the texture blocks and edge blocks, the Sigma filter is used to smooth the block boundaries in our algorithm. The experiment results show that the proposed algorithm reduces the blocking artifacts effectively and preserves the original edges faithfully.

2 Block Classification

Because human visual system has the activity-masking characteristic of blocking artifacts, the perception of blocking artifacts in smooth region is more sensitive than those of texture and edge region. The information of edge and texture is very important to the quality of image. Over smoothing will cause edge and texture information lost.

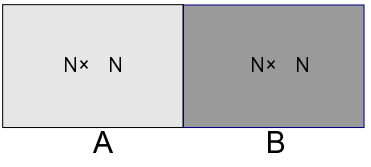


Fig. 1. Horizontally adjacent blocks

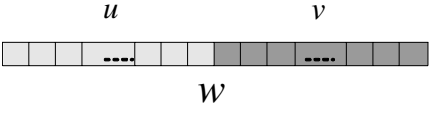


Fig. 2. The concatenated block w of u and v

2.1 Blocking Artifacts Modeling

For image coded by BDCT, the source image of size X×Y is segmented into the blocks of size N×N. Considering two horizontally adjacent blocks, there is a highly noticeable discontinuity at the boundary of blocks A and B, as shown in Fig.1. Fig.2 shows the combined 2N-point 1-D sequence $w(n)$, which is obtained by concatenating the N points from one row of the block A, $u(m)$, and the N points from the same row of the block B, $v(m)$. The blocking artifacts between blocks u and v can be modeled as a 1-D step function in block w. Define a 1-D step function in the block w as follows:

$$s(n) = \begin{cases} 1/4 & 0 \le n \le N-1 \\ -1/4 & \mathrm{N} \le n \le 2N-1 \end{cases} \tag{1}$$

Therefore, block w is modeled as follows:

$$w(n) = e(n) + \beta \bullet s(n) + \gamma(n), \qquad 0 \le n \le 2N-1 \tag{2}$$

Where e is the average value of the block w, γ is the residual block, which describes the local activity of block w, β is the amplitude of the 1-D step function $s(n)$. The value of $|\beta|$ shows the intensity of blocking artifacts.

Let us denote the 1-D DCT of u, v and w by $U(k)$, $V(k)$ and $W(k)$, respectively. The blocking artifacts are modeled by step function s in our paper. Let us denote the 1-D DCT of s by S, then we have

$$S(k)=\begin{cases}0, & \text{k=0, 2,4,...,2N-2}\\ (-1)^{(k-1)/2}\bullet\left(4\sqrt{N}\sin\dfrac{k\pi}{4N}\right)^{-1}, & \text{k=1,3,5,...,2N-1}\end{cases} \tag{3}$$

From (3), it can be found that the even-numbered DCT coefficients of S are zeros, however all the odd-numbered DCT coefficients of S are unequal to zeros. 1-D DCT is applied to the two sides of equation (2), then we have

$$W(k)=E(k)+\beta\bullet S(k)+R(k) \qquad 0\le k\le 2N-1 \tag{4}$$

From (3), (4), we can draw a conclusion that only the odd-numbered DCT coefficients of W are affected by the discontinuities caused by blocking artifacts.

Let us denote the location of the last nonzero DCT coefficient of $U(k)$ and $V(k)$ by $NZ(U)$ and $NZ(V)$ respectively. According to the relation between DCT and DFT [4], if inequation (5) is tenable, we conclude that $W(2N-1)$ is only affected by blocking artifacts, as shown in equation (6).

$$2\bullet\max(NZ(U),NZ(V))<2N \tag{5}$$

$$W(2N-1)=\beta\bullet S(2N-1) \tag{6}$$

2.2 The Classification of Block w

We define the edge block as the block w in which adjacent pixel's difference is larger than the difference of the block boundary. Thus, we have the following relations:

$$\left|w(m+1)-w(m)\right|>\left|w(N)-w(N-1)\right|,\ 0\le m\le 2N-2,\ m\ne N-1 \tag{7}$$

Where $w(N-1)$ and $w(N)$ are the nearest pixels of block boundary. In order to avoid the edge locating at the block boundary, w is also defined as edge block when it satisfies the inequation (8).

$$\left|w(N)-w(N-1)\right|\ge T \tag{8}$$

Where T is a constant. If block w does not satisfy both expression (7) and expression (8), it can be classified into texture block or smooth block according expression (5). Block w is defined as smooth block when it satisfies the expression (5), otherwise it is defined as texture block.

3 De-blocking Algorithm

3.1 De-blocking for Smooth Block

For smooth block w, the intensity of blocking artifacts can be computed according expression (6), as follow

$$\beta = \frac{W(2N-1)}{S(2N-1)} \tag{9}$$

The smooth block w may include some weak texture information. In order to preserve the texture information as perfect as possible, the intensity of de-blocking filter must be adjusted adaptively according to the degree of smoothness of block w. The degree of smoothness is defined as follows:

$$\rho = \log_2[1+\max(\sigma_u, \sigma_v)] \tag{10}$$

where $\sigma_u = \sum_{n=0}^{N-1} |u(n) - \mu_u|$, $\sigma_v = \sum_{n=0}^{N-1} |v(n) - \mu_v|$, where μ_u and μ_v are average value of u and v respectively. To reduce blocking artifacts effectively and preserve texture information faithfully, the step function $s(x)$ is replaced by adaptive function $f(x)$.

$$f(x) = \begin{cases} 1/4 & , \quad f(x) \geq 1/4 \\ -1/4 & , \quad f(x) \leq -1/4 \\ \dfrac{(1+\rho)\bullet x}{30} & , \quad otherwise \end{cases} \tag{11}$$

After replacing the step function with an adaptive function $f(x)$, the filtered block w' is:

$$w'(n) = w(n) + \beta\bullet[f(n) - s(n)] \qquad n = 0,1,2,...,2N-1 \tag{12}$$

3.2 De-blocking for Edge Block and Texture Block

If the concatenated block w belongs to edge blocks or texture blocks, the perception of blocking artifacts in it is not sensitivity for human visual system. Paying attention to both computation and texture information preserving, we use the 5×5 Sigma filter on the block boundaries.

4 Experimental Results

Computer simulations have been performed to demonstrate the proposed algorithm that reduces blocking effects on the JPEG decompressed images. These monochrome images Lena, Peppers and Barbara are used as test images. Three JPEG quantization tables [5] are used to give the images including blocking artifacts.

4.1 Subjective Quality Evaluation

The subjective quality of our algorithm is compared with Zeng's algorithm, Liu's algorithm, Luo's algorithm and Paek's algorithm. In Fig.3, we show the comparison of the results in an enlarged part of the image Lena quantified by Q1.

From Fig.3 (b), it can be observed that there is still exhibiting obvious blocking artifacts, and the new blocking artifacts can be noticed in the recovered image by Zeng's algorithm. In Fig.3 (e) and Fig.3 (f), the blocking artifacts in smooth region are removed effectively while the sharpness of the texture region are preserved faithfully by Paek's algorithm and our algorithm respectively. In Fig.3 (c), most of blocking artifacts in smooth region are removed by Liu's algorithm, however, the adorning region of the helmet is blurred. In Fig.3 (d), recovered by Luo's algorithm, the blocking artifacts are invisible in the face region, but the blocking artifacts are still visible in the texture region of the helmet.



(a) Decoded (b) Zeng's (c) liu's (d) Luo's (e)Luo's (f) Our's

Fig. 3. Comparison of the results in Lena image for Q1

4.2 Objective Quality Evaluation

In Table 1, we present the result of PSNR of the decoded image and the recovered image by different post-processing methods. It is observed that POCS method proposed by Paek [4] can achieve the highest PSNR performance among these five post-processing methods, and it can be seen that the PSNR of the proposed algorithm is better than Zeng [1], Luo[2] and Liu[3].

Simulation results show that the computational complexity of our algorithm is moderate among these five post-processing methods. Though the PSNR achieved by our algorithm is slightly lower than that obtained by Paek's algorithm, the computational complexity of Paek's algorithm is much higher than ours, and it costs the most time among these five post-processing methods.

Table 1. The comparison of block artifacts reduction with different algorithms

Image	Q	Decoded	Zeng's	Luo's	Liu's	Paek's	Our's
Lena	Q1	30.66	30.44	30.65	30.74	31.62	31.46
	Q2	30.08	29.95	30.31	30.34	31.03	31.10
	Q3	27.38	27.71	27.98	28.17	28.43	28.38
Peppers	Q1	30.40	30.22	30.62	30.68	31.22	31.20
	Q2	29.82	29.75	30.25	30.23	30.68	30.66
	Q3	27.22	27.61	27.94	28.02	28.25	28.16
Barbara	Q1	25.84	25.42	25.76	25.78	26.34	26.30
	Q2	25.52	25.19	25.53	25.49	25.99	25.91
	Q3	23.98	23.98	24.27	24.20	24.48	24.45

5 Conclusion

In this paper, an efficient adaptive de-blocking algorithm is proposed. For smooth blocks, the adaptive smooth filter according to amplitude of blocking artifacts and the smooth degree function is proposed to reduce blocking artifacts. For the texture blocks and edge blocks, the Sigma filter is used to smooth the block boundaries in our algorithm. The experimental results with the highly compressed images have shown significant improvement over the existing algorithms, both subjectively and objectively.

References

1. Zeng B.: Reduction of blocking effect in DCT-coded images using zero-masking techniques, Signal process., 79(2) (1999)205-211.
2. Ying Luo and Rabab K.ward.: Removing the blocking artifacts of block-based DCT compressed images, IEEE Trans. Image Processing, 12(7) (2003)838-842.
3. Shizhong Liu and Alan C. Bovik,: Efficient DCT-Domain blind measurement and reduction of blocking artifacts, IEEE Trans. Circuits Syst. Video Technol1, 2 (12)(2002)1139-1149.
4. H.Paek, R.C.Kin, and S.U.Lee.: On the pocs-based postprocessing technique to reduce the blocking artifacts in transform coded images, IEEE Trans. Circuites Sys. Video Technol., 3(8) (1998)358-367.
5. S.Wu, H.Yan, and Z.Tan.: An efficient wavelet based de-blocking algorithm for highly compressed images, IEEE Trans. Circuits Syst. Video Technol., 11(11) (2001)1193-1198.

Facial Features Location by Analytic Boosted Cascade Detector

Lei Wang[1], Beiji Zou[2], and Jiaguang Sun[3]

[1] School of Computer and Communication, Hunan University 410082, China
solehome@163.com
[2] School of Information Science and Engineering, Central South University 410083, China
bjzou@vip.163.com
[3] School of Software, Tsinghua University 100084, China
sunjg@tsinghua.edu.cn

Abstract. We describe a novel technique called Analytic Boosted Cascade Detector (ABCD) to automatically locate features on the human face. ABCD extends the original Boosted Cascade Detector (BCD) in three ways: (i) a probabilistic model is included to connect the classifier responses with the facial features; (ii) a features location method based on the probabilistic model is formulated; (iii) a selection criterion for face candidates is presented. The new technique melts face detection and facial features location into a unified process. It outperforms Average Positions (AVG) and Boosted Classifiers + best response (BestHit). It also shows great speed superior to the methods based on nonlinear optimization, e.g. AAM and SOS.

1 Introduction

The task of facial feature location is important for many tasks, such as quantitative measurement, feature tracking and face recognition. It has generally been addressed by model based algorithms that combine shape and texture modeling [1] [2] [3].

Active Shape Model (ASM) [2] and Active Appearance Model (AAM) [3] are efficient and popular. However a problem with both ASM and AAM is that a good initialization, close to the correct solution is required, otherwise both methods are prone to local minima. This problem has been addressed by two strategies. The first strategy is to use global optimization, e.g. the revised Pictorial Structures algorithm due to Felzenszwalb et.al. [4] [5]. This method is slower than the local optimization based methods and the pictorial structures are still too coarse to handle dozens of facial features points.

The second strategy for avoiding local minima is to combine face detection and facial features location, e.g. the SOS method introduced by Cristinacce et.al. [6] [7]. In such methods the face is first localized using a global face detector, which provides an approximate location and scale. Then shape fitting or features detection methods are applied to search in the approximate face area for the feature points. The chance of local minima is decreased by the good initialization by face detection. Such methods are faster and more robust than purely global optimization.

However, face detection and facial features location are nearly two separated processes in the above methods. Much of the results generated by the face detection have

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 959 – 964, 2005.
© Springer-Verlag Berlin Heidelberg 2005

been discarded without being made use of to locate the facial features. We think it computation waste.

In this paper we present a novel method to fuse face detection and facial features location. For global face detection we use the popular Boosted Classifier Detector (BCD), due to Viola and Jones [8]. We uncover the physical meanings of the feature classifiers in BCD by exploring their probabilistic relationship with the facial features. So we can make use of the feature classifier responses in the face detector to locate the facial features without any shape matching or features detection, which means great ease in both algorithm implementation and execution.

The paper is organized as follows. Section 2 describes the proposed algorithm in detail. Section 3 presents the experiment methods and the testing results. We summarize in section 4 and suggest some directions for future work.

2 Analytic Boosted Cascaded Detector (ABCD)

BCD [8] is a popular method for object detection. The Analytic Boosted Cascaded Detector (ABCD) introduced by this paper has extended the original BCD in three ways: (i) a probabilistic model is included to connect the classifier responses with the facial features; (ii) the features location method based on the probabilistic model is formulated; (iii) a selection criterion for face candidates is presented.

2.1 The Probabilistic Model for ABCD

We model the relationship between the feature classifiers in BCD and the facial features in a probabilistic way. Let the response of the feature classifier f be rsp(f). α1 and α2 are the two thresholds for a given decision feature. If rsp(f)=max(α1, α2), then fc=1, or else fc=0, which means whether the feature classifier responds positively. If its response is positive and the facial feature t lies in its rectangle region, then tc=1, or else tc=0, which means whether the feature classifier can be used to locate the facial feature. By this means, the facial feature and its position x correlate to the feature classifiers f in a probabilistic way, as shown in Equation 1.

$$P(x, fc=1, tc=1) = P(x_t \mid tc=1, fc=1)P(tc=1 \mid fc=1)P(fc=1) \tag{1}$$

Here P(x|tc=1, fc=1) is the probability distribution for the location of the facial feature who lies in a positive-responded feature classifier. We model the probability distribution as 2D normal distributions with the parameters μ and σ.

P(tc=1|fc=1) is the probability for the facial feature lying in a positive-responded feature classifier. It stands for the correlation between the facial feature t and the feature classifier f. The more often t lies in f, the bigger the probability should be. At the same time, the more dispersive t lies in f, the smaller the probability should be, because a dispersive distribution means uncertainty and independency. So we can define the probability as:

$$P(tc=1 \mid fc=1) = \frac{\sum_{i=1}^{Ns} ifIn(t, f) \cdot fc_i}{\sum_{i=1}^{Ns} fc_i} \cdot \frac{1}{\delta} \tag{2}$$

Here *Ns* is the number of training samples. *ifIn*() is a testing function that returns 1 when *t* lies in *f*, otherwise return 0. σ is the standard deviation of the normal distribution motioned above.

P(*fc*=1) in Equation 1 is the probability for the feature classifier to respond positively. This prior can be easily calculated by counting the training samples.

The joint distribution, as shown in Equation 1, describes the probabilistic relation between facial features and the feature classifiers in face BCD. In the following subsection we will present how to use the probabilistic relation to infer feature locations directly from the responses of the feature classifiers.

a) Facial Features Location Based on the Probabilistic Model

Facial features location aims to get the optimal location x* of the facial features. And x* corresponds with the maximization of the joint distribution in Equation 1. Although the optimization is hard to resolve, its logarithm representation can be approximated by the sum of weighted expectations:

$$x^* \approx \sum_{i=1}^{Nf} E[P(x_t \mid tc=1, fc=1)]P(tc=1 \mid fc=1)P(fc=1) \tag{3}$$

Here *Nf* is the number of the feature classifiers in the face BCD. *x** is determined by a series of features classifiers that respond positively, P(*fc*=1)>0, and correlated to it, P(*tc*=1|*fc*=1)>0.

b) Selection of Face Candidates

Generally speaking, the face BCD returns several face candidates *Fk*, *k*>1. So before locating the facial features with the method described above, we have to select the optimal face candidate. We present a selection criterions:

$$P(F_k) = \sum_{j=1}^{Nc} \frac{\sum_{i=1}^{Njf} rsp(f_i)}{Hsj} \tag{4}$$

Here *Nc* is the number of stages in the face BCD. *Njf* is the number of feature classifiers in the *j*'th stage. *rsp*(*fi*) is the feature classifier response. *Hsj* is the decision threshold for the stage. This equation originated from the fact: the more a classifier response exceeds its decision threshold, the surer the decision should be.

The Analytic Boosted Cascaded Detector (ABCD) described above makes face detection and facial features location a unified process.

3 Experiments and Results

We trained the probabilistic model of ABCD according to formulations in section 2.1. The training images were selected from the IMM Face Database[1] and The Caltech-101 Object Categories[2]. Each image has a reasonably large face. And 17 facial feature

[1] M. M. Nordstrom and M. Larsen and J. Sierakowski and M. B. Stegmann. The {IMM} Face Database. http://www2.imm.dtu.dk/pubdb/p.php?3160

[2] Fei-Fei Li, Marco Andreetto, Marc 'Aurelio Ranzato. The Caltech-101 Object Categories Dataset. http://www.vision.caltech.edu/feifeili/Datasets.htm

points were manually labeled for each image. We use the BCD offered by OpenCV[3], which contains 22 decision stages.

ABCD is tested on a public available image set know as the BIOID database[4]. The BIOID images consist of 1521 images of frontal faces taken in uncontrolled conditions using a web camera within an office environment. The face is reasonably large in each image, but there is background clutter and unconstrained lighting. The data set is now available with a set of 20 manually labeled feature points.

Some faces lie very close to the edge of the image in the BIOID data set, which prevents detection using BCD. To concentrate on the feature detection, we discard those examples as in [6].

To assess the accuracy of feature detection the predicted feature locations are compared with manually labeled feature points. The average point-to-point error (m_e) is calculated as follows.

$$m_e = \frac{1}{Nt \times s} \sum_{i=1}^{Nt} d_i \tag{5}$$

Where *di* is the point-to-point errors for each individual feature location and *s* is the known inter-ocular distance between the left and right eye pupils. *Nt* is the number of feature points modeled and it is 17 in this paper.

3.1 Comparison with Other Published Results

Four different features location algorithms are compared with ABCD:

1. Mean position predicted from full face match (AVG)
2. Best feature response for each detector (BestHit)
3. SOS using Nelder-meade Simplex
4. Active appearance model (AAM)

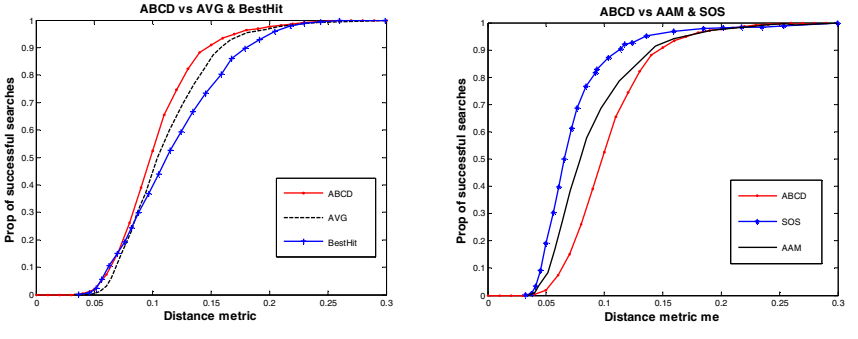


Fig. 1. Accuracy comparison with AVG, BestHit, AAM and SOS

AVG is the simplest algorithm in which the locations of feature points are predicted from the mean distribution without any search. BestHit searches for the feature

[3] Intel® Open Source Computer Vision Library. www.sourceforge.net/projects/opencvlibrary
[4] http://www.humanscan.de/support/downloads/facedb.php

points with many feature BCDs and choose the best feature responses without shape restriction. SOS[6] and AAM[3] are two algorithms based on nonlinear optimization and they locate feature points with shape restriction. In order to confine the search area, all of the algorithms are executed in the optimal face candidate returned by the face BCD.

The left graphs in Figure 1 show that ABCD is more accurate than both AVG and BestHit. With m_e <0.14 ABCD is successful for 88% of faces, whilst AVG works for 79% of faces and BestHit works for 70% of faces. The reason for the poor accuracy of BestHit is its shortage of shape restriction. The wrong locations may lead to big errors. However, Shape restriction is implied in ABCD, because it is only possible for the feature classifiers whose $P(tc=1|fc=1) >0$ to vote for the facial features. So ABCD beats AVG. On the other hand, only the positive-responded feature classifiers can vote for the facial features, which make ABCD a dynamic procedure rather than a static prediction as AVG. So ABCD also defeats AVG.

The right graphs in Figure 1 show that ABCD is less accurate than both AAM and SOS. It is reasonable because AAM and SOS are nonlinear optimization based. They search for the best fit in an iterative way; however, ABCD inferences the feature points in a relatively rough and determinate way. The accuracy disadvantage of ABCD compared with AAM and SOS is partially made up by its speed superior. The speed ratios among ABCD, BestHit, AAM and SOS are 1.00: 3.17: 4.23: 7.40.



Fig. 2. Example features location results with ABCD

Some example features location results are shown in Figure 2.

4 Summary and Conclusions

Analytic Boosted Cascaded Detector (ABCD) introduced by this paper melts face detection and facial features location into a unified process. We formulate the probabilistic relation between facial features and the feature classifiers in face BCD. And we present a features location scheme based on the probabilistic relation. Furthermore, we explore various selection criterions for face candidates. We test ABCD on a large face data set. The results show ABCD outperforms both AVG and BestHit. It also achieves a good balance between accuracy and speed compared with AAM and SOS.

Future work will involve revising the selection criterion for face candidates. Experiment shows the selection of face candidates deeply affects the accuracy of features location. However, the face BCD scans the image discretely in both position and scale, so even the optimal face candidate sometimes can not be good enough to locate feature points accurately. We plan to resample the face BCD around the optimal face

candidate to get the better face estimation. The technique is applied to faces; however ABCD could easily be applied to other features location tasks.

In conclusion ABCD is very fast, and relatively robust to locate facial features. The technique is very applicable to do automatic initialization for AAM or facial features tracking.

References

1. L. Wiskott, J.M. Fellous, N. Kruger, and C.von der Malsburg. Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):775–779, (1997)
2. T. F. Cootes, C. J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. Computer Vision and Image Understanding, 61(1):38–59, January(1995)
3. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H.Burkhardt and B. Neumann, editors, 5th European Conference on Computer Vision, volume 2, pages 484–498. Springer, Berlin, (1998)
4. PF. Felzenszwalb, Dan. Huttenlocher. Pictorial Structures for Object Recognition. International Journal of Computer Vision, Vol. 61, No. 1, January (2005)
5. PF. Felzenszwalb, Dan. Huttenlocher. Efficient Matching of Pictorial Structures. IEEE Conference on Computer Vision and Pattern Recognition, 2000onal Journal of Computer Vision, Vol. 61, No. 1, January (2005)
6. David Cristinacce, Tim Cootes. A Comparison of Shape Constrained Facial Feature Detectors. Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition. Pages 375-380. Seoul, South Korea, May, (2004)
7. David Cristinacce. Automatic Detection of Facial Features in Grey Scale Images.PhD thesis, University of Manchester, Accepted October (2004)
8. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition Conference 2001, volume 1, pages 511–518, Kauai, Hawaii, (2001)
9. Rainer Lienhart, Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. (2002)

New Approach for Segmentation and Pattern Recognition of Jacquard Images

Zhilin Feng[1], Jianwei Yin[2], Zhaoyang He[1,2], Wuheng Zuo[1,2], and Jinxiang Dong[2]

[1] College of Zhijiang, Zhejiang University of Technology, Hangzhou 310024, China
[2] State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou 310027, China
fengzl@zjc.zjut.edu.cn, zjuyjw@cs.zju.edu.cn

Abstract. Phase field models provide a well-established framework for the mathematical description of free boundary problems for image segmentation. In phase field models interfaces represent edges of jacquard images and the determination of the edges of jacquard images is the main goal of image segmentation. In this paper, the phase field model was applied to segment and recognize pattern structures of jacquard images. The segmentation was performed in two major steps. Firstly, a pattern extraction and representation was performed by an adaptive mesh generation scheme. For the conjugate gradient method has been successfully used in solving the symmetric and positive definite systems obtained by the finite element approximation of energy functionals, a novel conjugate gradient algorithm was adapted to the minimization of energy functional of discrete phase model. Experimental results show efficiency of our approach.

1 Introduction

In automatic visual inspection systems for real-time detection of jacquard defects, quantitative analyses require segmentation to identify jacquard pattern objects of interest [1]. Since manual segmentation is tedious, automatic or semi-automatic segmentation methods have been examined [2]. Unsupervised segmentation of jacquard images is a key challenge for any automated analysis of images and major goal of research in jacquard image processing.

Traditional jacquard images have regular periodic texture patterns produced during manufacturing. Hence, the segmentation process of such images can be formulated as a texture classification problem. To achieve that, autocorrelation function, local integration and gray level difference method have been used to extract statistical texture features for fabric segmentation [3,4].

Phase field models [5,6] have been well acknowledged as important methods for complex pattern segmentation. The first phase-field models were intended for the solidification of a pure substance or binary alloy into a single solid phase [7]. Chalupecky [8] presented a numerical scheme for solving the Cahn-Hilliard equation which models the phase separation in binary alloys, and showed some results demonstrating applications of the Cahn-Hilliard equation in image segmentation. Benes *et al.*[9] presented an algorithm of image segmentation based on the level set solution of the Allen-Cahn equation. The non-local Allen-Cahn equation and the constant-mobility Cahn-Hilliard equation have been successfully used for image segmentation [8,9].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 965 – 970, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we propose a two step algorithm to solve the phase field model. Pattern extraction and representation is performed in the first step using adaptive triangulation mesh adjustment. The discrete phase model is minimized in the second step using conjugate gradient algorithm.

The rest of this paper is organized as follows. Section 2 describes the phase field modeling of image system. Section 3 is devoted to discussing the implementation of the proposed method. Some experimental results and conclusions are presented in section 4.

2 Phase Field Modeling of Image System

The phase field model has recently been developed as a powerful tool to simulate and predict the complex microstructure developments in many fields of materials science [10]. The phase field is an auxiliary function of time and space, satisfying some appropriate equation. Let $\Omega \subset R^2$ be a bounded open set and $g \in L^\infty(\Omega)$ represent the original image intensity. The function g has discontinuities that represent the contours of objects in the image. Let $u = u(x,t) \in R$ be a image field, which stands for the state of the image system at the position $x \in \Omega$ and the time $t \geq 0$, and K be the set of discontinuity points of u.

In this case, the phase field energy of the image system is often given as follows:

$$E_\varepsilon(u,K) = \int_{\Omega \setminus K} \left(\frac{\varepsilon}{2} (\nabla u(x))^2 + \frac{1}{\varepsilon} F(u(x)) \right) dx \tag{1}$$

where $F(u) = (u^2 - 1)^2 / 4$ is a double-well potential with wells at -1 and $+1$. Here, the value $u(x)$ can be interpreted as a phase field (or order parameter), which is related to the structure of the pixel in such a way that $u(x) = +1$ corresponds to one of the two phases and $u(x) = -1$ corresponds to the other. The set of discontinuity points of u parameterizes the interface between the two phases in the corresponding configuration and is denoted by a closed set K.

Definition 1. Let $u \in L^1(\Omega; R^2)$, we say that u is a function with bounded variation in Ω, and we write $u \in BV(\Omega; R^2)$, if the distributional derivative Du of u is a vector-valued measure on Ω with finite total variation.

Definition 2. Let $u \in L^1(\Omega; R^2)$ be a Borel function. Let $x \in \Omega$, we define the approximate upper and lower limits of u as $y \to x$,

$$\text{ap} - \limsup_{y \to x} u(y) = \inf\{t : \{u > t\} \text{ has density 0 in } x\}$$

$$\text{ap} - \liminf_{y \to x} u(y) = \sup\{t : \{u > t\} \text{ has density 1 in } x\}$$

We define $u^+(x) = \text{ap} - \limsup_{y \to x} u(y)$ and $u^-(x) = \text{ap} - \liminf_{y \to x} u(y)$. We say that u is approximately continuous at x if $u^+(x) = u^-(x)$. In this case, we denote the common value by $\tilde{u}(x)$ or $\text{ap} - \liminf_{y \to x} u(y)$. Finally, we define the jump set of u by $S_u = \{x \in \Omega : \text{ap} - \lim_{y \to x} u(y) \text{ does not exist}\}$, so that $\tilde{u}$ is defined on $\Omega \setminus S_u$.

Definition 3. A function $u \in L^1(\Omega;R^2)$ is a special function of bounded variation on Ω if its distributional derivative can be written as $Du = fL_n + gH^{n-1}|_K$ where $f \in L^1(\Omega;R^2)$, K is a set of σ- finite Hausdorff measure, and g belongs to $u \in L^1(\Omega;R^2)$. The space of spectial functions of bounded variation is denoted by $SBV(\Omega)$.

By the above definitions, we can give the weak formulation of the original problem (1) as follows:

$$E_\varepsilon(u,K) = E_\varepsilon(u,S_u) \tag{2}$$

and easily prove that minimizers of the weak problem (2) are minimizers of the original problem (1). To approximate and compute solutions to Eq. (2), the most popular and successful approach is to use the theory of Γ- convergence [11].

Let $\Omega = (0,1)\times(0,1)$, let $T_\varepsilon(\Omega)$ be the triangulations and let ε denote the greatest length of the edges in the triangulations. Moreover let $V_\varepsilon(\Omega)$ be the finite element space of piecewise affine functions on the mesh $T_\varepsilon(\Omega)$ and let $\{T_{\varepsilon_j}\}$ be a sequence of triangulations with $\varepsilon_j \to 0$.

Modica [12] proved that

Theorem 1. Let $BVC(\Omega) = \{\psi \in SBV(\Omega) : \psi(\Omega) \subset \{-1,+1\}\}$, and let $W : R \to [0,+\infty)$ be a continuous function, then, the discrete functionals

$$E_\varepsilon(u,T) = \begin{cases} \int_\Omega \left(\frac{\varepsilon}{2}(\nabla u_T(x))^2 + \frac{1}{\varepsilon}F(u_T(x)) \right) dx & if \ \ u \in V_\varepsilon(\Omega), T \in T_\varepsilon(\Omega) \\ +\infty & otherwise \end{cases} \tag{3}$$

Γ- converge as $\varepsilon \to 0$ to the functional $E(u) = c_0 \int_\Omega \Phi(u)\,dx$ for every Lipschitz set Ω and every function $u \in L^1_{loc}(R^2)$, where $c_0 = \int_{-1}^{1} \sqrt{F(u)}du$, and

$$\Phi(u) = \begin{cases} H^1(S_u) & if \ u \in BVC(\Omega)) \\ +\infty & otherwise \end{cases} \tag{4}$$

3 Phase Field Segmentation Algorithm

In order to arrive at the joint minimum (u,T) of Eq. (3), an error strategy for the mesh adaptation is first enforced to refine triangular meshes to characterize contour structures of jacquard patterns. Then, a conjugate gradient algorithm is applied to solve the minimum of discrete version of the phase field functional.

Step 1. Initialize iteration index: $j \leftarrow 0$.

Step 2. Set initial ε_j and u_j.

Step 3. Generate an adapted triangulation T_{ε_j} by the mesh adaptation algorithm, according to u_j.

Step 3.1 Compute L^2 error $\eta_S^j = (\| u_{\varepsilon_j} - g) \|^2 + \frac{1}{\varepsilon} \cdot \sum_{e \subset \partial S \cap \Omega} \| \varepsilon^2 | \nabla u_{\varepsilon_{j-1}} |^2 + F(u_{\varepsilon_{j-1}})) \|^2_{L^2(e)})^{1/2}$, and $\eta(u_{\varepsilon_j}) = \left(\sum_{S \in T_\varepsilon} \eta_S^2 \right)^{1/2}$.

Step 3.2 If $\eta_S^j > \eta(u_{\varepsilon_j})$, then goto Step 4.

Step 3.3 Adjust the triangular mesh by error strategy, and return to Step 3.1.

Step 4. Minimize $E_{\varepsilon_j}(u_j)$ on the triangulation T_{ε_j} by the conjugate gradient minimizing algorithm.

Step 4.1. Initialize step index: $k \leftarrow 0$, define a initial descent direction p_k, define a subspace $W_k = \{p_k\}$ where a suitable admissible minimum of $E_{\varepsilon_j}(u_j)$.

Step 4.2. Compute the gradient $\nabla E_\varepsilon(u_k)$ and the Hessian approximation matrix $\widetilde{\nabla^2} E_\varepsilon(u_k)$, project $\nabla E_\varepsilon(u_k)$ and $\widetilde{\nabla^2} E_\varepsilon(u_k)$ on W_k.

Step 4.3. If $\nabla E_\varepsilon(u_k) = 0$, get a minimizer u_k and goto Step 9.

Step 4.4. Compute the incomplete Cholesky factorization HDH^t of $\widetilde{\nabla^2} E_\varepsilon(u_k)$, where H is lower triangular with unit diagonal entries, and D is positive diagonal.

Step 4.5. If $\widetilde{\nabla^2} E_\varepsilon(u_k)$ is sufficiently positive definite, then compute the descent direction d_k by solving the linear system $\widetilde{\nabla^2} E_\varepsilon(u_k) p_k = -\nabla E_\varepsilon(u_k)$ with the standard preconditioned conjugate gradient method; If $\widetilde{\nabla^2} E_\varepsilon(u_k)$ is only positive semi-definite or almost positive semi-definite, then compute the descent direction d_k by degenerate preconditioned conjugate gradient method.

Step 4.6. Compute the optimum step t_k along d_k by minimizing $E(u_k)$ to the interval $\{u_k + td_k\}$ for $t \in [0,1]$.

Step 4.7. Update the current index: $k \leftarrow k+1$.

Step 4.8. If $| u_k - u_{k-1} | < \gamma$, return to Setp 4.2. Otherwise, return to Step 4.1.

Step 5. Update the current index: $j \leftarrow j+1$.

Step 6. Generate a new ε_j.

Step 7. If $| \varepsilon_j - \varepsilon_{j-1} | > \mu$, return to Step 3. Otherwise, goto Step 8.

Step 8. Stop.

4 Experimental Results and Conclusions

In this section, we present results obtained by means of the above-described algorithms. We conduct experiments on several real jacquard images under noisy environments. Fig. 1 illustrates the segmentation results of noisy jacquard images by our algorithm. Fig. 1(a)-(c) gives three jacquard images with 40% noise. Fig. 1(d)-(f) and Fig. 1(g)-(i) show nodes of Delaunay triangulation and the final foreground meshes of Fig. 1(a)-(c). The segmented edge sets of Fig. 1(a)-(c) are shown in Fig. 1(j)-(l).

In this work a novel method for jacquard images segmentation was presented. Much better segmentation performance was obtained due t o the higher accuracy and

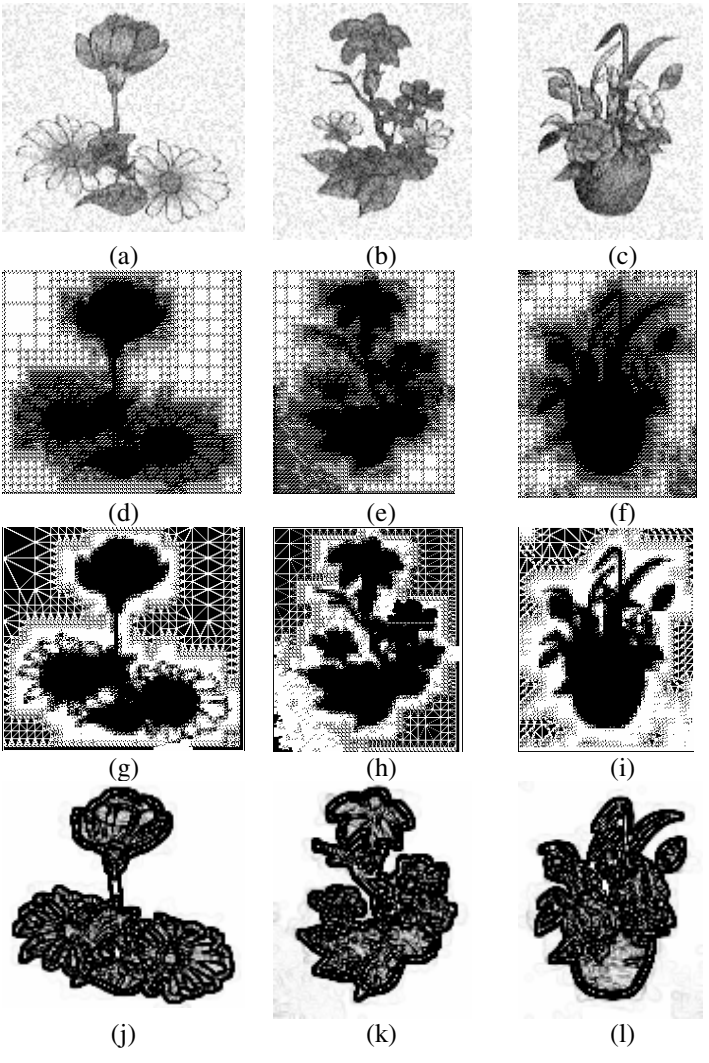


Fig. 1. Segmentation of noisy jacquard images

robustness against noise of phase field technique. The experimental results validate the efficiency of our approach.

Acknowledgement

The work has been supported by the Zhejiang Province Science and Technology Plan(jTang Middleware) and the National Natural Science Foundation, China (No. 60273056). The name and email address of corresponding author is Jianwei Yin and zjuyjw@cs.zju.edu.cn.

References

1. Abouelela, A., Abbas, H.M., Eldeeb, H., Wahdan, A.A., Nassar, S.M.: Automated vision system for localizing structural defects in textile fabrics. Pattern Recognition Letters. **26** (2005) 1435–1443
2. Kumar, A.: Neural network based detection of local textile defects. Pattern Recognition. **36** (2003) 1645–1659
3. Sezer, O.G., Ertuzun, A., Ercil, A.: Automated inspection of textile defects using independent component analysis. In: proceedings of the 12th IEEE Conference on Signal Processing and Communications Applications. (2004) 743–746
4. Yang, X.Z, Pang, G., Yung, N.: Fabric defect classification using wavelet frames and minimum classification error training. In: proceedings of the 37th IAS Annual Meeting on Industry Applications (2002) 290–296
5. Burger, M., Capasso, V.: Mathematical modelling and simulation of non-isothermal crystal-lization of polymers. Mathematical Models and Methods in Applied Sciences. **6** (2001) 1029–1053
6. Nestler, B., Wheeler, A.: Phase-field modeling of multi-phase solidification. Computer Physics Communications. **147** (2002) 230–233
7. Barles, G., Soner, H., Souganidis, P.: Front propagation and phase field theory, SIAM Journal on Control and Optimization. **31** (1993) 439–469
8. Chalupecky, V.: Numerical Studies of Cahn-Hilliard Equation and Applications in Image Processing. In: proceedings of Czech-Japanese Seminar in Applied Mathematics. (2004) 10–22
9. Benes, M., Chalupecky, V., Mikula, K.: Geometrical image segmentation by the Allen-Cahn equation. Applied Numerical Mathematics. **51** (2004) 187–205
10. Han, B.C., Van der Ven, A., Morgan, D.: Electrochemical modeling of intercalation processes with phase field models. Electrochimica Acta. **49** (2004) 4691–4699
11. Ambrosio, L., Tortorelli, V.M.: Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. Communications on Pure and Applied Mathematics. **43** (1990) 999–1036
12. Modica, L.: The gradient theory of phase transitions and the minimal interface criterion. Archive for Rational Mechanics and Analysis. **98** (1987) 123-142

Nonstationarity of Network Traffic Within Multi-scale Burstiness Constraint

Jinwu Wei and Jiangxing Wu

National Digital Switching System Engineering & Technology R&D Center (NDSC),
No. 783, P.O. Box:1001, Zhengzhou 450002, Henan province, China
wjw@mail.ndsc.com.cn

Abstract. The scaling behavior has been discovered in the past decade, which has provided hope that mathematical models can be found to describe the nature of the traffic. Similarly to long-range dependence (LRD), nonstationarity is also one of vital characteristics of network traffic. In this paper, a novel traffic model is proposed based on that the traffic aggregation behavior is abstracted in hierarchical way. The traffic model is focused on the burst traffic rate. Firstly, the burst size of output aggregated flow by edge device of a network domain is derived by pseudo queue system methods. And, the nonstationarity of input traffic is developed by a generalized fractal Gaussian noise process, which is constructed by a large number of train traffic series. They are Poisson arrival and their lifetime is exponential distribution. The model has a good performance of fitting to real traffic data within multi-scale for long time, which is illuminated by simulated results.

1 Introduction

Traffic modeling and understanding are imperative to network design and control. Until now, a large number of studies have been developed [1], however, most of them assumed that the network traffic process is stationary stochastic process. It is well known that traffic rates exhibit weekly and daily pattern because network usage has such pattern [2]. Jin Cao *et al.* show that the nonstationarity is important for Internet engineering [3]. Similarly to the LRD and heavy-tailed marginal distributions, the nonstationarity is also one of fundamental characteristics of network traffic. L. Cruz [4, 5] constructed a "burstiness constraint" network traffic model using non-statistical method. He shows the characteristics of traffic are in spirit somewhat related to the peakedness characterization developed by Eckberg [6]. Unfortunately, this analysis is a conservative worst-case approach because it neglects the statistical properties of traffic as well as their statistical independence.

The aim of this paper is to study the nonstationarity of network traffic in the burst traffic rate space. We propose a hierarchical burstiness constraint (HBC) network traffic model network traffic model, referring the upper bound of burstiness and average rate. The whole space of real network traffic is divided into two subspaces, one is *periodical subspace* and the other is *burstiness subspace*. Under a stability hypothesis, it can readily see that the periodical traffic rate is an almost invariant within an interval. So we focus on the burstiness parameters, which denote the

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 971 – 976, 2005.
© Springer-Verlag Berlin Heidelberg 2005

tendency of the traffic caused by the packet arrival time as well as some other factors. In HBC model, the burst traffic rate is fitted by a nonstationary stochastic process described by a generalized fractional Gaussian noise (fGn) process, and which is constrained by the (σ,ρ,π) regulator [5, 6] over the different time scales, where σ is a bucket size, ρ is the mean rate and π is peak rate.

The remainder of this paper is organized as follows. Section 2 presents the HBC model in detail. The simulated results are given in section 3. Finally, section 4 concludes the paper and presents our targets for future research directions.

2 The HBC Model

2.1 Problem Formulation

The cumulative quantity of information $x(t)=x(0,t)$ generated by a regulated stream in an interval of the length t is such that

$$x(0,t) \le \min\{\pi t, \sigma+\rho t\}. \tag{1}$$

Specially, we define stochastic burstiness of traffic $\sigma_r(\omega)$ considering a realization ω of the probability space Ω as

$$\sigma_r(\omega) = \max\{x(\omega,(s,t)) - \rho(t-s), 0\} \tag{2}$$

From the definition of the envelop of the input $I(t)$, the burst size σ defined by (1) is such that

$$\sigma = \sup_{\omega, s\le i} \sigma_r(\omega,(s,t)) = \max_{s\le I, \omega\in\Omega}\{x(\omega,(s,t)) - \rho(t-s)\} \tag{3}$$

In this paper, we are interested in the burstiness behavior of the traffic inside a network of many nodes as shown in Fig.1. The traffic aggregated by the network is just like what a queuing system does. Thus, we abstract the aggregated function of each edge node device shown in Fig.1 to network traffic as a pseudo queuing system in order to study the network traffic. We classify the possible input traffic processes into three cases shown as Fig.2 according to their different burst characteristics. The three cases under consideration are defined as follows:

Fresh Inputs: This type describes the input to the edge nodes of the network. The burst size is known at the edge node via subscription or a signaling procedure.

Single Output Flows: An output stream from some previous node. The burstiness parameters related to that of the corresponding input stream.

Aggregated Output Flows: One data flow composed of data from the output streams of the same previous node. Its traffic descriptor is a triplet $(\hat{\sigma}^j, \hat{\rho}^j, \hat{\pi}^j)$, where $\hat{\sigma}^j$ is related to the input burstiness parameter of each component.

In these three cases, we focus on the stochastic burstiness properties of the aggregated flows. We consider the following typical scenarios shown as Fig.3. At the k^{th} node, fresh inputs flow $\{x_k^j(t)\}$ and aggregated flows $\{\hat{x}_k^j(t)\}$ are multiplexed

together. All types of inputting data are aggregated together into one flow $\{\hat{x}_{k+1}(t)\}$ as the input to the $(k+1)^{\text{th}}$ node. The input process has $\hat{x}_{k+1}(t)$ parameters $(\hat{\sigma}, \hat{\rho}, \hat{\pi})$.

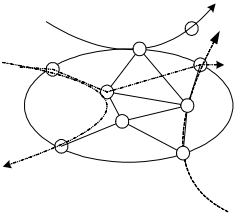


Fig. 1. Problem formulation

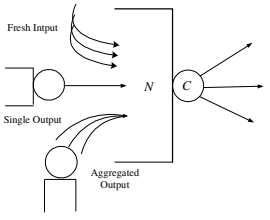


Fig. 2. Input categories

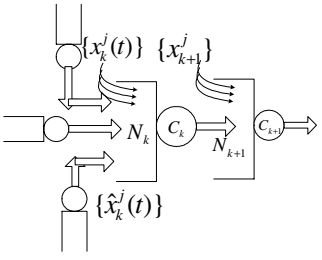


Fig. 3. A scenarios of aggregated flows

2.2 Burst Size for Aggregated Flow

Inputting flows to a general node k in the network are described in Fig.3. The aggregated output flow $\hat{x}_{k+1}^j(t)$ contains data both from M_k aggregated inputs $\{\hat{x}_k^j(t)\}$ with the average rate $\hat{\rho}_k^i$ and N_k independent and identically distributed (I.I.D) fresh inputs $\{\hat{I}_k^j(t)\}$ with the average rate ρ_1^i. Modeling $\hat{\sigma}_{k+1}$ can be achieved by analyzing the busy period of the following $M/G/1$ queue system, denoted by $\tilde{\tau}_k$. We define the burstiness parameter $\hat{\sigma}_{k+1}$ of aggregated flow $\hat{x}_{k+1}^j(t)$ as:

$$\hat{\sigma}_{k+1} = \tilde{\tau}_k (C_k - \hat{\rho}_{k+1}) \Big/ \hat{\rho}_{k+1} \rho_k , \tag{4}$$

where C_k is capacity of node k, and $\hat{\rho}_k$ denotes the average rate of aggregated flow.

The output flow $\hat{x}_{k+1}^j(t)$ is regulated with $(\hat{\sigma}_{k+1}, \hat{\rho}_{k+1}, C_k)$. The moments of burstiness parameter $\hat{\sigma}_{k+1}$ are directly related to those of the busy period $\tilde{\tau}_k$ of the $M/G/1$ queue. For $n \geq 1$,

$$E[(\hat{\sigma}_{k+1})^n] = \left(\hat{\rho}_{k+1}(C_k - \hat{\rho}_{k+1}) \Big/ \rho_k \right)^n E[\tilde{\tau}_k^n]. \tag{5}$$

The mean value $E[\tilde{\tau}_k^n]$ can be explicitly computed from the knowledge of Laplace transform $B_k^*(s)$ of the batch size distribution of the $M/G/1$ queue [7]:

$$B_k^*(s) = \frac{1}{\lambda_k} \sum_{j=1}^{N_k} \frac{\rho_k^j}{\sigma_k^j} e^{-\frac{s\sigma_k^j}{C_k}} + \frac{1}{\lambda_k} \sum_{j=1}^{M_k} \frac{\hat{\rho}_k^j}{E[\hat{\sigma}_k^j]} e^{-\frac{(\hat{\sigma}_k^j)^*(s)}{C_k}} . \tag{6}$$

The $M/G/1$ modeling for the distribution of fluid queue with regulated inputs is asymptotically tight if each source has an average rate very small compared with the capacity. A good approximation for $\hat{\sigma}_k$ requires that each node is not heavily loaded.

2.3 The Nonstationary fGn Process Construction

The nonstationary traffic process is constructed by the superposition of plenty of active fGn traffic train segments in HBC traffic model. The train traffic arrivals form a non-homogeneous Poisson process and train duration is exponential distribution.

Let x_{St}^k as aggregated traffic of the k^{th} output over the time interval $(St, S(t+1))$ with time scale S. Each of traffic trains is a fGn process with the common parameters $\hat{\rho}_k, \hat{\sigma}_k^{\,2}$ and H. So,

$$x_{St}^k \left| \{m_t = m\} = \hat{\rho}_k mS + S^H \sum_{i=1}^{m} \hat{\sigma}_k G_H^{(i)}(t). \right. \tag{7}$$

Now, $\hat{\rho}_k$ can be interpreted as the traffic rate per train and per unit time. $\hat{\sigma}_k$ is the standard deviation of the fGn process, which is constrained by (5), and denotes the traffic burst size. $G_H(t)$ is a stationary fGn process with mean zero. $m(t) = m$ is the counts of the arrival traffic trains, and $m = N_k + M_k$.

Next, consider a latent Poisson process n_t with intensity $\lambda_k(t)(\lambda_k(t) > 0)$ for the cumulative count of traffic train arrivals in the time interval $(0,t)$. Let train lifetimes or durations be independent exponentials with mean μ and $A(t)$ denote the set of active trains at time t. Then $m(t) = |A(t)| = n_t$ – number of train death in $(0,t)$.

Suppose $m(0) \sim$ Poisson $\{\Lambda(0)\}$ with given initial mean $\Lambda(0)$, then $m(t) \sim$ Poisson $\{\Lambda(t)\}$, $E\{m(t)\} = \Lambda(t)$ and autocovariance function

$$\operatorname{cov}\{m(s), m(t)\} = \Lambda(\min\{s,t\}) \exp\left\{ -{|t-s|}/{\mu} \right\}. \tag{8}$$

For time scale S, note that $E[x_{St}^k] = \hat{\rho}_k S\Lambda(tS)$ and variance of x_{St}^k is $[\rho_k^2 S^2 + S^{2H}\hat{\sigma}_k^2]\Lambda(tS)$, Since $E[m(tS)] = \operatorname{var}[m(tS)] = \Lambda(tS)$ and ${m(tS)}/{\Lambda(tS)}$ tends to one probability .The nonstationary fGn process is given by (7), and its variance is the function of the burst size $\hat{\sigma}_k$ of aggregated flows shown in (5).

3 The Results

This section illuminates results of fitting HBC model to real network traffic data collected by the National Laboratory for Applied Network Research (NLANR) network traffic packet header traces [8]. The real traffic data were collected at site of Texas universities GigaPOP at 0:07 in March 15th, 2003. In accordance with the hierarchical formulation of the HBC model in term of a train arrival process within train packet process, we use two-step procedure for fitting the HBC model. First, we estimate the daily pattern $\Lambda(\cdot)$ with a smooth curve method presented in [2, 9]. Second, we find the maximum likelihood estimate of the HBC parameters within the burst constraint by burst size given in (5) and (6). Given the daily pattern $\Lambda(\cdot)$, the

model for the packet counts is simply a variance-component normal model. We apply Quasi-Newton method to find the maximum likelihood estimates (MLE) of $\hat{\rho}_k, \hat{\sigma}, \mu$ and H.

The simulated environment parameters and estimated parameters are presented in table 1. Fig.4 displays the estimation of $\hat{\rho}_k \Lambda(\cdot)$ presented by a thick curve, expressed in packets per 300 seconds. By the estimated $\hat{\rho}_k \Lambda(\cdot)$, the parameters of HBC model is estimated by MLE method shown in table 1, those are constrained by the burst size of aggregated flow. By these parameters, we fit the real traffic data from noon to 23:00 p.m. by our HBC model, and the result is shown in Fig.5. The simulated results show that HBC model has a good performance on the long-term network traffic fitting.

Table 1. **S**imulated environment parameters and estimated parameters by MLE method

Environment Parameter	k	N_k	M_k	C_k	ρ_k^j	π_k^j	σ_k^j
	2	200	2	10	6/200	11	11
Estimated Parameter	$\hat{\rho}_2$	σ	μ	H			
MLE	0.0067	9.1	2.54	0.921			

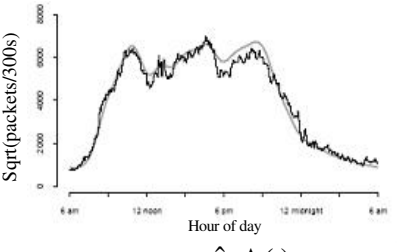


Fig. 4. Estimation of $\hat{\rho}_k \Lambda(\cdot)$ (think curve) along with the NLANR traffic data with the time scale of 300s

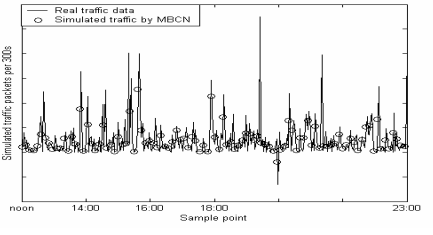


Fig. 5. HBC model fitting traffic data

4 Conclusions

A novel network traffic model has been proposed based on burstiness constraint and nonstationary process in this paper. The HBC model is constructed by two-steps. First, the traffic rate is divided into periodical rate and burstiness rate. Since the periodical traffic rate can be estimated from history network traffic records. Thus, we focus on the burst traffic rate, and present the burst size of aggregated flow in the way of traffic hierarchical aggregation, which can be as a burst constraint on network traffic. Second, as one of most important characteristics of traffic process, the nonstationarity is taken into account in HBC traffic model by a generalized fGn process.

We choose the real traffic data collected by NLANR, and analyze the performance of fitting to real traffic data by HBC traffic model within multi-scale for long time. The results indicate that HBC model has a good performance to analyze the network traffic which bursts seriously. Though the HBC traffic model can fit the Internet traffic effectively, the conclusion should be proved by the more data chosen from the core net.

References

1. Patrice Abry, Richard Baraniuk, Patrick Flandrin, Rudolf Riedi, Darryl Veitch.: The Multisclae Nature of Network Traffic: Discovery, Analysis, and Modelling. IEEE Signal Processing Magzine, Vol.19, No.3. IEEE, (2002) 28-46
2. Joseph L. Hellerstein, Fan Zhang, Perwez Shababuddin.: An Approach to Predictive Detection for Service Management. In Proc. of the Intel. Conf. on Systems and Network Management, (1999)
3. Jin Cao, Willianm S. Cleceland, Gong Lin, Don X. Sun.: On the Nonstationarity of Internet Traffic. Proc. SIGMETRICS/Performance 2001, Joint International Conference on Measurements and Modeling of Computer Systems, ACM, Cambridge, MA, USA. (2001)102-112
4. Rene L. Cruz.: A Calculus for Network Delay, Part I: Network Elements in Isolation. IEEE Trans. on Information Theory, Vol. 37, No. 1. IEEE, (1991)114-131
5. Rene L. Cruz.: A Calculus for Network Delay, Part II: Network Analysis. IEEE Trans. on Information Theory, Vol. 37, No. 1. IEEE, (1991)132-141
6. A.E.Eckberg.: Approximations for Burst (and Smoothed) Arrival Queuing Delay Based on Generalized Peakedness. In Proc. 11th International Teletraffic Congress, Kyoto, (1985)
7. Eick.S.G., Massey, W.A., Whitt, W.: The Physical of the $M_t / G / \infty$ Queue. Operations Research, Journal of the Operations Research Society, (1993)731-741
8. NLANR network traffic packet header traces, http://pma.nlanr.net/Traces/
9. Chuanhai Liu, Scott Vander Wiel, Jiahai Yang.:A Nonstationary Traffic Train Model for Fine Scale Inference From Coarse Scale Counts. IEEE Journal on Selected Areas in Communications: Internet and WWW Measurement, Mapping, and Modeling, Vol.21, No.6. IEEE, (2003)895-907

Principle of Image Encrypting Algorithm Based on Magic Cube Transformation

Li Zhang[1], Shiming Ji[1], Yi Xie[2], Qiaoling Yuan[1], Yuehua Wan[1], and Guanjun Bao[1]

[1] Zhejiang University of Technology, 310032 Hangzhou, China
Zhangli@zjut.edu.cn
[2] Zhejiang Gongshang University, 310035 Hangzhou, China
xieyi@mail.hzic.edu.cn

Abstract. A new method for digital image scrambling transformation, namely Magic Cube Transformation, was proposed and its principle was introduced. Combined with the Logistic mapping in non-linear dynamics system, an image encrypting and decrypting algorithm based on Magic Cube Transformation was designed. A natural chaotic sequence was created with the key. The image matrix was transformed with this chaotic sequence, using the Magic Cube Transformation. Then an encrypted image was resulted. The decrypting operation was the reverse process of encryption. The experimental results indicate that the proposed algorithm can get satisfying effect. Finally, the characteristics of the algorithm were summarized and the aspects of the subsequent work were prospected.

1 Introduction

With the rapid development of multimedia and Internet technology, people have more and more opportunities to exchange information and do e-business via the Internet. The security and confidentiality of the information on Internet are becoming more and more important. As for digital image security and confidentiality, the main methods are information hiding and camouflage. At present, there are mainly four research aspects in this field: digital image scrambling transformation [1], digital image sharing [2], digital image hiding [3] and digital image watermarking [4]. In recent years, digital image scrambling transformation is popularly researched and widely used. It takes into account the characteristics of the image data, while the traditional cryptogram technology simply takes the image as ordinary data flow and encrypts it.

At present, the already existing image scrambling transformations are: Arnold Transformation, FASS Curve, Gray Code Transformation, Conway Game, IFS Model, Tangram Algorithm, Magic Square Transformation, Hibert Curve, Ellipse Curve, Generalized Gray Code Transformation and so on [5],[6]. After studying the already existing image scrambling transformations, a new method for image scrambling transformation, namely Magic Cube Transformation, is proposed in this paper.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 977–982, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Define of Magic Cube Transformation

Magic Cube is a kind of toy which is divided into several sub-cubes. By turning the sub-cubes, a certain picture can be pieced together on the surface of the Magic Cube. And also, the picture pieced together can be thrown into confusion. This kind of transformation can be used to encrypt digital image. A digital image can be regarded as a 2D matrix $I_{M\times N}$. M and N are the height and width of the image respectively.The elements $i_{kl}(k = 1, 2, ..., M; l = 1, 2, ...N)$ of matrix $I_{M\times N}$ are the gray-levels of the pixel (k,l)of the image. A digital image can be looked as the picture on the surface of the Magic Cube and every pixel of the image can be taken as a sub-cube. According to the rules of Magic Cube toy, each row and column of the image matrix can be"turned", just as the sub-cube. Because digital image is planar, circular convolution is adopted to realize the"turning" of the image pixels."Turning" the row $i_k$can be regarded as shifting the row in a certain direction with step h_k.In this paper, h_k is called shifting parameter, which can be obtained by given algorithm. The circular convolution can also be implied to the columns of the image. When every row and column of the original image is circular convoluted a time, the transformation is done and a new encrypted image $I'_{M\times N}$ is produced. That is

$$I'_{M\times N} = p(I_{M\times N}) \tag{1}$$

Where p denotes the mapping from $I_{M\times N}$ to $I'_{M\times N}$. This transformation is defined as Magic Cube Transformation. To get expected encrypting result, the Magic Cube Transformation can be done several times.

3 The Image Encrypting Algorithm Based on Magic Cube Transformation

3.1 Describing the Algorithm

In the traditional encrypting methods, the parameters used to scramble the plaintext are given in advance. In this paper, a sequence produced by the key is used to encrypt digital image, which enhances the security of the encrypted data.

Firstly, with the input number x_0 , a chaotic sequence is created by Eq.2.The sequence is properly processed to obtain a natural chaotic sequence. Then the image matrix is transformed with the elements in this natural chaotic sequence as the shift-parameters, using the Magic Cube Transformation. The transforming operation is repeated for n times. Here x_0 and n form the key together.

The proposed algorithm only rearranged the location of the image pixels, without changing the values of their gray-level. So the original image and the encrypted one have the same gray-level histogram. Considering the encrypting and decrypting efficiency, the proposed algorithm is only applied in the spatial domain.

3.2 Producing the Shifting Parameters Using Chaotic Sequence

To scramble a digital image $I_{M\times N}$ with Magic Cube Transformation, the shifting parameter h_k must be given firstly. This is also an important thing in the

image encrypting. Researches imply that non-linear dynamics system will become chaotic under a certain range of control parameter. The chaotic sequence produced by non-linear dynamics system has characteristics such as certainty, pseudo-randomicity, non-periodicity, non-convergence and sensitively depends on the initial value [7],[8]. All these traits are very useful to image encrypting. The initial value can be part of the key and the chaotic sequence can be used as shifting parameter set H after properly processed.

A one-dimensional discrete-time non-linear dynamics system can be defined as follows [8],[9]:

$$x_{k+1} = \tau(x_k) \tag{2}$$

where $x_k \in V, k = 0, 1, 2...$, is called state. $\tau : V \to V$ is a mapping, which maps the current state x_k to the next state x_{k+1} . A chaotic sequence$\{x_k \mid k = 0, 1, 2, ...\}$ can be gained by repeating Eq.2 with the initial value x_0. This sequence is called a track of discrete-time dynamics system.

Logistic mapping is a kind of simple dynamics system, which is widely investigated and applied. It is defined as follows:

$$x_{k+1} = \mu x_k(1 - x_k) \tag{3}$$

where μ is called ramus parameter, $0 \leq \mu \leq 4$. When $\mu = 4$, the probability distribution function of Logistic mapping is:

$$\rho(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}}, -1 < x < 1 \\ 0, \qquad \text{else} \end{cases} \tag{4}$$

Then the mean value of the chaotic sequence's track points can be calculated using following equation:

$$\overline{x} = \lim_{N\to\infty} \frac{1}{N} \sum_{i+0}^{N-1} x_i = \int_0^l \rho(x)dx = 0 \tag{5}$$

Suppose x_0 and y_0 are two independent initial values, then the correlation function is:

$$c(l) = \lim_{N\to\infty} \frac{1}{N} \sum_{i+0}^{N-1} (x_i - \overline{x})(y_{i+l} - \overline{y}) = \int_0^l \int_0^l \rho(x, y)(x - \overline{x})(\tau^l(y) - \overline{y})dxdy = 0 \tag{6}$$

The united probability distribution function is:

$$\rho(x, y) = \rho(x)\rho(y) \tag{7}$$

And the self-correlation function is delta function$\delta(l)$.

All these statistical characteristics indicate that the traversal statistical characteristics of chaotic sequence are the same as white noise. This is just what image processing needs.

By simple variable replacement, the Logistic mapping can be defined within $(-1, 1)$:

$$x_{k+1} = 1 - \lambda x_k^2 \tag{8}$$

Where $\lambda \in [0, 2]$.When $\lambda = 2$, it is called full mapping.

3.3 Steps of Encrypting Algorithm

The detailed steps of encrypting algorithm using Magic Cube Transformation are as follows:

Step 1. Input parameter: (1) Input the image file InImage $I_{M\times N}$. M and N are the height and width of the image respectively. (2) Input the initial value x_0 and iterative time n.

Step 2. Calculate parameter: (1) Using Eq.2 to get a chaotic sequence $\{x_k \mid k = 0, 1, 2, ..., (M + N) \times n - 1\}$, with $\lambda = 2$. (2) Processing the sequence above properly to get a natural chaotic sequence. That is, the position values of the track points of the chaotic mapping compose the new sequence $\{h_k \mid k = 0, 1, 2, ..., (M+N) \times n-1\}$. This natural chaotic sequence is used as the shifting parameter set H for Magic Cube Transformation.

Step 3. Encrypt image: (1)Circularly shift every row of the image matrix $I_{M\times N}$ with shifting parameter h_k. Then circularly shift every row of the image matrix. (2) Repeat for n times.

Step 4. Get the encrypted image OutImage and output it.

3.4 Steps of Decrypting Algorithm

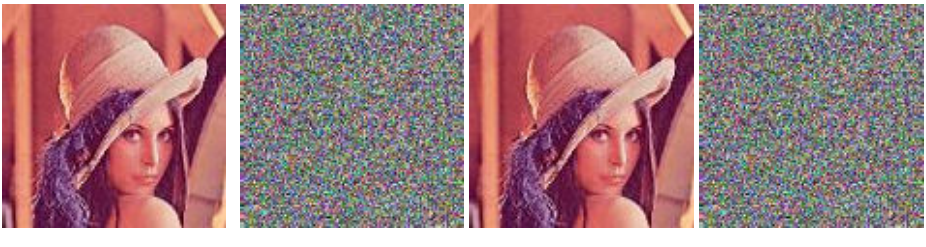
When the correct key x_0 and n are obtained, the original image can be regenerated by imply the reverse course of the encrypting algorithm.

4 Experimental Results

Several digital images are encrypted and decrypted using the Magic Cube Transformation proposed in this paper.

One of the experimental results is shown in Fig.1. The Lena image is a 24-bit true color image. Its size is 136×136. The encrypted image is shown in Fig.1(b), with the initial value $x_0 = 0.75436$ and iterative time n=1. The decrypted image with the correct key $x_0 = 0.75436$, n=1 is shown in Fig.1(c). The decrypted image with the wrong key $x_0 = 0.754359$, n=1 is shown in Fig.1(d).

The encrypted image, which is blurred by noise or geometrically distorted, can also be decrypted. The experimental result is shown in Fig.2. Random noise,



(a) Original image (b) Encrypted image (c) Decrypted image with correct key (d) Decrypted image with wrong key

Fig. 1. The encrypting and decrypting of Lena image



(a) Random noise (b) Salt and pepper noise (c) Gauss noise (d) Geometrically distorted

Fig. 2. The decrypted images of blurred by noise or geometrically distorted data

Table 1. The times for encrypting (unit: second)

	n					
size	10	20	40	60	80	100
136×136	2	4	15	31	54	79
256×256	5	16	86	131	168	198

salt and pepper noise and Gauss noise with intensity of 5% are added to the encrypted image. The decrypted images are shown as Fig.2(b),2(c). Another encrypted image is geometrically distorted, that is, the 5% area of it is rubbed to white and the decrypted image is shown as Fig.2(d).

To discuss the efficiency of the algorithm, different iterative time n is used for experiment. Two 24-bit true color images, which sizes are 136×136 and 256×256 respectively, are encrypted, with the initial chaotic value $x_0 = 0.75436$. The times are shown in Table 1.

5 Conclusions

A new method for image scrambling transformation, Magic Cube Transformation, is proposed in this paper. Its designing thought comes from the magic cube toy and it employs Logistic mapping to generate the needed parameter set. Experiments are done to prove the algorithm.

From the theory analysis and experiment results, the characteristics of the algorithm proposed can be summed up as follows:

- Good encrypting efficiency and security. In Fig.1, the correct key is $x_0 = 0.75436$ and the wrong key is $x_0 = 0.754359$. The difference is only 0.000001. But the decrypted image with the wrong key is far different from the original one.
- Quite well robustness. From Fig.2, it can be concluded that the proposed algorithm can resist noise and geometrically distortment in certain intensity. The images, which decrypted from attacked data, are quite clear and ideal.

- High efficiency. The time needed to encrypt a 136×136 32-bit true color image with iterative time n=80 is less than 1 minute, as shown in Table 1. The less the iterative time n is, the faster it will be.
- Flexible adaptability. The algorithm can be used to encrypt binary image, gray-level image, 256 color image and true color image.

In the future, the periodicity and other characteristics of Magic Cube Transformation will be investigated. The algorithm can also be combined with other encrypting technologies to enhance its encrypting effect and security.

References

1. Yen, J.C.,Guo,J.I.: Efficient hierarchical chaotic image encryption algorithm and its VLSI realization. IEE proceedings-vision image and signal processing, Vol. 147 (2)(2000)167–175
2. Ujvari, T.,Koppa,P.,Lovasz, M.,et al.: A secure data storage system based on phase-encoded thin polarization holograms. Journal of optics A-pure and applied optics, Vol. 6(4)(2004)401–411
3. Wang, R.Z.,Lin,C.F.,Lin,J.C.: Image hiding by optimal LSB substitution and genetic algorithm. Pattern recognition, Vol. 34(3)(2001)671–683
4. Cai, L.Z.,He,M.Z.,Liu.,Q.: Digital image encryption and watermarking by phase-shifting interferometry. Applied optics, Vol. 43(15)(2004)3078–3084
5. Qi, D.X.,Zou,J.C.,Han,X.Y.: A new class of scrambling transformation and its application in the image information covering. Chinese in Science(Series E), Vol. 43(3)(2000)304–312
6. Finher, Y.: Fractal image compression. Fractals, Vol. 2(3)(1994)347–361
7. Scharinger,J.: Fast encryption of image data using chaotic Kolmogorov flows. Proceedings of the International Society for Optical Engineering, San Joe, California, Vol. 3022(1997)278–289
8. Dedieu, H.,Ogorzalek.M.J.: Identifiability and identification of chaotic systems based on adaptive synchronization. IEEE Trnas Circuits & Sys I, Vol. 44(10)(1997)948–962
9. Xiang, H.: Digital watermarking systems with chaotic sequences. In: Proceedings of Electronic Imaging'99. Security and Watermarking of Multimedia Contents, SPIE, San Jose, Vol. 3659(1999)449–457

A Study on Motion Prediction and Coding for In-Band Motion Compensated Temporal Filtering*

Dongdong Zhang, Wenjun Zhang, Li Song, and Hongkai Xiong

The Institute of Image Communication & Information Processing, Shanghai Jiao Tong Univ., Haoran Hi-tech Building, No.1954 Huashan Road, Shanghai 200030, China
{cessy, zhangwenjun, songli, xionghongkai}@sjtu.edu.cn

Abstract. Compared with spatial domain motion compensated temporal filtering (MCTF) scheme, in-band MCTF scheme needs more coding bits for motion information since the motion estimation (ME) and motion compensation (MC) are implemented on each spatial subband. Therefore, how to employ motion prediction and coding is a key problem to improve the coding efficiency of in-band MCTF. In this paper, we proposed an efficient level-by-level mode-based motion prediction and coding scheme for in-band MCTF. In our scheme, three motion prediction and coding modes are introduced to exploit the subband motion correlation at different resolution as well as the spatial motion correlation in the high frequency subband. To tradeoff the complexity and the accuracy of block-based motion search, a jointly rate-distortion criterion is proposed to decide a set of optimized motion vector for three spatial high frequency subbands at the same level. By the rate-distortion optimized mode selection engine, the proposed scheme can improve the coding efficiency about 0.6db for 4CIF sequence.

1 Introduction

As many video communication applications take place in heterogeneous environment, the scalability of a video codec becomes an important feature besides coding efficiency. 3D Wavelet video coding provides an elegant solution for scalable video coding due to its multi-resolution nature. Of various 3D wavelet video coding schemes, most can be classified into two categories: spatial domain MCTF (SDMCTF) scheme [1], [6] and in-band MCTF (IBMCTF) scheme [3], [5]. The major difference between them is whether temporal transform is implemented before spatial decomposition or not.

Compared with SDMCTF scheme, IBMCTF scheme can achieve a competitive performance for lower resolution coding, while it suffers performance loss for higher resolution coding. There are two primary reasons for this: (1) the shift-variance of wavelet makes ME and MC with critical sampled wavelet coefficients not very efficient. (2). IBMCTF scheme needs more bits for coding motion information since

* This work was supported in part by the National Natural Science Foundation Program of China (No. 60332030) and the Shanghai Science and Technology Program (04ZR14082).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 983 – 988, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the ME and MC are implemented at each subband, which leads to decreased coding bits for residual coefficients so that the coding performance is deteriorated.

To overcome the effect of shift-variance of wavelet transform on ME and MC in wavelet domain, Park and Kim proposed low band shift (LBS) method, which can do ME and MC more efficiently with the overcomplete form of reference frames [2]. Based on LBS method, Schaar et al employed the interleaving algorithm for the overcomplete wavelet coefficients to get an optimum sub-pixel interpolated reference for in-band ME/MC, which leads to an improved performance [3].

With overcomplete in-band ME and MC, a satisfied coding efficiency can be achieved for spatial low frequency subband. However, for spatial high frequency subbands, the individual band-by-band ME does not work satisfactory because of the lack of the low frequency signals. Recently, there have been many attempts to further improve the macroblock-based ME/MC for spatial high bands. In [4], the interpolated motion vectors of spatial low frequency band were directly used for the motion compensation of high frequency bands. Reference [5] presents several motion vector prediction and coding schemes for level-by-level in-band ME and MC. Reference [7] focuses on the close-loop in-band scheme with h.264 compatibility and presents several promising inter motion prediction modes and intra prediction modes for spatial high frequency subbands. This method can improve coding performance efficiently. However, it is very time-consuming because each block of each spatial high frequency subband needs many times of motion search.

In this paper, we focus on the in-band MCTF scheme and investigate inter-subband motion correlations at different resolution and the spatial motion correlation. Three motion prediction modes are introduced to exploit these correlations. To tradeoff the complexity and the accuracy of block-based motion search, a jointly rate-distortion criterion is proposed to decide a set of optimized motion vector for three spatial subbands at the same level. The rest of this paper is organized as follows. In section 2, the in-band MCTF scheme is introduced. Section 3 presents our proposed motion prediction and coding scheme. In section 4, the experimental results are presented and discussed. Section 5 concludes this paper.

2 In-Band MCTF Scheme

Fig. 1 shows the generalized in-band MCTF scheme block diagram. The spatial wavelet transform is applied to original video sequence prior to temporal decomposition. Then temporal decomposition performs a multi-level MCTF operation which decomposes each spatial subband into several temporal subbands. For each temporal band of a certain spatial band, the spatial transform can be further employed to remove the spatial correlation. Then the residual coefficients of each spatio-temporal band and the motion vector information are coded with entropy coding. To exploit the temporal correlation more efficiently, the LBS method in [2] and the interleaving algorithm in [3] can be used to overcome the wavelet shift-variance. For each temporal filtering, the adaptive block-based motion alignment technique in [6] was adopted for motion estimation of each spatial subband. Except for the temporal correlation between the subband frames, there is strong motion correlation among the subbands at different resolution because motion activities at different spatial

resolution actually characterize the same motion with different scales and different frequency ranges. In the next section, we discuss several motion vector prediction and coding modes to exploit these correlations.

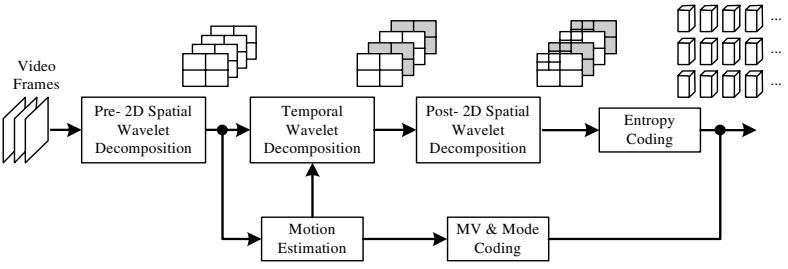


Fig. 1. Generalized In-band MCTF scheme block diagram

3 Level-by-Level Mode-Based Motion Prediction and Coding

There are two methods for the inter-subband de-correlation. The first one is to predict the motion vectors of each subband at higher resolution from the ones at lower resolution with band-by-band manor. The second one is to jointly predict a set of motion vector for three high frequency bands at the same resolution from the ones at lower resolution with level-by-level manor. Although band-by-band manor can exploit more accurate motion vectors for each high frequency band, it is very time-consuming and needs more bits for motion information compared with level-by-level prediction manor. Our experimental results in Fig. 2 demonstrate that these two manors have similar coding performance. To reduce the complexity of motion search, we adopted level-by-level manor to investigate the motion correlation of subbands. In the following, three modes are introduced and a jointly rate-distortion criterion is presented to decide a set of optimized motion vector for three spatial subbands at the same level.

3.1 Mode Types

Assume that N-level spatial transform is used before the temporal transform. Each mode is described as follows:

- **DirL:** in this mode, the motion vectors of all high frequency bands are inherited from the low frequency band at the lowest spatial resolution. This mode considers that all spatial subbands have same motion tend so the motion vectors of low frequency band can be used for the high frequency bands to greatly save motion bits. The motion correlation from the coarse to fine resolutions can be described as

$$\mathrm{LL}_1 -> LH_i, HL_i, HH_i \qquad i = 2,...,N \tag{1}$$

  Where, the motion vectors in the right subbands are obtained by scaling the motion vectors in the left subband according to the resolution level. That is, all motion vectors have the same direction but different size.

- **LPR:** in this mode, motion vectors of subbands at higher resolution are predicted by motion vectors at lower resolution and are refined level-by-level. Based on the predictor of motion vector, more accurate motion vectors can be obtained for the high frequency subbands. This mode greatly exploits the correlation among the subbands at different resolution. At the same time it considers the motion similarity among the subbands at the same level. Expression (2) presents the correlation between subbands at different resolution.

$$\begin{aligned} &\mathrm{LL}_1 -> LH_1, HL_1, HH_1 \\ &LH_{i-1}, HL_{i-1}, HH_{i-1} -> LH_i, HL_i, HH_i \qquad i = 2,...,N \end{aligned} \tag{2}$$

  Where, the motion vectors in the right subbands are refined by the motion vectors in the left subbands.
- **SPR:** in this mode, the motion predictor of each macroblock in spatial high frequency subbands is the median of the three previous coded motion vectors from the left block, the above block and the above-right block adjacent to the current macroblock. If either one motion vector is not available (outside the picture), it is not as the candidate motion vector. This mode can further reduce motion coding bits when the neighboring macroblocks have close motion vectors after refining. It represents the spatial motion relevance inside high frequency subband.

3.2 R-D Optimized Mode Decision

To tradeoff the complexity and the accuracy of block-based motion search, one of the pre-mentioned three prediction modes will be selected based on the following jointly R-D optimized criterion.

$$RD_Cost = \lambda_{motion} \cdot (R_{mv} + R_{\mathrm{mode_pr}} + R_{\mathrm{mode_pa}}) + \sum_{j=1}^{3} W^j \cdot D_{SAD}^j(\mathrm{mode_pr}, \mathrm{mode_pa}, mv) \tag{3}$$

Where, $\mathrm{mode_pr}$ and $\mathrm{mode_pa}$ denotes the proposed motion prediction mode and the partition mode of one macroblock, respectively. R_{mv}, $R_{\mathrm{mode_pr}}$ and $R_{\mathrm{mode_pa}}$ are the coding bits for coding predicted motion vectors, motion prediction mode and partition mode of one macroblock, respectively. λ_{motion} is the Lagrange multiplier. It is set to 16 in our experiments. $D_{SAD}^j(\mathrm{mode_pr}, \mathrm{mode_pa}, mv)$ is the sum absolute difference between the current macroblock and the motion compensated matching block. $j \in (1, 2, 3)$ denotes the high frequency subbands LH, HL and HH at the same level. W^j is a normalized weight assigned to show the importance of the coefficients in each subband. It is calculated by the normalized synthesis gain of each high frequency subband at the same level. For the spatial 9/7 filter, we have $W^1 = W^2 \approx 0.4$ and $W^3 \approx 0.2$.

For each macroblock, the mode with the smallest cost value is selected and the mode symbols are coded with Universal Variable Length Coding (UVLC) according to the probability distribution of each mode.

4 Experimental Results

Based on MPEG scalable video coding (SVC) reference software for wavelet ad-hoc group [8], we test our proposed mode-based motion vector coding method. Two standard sequences: City 4CIF 60Hz and Soccer 4CIF 60Hz are used for the test. In the experiments, two-level spatial transform are first applied to the original sequences, respectively. Orthogonal 9/7 filter is used for spatial transform. Then four-level temporal filtering with 5/3 filter is used for each spatial band. Macroblock size is set to 16×16 at the lowest resolution. Macroblock size for other spatial subband is scaled according to the resolution level.

Table 1. Motion Prediction and Coding schemes

Scheme	Modes			motion prediction manor
	DirL	*LPR*	*SPR*	
I	Off	Off	On	Band-by-band
II	On	Off	Off	Level-by-level
III	On	On	Off	Level-by-level
IV	On	On	On	Level-by-level
V	On	On	On	Band-by-band

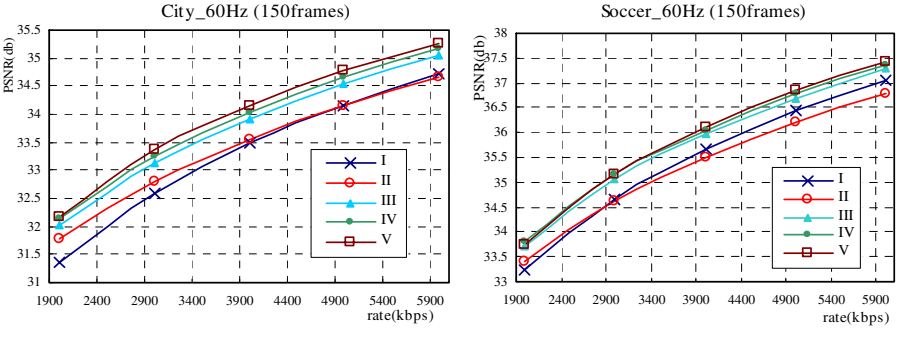


Fig. 2. The coding performance for different schemes

Five schemes are designed to test the coding performance of our proposed motion prediction and coding scheme. Table I presents the configurations of these schemes. Fig. 2 shows that the performance of these schemes. Compared with scheme I in which only SPR mode is used for motion prediction coding and scheme II in which only DirL mode is used, scheme III has achieved a better improved coding performance. The reason for this is that independent motion search in high frequency subband can not get the efficient motion vectors only with SPR mode. And more accurate motion vectors can not be obtained only with DirL mode when the coding bits increase. DirL mode incorporated with LPR modes in scheme III can get more accurate motion vectors at the same time exploiting the similarity of subband motion. Scheme IV shows that SPR mode can still continue to reduce the motion coding bits

and improve the coding efficiency based on the DirL and LPR modes. Comparing scheme IV and scheme V, we can see that the mode-based motion prediction and coding schemes with level-by-level manor and band-by-band manor have similar performance. However, the band-by-band method needs one motion search for each frequency subband. It is very time-consuming. While the level-by-level method only needs one motion search for three high frequency subbands at the same level based on the jointly rate distortion criterion in expression (3).

5 Conclusion

This paper proposed an efficient level-by-level mode-based motion prediction and coding scheme for in-band MCTF. Three motion prediction modes are presented to exploit subband motion correlation. In addition, a jointly rate-distortion criterion is proposed to tradeoff the complexity and the accuracy of block-based motion search. By the rate-distortion optimized mode selection engine, the proposed scheme can improve the coding efficiency about 0.6db for 4CIF sequence.

References

1. Chen, P., Hanke, K., Rusert, T., Woods, J. W.: Improvements to the MC-EZBC scalable video coder. Proc. IEEE International Conf. on Image Processing, Vol.2 (2003) 14-17.
2. Park, H.W., Kim, H.S.: Motion Estimation Using Low-Band-Shift Method for Wavelet-Based Moving-Picture Coding. IEEE Trans. on Image Processing, Vol. 9 (2000) 577-587
3. van der Schaar, M., Ye, J.C.: Adaptive Overcomplete Wavelet Video Coding with Spatial Transcaling. Proc. SPIE Video Communications and Image Processing (VCIP 2003) 489-500
4. Mayer, C.: Motion compensated in-band prediction for wavelet-based spatially scalable video coding. Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing, Vol 3 (2003) 73-76
5. Barbarien, J., Andreopoulos, Y., Munteanu, A., Schelkens, P., Cornelis, J.: Motion vector coding for in-band motion compensated temporal filtering. Proc. IEEE International Conf. on Image Processing, Vol. 2 (2003) 783-786
6. Xiong, R.Q., Wu, F., Li, S.P., Xiong, Z.X., Zhang, Y.Q.**:** Exploiting temporal correlation with adaptive block-size motion alignment for 3D wavelet coding. Proc. SPIE Visual Communications and Image Processing, (2004) 144-155
7. Jin, X., Sun, X.Y., Wu, F., Zhu, G.X., Li, S.P.: H.264 compatible spatially scalable video coding with in-band prediction. Proc. IEEE International Conf. on Image Processing (2005), to be appear.
8. Xiong, R.Q., Ji, X.Y., Zhang, D.D., Xu, J.Z., Maria Trocan, G.P., Bottreau, V.: Vidwav Wavelet Video Coding Specifications. Int. Standards Org./Int.Electrotech. Comm.(ISO/IEC) ISO/IEC JTC1/SC29/WG11 Document M12339, (2005)

Adaptive Sampling for Monte Carlo Global Illumination Using Tsallis Entropy

Qing Xu[1], Shiqiang Bao[1], Rui Zhang[1], Ruijuan Hu[1], and Mateu Sbert[2]

[1] Department of Computer Science and Technology, Tianjin University, China
qingxu@tju.edu.cn
[2] Institute of Informatics and Applications, University of Girona, Spain

Abstract. Adaptive sampling is an interesting tool to eliminate noise, which is one of the main problems of Monte Carlo global illumination algorithms. We investigate the Tsallis entropy to do adaptive sampling. Implementation results show that adaptive sampling based on Tsallis entropy consistently outperforms the counterpart based on Shannon entropy.

1 Introduction

Monte Carlo methods are quite usable for the global illumination when highly complex scenes with very general and difficult reflection models are rendered. Especially, Monte Carlo is applied as a last method when all other analytical or numerical methods fail [2]. But the Monte Carlo synthesized images are noisy.

Adaptive image sampling tries to use more samples in the difficult regions of the image where the sample values vary obviously. That is, each pixel is firstly sampled at a low density, and then more samples are taken for the complex part based on the initial sample values.

Many related works are within the adaptive sampling. Based on the RMS SNR, Dippe and Wold proposed an error estimate for the stopping condition [5]. Mitchell utilized the contrast to measure pixel quality [11]. Lee et al. sampled pixel based on the variance of sample values [10]. Purgathofer proposed the use of confidence interval [14]. Tamstorf and Jensen refined Purgathofers approach by using tone operator [25]. Kirk and Arvo demonstrated a correction scheme to avoid the bias [8]. Bolin and Meyer got a perceptually based method using human vision models [4]. Rigau, Feixas and Sbert introduced Shannon entropy and f-divergences to conduct adaptive sampling [16], [17], [18], [19]. We concentrate on the pixel based adaptive sampling for path tracing to take more samples for pixels with low qualities based on the examination of the sample values within a pixel. Progressive rendering by Painter, Guo, Scheel, Farrugia, etc [6], [7], [13], [21] is valuable, but it is not our topic.

We explore Tsallis entropy by means of getting its available entropic indices to perform adaptive sampling in the context of stochastic ray tracing. Experimental results indicate that adaptive sampling based on Tsallis entropy with appropriate indices can accomplish better than classic Shannon entropy based method.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 989–994, 2005.
© Springer-Verlag Berlin Heidelberg 2005

This paper is the organized as follows. Section 2 is the description of Tsallis entropy. Details of the Tsallis entropy based adaptive sampling are depicted in section 3. The fourth portion discusses the implementation and experimental results. Finally, conclusion and future works are presented in last section.

2 Tsallis Entropy

Let $X = \{x_1, \ldots, x_n\}$ be a discrete random variable with n different values, and let $P = \{p_1, \ldots, p_n\}$ be the corresponding probability distribution, and $n = |x|$. Many general entropy measures such as Tsallis entropy [9], trigonometric measures [22], polynomial B-entropy [20] and Renyi entropy [15], have been proposed. Tsallis entropy is selected in this paper as it is widely used in many practical applications [23]. Tsallis entropy is defined as [26]:

$$H_q^T = \frac{1}{q-1}(1 - \sum_{i=1}^{n} p_i^q) \tag{1}$$

Here q is called as entropic index, and $q \neq 1$ .When $q \to 1$, using L'Hopital's rule, Tsallis entropy tends towards Shannon entropy. $H_q^T(X)$ is a concave function of p for $q > 0$. $H_q^T(X)$ takes its maximum value for the equiprobability: $H_{q,max}^T = \log_q n$. Here $\log_q x = \frac{1}{1-q}(x^{1-q} - 1)(x > 0)$ stands for the q-logarithmic function tending to the ordinary logarithmic function in the limit $q \to 1$.

3 Adaptive Sampling Based on Tsallis Entropy

The key to adaptive sampling is to evaluate pixel quality appropriately. Since entropy is a fine measure of information, it can be exploited to be as the measure of pixel homogeneity that is a good indicator for the pixel quality. Actually, we can take advantage of entropy to measure homogeneity of the information provided by the set of ray samples passing through a pixel.

We use RGB color space to describe sample and pixel value. According to the definition of Tsallis measure, the pixel Tsallis entropy for each RGB channel is defined as follows:

$$H_k^T = \frac{1}{q-1}(1 - \sum_{i=1}^{n} p_{k,j}^q) \tag{2}$$

Here k is the index of spectral channel, n is the number of rays through the pixel, and $p_{k,j}$ refers to the channel fraction of a ray with respect to the sum of the values of the same channel of all the rays passing through the pixel. The pixel entropy defined here describes the homogeneity or quality of the pixel. Because pixel value may be estimated from any number of passing rays in the course of rendering, the pixel value entropy should be normalized with its maximum for the sake of scoring the quality of the pixel correctly. The pixel Tsallis quality for each spectral channel is defined as:

$$Q_k^T = \frac{H_k^T}{H_{k,max}^T} \tag{3}$$

As a result, the pixel value quality defined with Tsallis entropy is

$$Q^T = \frac{\sum_{k=1}^{n_s} w_k Q_k^T \overline{s_k}}{\sum_{k=1}^{n_s} w_k} \tag{4}$$

Here n_s is the number of spectral channels, w_k is the weight coefficient for RGB channel and set to 0.4, 0.3 and 0.6 based on the relative sensitivity of the visual system [11], and $\overline{s_k}$ is the average of values of all the pixel rays for each channel.

4 Implementation and Results

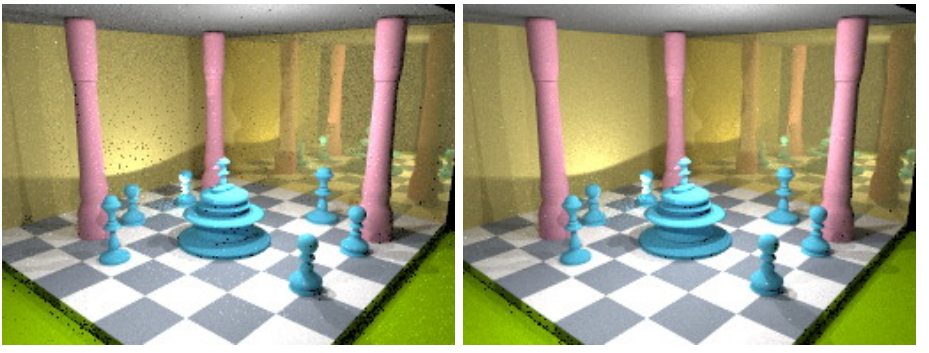
We have fulfilled the adaptive sampling scheme developed in this paper for a pure stochastic ray tracing algorithm constructed on a global illumination renderer, pbrt [12]. All the experimental results are obtained by running algorithms on a P4 1.8 GHz machine with 256 MB RAM.

Since the appropriate selection of the entropic index to Tsallis entropy for practical applications is an open issue [1], [24], [27] experiments are carried out to set entropic indices manually to do adaptive sampling for the different test scenes. Table 1 shows the results using different indices for the test scene 1, "none" means that the image cannot be obtained with the given index around the required average samples per pixel. The images are produced with a large (302) and a small (51) average samples per pixel, for a high quality image and a little noisy image. We have found empirically that the Tsallis method with entropic index within a large range, from 1.5 to 3.0, can do better than Shannon approach. The two images spending 450s and 404s produced by using Shannon and Tsallis based approaches with entropic index 3.0 at average 51 and 51 samples per pixel are shown in Figure 1 . An overall superior image quality resulted from the new method is evident. The new method is more discriminating and sensitive, and behaves better in difficult parts.

Table 1. Different entropic indices and the corresponding RMS errors of the produced images with different avg. number of samples per pixel for the test scene 1

RMS	Shannon	Tsallis entropy						
	entropy	Entropic index q						
avg		0.0001	0.001	0.01	2.0	3.0	4.0	6.0
51	7.594	none	7.767	7.612	7.141	6.223	9.537	none
302	5.893	5.932	5.94	5.945	5.815	5.754	none	none

Table 2 gives different entropic indices and the corresponding RMS errors of the images for the test scene 2. The two images spending 2070s and 2074s are generated with a large (526) and a small (70) average samples per pixel. It can be derived experimentally that the Tsallis method with entropic index within a large range, from 1.5 to 3.0, can achieve better than Shannon method.

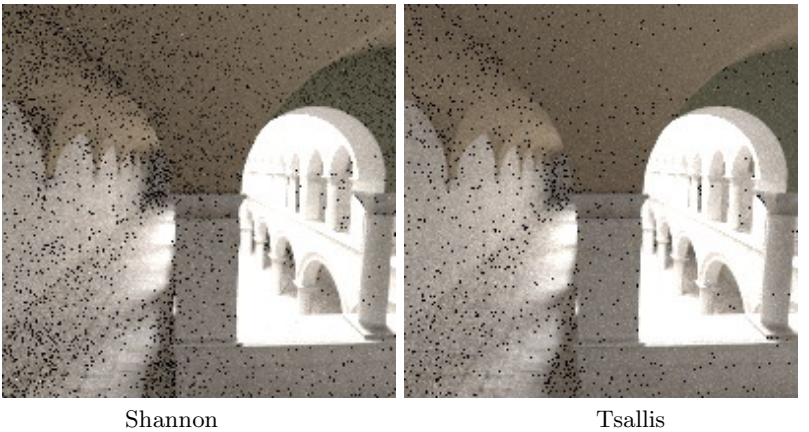


Shannon Tsallis

Fig. 1. Resultant images for the test scene 1

Table 2. Different entropic indices and the corresponding RMS errors of the produced images with different avg. number of samples per pixel for the test scene 2

RMS	Shannon	Tsallis entropy						
	entropy	Entropic index q						
avg		0.0001	0.001	0.01	2.0	3.0	4.0	6.0
70	2.688	2.478	2.404	2.379	2.077	1.584	1.571	none
526	0.126	0.138	0.138	0.137	0.099	0.087	none	none



Shannon Tsallis

Fig. 2. Resultant images for the test scene 2

The images generated by using Shannon and Tsallis based approaches with entropic index 3.0 at average 70 and 70 samples per pixel are compared in Figure 2. The numbers of initially testing samples and added samples in each refinement step are 9 and 9. Our method can reduce the noisy spots more and lead to a more homogeneous image.

5 Conclusion and Future Work

We have presented a new adaptive sampling scheme based on Tsallis entropy with experimentally selected entropic indices. The results obtained by our method show an improvement on error reduction over the Shannon entropy based approach. We will research on choosing entropic indices automatically or systematically.

References

1. Portes, M., Esquef, I.A., Gesualdi, A.R.: Image thresholding using Tsallis entropy. Pattern Recognition Letters. 25(2004)1059-1065
2. Bekaert, P.: Hierarchical and Stochastic Algorithms for Radiosity. Ph.D. Dissertation, Katholieke Universiteit Leuven, December 1999.
3. Blahut, R.E.: Principles and Practice of Information Theory. Addison-Wesley, Boston(1987)
4. Bolin, M.R., and Meyer, G.W.: A perceptually based adaptive sampling algorithm. In: M. Cohen(eds.): SIGGRAPH 98 Conference Proceedings. Orlando, FL, USA(1998)299-310
5. Dippe, M.A.Z., and Wold, E.H.: Antialiasing through Stochastic Sampling. Computer Graphics. 19(1985)69-78
6. Philippe, J., and Peroche, B.: A Progressive Rendering Algorithm Using an Adaptive Perceptually Based Image Metric. In: Cani, M.-P. and M. Slater(eds.): Proceedings of Eurographics'2004, INRIA and Eurographics Association(2004)
7. Guo, B.: Progressive Radiance Evaluation using Directional Coherence Maps. In: Cohen, M.(eds.): SIGGRAPH 98 Conference Proceedings. Orlando, FL, USA(1998)255-266
8. Kirk, D., and Arvo, J.: Unbiased variance reduction for global illumination. In:Brunet, P., Jansen, F.W(eds.):Proceedings of the 2nd Eurographics Workshop on Rendering. Barcelona(1991)153-156
9. Kapur, J.N., Kesavan, H.K.: Entropy Optimization Principles with Applications. Academic Press, New York(1992)
10. Lee, M.E., Redner, R. A., and Uselton, S.P.: Statistically Optimized Sampling for Distributed Ray Tracing. Computer Graphics. 19(1985)61-65
11. Mitchell, D.P.: Generating Antialiased Images at Low Sampling Densities. Computer Graphics. 21(1987)65-72
12. Pharr, M., and Humphreys, G.: Physically Based Rendering : From Theory to Implementation. Morgan Kaufmann, San Fransisco(2004)
13. Painter, J.,and Sloan, K.: Antialiased Ray Tracing by Adaptive Progressive Refinement. Computer Graphics. 23(1989)281-288
14. Purgathofer, W.: A Statistical Method for Adaptive Stochastic Sampling. Computers Graphics. 11(1987)157-162
15. Renyi, A.: On measuresof entropy and information. In Selected Papers of Alfred Renyi. 2(1976)525-580
16. Rigau, J., Feixas, M., and Sbert, M.: Entropy-based Adaptive Supersampling. In: Debevec, P.and Gibson, S.(eds.): Proceedings of Thirteenth Eurographics Workshop on Rendering. Pisa, Italy, June 2002
17. Rigau, J., Feixas, M., and Sbert, M.: New Contrast Measures for Pixel Supersampling. Springer-Verlag London Limited, London(2002)439-451

18. Rigau, J., Feixas, M., and Sbert, M.: Entropy-based Adaptive Sampling. In: Proceedings of CGI'03,IEEE Computer Society Press, Tokyo(2003)
19. Rigau, J., Feixas, M., and Sbert,M.: Refinement Criteria Based on f-Divergences. In: Proceedings of Eurographics Symposium on Rendering 2003.Eurographics Association(2003)
20. Sharma, B.D., Autar, R.: An inversion theorem and generalized entropies for continuous distributions. SIAM J.Appl.Math. 25 (1973) 125-132
21. Scheel, A., Stamminger, M., Putz, J., and Seidel, H.: Enhancements to Directional Coherence Maps. In :Skala, V.(eds):WSCG 01(Proceedings of Ninth International Conference in Central Europeon Computer Graphics and Visualization. Plzen, Czech Republic,05-09 February 2001.).Plzen(2001)
22. Santanna, A.P., Taneja, I.J.: Trigonometric entropies, Jensen difference divergence measures, and error bounds. Inf. Sci. 35 (1985) 145-156
23. Smolikova, R., Wachowiak, M.P., Zurada, J.M.: An information-theoretic approach to estimating ultrasound backscatter characteristics. Computers in Biology and Medicine. 34 (2004)355-370
24. Tsallis, C., Albuquerque, M. P.: Are citations of scientific paper a case of nonextensivity. Euro. Phys. J. B 13, 777-780
25. Tamstorf, R., and Jensen, H. W.: Adaptive Sampling and Bias Estimation in Path Tracing. In: J. Dorsey and Ph. Slusallek.(eds.): Rendering Techniques '97. Springer-Verlag, (1997) 285-295
26. Tsallis, C.: Possible generalization of Boltzmann-Gibbls statistics. Journal of Statistical Physics. 52(1988)480-487
27. Tatsuaki, W., Takeshi, S.: When nonextensive entropy becomes extensive. Physica A 301, 284-290.

Incremental Fuzzy Decision Tree-Based Network Forensic System

Zaiqiang Liu and Dengguo Feng

State Key Laboratory of Information Security,
Institute of Software, Chinese Academy of Sciences,
Beijing 100080, China
{liuzq, feng}@is.iscas.ac.cn

Abstract. Network forensic plays an important role in the modern network environment for computer security, but it has become a time-consuming and daunting task due to the sheer amount of data involved. This paper proposes a new method for constructing incremental fuzzy decision trees based on network service type to reduce the human intervention and time-cost, and to improve the comprehensibility of the results. At the end of paper, we discuss the performance of the forensic system and present the result of experiments.

1 Introduction

Network forensic is the act of capturing, recording and analyzing network audit trails in order to discover the source of security breaches or other information assurance problems[1]. The biggest challenge in conducting network forensics is the sheer amount of data generated by the network[2]; The comprehensibility of evidence that is extracted from collected data is also an important aspect for forensic experts. Beside these, it is also impossible to collect enough clean training data sets before the process of investigation, so we need an effective method to update knowledge bases. Incremental fuzzy decision tree based-data mining technology is an efficient way to resolve the above problems by creating a classification model through extracting key features from network traffic by providing the resulting fuzzy decision tree with better noise immunity and increasing applicability in uncertain or inexact contexts while proliferating much praised comprehensibility of the resulting knowledge[3]. In this paper, we develop an incremental fuzzy decision tree based system for network forensics system(IFDTNFS) that can analyze computer crime in networked environments, and collect digital evidences automatically, and update rule base incrementally.

The remainder of the paper is organized as follows: Section 2 describes the proposed Incremental Fuzzy Decision Tree-based system for network forensics. Section 3 explains the experiment data set which is used in this paper and shows the experiment results. Section 4 discusses related work in network forensics and fuzzy decision tree system. Finally, a discussion of conclusions and further issues in network forensics are given in section 5.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 995–1002, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Incremental Fuzzy Decision Tree-Based Network Forensic System

We develop a Network Forensic System based on Incremental Fuzzy Decision Tree technology (IFDTNFS). IFDTNFS consists of three components: Network Traffic Separator, Traffic Detector, Forensic Analyzer. Figure 1 shows the architecture of the proposed system.

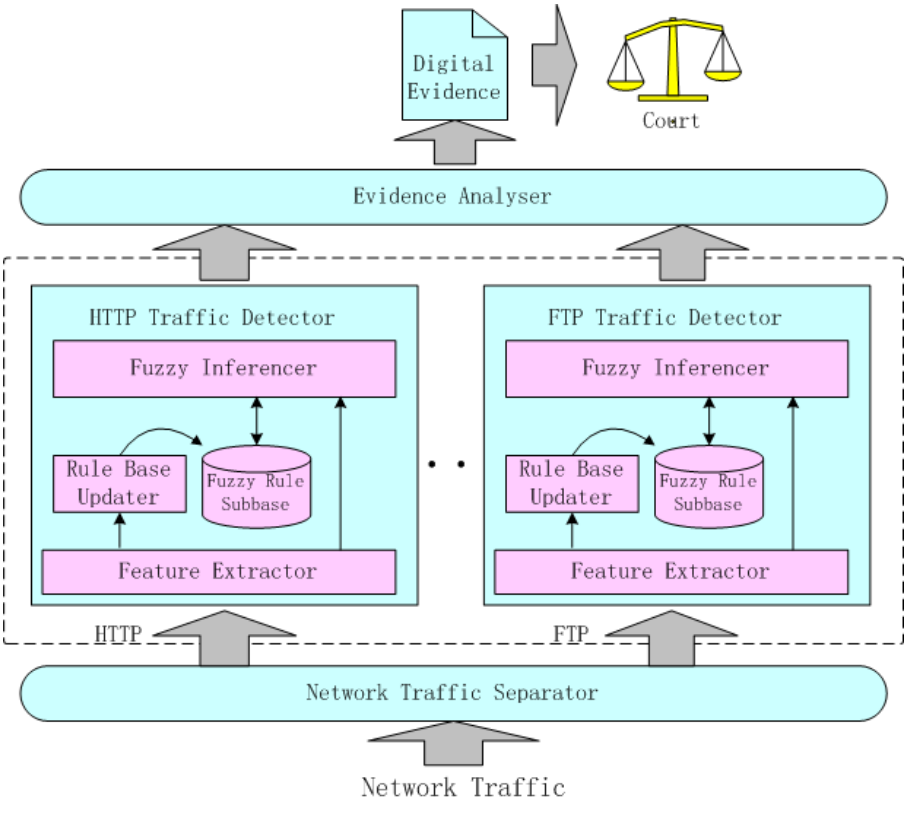


Fig. 1. IFDTNFS Forensic System

2.1 Network Traffic Separator

The Network Traffic Separator component is responsible for capturing network traffic and separating the network traffic according to the service type, and directing the separated traffic to corresponding traffic detector. The process of traffic capture is the first step of the proposed forensic system. While the capturing and separating function is simple and straightforward, it provides the data sources needed to be analyzed by other components of the forensic system. For improving the effectiveness of the system, the network forensic system should maintain a complete record of network traffic and consider packet capture rate.

Currently the traffic separator is based on the modified packet capture program-TcpDump [4]. In order to reduce the system burden and the rate loss of capturing packet, users can turn the traffic separator into a filter and integrate it with the corresponding detector respectively.

2.2 Traffic Detector

The Traffic Detector component is the core component of the IFDTNFS system, and consists of four components: Feature Extractor, Fuzzy Rule Base, Rule Base Updater, Fuzzy Inferencer.

Feature Extractor. Feature Extractor extracts features from the network traffic captured by Traffic Separator component. Feature extraction and selection from the available data is one of the most important steps to the effectiveness of the methods employed. The Feature Extractor uses the method in [5] to extract 41 different features consisting of essential attributes, secondary attributes and calculated attributes. More detail information, please refer to [5][6][7][8].

Fuzzy Rule Bases. Fuzzy Rule Bases are also the knowledge bases of the IFDTNFS system. For reducing the size of rules and improving the efficiency of inference, we create independent subrule bases for different service type traffic. The process of building fuzzy rule bases is also the process of building fuzzy decision trees, which includes the following steps:

- Step 1: Assume all the training samples $E = \{e_1, e_2, \cdots, e_i, \cdots, e_N\}$ (N denotes the count of training samples). According to the service type attribute of the training data set, we first use the traffic separator component to divide the training data set into 66 subsets, and then use the following algorithm to construct the fuzzy decision tree $T = t_1, t_2 \cdots, t_i \cdots, t_W$(fuzzy rule base) respectively (W denotes the count of service types). For easy description, we regard all of the fuzzy decision tree as a whole tree T and the fuzzy decision tree of each service type as a subtree t_i.
- Step 2: Sort the training samples according to the values of the chosen attribute at the given node under the sub-tree t_i, and generate an ordered sequence.
- Step 3: Calculate the cut points C using the method in [9].
- Step 4: Calculate the information gain of the current node v,

$$G^v_{A_q} = I(A_q; E^v, t_i) - I(A_q, C; E^v, t_i),$$

where

$$I(A_q; E^v, t_i) = -\sum_{j=1}^{k} p(c_j; E^v, t_i) \log p(c_j; E^v, t_i)$$

$$I(A_q, C; E^v, t_i) = \sum_{j=1}^{m+1} \frac{|E^v_j|}{|E^v|} I(E^v_j; t_i),$$

$$I(E_j^v; t_i) = -\sum_{j=1}^{k} p(c_j; E_j^v, t_i) \log p(c_j; E_j^v, t_i)$$

Where E_j^v denotes the sample subset of the jth child node of node v , E^v denotes the sample of node v, k is the count of class, m denotes the count of cut points.

- Step 5: Repeat step.2-step.4 to generate the information gain of other attributes.
- Step 6: Select the attribute A_q , which has the maximum value of $G_{A_q}^v$ to generate child nodes.
- Step 7: Repeat step.2-step.6 until:
  - All data belongs to the same class, or
  - The proportion of a data set of a class is greater than or equal to a given threshold, or
  - The number of elements in a data set is less than a given threshold, or
  - There are no more attributes for classification.
- Step 8: Repeat step.2-step.7 until all the sub-trees are built.

Rule Base Updater. The Rule Base Updater consists of two components: SampleAdder, Adjuster. SampleAdder component is responsible for adding new samples to the fuzzy decision tree that has been constructed using the method in the Fuzzy Rule Bases section. The function of the SampleAdder is similar to that of the AddExample in [10]. Usually calling SamplerAdder component will cause two aspects of result: one aspect is that class distribution may change; another is the structure of tree may change[10]. In the algorithm, the SampleAdder does not change the structure of the tree, but only changes the memorized values that are used for recalculating the information gain. Certainly, the decision tree may grow due to the new information.

With the adding of lots of new samples, the decision tree will become bigger and bigger, and even worse is that the newly constructed decision tree will not satisfy the need for optimization. Therefore we design an Adjuster component to adjust the new constructed tree in the course of adding new samples to make it optimal or nearly optimal. The SampleAdder will call the Adjuster to optimize the decision tree if the last adjusting was at least "w" samples ago. For simplicity, the system uses the subtree replacement for postpruning the decision tree.

Fuzzy Inferencer. The Fuzzy Inferencer functions as a fuzzy preprocessor and a fuzzy decision maker. For improving the efficiency of inference, we adopt a top-down strategy that searches for a solution in a part of the search space to construct the fuzzy decision tree. It guarantees that a simple tree will be found. There exists two different kinds of domains for features extracted by the Feature Extractor: continuous and discrete. Each input variable's crisp value needs to be first fuzzified into linguistic values before the Fuzzy Inferencer or the Fuzzy Decision-maker processes them with the Rule Base. The Fuzzy Inferencer uses two different ways to fuzzify the continuous and the discrete domains respectively. For the discrete features, the Fuzzy Inferencer component uses the same

method as the classical set. For continuous features, we use the trapezoidal function as their membership function.

To decide the classification assigned to a test sample, we have to find leaves from the fuzzy decision tree (that is, Fuzzy Rule Base) whose restrictions are satisfied by the sample, and combine their decisions into a single crisp response. For simplicity the Fuzzy Inferencer uses the Center-of-Maximum method in [11].

2.3 Evidence Analyzer

The main functions of the Forensic Analyzer include: collecting relative event data, analyzing correlated information relating with the event, and establishing digital evidences. The component receives the inference results of the Fuzzy Network Detectors, and decides the level and range of collecting data for network forensic. For example, if the output value of the Fuzzy Inferencer is between 0.7 and 1.0, then it decides that an attack has occurred and automatically collects information from current network connection and system logs to establish digital evidences. Currently the Evidence Analyzer component is still being developed.

3 Experiments and Results

The data for our experiments was prepared by the 1998 DARPA intrusion detection evaluation program by MIT Lincoln Labs.The following experiment is based on the 10% training data subset with 494,021 data records. There are lots of ways to evaluate the performance and efficiency of a classifier. Usually TP Rate (True Positive Rate) and FP Rate (False Positive Rate) are employed. In our experiment we also use F-Measure measurements to characterize the performance of IFDTNFS system.

Table 1 shows the result of IFDTNFS using the dataset with different measurements. From Table 1 we can see that the correctly classified instance rate of IFDTNFS reaches 91.62% on average(Actually the true positive rate will be even better due to the uneven distribution of the training dataset). But just like other classifiers, the performance of the IFDTNFS still depends on the quality of training samples to some degree, so we choose the ten-fold cross validation evaluation method to test the IFDTNFS system.

Figure 2 shows the computation time in seconds needed to learn from the examples over the number of training samples. From the diagram we can see that the method proposed in the paper is less cost-consuming both in the constructing phase of the decision tree and in the incremental phase.

4 Related Works

The term "Network Forensics" was introduced by the computer security expert Marcus Ranum in the early 90's [12], and is borrowed from the legal and criminology field where "forensics" pertains to the investigation of crimes. In [13], network forensic was well defined. Network forensic systems are designed

Table 1. Experiment Result with ten-fold cross validation evaluation method

Class	TP rate	FP rate	Precision	Recall	F-Measure
back	1	0	1	1	1
buffer_overflow	0.75	0.009	0.667	0.75	0.706
ftp_write	0.75	0.003	0.857	0.75	0.8
guess_passwd	0.964	0.024	0.883	0.964	0.922
imap	0.636	0.003	0.875	0.636	0.737
ipsweep	1	0	1	1	1
land	0.833	0	1	0.833	0.909
loadmodule	0.333	0.006	0.5	0.333	0.4
multihop	0.786	0.009	0.786	0.786	0.786
neptune	0.941	0.003	0.941	0.941	0.941
nmap	0.947	0.003	0.947	0.947	0.947
normal	1	0	1	1	1
perl	0.667	0	1	0.667	0.8
phf	1	0.003	0.8	1	0.889
pod	1	0.003	0.8	1	0.889
portsweep	1	0	1	1	1
rootkit	0.5	0	1	0.5	0.667
satan	1	0	1	1	1
smurf	1	0	1	1	1
spy	0	0	0	0	0
teardrop	1	0.003	0.75	1	0.857
warezclient	1	0.006	0.933	1	0.966
warezmaster	0.875	0.016	0.808	0.875	0.84

to identify unauthorized use, misuse, and attacks on information. Usually, network forensics which is based on audit trails is a difficult and time-consuming process. Recently artificial intelligence technologies, such as artificial neural network (ANN) and support vector machine (SVM) [1], were developed to extract significant features for network forensic to automate and simplify the process. These techniques are effective in reducing the computing-time and increasing the intrusion detection accuracy to a certain extent, but they are far away from being perfect. Particularly, these systems are complex, and the results produced by these methods lack good comprehensibility. Besides this, fuzzy expert systems have also been proposed for network forensic [14], but it still requires experts to build knowledge bases and they lack the capability of self-learning. Incre-

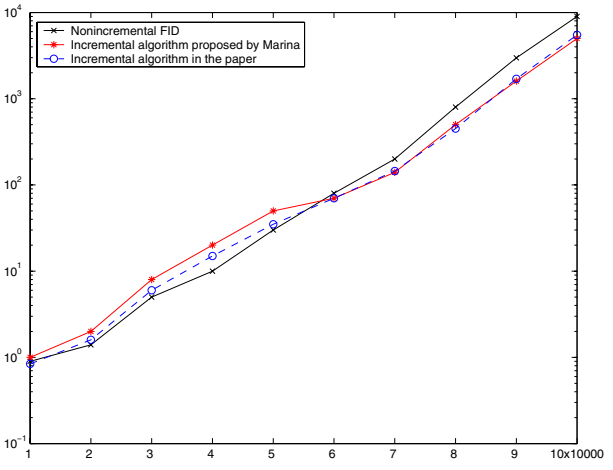


Fig. 2. Time behavior comparison between the nonincremental algorithm, the incremental method in [10] and the methods proposed in the paper

mental Fuzzy decision tree-based forensic system as proposed in this paper, can effectively solve the above problems while keeping higher detection accuracy.

5 Conclusions and Future Works

In this paper, we developed an automated network forensic system which can produce interpretable and accurate results for forensic experts by applying a fuzzy logic based decision tree data mining system. The experiment proves that the IFDTNFS system can effectively analyze the network traffic under the current high-speed network environment. Our ongoing and future tasks will focus on the correlation and semantic analysis of evidence.

References

1. Srinivas Mukkamala, Andrew H.Sung.: Identifying Significant Features for Network Forensic Analysis Using Artificial Intelligent Techniques. International Journal of Digital Evidence, Winter 2003, Volume 1, Issue 4(2003)
2. Jay Heiser: Must-haves for your network forensic toolbox, 11 Feb 2004. http://searchsecurity.techtarget.com/tip/
3. C. Z. Janikow.: Fuzzy decision trees: Issues and methods. IEEE Transactions on Systems, Man and Cybernetics, 28(1)(1998)1-14
4. http://www.tcpdump.org
5. Salvatore J. Stolfo, Wei Fan, Wenke Lee, etc.: Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project.
6. Carbone,P. L.: Data mining or knowledge discovery in databases: An overview. In Data Management Handbook. New York: Auerbach Publications(1997)

7. Lee, W. and S. J. Stolfo.: Data mining approaches for intrusion detection.In Proc. of the 7th USENIX Security Symp., San Antonio, TX. USENIX(1998)
8. Lee, W., S. J. Stolfo, and K. W. Mok.: Mining in a data-flow environment:Experience in network intrusion detection. In S. Chaudhuri and D. Madigan (Eds.),Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA, ACM (1999)pp.114-124
9. U.M. Fayyad and K.B. Irani.: Multi-interval discretization of continuous valued attributes for classification learning. In Proc of the 13th IJCAI, France(1993)1022-1027
10. Marina Guetova, Steffen Holldobler, Hans-Peter Storr.: Incremental Fuzzy Decision Trees, 25th German Conference on Artificial Intelligence (KI2002)
11. Zimmermann, H. J.: Fuzzy Set Theory and Its Applications, Kluwer Academic Publishers(1996)
12. Marcus Ranum.: Network Flight Recorder, http://www.ranum.com/
13. Digital Forensic Research Workshop.: A Road Map for Digital Forensic Research,(2001)
14. Jung-Sun Kim, Minsoo Kim, and Bong-Nam Noh.: A Fuzzy Expert System for Network Forensics, ICCSA 2004, LNCS 3043(2004)175-182

Robust Reliable Control for a Class of Fuzzy Dynamic Systems with Time-Varying Delay

Youqing Wang and Donghua Zhou*

Department of Automation, Tsinghua University,
Beijing 100084, China
yq-wang03@mails.tsinghua.edu.cn

Abstract. This paper deals with the problem of robust reliable control design for a class of fuzzy uncertain systems with time-varying delay. The system under consideration is more general than those in other existent works. A reliable fuzzy control design scheme via state feedback is proposed in terms of linear matrix inequality (LMI). The asymptotic stability of the closed-loop system is achieved for all admissible uncertainties as well as actuator faults. A numerical example is presented for illustration.

1 Introduction

Since proposed by Zadeh [1], fuzzy logic control has been developed into a conspicuous and successful branch of automation and control theory. In 1985, Tankagi and Sugeno proposed a design and analysis method for overall fuzzy systems, in which the qualitative knowledge of a system was first represented by a set of local linear dynamics [2]. This allows the designers to take advantage of conventional linear systems to analyze and design the nonlinear systems [3]-[4]. Due to the universal existence of model uncertainties in practice, robust fuzzy control of uncertain systems has received much more attention in recent years [5]-[6]. Though T-S fuzzy model with uncertainties can represent a large class of nonlinear systems, however, even more nonlinear systems can not be described in this form.

Owing to the increasing demand for high reliability of many industrial processes, reliable control is becoming an ever increasingly important area [7]-[9]. In this paper, we make use of the system formulation in [8] as the local dynamic model. Obviously, this formulation is more general than those in other existent works. The proposed design method is based on the result in [6]. And it is proved that our controller can work for more general systems than the state feedback control proposed in [6].

2 Problem Formulation

The continuous fuzzy dynamic model, proposed by Takagi and Sugeno, is presented by fuzzy IF-THEN rules. Consider an uncertain nonlinear system with time delay described by the following T-S fuzzy model with time delay.

* Senior Member (IEEE).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1003 – 1010, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Plant Rule i:
IF $\theta_j(t)$ is N_{ij} for $j=1,2,\cdots,p$, THEN

$$\begin{aligned}&\dot{x}(t)=(A_i+\Delta A_i)x(t)+(A_{1i}+\Delta A_{1i})x(t-\sigma(t))+(B_i+\Delta B_i)u(t)+D_i f_i(x(t)),\\&x=\varphi(t),\quad t\in[-\sigma_0,0],\quad i=1,2,\cdots,k,\end{aligned}\tag{1}$$

where N_{ij} are the fuzzy sets, $x(t)\in R^n$ is the state vector, $u(t)\in R^m$ is the control input, $f_i(\cdot):R^n\to R^{n_f}$ is an unknown nonlinear function ($\forall i=1,\cdots k$). $A_i, A_{1i}\in R^{n\times n}$, $B_i\in R^{n\times m}$, $D_i\in R^{n\times n_f}$. Scalar k is the number of IF-THEN rules; and $\theta_1(t),\cdots,\theta_p(t)$ are the premise variables. It is assumed that the premise variables do not depend on the input $u(t)$.

Assumption 1. The real valued functional $\sigma(t)$ is the time-varying delay in the state and satisfies $\sigma(t)\le\sigma_0$, a real positive constant representing the upper bound of the time-varying delay. It is further assumed that $\dot{\sigma}(t)\le\beta<1$ and β is a known constant.

Assumption 2. Matrices ΔA_i, ΔA_{1i} and ΔB_i denotes the uncertainties in system (1) and take the form of

$$\begin{bmatrix}\Delta A_i & \Delta A_{1i} & \Delta B_i\end{bmatrix}=MF(t)\begin{bmatrix}E_i & E_{di} & E_{bi}\end{bmatrix}\tag{2}$$

where M, E_i, E_{di} and E_{bi} are known constant matrices and $F(t)$ is an unknown matrix function satisfying

$$F^T(t)F(t)\le I\tag{3}$$

Given a pair of $(x(t),u(t))$, the final output of the fuzzy system is inferred as follows:

$$\dot{x}(t)=\sum_{i=1}^{k}h_i(\theta(t))[(A_i+\Delta A_i)x(t)+(A_{1i}+\Delta A_{1i})x(t-\sigma(t))+(B_i+\Delta B_i)u(t)+D_i f_i(x(t))]\tag{4}$$

where $h_i(\theta(t))=\mu_i(\theta(t))\Big/\sum_{i=1}^{k}\mu_i(\theta(t))$, $\mu_i(\theta(t))=\prod_{j=1}^{p}N_{ij}(\theta_j(t))$ and $N_{ij}(\theta_j(t))$ is the degree of the membership of $\theta_j(t)$ in N_{ij}. In this paper, we assume that $\mu_i(\theta(t))\ge 0$ for $i=1,2,\cdots,k$ and $\sum_{i=1}^{k}\mu_i(\theta(t))>0$ for all t. Therefore, $h_i(\theta(t))\ge 0$ for $i=1,2,\cdots,k$ and $\sum_{i=1}^{k}h_i(\theta(t))=1$.

Assumption 3. There exist some known real constant matrixes G_i such that the unknown nonlinear vector functions $f_i(\cdot)$ satisfy the following boundedness conditions:

$$\left\|f_i(x(t))\right\|\le\left\|G_i x(t)\right\|\tag{5}$$

for any $x(t)\in R^n$.

Remark 1. The formulation (1) can describe a larger class of systems than that in [6].
The actuator fault model is described as follow.

$$u_\omega(t) = \Phi_\omega u(t) \tag{6}$$

where

$$\Phi_\omega = diag[\delta_\omega(1), \delta_\omega(2), \cdots, \delta_\omega(m)], \quad \delta_\omega(i) \in [0,1], \quad i = 1,2,\cdots,m. \tag{7}$$

Matrix Φ_ω describes the fault extent. $\delta_\omega(i) = 0$ means that the ith system actuator is invalid, $\delta_\omega(i) \in (0,1)$ implies that the ith system actuator is at fault in some extent and $\delta_\omega(i) = 1$ denotes that the ith system actuator operates properly. For a given diagonal matrix Φ_Ω, the set $\Omega = \{u_\omega = \Phi_\omega u, \, and \, \Phi_\omega \geq \Phi_\Omega\}$ is named an admissible set of actuator fault. Our design object is to design a fuzzy reliable control law $u(t)$ such that system (1) is asymptotically stable not only when all controller components operate well, but also in the case of any fault $u_\omega \in \Omega$ occurring.

3 Robust Reliable Controller Design

In this section, we present the design of fuzzy reliable controllers. Suppose the following fuzzy controller is used to deal with the fuzzy control system (1).

Control rule i*:*

IF $\theta_j(t)$ is N_{ij} for $j = 1,2,\cdots,p$, THEN

$$u(t) = K_i x(t), \quad i = 1,2,\cdots,k \tag{8}$$

Then, the overall fuzzy controller is given by

$$u(t) = \sum_{i=1}^{k} h_i(\theta(t)) K_i x(t) \tag{9}$$

where K_i ($i = 1,2,\cdots,k$) are the control gains. The design goal of this paper is to determine the feedback gains K_i ($i = 1,2,\cdots,k$) such that the resulting closed-loop system is asymptotically stable even when actuator failure $u_\omega \in \Omega$ occurs.

For the case of $u_\omega \in \Omega$, the closed-loop system is given by

$$\begin{aligned} \dot{x}(t) = & \sum_{i=1}^{k}\sum_{j=1}^{k} h_i(\theta) h_j(\theta)[(A_i + \Delta A_i)x(t) + (A_{1i} + \Delta A_{1i})x(t - \sigma(t)) \\ & + (B_i + \Delta B_i)\Phi_\omega K_j x(t) + D_i f_i(x(t))], \\ x = & \varphi(t), \quad t \in [-\sigma_0, 0], \quad i = 1,2,\cdots,k, \end{aligned} \tag{10}$$

Theorem 1. Consider system (2). Assume there exist matrices $X > 0, T > 0$ and scalars $\lambda_i, \varepsilon_i > 0$ and $\varepsilon_{ij} > 0$ such that the following LMIs are satisfied**:**

$$\begin{bmatrix} A_i X + XA_i^T + \varepsilon_i MM^T + \lambda_i D_i D_i^T - B_i \Phi_\Omega B_i^T & A_{1i}T & 0 & XE_i^T & X & XG_i^T \\ * & -T & 0 & TE_{di}^T & 0 & 0 \\ * & * & -I & E_{bi}^T & 0 & 0 \\ * & * & * & -\varepsilon_i I & 0 & 0 \\ * & * & * & * & -(1-\beta)T & 0 \\ * & * & * & * & * & -\lambda_i I \end{bmatrix} < 0, \quad (11)$$

$i = 1, 2, \cdots, k,$

$$\begin{bmatrix} \dfrac{A_i + A_j}{2} X + X \dfrac{(A_i + A_j)^T}{2} + \varepsilon_{ij} MM^T + \dfrac{\lambda_i D_i D_i^T + \lambda_j D_j D_j^T}{2} + (B_i - B_j)(B_i - B_j)^T - \dfrac{B_i \Phi_\Omega B_i^T}{2} - \dfrac{B_j \Phi_\Omega B_j^T}{2} & \dfrac{A_{1i} + A_{1j}}{2} T & 0 & X \dfrac{(E_i + E_j)^T}{2} & X & XG_i^T & XG_j^T \\ * & -T & 0 & T \dfrac{(E_{di} + E_{dj})^T}{2} & 0 & 0 & 0 \\ * & * & -I & \begin{bmatrix} \dfrac{E_{bi}}{2} & \dfrac{E_{bj}}{2} \end{bmatrix}^T & 0 & 0 & 0 \\ * & * & * & -\varepsilon_{ij} I & 0 & 0 & 0 \\ * & * & * & * & -(1-\beta)T & 0 & 0 \\ * & * & * & * & * & -2\lambda_i I & 0 \\ * & * & * & * & * & * & -2\lambda_j I \end{bmatrix} \leq 0, \quad (12)$$

$i, j = 1, 2, \cdots, k; \quad i \neq j$

where $*$ denotes the transposed elements in the symmetric positions. Then, the local control gains are given by

$$K_i = -B_i^T X^{-1}, i = 1, 2, \cdots, k \quad (13)$$

Proof: Consider the Lyapunov function

$$V(x,t) = x^T(t) P x(t) + \frac{1}{1-\beta} \int_{t-\sigma(t)}^{t} x^T(\alpha) S x(\alpha) d\alpha \quad (14)$$

where the weighting matrices P and S are given by $P = X^{-1}$ and $S = T^{-1}$. In the presence of the actuator failure $u_\omega \in \Omega$, the resulting closed-loop system is given by (10). Differentiating $V(x,t)$ along the trajectory of system (10) gives

$$\begin{aligned}\dot{V}(x,t) &= \dot{x}^T Px + x^T P\dot{x} + \frac{1}{1-\beta}x^T Sx - \frac{1-\dot{\sigma}(t)}{1-\beta}x^T(t-\sigma(t))Sx(t-\sigma(t)) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} h_i(\theta(t))h_j(\theta(t))x^T[P(A_i+\Delta A_i+(B_i+\Delta B_i)\Phi_\omega K_j) \\ &+(A_i+\Delta A_i+(B_i+\Delta B_i)\Phi_\omega K_j)^T P]x \\ &+\sum_{i=1}^{k} h_i(\theta(t))(x^T PD_i f_i + f_i^T D_i^T Px) \\ &+2\sum_{i=1}^{k}\sum_{j=1}^{k} h_i(\theta(t))h_j(\theta(t))x^T P(A_{1i}+\Delta A_{1i})x(t-\sigma(t)) \\ &+\frac{1}{1-\beta}x^T Sx - \frac{1-\dot{\sigma}(t)}{1-\beta}x^T(t-\sigma(t))Sx(t-\sigma(t))\end{aligned} \tag{15}$$

Note that for any $\lambda_i > 0$

$$\begin{aligned}&x^T PD_i f_i + f_i^T D_i^T Px \le \lambda_i x^T PD_i D_i^T Px + \lambda_i^{-1} f_i^T f_i \\ &\le \lambda_i x^T PD_i D_i^T Px + \lambda_i^{-1} x^T G_i^T G_i x = x^T(\lambda_i PD_i D_i^T P + \lambda_i^{-1} G_i^T G_i)x\end{aligned} \tag{16}$$

which implies that

$$\begin{aligned}\dot{V}(x,t) &\le \sum_{i=1}^{k}\sum_{j=1}^{k} h_i h_j \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix}^T \begin{bmatrix} \Theta_{ij} & P(A_{1i}+\Delta A_{1i}) \\ * & -S \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix} \\ &= \sum_{i=1}^{k} h_i^2 \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix}^T \begin{bmatrix} \Theta_{ii} & P(A_{1i}+\Delta A_{1i}) \\ * & -S \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix} \\ &+2\sum_{i=1}^{k}\sum_{j>i}^{k} h_i h_j \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix}^T \begin{bmatrix} \dfrac{\Theta_{ij}+\Theta_{ji}}{2} & P\dfrac{(A_{1i}+\Delta A_{1i}+A_{1j}+\Delta A_{1j})}{2} \\ * & -S \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-\sigma(t)) \end{bmatrix}\end{aligned} \tag{17}$$

where

$$\begin{aligned}\Theta_{ij} &= P(A_i+\Delta A_i+(B_i+\Delta B_i)\Phi_\omega K_j)+(A_i+\Delta A_i+(B_i+\Delta B_i)\Phi_\omega K_j)^T P \\ &+\lambda_i PD_i D_i^T P + \lambda_i^{-1} G_i^T G_i + \frac{1}{1-\beta}S\end{aligned} \tag{18}$$

The remaining part of the proof is similar to the proof of Theorem 1 in [6], so it is omitted. We can prove that $\dot{V} < 0$ and system (10) is asymptotically stable for any actuator failure if LMIs (11) and (12) are satisfied. This completes the proof. □

4 Simulation

Consider an example of DC motor controlling an inverted pendulum via a gear train [6]. The fuzzy model is as follows:

$$A_1=\begin{bmatrix}0 & 1 & 0\\ 9.8 & 0 & 1\\ 0 & -10 & -10\end{bmatrix}, A_2=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & -10 & -10\end{bmatrix}, B_1=\begin{bmatrix}0\\ 0\\ 10\end{bmatrix}, B_2=\begin{bmatrix}0\\ 0\\ 10\end{bmatrix} \tag{19}$$

$$A_{11}=A_{12}=\begin{bmatrix}0.125 & 0 & 0.15\\ 0 & -0.25 & 0\\ 0 & 0.15 & 0.25\end{bmatrix} \tag{20}$$

The F_1 and F_2 are fuzzy sets defined as

$$\begin{aligned}F_1(x_1(t))&=\begin{cases}\dfrac{\sin\left(x_1(t)\right)}{x_1(t)}, & x_1(t)\neq 0\\ 1, & x_1(t)=0\end{cases}\\ F_2(x_1(t))&=1-F_1(x_1(t))\end{aligned} \tag{21}$$

To illustrate the application of Theorem 1, the uncertainties are introduced.

$$\begin{aligned}\Delta A_1=\Delta A_2&=\begin{bmatrix}0.15\\ 0\\ 0.1\end{bmatrix}F(t)\begin{bmatrix}0.1 & -0.15 & 0.15\end{bmatrix},\\ \Delta A_{11}=\Delta A_{12}&=\begin{bmatrix}0.15\\ 0\\ 0.1\end{bmatrix}F(t)\begin{bmatrix}0.15 & -0.15 & 0\end{bmatrix},\\ \Delta B_1=\Delta B_2&=\begin{bmatrix}0.15\\ 0\\ 0.1\end{bmatrix}F(t)\times 0.1,\\ D_1=D_2&=0.1*I, f_1=\begin{bmatrix}0.1\sin x_1\\ 0.1\sin x_2\\ 0.1\sin x_3\end{bmatrix}, f_2=\begin{bmatrix}0.1\sin x_3\\ 0.1\sin x_2\\ 0.1\sin x_1\end{bmatrix}, G_1=G_2=0.1*I_3\end{aligned} \tag{22}$$

Solving the LMIs (11)-(12) for a reliable fuzzy controller with $\Phi_\Omega=0.1*I_1$ produces

$$X=\begin{bmatrix}0.0010 & -0.0028 & -0.0020\\ -0.0028 & 0.0091 & -0.0268\\ -0.0020 & -0.0268 & 0.7234\end{bmatrix}, T=\begin{bmatrix}0.2070 & 0.0621 & -0.1306\\ 0.0621 & 0.0910 & -0.0127\\ -0.1306 & -0.0127 & 0.1245\end{bmatrix}, \tag{23}$$

$$K_1=K_2=\begin{bmatrix}-2.7681 & -0.9903 & -0.0456\end{bmatrix}\times 10^4 \tag{24}$$

Simulations are run with the delay $\sigma(t)=3, F(t)=\sin(t)$. When actuator faults with $\Phi_\omega=0.1*I_1$ occurred after the system has been operating properly for 5 seconds, the

state responses are shown in Fig. 1. When actuator faults occur, the closed-loop system using the proposed reliable fuzzy controller still operates well and maintains an acceptable level of performance.

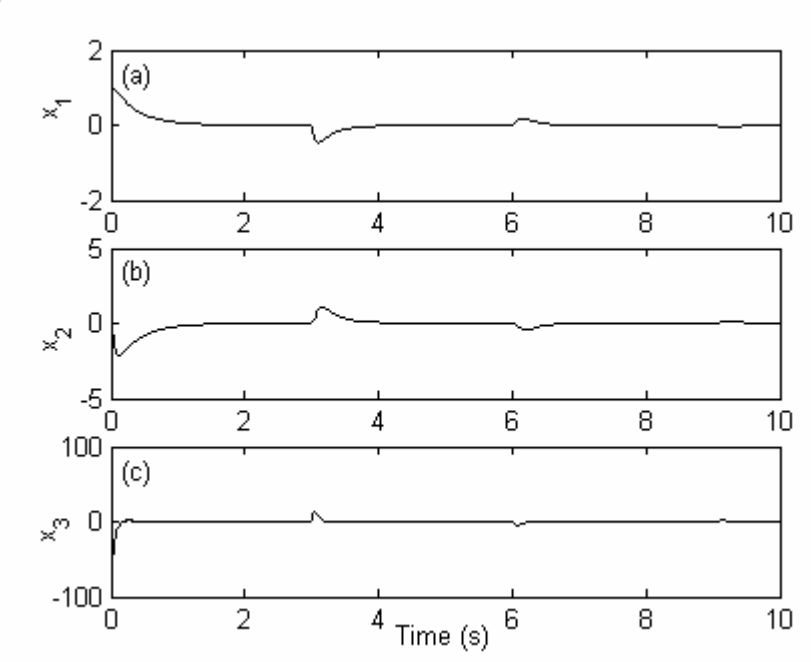


Fig. 1. (a), (b) and (c) show the tracks of x_1, x_2 and x_3, respectively. The closed-loop system with faults is still asymptotically stable.

5 Conclusion

In this paper, robust reliable fuzzy control for a class of nonlinear uncertain systems with state-delay has been proposed. The closed-loop system is asymptotically stable, not only when the system is operating properly, but also in the presence of certain actuator faults. The state-delay is assumed to be time-varying. The construction of the desired state feedback gains is given in terms of positive-definite solutions to LMIs. It has been proved that our design method is more general than the state feedback control law proposed in [6].

Acknowledgements

This work was mainly supported by NSFC (Grant No., 60234010), partially supported by the Field Bus Technology & Automation Key LAB of Beijing at North China and the national 973 program (Grant No. 2002CB312200) of China.

References

1. Zadeh, L.A.: Fuzzy Set Information. Control 8 (1965) 338-353
2. Takagi, T., Sugeno, M.: Fuzzy Identification of Systems and Its Applications to Modeling and Control. IEEE Trans. System, Man and Cybernetic 15 (1985) 116-132
3. Cuesta, F., Gordillo, F., Aracil, J.: Stability Analysis of Nonlinear Multivariable Takagi-Sugeno Fuzzy Control Systems. IEEE Trans. Fuzzy Systems 7 (1999) 508-520
4. Li, N., Li, S.-Y.: Stability Analysis and Design of T-S Fuzzy Control System with Simplified Linear Rule Consequent. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics 34 (2004) 788-795
5. Luoh, L.: New Stability Analysis of T-S Fuzzy System with Robust Approach. Mathematics and Computers in Simulation 59 (2002) 335-340
6. Chen, B., Liu, X.: Reliable Control Design of Fuzzy Dynamic Systems with Time-Varying Delay. Fuzzy Sets and Systems 146 (2004) 349-374
7. Veillette, R.J., Medanić, J.V., Perkins, W.R.: Design of Reliable Control Systems. IEEE Trans. Automatic Control 37 (1992) 290-304
8. Wang, Z., Huang, B., Unbehauen, H.: Robust Reliable Control for a Class of Uncertain Nonlinear State-Delayed Systems. Automatica 35 (1999) 955-963
9. Wu, H.-N.: Reliable LQ Fuzzy Control for Continuous-Time Nonlinear Systems with Actuator Faults. IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics 34 (2004) 1743-1752

Using Concept Taxonomies for Effective Tree Induction

Hong Yan Yi, B. de la Iglesia, and V.J. Rayward-Smith

School of Computing Sciences, University of East Anglia,
Norwich, NR4 7TJ, UK
{hyy, bli, vjrs}@cmp.uea.ac.uk

Abstract. Taxonomies are exploited to generate improved decision trees. Experiments show very considerable improvements in tree simplicity can be achieved with little or no loss of accuracy.

1 Introduction

Decision tree induction algorithms, such as ID3, C4.5, and C5, have been widely used for machine learning and data mining with encouraging results (see, e.g. [7]). These algorithms can also be the basis of rule induction. The leaves of a decision tree predict the value of an output variable, while the internal nodes represent tests on input variables. Given a tuple of input values, we start at the root of the tree. Then, at each internal node visited, a test is applied and, depending upon the result of the test, one of the branches from the internal node is followed. A tuple of input values thus results in a path from the root of the tree to some leaf and when the leaf is reached, a corresponding output is predicted. The accuracy of the tree, usually measured against a test set, is the percentage of those records in the test set for which, given the values of the input variables, the tree predicts the correct output value. Pursuing high accuracy usually results in a complex tree and overly detailed rules. In practice, this is often not what is required by the end users; a simpler and more understandable tree with slightly less accuracy will often prove more useful. In this paper, a form of attribute (or feature) construction will be introduced to enhance the simplicity of the tree being generated. It will explore the use of concept taxonomies on the values of categorical attributes to create new attributes based on existing ones. Concept taxonomies have been used to simplify rules in rule induction [2] during the heuristic selection process. We will use them for tree induction and as a preprocessing activity. Some experimental results are presented to show the improvements that can be obtained by using different combinations of the new attributes.

2 Attribute Construction and Selection

During data preprocessing in data mining, attribute construction aims to improve the concept representation space. It can be done by combining attributes

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1011–1016, 2005.
© Springer-Verlag Berlin Heidelberg 2005

to generate new attributes, and/or abstracting values of given attributes [3, 4]. This paper describes a knowledge-driven method of attribute construction that uses concept taxonomies to help generate a simpler and more meaningful decision tree. In this paper, only categorical attributes are considered. The construction of concept taxonomies can either be done manually or can be derived from previously defined ontologies [6].

When exploiting the taxonomies to produce a simpler decision tree, we assume that the taxonomy is presented as a rooted tree and all the nodes in the taxonomy are either leaf nodes or have more than one child. The root node represents the class definition, other internal nodes represent subclasses. For this initial study we assume leaf nodes represent values that will appear in our database.

A new attribute can be constructed by any cut on a concept taxonomy tree, except the cut composed by all the leaf nodes, which represents the original attribute-value space. A *cut*, C, for a tree, T, with a set of nodes, N, is a nonsingleton set of nodes, $C \subset N$, s.t.

1. every leaf in the tree is either in C or has an ancestor in C.
2. if $n \in C$, then either n is a leaf or no node in the subtree rooted at n is in C.

Thus every path from the root to a leaf contains a single node in C.

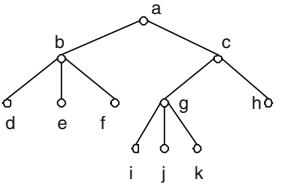


Fig. 1. Taxonomy Tree

Let $C(n)$ be the number of cuts for the subtree rooted at n. Then $C(n) = 1$ if n is a leaf, and if n has children $n_1, n_2, \ldots, n_k$ then

$$C(n) = \begin{cases} C(n_1) \times C(n_2) \ldots \times C(n_k) & \text{if } n \text{ is the root,} \\ C(n_1) \times C(n_2) \ldots \times C(n_k) + 1 & \text{if } n \text{ is not the root.} \end{cases}$$

For the tree in figure 1, we have $C(b) = 2,\ C(g) = 2,\ C(c) = 3$, and $C(a) = 6$. These 6 cuts are $\{b, c\}$, $\{d, e, f, c\}$, $\{b, g, h\}$, $\{d, e, f, g, h\}$, $\{b, i, j, k, h\}$, and $\{d, e, f, i, j, k, h\}$.

The cut comprising all the leaf nodes of a concept taxonomy corresponds to a column of data in the original database. We call this column of data, the *primitive* column for that concept. Any other cut, C, for this concept can be used to generate a new column of data for the same concept. This is constructed from the primitive column by applying the following rule. For each $n \in C$, all leaf nodes in the the subtree rooted at n are replaced by n. The attribute associated with the new column is the same as that associated with the primitive column, albeit with a suitable index. Though the use of these indices is necessary for use within the tree induction algorithm, they can be ignored when the tree is finally delivered.

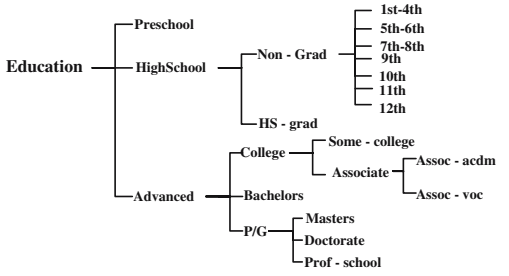
The above construction can result in a large number of additional attributes. This will be the case if the attribute taxonomy is large and complex; trees with multiple branches and many nodes will lead to a great number of different cuts. However, for many cases, the attribute taxonomy is not large and the number of new attributes is manageable. Where this is not the case, the selection of the "best" subset of these new attributes becomes important. In ID3, the *information gain* measure is used when building the decision tree, whilst C4.5 uses a modification known as *gain ratio*. Both of these can be used to limit the number of new attributes, considered by a tree induction algorithm. However, since either measure will be determined in advance of developing a tree, it can only be calculated with respect to the whole database and thus might prove ineffective. It is better to include all new attributes whenever possible.

3 Case Study

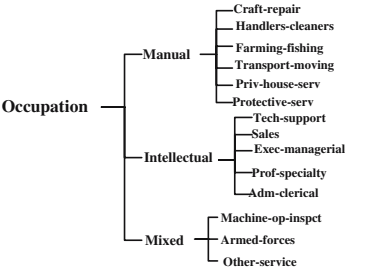
We conducted a case study to demonstrate the impact on the generated decision tree of using concept taxonomies for categorical attributes. A real world data set, the *Adult* database, obtained from the UCI data repository [1], was used. In the *Adult* data set, there are 30,162 instances of training data and 15,060 instances of test data, once all missing and unknown data are removed. The task is to predict whether a person's income is more or less than $50K (approximately 75% people belong to the latter class). There are one target class, seven categorical attributes, and six numeric attributes in the data set. Four categorical attributes, viz. education, occupation, marital-status, and workclass, were given attribute-value taxonomies. Since there was no suitable ontology for us to use, we manually defined the taxonomies for them based on our background knowledge, see figure 2. The number of cuts calculated for these four taxonomies are 21, 5, 5, and 8 respectively. Hence 35 new columns of data are added to the original database.

We also considered some data preprocessing issues. The numeric-valued attributes *Capital-gain* and *Capital-loss* have over one hundred unique values. Discretisation of such data can lead to better search trees [5]. For our experiments, we chose a simple discretisation using values "Low", "Medium", and "High", which are decided by using the range points 5000 and 50000 for the values in *Capital-gain*, and 1000 and 2000 for *Capital-loss*. We also did some experiments by removing these two attributes altogether to see how much it will affect the mining results. This is an interesting exercise since it shows that income can still be successfully predicted without information relating to personal financial affairs. In addition, the values of a categorical attribute, *Native-country*, are abstracted by replacing them with either "United-States" or "non-US", since 91.2% of records refer to Americans. The data after such preprocessing will be called *preprocessed data* in the later experiments.

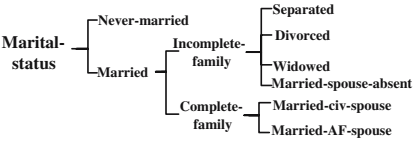
We conducted a series of experiments to compare trees built by using C4.5 and C5 on the following data with or without *Capital-gain* and *Capital-loss* attributes:



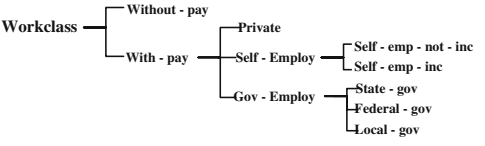
(a) Education



(b) Occupation



(c) Marital-status



(d) Workclass

Fig. 2. Concept Taxonomies

1. the original data,
2. the original data together with all the possible new attributes constructed from the taxonomies (i.e. 35 new fields),
3. the original data plus just the five most promising new attributes according to gain ratio for each of the concepts,
4. the preprocessed data plus all the possible new attributes, or just the five most promising ones.

Table 1 and table 2 show all the experimental results on the *Adult* data. In the "Test Data" column, "C" represents the attributes "*Capital-gain*" and "*Capital-loss*"; "All" means all the possible attributes; "5Max" means the best five attributes according to gain ratio for each concept; '+" and "–" are used to denote whether the data includes or does not include the various features.

The software packages for C4.5 and C5 provide us with different facilities to manipulate the tree induction algorithms. In the experiments on C5, as illustrated in table 1, only default settings can be used. But with C4.5, the situation is changed. Some thresholds can be set to generate a tree, say, with prescribed depth maximum, or some number of minimum objects in one node, etc.

Aiming at a simpler tree, we fix the tree to be generated by C4.5 to have a tree depth of 4, and a minimum of 15 objects in one node, see table 2 below.

We observe the following:

1. Although we might sometimes have to sacrifice the accuracy slightly, a much simpler tree in terms of the number of nodes can be generated by introducing the new attributes constructed from the semantic taxonomies. When using C4.5 with settings designed to generate simple tree, a much simpler tree of

Table 1. Experimental Comparisons Using C5 with Default Settings

Test Data	Accuracy	Tree Depth	No. of Tree Nodes
Original + C	85.55%	11	90
Original – C	81.51%	2	29
Original + All/5Max + C	85.23%	9	70
Original + All/5Max – C	82.22%	3	9
Preprocessed + C	83.92%	5	61
Preprocessed + All + C	84.03%	6	46
Preprocessed + 5Max + C	84.04%	5	34

Table 2. Experimental Comparisons Using C4.5 with Manual Settings

Test Data	Accuracy	Tree Depth	No. of Tree Nodes
Original + C	81.04%	4	21
Original – C	81.82%	3	27
Original + All/5Max + C	81.04%	3	9
Original + All/5Max – C	81.02%	3	8
Preprocessed + C	80.73%	4	24
Preprocessed + All/5Max + C	79.38%	3	13

equal accuracy was generated using the new attributes. In several cases, however, we obtain both a simpler and a more accurate tree. By removing the attributes "*Capital-gain*" and "*Capital-loss*" and adding all the possible new attributes or even just the best five of them according to gain ratio to the original attribute list, we still can get a very simple but highly predictive tree.

2. Feature selection by using *Gain Ratio* measure appears to be efficient. From almost all our experiments, using all attributes or only using the best five attributes generated trees of the same accuracy. For a data set with a large attribute-value space and a large size, this kind of selection can not only save the memory occupation required by the algorithm, but also greatly reduce the size of the data file expanded with the new attributes.

3. Data preprocessing can be very important. Sometimes by only removing some features or doing the discretization on numeric attributes, a simpler tree can be generated. In our experiments using C5, the number of tree nodes reduced by one third simply by undertaking the preprocessing we described earlier. However, even with preprocessed data, adding the new attributes considerably simplifies the tree without undue sacrifice of accuracy.

4 Conclusion and Ideas for Future Research

We have shown how taxonomies can be used to simplify decision tree quite dramatically without significant loss of accuracy. We believe their use will often prove more effective than established pruning techniques but this claim will need to be justified by further experimentation.

With increasing use of ontologies and the merging of data from various depositories, data mining is becoming an increasingly complex task. Since the availability of embedded taxonomies can be so usefully exploited by researchers,

we need to investigate how the extraction of these taxonomies from ontologies can be best automated. When databases are constructed from various sources, not all entries need necessarily be leaves of the taxonomy; the internal nodes (subclasses) can also appear. We need to develop ways of handling such data in tree induction.

We have observed that as taxonomies increase in complexity, so there is an exponential rise in the number of cuts and hence of potentially new attributes. This growth will ultimately damage the performance of any tree induction algorithm. We have discussed how a simple test such as gain ratio might be used to limit the number of new attributes created, but further research is required to determine the true effectiveness of this approach.

This study has focussed on the use of the C4.5 and C5 tree induction algorithm. Further research is needed into the exploitation of taxonomies with alternative tree induction algorithms.

References

1. C. L. Blake and C. J. Merz. UCI repository of machine learning database. University of California, Department of Information and Computer Science, Irvine, CA, 1998.
2. V. Kolluri, F. Provost, B. Buchanan, and D. Metzler. Knowledge discovery using concept-class taxonomies. In *Proc. of the 17th Australian Joint Conference on Artificial Intelligence*, pages 450–461. 2004.
3. H. Liu and H. Moroda. *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic publishers, Boston, 1998.
4. H. Liu and H. Moroda. Feature selection for knowledge discovery and data ming. In *The Kluwer International Series in Engineering and Computer Science*, volume 454, pages 161–168. Kluwer Academic publishers, Boston, 1998.
5. H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proc. of the 7th IEEE. International Conference Tools with AI (ICTAI'95)*, pages 388–391. IEEE Computer Society, Los Alamitos, CA, 1995.
6. D. L. McGuinness. Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2002.
7. P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Education, Boston, 2006.

A Similarity-Based Recommendation Filtering Algorithm for Establishing Reputation-Based Trust in Peer-to-Peer Electronic Communities*

Jingtao Li[1], Yinan Jing[1], Peng Fu[2], Gendu Zhang[1], and Yongqiang Chen[3]

[1] Department of Computer and Information Technology,
Fudan University, Shanghai 200433, China
{lijt, Jingyn}@Fudan.edu.cn
[2] Network Institute, School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an 710049, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China

Abstract. The issues of trust are especially of great importance in peer-to-peer electronic online communities [5]. One way to address these issues is to use community-based reputations to help estimate the trustworthiness of peers. This paper presents a reputation-based trust supporting framework which includes a mathematical trust model, a decentralized trust data dissemination scheme and a distributed implementation algorithm of the model over a structured P2P network. In our approach, each peer is assigned a unique trust value, computed by aggregating the similarity-filtered recommendations of the peers who have interacted with it. The similarity between peers is computed by a novel simplified method. We also elaborate on decentralized trust data management scheme ignored in existing solutions for reputation systems. Finally, simulation-based experiments show that the system based on our algorithm is robust even against attacks from groups of malicious peers deliberately cooperating to subvert it.

1 Introduction

Peer-to-peer (P2P) electronic communities, such as Gnutella [6], eBay [4], are often established dynamically with peers (members) that are unrelated and unknown to each other. Peers have to manage the risk involved with the transactions without prior experience and knowledge about each other's reputation. For example, inauthentic file attacks by malicious peers are common on today's popular P2P file-sharing communities. Malicious peer *u* provides a fake resource with the same name as the real resource peer *v* is looking for. The actual file could be a Trojan Horse program or a virus like the well-known VBS.Gnutella worm [3]. The recent measurement study of KaZaA [1] shows that pollution is indeed pervasive in file sharing, with more than 50% of the copies of many popular recent songs being polluted [2].

One way to address this uncertainty problem is to develop strategies for establishing reputation-based trust that can assist peers in assessing the level of trust they

* This work is supported by the National Natural Science Foundation of China (No. 60373021).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1017 – 1024, 2005.
© Springer-Verlag Berlin Heidelberg 2005

should place on a transaction. The very core of the reputation mechanism in a P2P community is to build a distributed reputation management system that is efficient, scalable, and secure in both trust computation and trust data storage and dissemination. The main challenge of building such a reputation mechanism is how to effectively cope with various malicious collectives of peers who know one another and attempt to collectively subvert the system by providing fake or misleading ratings about other peers [8, 9]. In recent years, many people have come up with reputation mechanisms for various applications [4, 8, 9, 7].

However, much of this work has been creating specific method to compute trust values of peers based on the assumption that the peers with high trust value will give the honest recommendations, so recommendations of the peers are only filtered by their trust values. We argue that assumption and recommendations are filtered by the similarity between the peers in our trust model (section 2). Much of this work also omitted, or only briefly mentioned how to securely store and look up trust data that are needed to compute the trust values. In contrast, we would like to explore the mechanism for trust data storage and dissemination. We present a trust data management scheme based on a variant of the Chord [10] algorithm in section 3. In section 4, the **Sim**ilarity-based **R**ecommendation **F**iltering **A**lgorithm (SimRFA) for supporting reputation-based trust in a P2P community is presented. In section 5, a series of simulation-based experiments show that our algorithm is robust and efficient to cope with various malicious collectives.

2 Notations and Definitions

In this section, we present a general trust metric and describe the formulas we use to compute the trust value for each peer in a P2P electronic community.

Each peer i rates another peer j with which it tries to make transaction by rating each transaction as either positive or negative, depending on whether i was able to accomplish a transaction with j or not. The sum of the ratings of all of i's interactions with j is called a **satisfaction value** $\boldsymbol{S_{ij}}$. $S_{ij}=G_{ij}-F_{ij}$ wherein G_{ij} denotes the number of positive transactions which i has made with j and F_{ij} denotes the number of negative transactions. We let n denotes the number of all peers in a network.

Definition 1. A **trust value vector**, $\boldsymbol{T} = [T_1, T_2, \ldots, T_n]$, is given by

$$T_i = \sum_{k \in U_i} (L_{ki} \cdot C_{ki} \cdot T_k) \quad , \tag{1}$$

where T_i is the **trust value** of peer i. we filter the recommendations by weight each recommender k's opinion by the similarity between peers k and i. And the trust value of peer i is the sum of the weighted recommendations of the peers that have interacted with i in a single iteration. L_{ij}, U_i and C_{ij} are defined as follows:

Definition 2. A **recommendation** $\boldsymbol{L_{ij}}$ is defined as follows:

$$L_{ij} = \frac{\max(S_{ij}, 0)}{\sum_j \max(S_{ij}, 0)} \quad . \tag{2}$$

L_{ij} is a real value between 0 and 1, and $\sum_j L_{ij} = 1$. If $\sum_j \max(S_{ij}, 0) = 0$, let $L_{ij} = T_i / n$.

We let L_{ii}=0 for each peer i; otherwise, peer i can assigns an arbitrarily high recommendation value to itself. L_{ij} is the normalized satisfaction value that peer i has in peer j, used as the recommendation of peer i to j.

Definition 3. For any peer i and peer j, **the similarity** between peers i and j, denoted by $\boldsymbol{C_{ij}}$, is given by

$$C_{ij} = \frac{B_i * B_j}{\|B_i\| \cdot \|B_j\|} , \tag{3}$$

where " * " denotes the dot-product of the two vectors. Here $\boldsymbol{B_i}$ denotes the **rating opinion vector** of peer i, defined as $B_i = [B_{i1}, B_{i2}, ..., B_{in}]$ where B_{ik} (k=1, ...,n) is the rating opinion of peer i on peer k. B_{ik} is defined as follows:

$$B_{ik} = \begin{cases} 1 & G_{ik} \geq F_{ik} \\ -1 & G_{ik} < F_{ik} \end{cases} . \tag{4}$$

One critical step in our model is to compute the similarity between rating opinions of peers and then to filter the recommendations by that value. Actually we do not use (3) to compute C_{ij} which involves too many multiplications. In Algorithm 1, we design such a method for similarity computation, called simplified cosine-based similarity. The similarity between them is measured by calculating the cosine of the angle between these two vectors, each element of which can only be 1 or -1.

Algorithm 1. Simplified similarity computation.

Input: B_i, B_j, n, **Output:** C_{ij}

```
SimplifiedSimilarity(Bi, Bj, n)
{    Sum=0;
     For (k=1, k<n, k++)
     {      If (Bik + Bjk ==0) Sum+=-1;
            Else Sum+=1;      }
     Return Sum / n;
}
```

Definition 4. For each peer i, the **upstream set** of peer i, denoted by U_i, is given by:

U_i = {j | peer j has interacted with peer i, and rated i by a recommendation, L_{ji} }.

3 The Decentralized Trust Data Management Scheme

Although the trust model for a P2P community is independent of its implementation, the effectiveness of supporting reputation in a community depends not only on the factors and metric for computing trust values, but also on the implementation and usage of the trust model in a P2P network. Typical issues in implementing a P2P trust model in a decentralized P2P network include decentralized and secure trust data management, distributed trust value computation algorithm, and trust-based peer

selection scheme. This section discusses the trust data management for implementing our model in a decentralized P2P network.

One approach to manage the trust data in a decentralized way is to let the peers, by themselves, collect the data to be used in their trust computation. It is easy to implement, but each peer is allowed to compute and report its own trust value to the other peers on the network. A problem with this approach is that malicious peers can easily report false trust values, thereby subverting the system. To prevent this from happening, **distributed hash table** (DHT) mechanism is used to assign **trust agents**, which are peers that calculate the trust values of other peers, to each peer in a community.

Here we use the DHT algorithm, Chord [10], to assign trust agents to each peer in a network. At its heart, Chord provides fast distributed computation of a hash function mapping keys to peers (here we map peers to its trust agents) responsible for them. The consistent hash function assigns each peer and key an *m*-bit identifier using a base hash function such as SHA-1. A peer's identifier is chosen by hashing the node's IP address, while a key identifier is produced by hashing the peers' identifier (the two hash functions may be the same one or two different ones). We will use the term "key" to refer to the image of a peer's identifier under the hash function. Consistent hashing assigns keys to peers as follows. Identifiers are ordered in an identifier circle modulo 2^m. Key k is assigned to the first peer whose identifier is equal to or follows (the identifier of) k in the identifier space. This peer is called the successor peer of key k, denoted by *successor*(k). If key k (k's identifier) is produced by hashing the peer u's identifier, then the *successor*(k) is called a trust agent of peer u, denoted by M_u.

Each peer can have a number L of trust agents in the identifier space of Chord, which are determined by a number L of keys by applying a set of one-way hash functions $h_1, h_2, ..., h_L$ to the peer's identifier. That ensures the repetition of reputation data across the trust agents in case one of them fails.

Each peer is assigned several trust agents by Chord and each peer also serves as a trust agent for some other peers to store trust data and compute trust value. A trust agent of peer i, M_i, shall be at least responsible for the following duties: (1)Storing the trust data of peer i such as G_{ij}, F_{ij}, L_{ij}, etc.; (2)Verifying the consistency of the stored data (i.e., comparing the G'_{ij}, newly submitted by peer i, to G_{ij}, now stored by M_i, if $G'_{ij} - G_{ij} > 1$, it is very possible that peer i has submitted the malicious data); (3)Computing the trust value of peer i; (4)Submitting the trust value, rating opinion vector, etc. of peer i as responses to queries of other peers in the network.

4 Algorithm Description

Here we describe the SimRFA algorithm which includes two algorithms to compute a trust value vector. We will use these primitives:

Submit (ID_i, (ID_j, ID_k), G_{jk}, F_{jk}): Submitting G_{jk} and F_{jk} to the trust agents of peer i;

Query (ID_j, T_j, L_{ji}, B_j): Querying the trust value, recommendation to peer i, and rating opinion vector of peer j in the identifier space of Chord.

Each peer i plays two roles in a community: an ordinary peer who rates, or rated by, other peers, and a trust agent of certain peers. As an ordinary peer, peer i uses the Algorithm 2; as a trust agent of peer u, i uses the Algorithm 3.

Algorithm 2. ReportTrustData, peer *i* as an ordinary peer.

```
ReportTrustData( )          // submits its ratings (G_ij, F_ij) after each transaction with j.
{
   If ( a successful transaction ) G_ij←G_ij+1;
   else                F_ij←F_ij+1;
   Submit(ID_i, (ID_i, ID_j ), G_ij, F_ij);
}
```

Algorithm 3. ComputeTrustValue, peer *i* as a trust agent of peer *u*.

```
ComputeTrustValue( )                          // computes the trust value of peer u.
{
   for ( every j∈U_u ( j≠u) )
    {
        Query (ID_j, T_j, L_ju, B_j );
        C_ju←SimplifiedSimilarity( );        // computes the similarity between u and j.
        T_u←T_u + L_ju · C_ju · T_j ;
    }
    return T_u;
}
```

As a trust agent of peer *u*, peer *i* also need to verify the consistency of G_{uv} and F_{uv} and to compute and update S_{uv}, L_{uv}, B_{uv} when receiving a *Submit*(ID_u, (ID_u, ID_v), G_{uv}, F_{uv}) primitive from *u*.

5 Experiments

In this section, we will assess the performance of our algorithm as compared to a P2P community where no reputation system is implemented, called non-trust based algorithm. We shall demonstrate the performance under a variety of threat models.

5.1 Simulation

As a test bed for our experiments, we use the Query Cycle Simulator [12] which simulates a typical peer-to-peer file-sharing network, an instance of an electronic community. Trust values in this community are used to bias download sources.

Each simulation is divided into query cycles. In each query cycle, any given peer on the network could be issuing a query, inactive, or down and not responding to queries. After issuing a query, a peer waits for incoming responses from peers that have the file it is looking for, selects a download source whose trust value is the highest among those peers, and downloads the file. The last two steps are repeated until the peer receives a complete, authentic copy of the file. Then the simulation comes into the next query cycle. Each experiment was run several times and the results of the runs were averaged.

We set up a network consisting of 500 peers that divided into good peers (nodes who are participating in the network in order to share files) and malicious peers (nodes who are participating in the network in order to spread inauthentic files and

undermine its performance). The proportion of malicious peers will be given in the descriptions of the experiments. When they join the network, malicious peers connect to the 6 most highly-connected peers in the network in order to receive as many queries traveling though the network as possible. And the good peers only connect to 3 neighbors. The initial distribution of the trust values is a uniform probability distribution over all n peers, that is, $T_i^{(0)}=1/n$ (i=1, … , n). The other detailed settings are described in [12].

5.2 Threat Models

We will consider several strategies of malicious peers to cause inauthentic uploads even when our scheme is activated.

Threat Model IM: Individual malicious peers, called IM peers, always provide an inauthentic file when selected as a download source and they always set their satisfaction values to $S_{ij}= F_{ij} -G_{ij}$.

Threat Model CM: Malicious peers, called CM peers, always provide an inauthentic file when selected as a download source and each CM peer i let $L_{ij}=1/\|CMs\|$ for any $j\in$CMs where CMs denotes the set of all CM peers, and $\|CM\|$ denotes the total of all CM peers.

5.3 IM Experiments and Discussion

IM experiments are carried out in the presence of IM peers. Fig. 1 shows the total of inauthentic file downloads of each cycle changing when the simulation of a network proceeds from one query cycle to the next. We have IM peers make up 40% of all peers in the network. In the absence of a reputation system, malicious peers succeed in inflicting many inauthentic downloads on the network. Yet, if our scheme is activated, the inauthentic files downloaded are almost eliminated after the 5th cycle. That means malicious peers can not get high trust values when our scheme is activated, and because of their low trust values, malicious peers are rarely chosen as download source which minimizes the number of inauthentic file downloads in the network.

In the next experiment, we have between 0% and 50% of the peers on the network be malicious peers, increasing this percentage in steps of 10% for each run of the experiment. The change of the number of inauthentic downloads (using our algorithm, inauthentic downloads are almost eliminated after 3-6 cycles) are similar to that showed in Fig. 1. **Proportion of Authentic Downloads (PAD)**, the ratio of the number of authentic downloads to the number of all downloads viewed by all good peers, of two algorithms are showed in Fig. 2. The PAD is higher than 80% even if IM peers make up a half of all peers in the network when our scheme is activated.

5.4 CM Experiments and Discussion

CM peers are more destructive than IM peers in that they form a malicious collective and boost the trust values of all peers in the collective. In our algorithm, a peer i trusts the peers whose rating opinions are similar to that of i, as if the good peers also form a

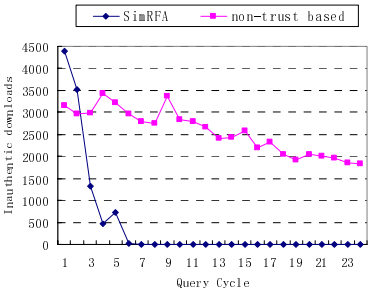


Fig. 1. Change of inauthentic downloads of each cycle when the simulation proceeds

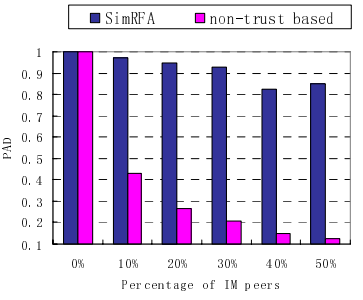


Fig. 2. At different percentage of IM peers, the PADs of the two algorithms

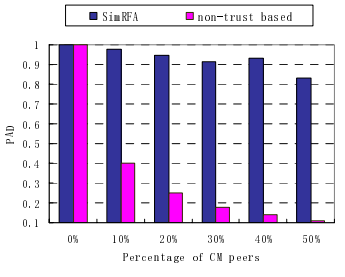


Fig. 3. At different percentage of CM peers, the PADs of the two algorithms

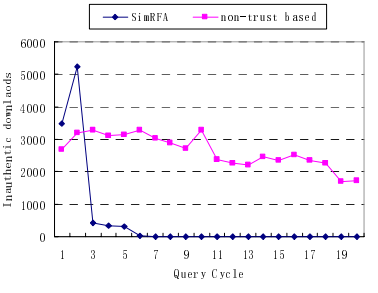


Fig. 4. Change of the number of inauthentic downloads of each cycle when the simulation proceeds. We have CM peers make up 40% of all peers in the network

collective to suppress the CM peers. Therefore our algorithm performs well even if CM peers make up of a half of peers in the network. The experiment clearly shows that CM peers are tagged with a low trust value and thus rarely chosen as download source. Fig. 3 shows the PAD of each algorithm when we add a number of malicious peers to the network such that malicious peers make up between 0% and 50% of all peers in the network. Fig. 4 shows the total of inauthentic file downloads of each cycle changing when the simulation of a network proceeds from one query cycle to the next. When our scheme is activated, the inauthentic files downloaded are almost eliminated after the 4th cycle.

6 Conclusion

We have presented a reputation-based trust supporting framework to minimize the impact of malicious peers on the performance of a P2P community. In our approach, the SimRFA algorithm computes a trust value for a peer by filtering the recommendations of the peers who has interacted with the peer. In simulations, these trust values are used to bias downloads and this method is successful to reduce the number of inauthentic files on the network under a variety of threat scenarios.

References

1. Kazaa.: http://www.kazaa.com/
2. Liang, J., Kumar, R., Xi, Y., Ross, K.: Pollution in p2p file sharing systems. Proceedings of IEEE Infocom 2005. IEEE Press. (2005) Volume 2, 1174 - 1185
3. VBS.GnutellaWorm. http://securityresponse.symantec.com/avcenter/venc/data/vbs.gnutella.html.
4. eBay Feedback Forum.: http://pages.ebay.com/services/forum/feedback.html?ssPage Name =STRK
5. Preece, J.: Online Communities: Designing Usability, Supporting Sociability. John Wiley & Sons, New York. (2000)
6. Gnutella. http://www.gnutella.com/
7. Gupta, M., Judge, P., and Ammar, M.: A Reputation System for Peer-to-Peer Networks. Proceedings of 13th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'03). ACM Press. (2003) 144–152
8. Xiong, L., Liu, L.: PeerTrust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16(7). IEEE Press. 843–857
9. Kamvar, S.D., Schlosser, M.T.: EigenRep: Reputation management in P2P networks. Proceedings of the 12th International World Wide Web Conference. ACM Press. (2003) 123–134
10. Stoica, I., Morris, R., Karger, D.: Chord: A scalable peer-to-peer lookup service for Internet applications. Technical Report. MIT. (2002). http://pdos.csail.mit. edu/chord/
11. Shi, WM., Yang HF., Wu, YS., Sun, X.: Numerical Analysis. 2nd edition, Beijing. BEIJING INSTITUTE OF TECHNOLOGY PRESS. (2004) 91–93
12. Schlosser, M.T., Condie, T.E., Kamvar, S.D.: Simulating a File-Sharing P2P Network. Proceedings of First Workshop on Semantics in P2P and Grid Computing (SemPGRID'03). (2003) 69–80

Automatic Classification of Korean Traditional Music Using Robust Multi-feature Clustering

Kyu-Sik Park, Youn-Ho Cho, and Sang-Hun Oh

Dankook University Division of Information and Computer Science,
San 8, Hannam-Dong, Yongsan-Ku, Seoul, Korea 140-714
{kspark, adminmaster, taru74}@dankook.ac.kr

Abstract. An automatic classification system of Korean traditional music is proposed using robust multi-feature clustering method. The system accepts query sound and automatically classifies input query into one of the six Korean traditional music categories. This paper focuses on the feature optimization method to alleviate system uncertainty problem due to the different query patterns and lengths, and consequently increase the system stability and performance. In order to fit this needs, a robust feature optimization method called multi-feature clustering (MFC) based on VQ and SFS feature selection is proposed. Several pattern classification algorithms are tested and compared in terms of the system stability and classification accuracy.

1 Introduction

Most of content-based classification and retrieval of audio sound has three common stages of a pattern recognition problem: *feature extraction, classification* based on the selected feature, and *retrieval* based on the similarity measure. Depending on the different combinations of these stages, several strategies are employed. Wold *et al* [1] developed a system called "Muscle Fish" where various perceptual features such as loudness, brightness, pitch, timbre are used to present a sound. As a direct at-tempt of acoustical features to model musical signals, Tzanetakis and Perry [2] com-bined standard timbral features with representations of rhythm and pitch content, and they achieved about 60% for ten musical genres. Other good works are described in [3-4].

This paper focuses on the following issues on music classification. First, the proposed system accepts query sound and automatically classifies a query into one of the six Korean traditional music categories such as Court music, Classical chamber music, Folk song, Folk music, Buddhist music, and Shamanist music. Court music refers to ritual temple music which includes Confucian music and royal shrine music, while banquet music is of course music for courtly banquets. It has elegant rhythm features with various timbral instruments. Classical chamber music refers to a type of ensemble music for the nobility which consists mainly of stringed instru-ments. Folk song is defined as song in drama such as the spoken description of the dramatic content between songs. Folk music is popular for the common people. It has fast rhythm structure that feels like live performance.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1025–1029, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Buddhist music is in the form of two types. One is a syllabic style and the other is song at special rites to praise to Buddha. Shamanist Music has been held to promote the well being and prosperity of individual homes and village people, or for the benefit of a dead person's soul. Sha-man rituals are performed with songs and dances as well as theatrical and acrobatic actions by shamans. Secondly, In order to overcome the uncertainty of the system per-formance due to different query patterns and lengths, a robust feature extraction method called MFC (multi-feature clustering) combined with SFS (sequential forward selection) is proposed based on VQ (Vector Quantization).

2 Proposed System

The proposed system is illustrated in Fig. 1. The system consist 3 stages - feature extraction, feature optimization, and classification. Firstly, in the feature extraction stage, feature sets for training music are produced and passed to feature optimization stage. Secondly, in feature optimization stage, we normalize the training feature set and then applied MFC-VQ feature clustering with SFS feature selection to provide optimal decision model for the classification of each music patterns.

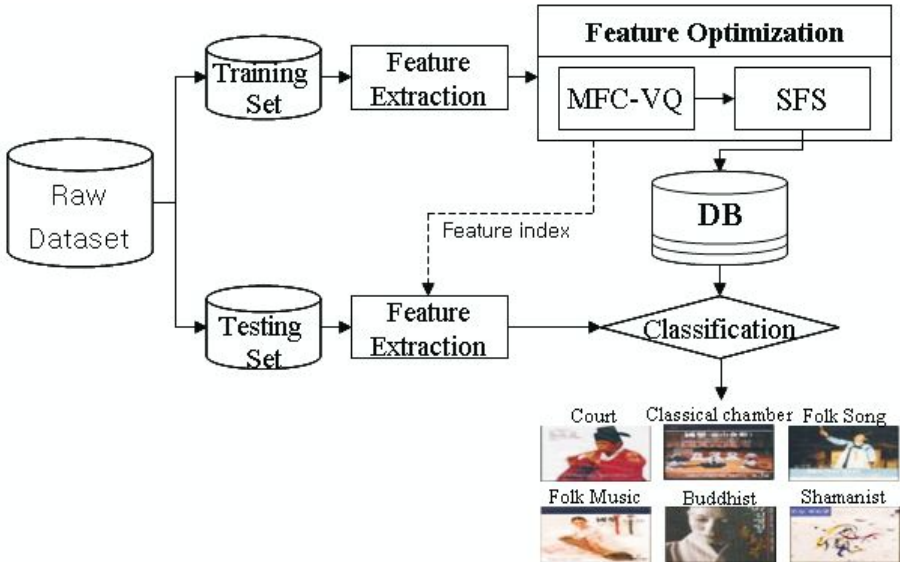


Fig. 1. Overall Framework of Proposed Classification System

We note that the feature sets for testing music are extracted according to the feature index from the feature optimization stage. Finally, k-NN and SVM pattern classifier [5] are tested for the classification of the test music.

3 Feature Extraction, Selection and Optimization

The music signals are divided into 23ms frames with 50% overlapped hamming window at two adjacent frames. Two types of features are computed from each frame. One is the timbral features such as spectral centroid, spectral rolloff, spectral flux and zero crossing rates. The other is coefficient domain features of thirteen mel-frequency cepstral coefficients (MFCC) and twelve linear predictive coefficients (LPC). The mean and standard deviation of these six original features are computed over each frame for each music file to form a total of 58-dimensional feature vector. In feature optimization stage, an efficient feature selection method is required to reduce the computational burden and so speed up the search process. As in ref. [6], we used a sequential forward selection (SFS) method for feature selection. The classification results corresponding to different input query patterns and lengths within the same music file or same class may be much different. It may cause serious uncertainty of the system performance. In order to overcome this problem, a robust feature extraction method called multi-feature clustering (MFC) combined with SFS is implemented based on VQ. Key idea is to extract features over the full-length music signal in a step of 15 sec large window and then cluster these features in four disjoint subsets (centroids) using LBG-VQ technique . At this moment, SFS feature selection is applied to reduce the feature dimensions of each centroid. After the MFC-VQ and SFS feature optimization procedure, the system is then ready to classify the genre of the queried music using pattern classification algo-rithms. We now recall that the proposed method has four sets of centroids for each music signal in trained database instead of one as usual. Because of this difference, the pattern classification algorithm needs to be slightly modified. For example, in modified k-NN, a query to be classified is compared to each one of the four feature sets in training data vectors and classification is performed according to the distance to the k nearest feature points with Euclidean distance measure. Finally, the classifica-tion is done by picking the k points nearest to the current test query point, and the class most often picked is chosen as classification result. SVM pattern classifier follows same strategies.

4 Experimental Results

The proposed algorithm has been implemented to classify the genre of Korean traditional music from a database of 180 music files. 30 music samples were collected for each of the six categories in "Court music", "Classical chamber music", "Folk song", "Folk music", "Buddhist music", and "Shamanist music" resulting in resulting in 180 music files in database. The 180 music files are partitioned randomly into a training set of 126 (70%) sounds and a test set of 54 (30%) sounds. This random division was iterated one hundred times. The overall accuracy was obtained as the arithmetic mean of the success rate of the individual iterations. The classification results corresponding to different input query patterns (or portions) may be much different and so the serious uncertainty of the system perform-ance. In order to overcome this problem,

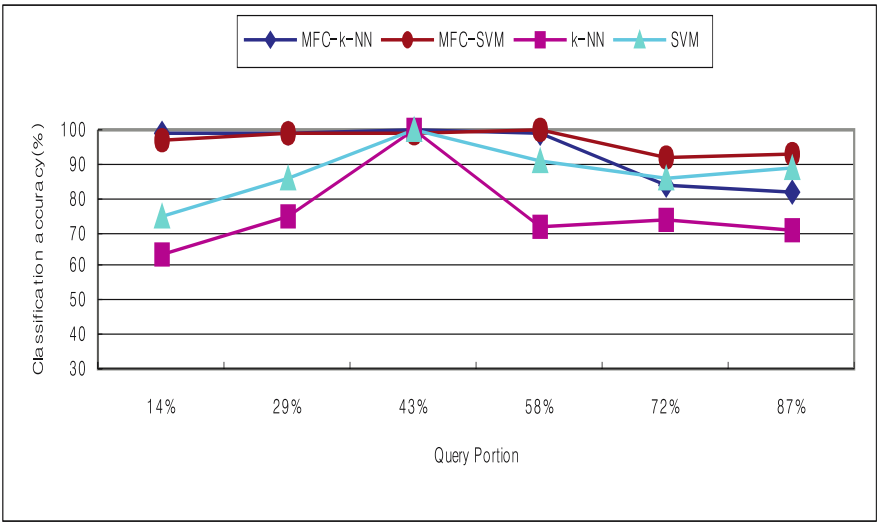


Fig. 2. Average Classification results with proposed MFC-VQ and without MFC-VQ

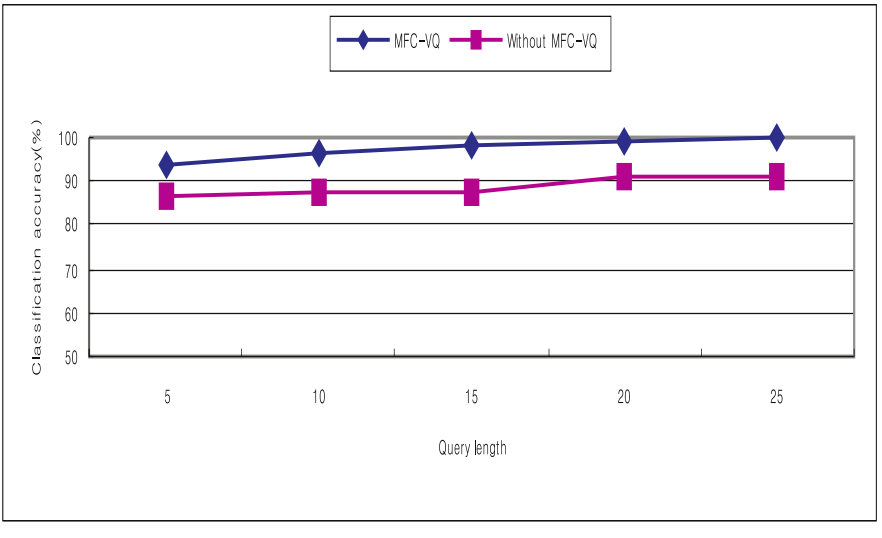


Fig. 3. Genre classification results at different query lengths with MFC-VQ

MFC-VQ feature optimization technique is used as explained in section 3. Six excerpts with duration of 15 sec were extracted from every other position in a same query music file. Fig. 2 shows the classification results with proposed MFC-VQ and without MFC-VQ for the case of k-NN and SVM classifiers.

As we expected, the classification results without MFC-VQ greatly depends on the query positions and its performance is getting worse as query portion

towards to two extreme cases of beginning and ending position of the music signal. On the other hand with the proposed MFC-VQ method, we can find quite stable classification performance and it even yields higher accuracy rate in the range of 83% 98%. Across the every query position, SVM seems to be a good choice because of the robust classification accuracy and the high speed.

Fig. 3 shows an importance of the query length to the overall system performance. Five excerpts with duration of 5sec, 10sec, 15 sec, 20sec and 25sec are used as a test query positioned at 20% of the music. Only the result of SVM classifier is listed. Again, we see the desirable characteristics of MFC-VQ with stable classification performance and more than 20% improvement over the one without MFC-VQ.

5 Conclusion

In this paper, we propose an automatic classification system that classifies the Ko-rean traditional music into one of six categories. In order to overcome system uncer-tainty due to the different query patterns, a robust feature optimization method called MFC-VQ with SFS feature selection is proposed. Several pattern classifiers are tested and compared in terms of the system stability and the classification accuracy. From the comparison statistics on genre classification, we verify the stable and successful classification performance over 91%, and SVM classifier has advantage of the robust classification accuracy and faster classification speed.

Acknowledgment

This work was supported by grant No. R01-2004-000-10122-0 from the Basic Re-search Program of the Korea Science & Engineering Foundation.

References

1. E. Wold, T. Blum, D. Keislar, J. Wheaton: Content-based classification, search, and retrieval of audio, IEEE Multimedia, 3(1996)
2. G. Tzanetakis, P. Cook: Musical genre classification of audio signals," IEEE Trans. on Speech and Audio Processing, 10(2002) 293-302
3. T. Li, M. Ogihara, Q. Li: A comparative study on content-based music genre classification, in Proc. of the 26th annual internal ACM SIGIR, ACM Press 282-289 (2003)
4. J. Foote et. al. An overview of audio information retrieval, ACM-Springer Multimedia Systems, 7 (1999) 2-11
5. R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd Ed., Wiley-Interscience Publication (2001)
6. M. Liu, C. Wan: A study on content-based classification retrieval of audio database, Proc. of the International Database Engineering & Applications Symposium (2001) 339-345

A Private and Efficient Mobile Payment Protocol

Changjie Wang and Ho-fung Leung

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
{cjwang, lhf}@cse.cuhk.edu.hk

Abstract. Many secure electronic payment protocols have been proposed, most of which are based on public key cryptograph. These protocols, however, are not suitable for mobile network due to the limitations of mobile environments, such as limited computation capability of mobile devices, limited bandwidth, etc. In this paper, we propose a private and efficient payment protocol for mobile network, which only involves symmetric key algorithm, such as symmetric encryption, hash function and keyed hash function. All these operations can be implemented on mobile devices feasibly. The proposed protocol also achieves completely privacy protection of buyers, which is one of the important requirements in mobile commerce. First, the identity of the buyer is protected from the merchant. Second, the transaction privacy of the buyer, such as what the buyer buys, and whom the buyer buys from, are also protected from any other parties and financial institutions. By giving a security analysis, we show that our protocol satisfies all security requirements in electronic payment.

1 Introduction

As the key technology to electronic commerce, electronic payment protocols have been studied in the past several years as a substitute method for traditional cash used in the virtual world. Although extensive work has been done on the field of securing electronic payment in fix networks, say, the Internet, there are few studies concerning the electronic payment in the mobile network. Due to the characteristics of the mobile devices, such as lower computational capability, limitation of storage, limited bandwidth, etc., most available electronic payment protocols using public key cryptographic technology cannot be applied to mobile network directly.

Some electronic payment schemes for mobile network can be found in recent literatures [2, 3]. These schemes, however, employ pubic key cryptography to certain degrees, which make them not practically feasible in mobile devices. There are also some arguments on securing mobile payment with symmetry cipher mechanisms. In [4], Kungpisdan presents a secure account-based mobile payment protocol, which, however, still need a public key cryptograph based key negotiation when customers register with merchants at the first time. Another main problem of this protocol is that the privacy of the customer is not

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1030–1035, 2005.
© Springer-Verlag Berlin Heidelberg 2005

protected at all during the transactions.[1] That is, the customer's identity and the transaction details are revealed not only to the merchant, but also to the payment gateway and the banks. In this paper, we propose a novel electronic payment protocol, which can be implemented in mobile device efficiently. The proposed protocol only involves symmetric key cryptographic operation, which is much faster than public key cryptography. With an analysis, we show that our protocol satisfies not only the security requirements of electronic commerce applications, but also the privacy protection of the customer.

This paper is organized as follows. In section 2, we propose an efficient and private electronic payment protocol suitable for mobile network. Some discussions and analysis on the proposed protocol are made in section 3. Finally, we conclude in Section 4.

2 An Efficient Mobile Payment Protocol

In this section, we propose a secure and private mobile payment protocol based on the credit card mode which is the most popular option for electronic payment system at present. It should be noted that other charging models, such as e-cash model, debit model etc., are also supported in our protocol.

There are four roles in our electronic mobile payment protocol, named Buyer, Merchant, Issuer (i.e. financial institution of Buyer) and Acquirer (i.e. financial institution of Merchant). The basic idea of our protocol is to divide the transaction payment details into two parts. On Issuer's side, the only available information is that Buyer authorizes a payment to Acquirer, while no merchant identity or account information is revealed, so that the Issuer has no idea of what the Buyer buys. Similarly, Acquirer should only know there is a payment waiting to be credited to the merchant's account, while knowing nothing about who makes the payment. Based on the assumption that Issuer and Acquirer never collude to sell out the private information of Buyer, we can achieve the privacy protection of Buyer during the electronic transactions through mobile network.

2.1 Notations

In the context, $\{B, M, I, A\}$ is the set standing for *Buyer*, *Merchant*, *Issuer* and *Acquirer*, respectively. ID_X denotes the identity of party X. TID_B means the temporary identity of the B, *i.e.*, B's nickname. $E_K(m)$ denotes the message m encrypted by the secret key K. TS_X represents the time stamp generated by party X. $Price_{Item}$ means the price of an item offered by *Merchant*, and V is the payment value involved. $H(m)$ is a collision free one-way hash function on

[1] Privacy protection is considered one of the most important factors in electronic commerce. Basically, people expect to protect their private information during the electronic transactions, just as what happens in real world. For example, when someone wants to buy a piece of chocolate, he can then just pay cash and take the piece of chocolate without having to provide his name, address and other information to the shopkeeper or his bank.

message m. $H(m,k)$ is a keyed collision free one-way hash function on message m, which works as a message authentication code (MAC) on message m with the secret key k. The difference between $H(m,k)$ and $H(m)$ is that $H(m,k)$ not only checks the integrity of the message m, but also achieves mutual authentication between two parties who share the secret key k, since no one else can generate $H(m,k)$ without knowing k. Furthermore, notation x,y denotes the concatenation of messages x and y.

2.2 Session Key Generation

As we present before, B and I should share a secret key, which is denoted by X. Similarly, M and A share a secret key, denoted by Y. X and Y are used to generate a series of random session keys for mutual authentication between B, I and M, A. Given the shared secret key X, both B and I execute the following key generation algorithm to generate a session key sequence: $X_i, i = 1, 2, ..., n$, where

$$X_1 = H(CI_B||X),\ X_2 = H(X \oplus X_1),\ ..., X_n = H(X \oplus X_{n-1})$$

Here, CI_B is the credit card information of B, which is known to B and I only. A random integer i is chosen in different transaction and the corresponding session key X_i is then generated accordingly for authentication between B and I. In such a way, B and I can generate different keys for multiple secure communications. Using the same method, M and A can also generate a session key sequence $Y_j, j = 1, 2, ..., n$ as (AI_M is the account information code of M at A):

$$Y_1 = H(AI_M||Y),\ Y_2 = H(Y \oplus Y_1), ..., Y_n = H(Y \oplus Y_{n-1})$$

2.3 Secure Mobile Payment Protocol

Fig. 1 shows the details of the proposed protocol, and we explain it step by step as follows.

Step 1: Negotiation of price between B and M

B initiates the transaction by sending an inquiring message to M to ask for price of certain Item. Note that the nickname TID_B should be different in different transactions. M then responds to B, with his identity ID_M, price of the Item $price_{Item}$, and ET_{Item} that stands for the expired time of the price for this item. If the price is accepted by B, he confirms the transaction by sending a order acceptance notice to M.

Step 2: Order generation and money claim

M generates an *Order_Form* and a *Money_Claim_Form* which are sent to B and A separately. It should be noted that the *Credit_Form* included in *Order_Form* cannot be generated by any one else, except M, since only he knows the secret session key Y_i. On receiving the *Money_Claim_From*, A records it in his database for future use.

Step1: Price negotiation
$B \rightarrow M$: TID_B, *enquiry of the price for ID* $_{Item}$
$M \rightarrow B$: ID_M, TID_B, ID_{Item}, $price_{Item}$, ET_{ITEM}
$B \rightarrow M$: *Order Acceptance* , ID_M, TID_B, ID_{Item}, $price_{Item}$
Step2: Order generation and money claim
$M \rightarrow B$: *Order_From* , $H(Order_Form)$, where
$Order_Form = \{ Cridet_From , ID_M, ID_A, TID_B, ID_{Item}, price_{Item}, TN, TS_M, i, H(Y_i)\}$
$Credit_From = H((ID_M, ID_A, V, TN, TS_M, ET_{Form}, i), Y_i)$
$M \rightarrow A$: $E_{Y_{i+1}}(Money_Claim_Form , H(Money_Claim_Form)), i, ID_M, ID_A$, where
$Money_Claim_Form = \{ TN, price_{Item}, i, TS_M, ET_{Form} \}$
Step3: Billing Process
$B \rightarrow I$: $E_{X_j}(Billing_Form , H(Billing_Form)), ID_B, ID_I, j$, where
$Billing_Form = \{ Credit_Form , ID_B, ID_I, ID_A, V, TN, TS_B, j, H(Y_i)\}$
Step4: Fund transfer
$I \rightarrow A$: $E_{K_{I-A}}(Transfer_Form , H(Transfer_Form))$, where
$Transfer_Form = (ID_I, ID_A, V, TN, TS_I, Credit_Form)$
Step5: Payment confirmation
A verifies equation: $Credit_Form = H((Money_Claim_Form , ID_M, ID_A), Y_i)$?
$A \rightarrow I$: $Transfer_response = E_{K_{I-A}}(Yes/No, TN, V, K_{B-M}, h(Y_i), h(.))$
$A \rightarrow M$: $Credit_response = (E_{Y_i}(Yes/No, TN, V, K_{B-M}, h(.)), TN)$
$I \rightarrow B$: $Billing_response = (E_{X_j}(Yes/No, TN, price_{Item}, h(Y_i), h(.)), TN)$

Fig. 1. Secure Mobile Payment Protocol

Step 3: Billing process

B first checks the validity of *Order_Form*, and then generates a *Billing_Form* sent to I, encrypted by secret session key X_j, with ID_B, ID_I, j attached. In the *Billing_Form*, ID_B and ID_I identify B and I, respectively. j is used for I to calculate the session key X_j. ID_A identifies A, so that I knows whom (financial institution) to transfer the due payment. There is also a time stamp TS_B to resist replying attack. On decrypting this *Billing_Form* successfully, I believes that B authorizes this payment, since only B can generate this encrypted *Billing_Form.*

Step 4: Fund transfer

Note that even I does not know the identity of M, he does know to make the payment to financial institution A. Here, we assume that a secret key K_{I-A} has been shared between I and A through some offline key exchange protocol. Actually, they can also achieve the authentication and key exchange using any public key negotiation protocol, because both I and A should be servers with sufficient computation power and they can perform public key cryptographic operation fast enough. I then sends A an encrypted *Transfer_Form*, which includrd the *Credit_Form* forwarded from B, and some other data, say, ID_I, ID_A, V, TN, and TS_I.

Step 5: Payment confirmation

On receiving the *Transfer_Form*, A should check his database for the recorded *Money_Claim_Form* to determine to which account owned by M to credit this fund. A first finds the corresponding *Money_Claim_Form* with the same TN as *Transfer_Form*, and checks whether the following equation holds:

$Credit_Form = H((Money_Claim_Form, ID_M, ID_A), Y_i)$ If the equation holds, A credits the due fund to M's account, and returns payment confirmation messages, including a secret session key K_{B-M}, to both M and$I$. The *Transfer_Response* to I also includes $H(Y_i)$, which is a non-repudiation evidence for

payment acceptance. Since only A and M know the secret session key Y_i. B and I can prove to any authorities that the payment has been made to$M$ successfully with the help of A. Furthermore, A sends a *Credit_Response* back to M, including the secret session key K_{B-M}. Note that in all response message, $H(\cdot)$ means a hash value on all above information in the form.

Finally, both B and M know that the payment is made successfully and they then start the secure transactions using the secret session key K_{B-M}. We suggest that each X_i and Y_i should be discarded after each transaction. Combining this point with the time stamp we applied, our protocol is immune to replying attack and keys guessing attack.

3 Security Analysis

3.1 Privacy Protection

Achievement of privacy protection of Buyer is the most significant property of our protocol, compared with other work in the literature. In our protocol, the privacy of Buyer is protected from not only Merchant, but also any third (trusted) parts, say, banks or other financial institutions. In Table 1, we give a comparison on privacy protection of our protocol with those other available work, including SET-based mobile payment protocol [2], iKP [1] and some other online mobile payment protocols [4, 5], and it is easy to see that only our protocol satisfies all privacy protection requirements.

Table 1. Comparison of privacy protection of our protocol and other work

Privacy Protection of buyer	SET-based protocol[2]	iKP[1]	[4]	[5]	Ours
ID privacy protection from merchant	No	No	No	Yes	Yes
ID privacy protection from eavesdropper	Yes	Yes	Yes	Yes	Yes
Transaction privacy protection from eavesdropper	Yes	Yes	Yes	Yes	Yes
Transaction privacy protection from TTP or related financial institution	No	No	No	No	Yes

3.2 Other Security Properties

We consider other security properties of our protocol in this section. Confidentiality is achieved since that all important information communicated in payment and transaction is encrypted using the secret session keys X_i, Y_i that are shared only among the communicating parties. We assure the integrity of data by including a message digest in *Transfer_Form*, *Money_Claim_Forms* and *Billing_Form*, so that the receiver can check whether the data is corrupted. To achieve mutual authentication between *Buyer* and *Merchant*, we use a symmetric key-based authentication mode, similar to Kerberos protocol [7]. Here, we assume that *Issuer* and *Acquirer* are trusted third parties and they authenticate

Buyer and *Merchant*, respectively. It should be noted that *Buyer* and *Merchant* may not be honest sometime, say, *Merchant* may repudiate that he has received the payment and may deny providing service or electronic goods to *Buyer*. In our protocol, however, such case can be resisted since all sensitive messages (say, *Order_Form*, *Money_Claim_Form*, *Credit_Form*) are included a MAC code which can only be generated by message sender.

4 Conclusion

In this paper, we propose an efficient online payment protocol for mobile network. We employ only efficient symmetric key algorithms in our protocol, including symmetric encryption, hash function and keyed hash function. All these operations can be implemented on mobile device feasibly. We also achieve the completely privacy protection of buyers, which has not been addressed in previous work. First, the identity of the buyer is protected from the merchant. Second, the transaction privacy of the buyer, such as what the buyer buys, and from whom the buyer buys, are protected from even TTP or related financial institutions. We analyze the protocol, and show that it satisfies all security requirements for mobile electronic payment.

References

1. Bellare, M., Garay, J., Hauser, R., Herzberg, A., Steiner, M., Tsudik, G., Van Herreweghen, E., and Waidner, M.: Design,Implementation, and Deployment of the iKP Secure Electronic Payment system. *IEEE Journal of Selected Areas in Communications.* **18** (2000) 611-627
2. Kungpisdan, S., Srinivasan, B., and Phu Dung, L.: A Practical Framework for Mobile SET Payment. In *Proceedings of 2003 International E-Society Conference* . (2003) 321-328
3. Huang, Z., and Chen, K.F.: Electronic Payment in Mobile Environment. In *Proceedings of the 13th Internatinal Workshop on DEXA.* (2002) 413-417
4. Kungpisdan, S., Srinivasan, B., and Phu Dung, L.: A Secure Account-Based Mobile Payment Protocol. In *Proceedings of International Conference on Information Technology: Cod-ing and Computing.* (2004) 321-328
5. Hu, Z.Y., Liu, Y.W., Hu X., and Li, J.H.: Anonymous Micropayments Authentication (AMA) in Mobile Data Network. In *Proceedings of IEEE INFOCOM 2004.* (2004) 46-53
6. Fourati, A., Ayed, H.K.B., and Benzekri, A.: A SET Based Approach to Secure the Payment in Mobile Commerce. In *Proceedings of 27th IEEE Conference on Local Computer Networks.* (2002) 136-137
7. Kohl, J., and Neuman, C.: The Kerberos Network Authentication Service (V5). *RFC 1510.* (1993)

Universal Designated-Verifier Proxy Blind Signatures for E-Commerce

Tianjie Cao[1,2,3], Dongdai Lin[2], and Rui Xue[2]

[1] School of Computer Science and Technology,
China University of Mining and Technology, Xuzhou 221008, China
[2] State Key Laboratory of Information Security,
Institute of Software of Chinese Academy of Sciences, Beijing 100080, China
[3] Graduate School of Chinese Academy of Sciences, Beijing 100039, China
{tjcao, ddlin, rxue}@is.iscas.ac.cn

Abstract. To protect the privacy of proxy blind signature holders from dissemination of signatures by verifiers, this paper introduces universal designated-verifier proxy blind signatures. In a universal designated-verifier proxy blind signature scheme, any holder of a proxy blind signature can convert the proxy blind signature into a designated-verifier signature. Given the designated-signature, only the designated-verifier can verify that the message was signed by the proxy signer, but is unable to convince anyone else of this fact. This paper also proposes an ID-based universal designated-verifier proxy blind signature scheme from bilinear group-pairs. The proposed scheme can be used in E-commerce to protect user privacy.

1 Introduction

Steinfeld et al. proposed a formalization of universal designated-verifier signatures [8]. These signatures are ordinary digital signatures with the additional functionality that any holder of a signature is able to convert it into a designated-verifier signature specified to any designated verifier of his choice. Given the designated-signature, the designated-verifier can verify that the message was signed by the signer, but is unable to convince anyone else of this fact. Steinfeld et al. also showed how to construct efficient deterministic universal designated-verifier signature schemes from bilinear group-pairs.

Proxy signatures was introduced by Mambo, Usuda and Okamoto [5, 6]. In a proxy signature scheme, original signer delegates her/his signing capability to proxy signer, and then the proxy signer can sign messages on behalf of the original signer.

The concept of blind signature scheme was introduced by Chaum [1]. A blind signature scheme is a protocol played by two parties in which a user obtains a signer's signature for a desired message and the signer learns nothing about the message.

Combining proxy signatures and blind signatures Lin and Jan introduced the first proxy blind signature scheme [4]. In a proxy blind signature scheme, the proxy signer is allowed to generate a blind signature on behalf of the original signer.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1036–1041, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In electronic commerce, private credential systems can be implemented through the use of blind signatures. A holder of private credential can simply send a copy of his/her credential to any interested verifier. On the other hand, the verifier can easily disseminate this credential and convince an unlimited number of third-party verifiers about the truth of the statements contained in the credential. This possibility poses a serious threat to the privacy of the holder. In this paper, we introduce universal designated-verifier proxy blind signatures. In a universal designated-verifier proxy blind signature scheme, any holder of a proxy blind signature can convert the proxy blind signature into a designated-verifier signature. Given the designated-signature, only the designated-verifier can verify that the message was signed by the proxy signer, but is unable to convince anyone else of this fact.

There are four entities in a universal designated-verifier proxy blind signature scheme: original signer, proxy signer, requester (holder) and designated verifier.

Original signer: Alice. The original signer can delegate her signing capability to a proxy signer.

Proxy signer: Bob. The proxy signer can create a valid blind signature on behalf of the original signer.

Requester (Holder): Cindy. The requester can acquire a signature on a message m without revealing anything about the message m to the proxy signer. When the requester obtains a blind signature on the message m, she can generate a designated verifier signature to prove to a designated verifier that it's she who actually possesses a blind signature on the message m.

Designated verifier: Dave. Only the designated verifier nominated by the holder of the proxy blind signature can verify the designated verifier signatures, but is unable to convince anyone else of the validity of the signature even the designated verifier reveals his private key.

A secure universal designated-verifier proxy blind signature scheme should satisfy the following requirements:

Unforgeability: Only the legitimate proxy signer can generate a valid proxy blind signature, even the original signer cannot. Only the holder of a legitimate proxy blind signature can generate a valid designated verifier signature.

Restrictive verifiability: Only the designated verifier nominated by the holder of the proxy blind signature can verify the designated verifier signatures.

Non-transferability: A designated-verifier can not produce evidence which convinces a third party that a message was signed by the proxy signer even if the designated verifier would like to reveal his private key.

Identity based public key cryptography is a paradigm proposed by Shamir in 1984 [7] to simplify key management and remove the necessity of public key certificates. In this paper, we also propose an ID-based universal designated-verifier proxy blind signature scheme from bilinear group-pairs.

An ID-based universal designated-verifier proxy blind signature scheme is usually comprised of the following phases.

Setup: An algorithm that takes a security parameter k as input and returns **params** (system parameters) and **master-key**. Intuitively, the system parameters will be publicly known, while the **master-key** will be known only to the "Private Key Generator" (PKG).

Extract: An algorithm that takes as input **params**, **master-key**, and an arbitrary $ID \in \{0,1\}^*$, and returns a private key d_{ID}. Here ID is an arbitrary string that will be used as a public key, and d_{ID} is the corresponding private key. The Extract algorithm extracts a private key from the given public key.

Proxy Key Pair Generation: An algorithm that takes as input the original signer and proxy signer's private keys, a warrant, outputs a proxy key pair for the proxy signer. Only the proxy signer knows the proxy private key, while proxy public key is public or publicly recoverable. The warrant specifies the original signer, the proxy signer and other application dependent delegation information explicitly.

Proxy Blind Signature Generation: An interactive protocol between the proxy signer and the requester. The public input contains proxy public key. The private input of the proxy signer contains the proxy private key, and that for the requester contains message m. When they stop, the public output of the requester contains either completed or not completed, the private output of the requester contains either "fail" or a proxy blind signature on message m.

Proxy Blind Signature Verification: An algorithm that takes a proxy blind signature as input, returns 1 or 0 meaning accept or reject the information that the signature is valid with respect to a specific original signer and a proxy signer.

Designation: An algorithm that takes as input a warrant, a verifier's identity, a message and a proxy blind signature, outputs a designated verifier signature.

Designated Verification: An algorithm that takes as input a warrant, a verifier's identity, a message, a designated verifier signature, designated verifier's secret key, returns 1 or 0 meaning accept or reject.

In this paper, our main contribution is that we introduce the notion of universal designated-verifier proxy blind signatures and design an ID-based universal designated-verifier proxy blind signature scheme from bilinear group-pairs.

2 ID-Based Universal Designated-Verifier Proxy Blind Signature Scheme

Let G_1 be a cyclic additive group generated by P, whose order is a prime q, and G_2 be a cyclic multiplicative group with the same order q: Let $e : G_1 \times G_1 \to G_2$ be a bilinear pairing with the following properties:

Bilinearity: $e(aP, bQ) = e(P, Q)^{ab}$ for all $P, Q \in G_1$, $a, b \in Z_q$.

Non-degeneracy: There exists $P, Q \in G_1$ such that $e(P, Q) \neq 1$, in other words, the map does not send all pairs in $G_1 \times G_1$ to the identity in G_2.

Computability: There is an efficient algorithm to compute $e(P, Q)$ for all $P, Q \in G_1$.

The proposed ID-based universal designated-verifier proxy blind signature scheme can be built from any bilinear map $e : G_1 \times G_1 \to G_2$ between two

groups G_1, G_2 as long as computational Diffie-Hellman problem (CDHP) in G_1 is hard.

Setup. Let $(G_1, +)$ and $(G_2, \cdot)$ denote cyclic groups of prime order q, let P be a generator of G_1 and the bilinear pairing is given as $e : G_1 \times G_1 \rightarrow G_2$. Pick a random $s \in Z_q^*$ and set $P_{pub} = sP$, choose three cryptographic hash function $H_1 : \{0,1\}^* \rightarrow G_1^*$, $H_2 : \{0,1\}^* \times G_1^* \rightarrow Z_q^*$, $H_3 : \{0,1\}^n \times G_1^* \rightarrow Z_q^*$, $H_4 : G_2^* \rightarrow Z_q^*$. The message space is $M = \{0,1\}^n$. The system parameters are $params =< q, G_1, G_2, e, n, P, P_{pub}, H_1, H_2, H_3, H_4 >$. The master-key is $s \in Z_q^*$.

Extract. For a given string $ID \in \{0,1\}^*$ the PKG computes $Q_{ID} = H_1(ID)$, and sets the private key d_{ID} to be $d_{ID} = sQ_{ID}$ where s is the master key.

Proxy Key Pair Generation. To delegate the blind signing capability to Bob, Alice and Bob have agreed on a warrant m_w, which states application dependent delegation information such as the identity information of the original signer and the proxy signer. In this phase, the original signer Alice can use Dong et al.'s proxy key pair generation algorithm directly [3]. Here based on Cheng et al.'s ID-based signature scheme [2] we construct a more efficient algorithm to generate proxy key pair.

Step 1. Alice chooses a random number $r_1 \in Z_q^*$ and computes $U_1 = r_1 P$, $V_1 = r_1 P_{pub} + H_2(m_w, U_1) d_{ID_A}$. Alice sends (m_w, U_1, V_1) to Bob.

Step 2. Bob checks the validity of the warrant m_w. Bob accepts (m_w, U_1, V_1) by checking whether the following equation holds:

$$e(P_{pub}, U_1 + H_2(m_w, U_1) Q_{ID_A}) = e(P, V_1)$$

Step 3. If (m_w, U_1, V_1) passes the checking, Bob computes his proxy key pair as $d_p = V_1 + H_2(m_w, U_1) d_{ID_B}$ and $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$. Bob can use d_p to sign any message which conforms to m_w on behalf of the original signer Alice.

Proxy Blind Signature Generation. In our scheme, the proxy signer Bob uses Zhang and Kim's ID-based blind signature scheme [9] to generate the proxy blind signature for the user Cindy.

Step 1. The proxy signer Bob randomly chooses a number $r_2 \in Z_q^*$, computes $U_2 = r_2 Q_p$, and sends U_1, U_2 and m_w to the user Cindy.

Step 2. Cindy computes the proxy public key $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$. Cindy randomly chooses $\alpha, \beta \in Z_q^*$ as blinding factors. She computes $U_2' = \alpha U_2 + \alpha\beta Q_p$ and $h = \alpha^{-1} H_3(m, U_2') + \beta$, sends h to Bob.

Step 3. Bob sends back V_2, where $V_2 = (r_2 + h) d_p$.

Step 4. Cindy computes $V_2' = \alpha V_2$. Then (m_w, U_1, U_2', V_2') is the proxy blind signature of the message m.

Proxy Blind Signature Verification. Cindy accepts the signature if and only if

$$e(P_{pub}, U_2' + H_3(m, U_2') Q_p) = e(P, V_2')$$

Designation. If Cindy wants to prove to a designated verifier Dave that the possession of a blind signature $(m_w, U_1, U_2', V_2', m)$, she can do the follows. Cindy computes $\delta = H_4(e(V_2', T Q_{ID_D}))$ where $T \in Z_q^*$ is a timestamp. Cindy sends

$(m_w, U_1, U_2', \delta, T, m)$ to Dave. 7-tuple $(m_w, U_1, U_2', \delta, T, m, ID_D)$ is a designated verifier signature.

Designated Verification. Dave computes the proxy public key $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$ and checks the timestamp T. Dave accepts the proof if and only if

$$\delta = H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D}))$$

The following theorems show that our scheme gives expected result if all the parties perform honestly.

Theorem 1. *If 5-tuple $(m, m_w, U_1, U_2', V_2')$ is a proxy blind signature produced by the proposed scheme, the requester Cindy will accept it.*

Theorem 2. *If 7-tuple $(m_w, U_1, U_2', \delta, T, m, ID_D)$ is a designated verifier signature produced by the holder of a proxy blind signature $(m_w, U_1, U_2', V_2', m)$ in the proposed scheme, the designated verifier Dave will accept it.*

3 Security

Unforgeability: (1) Only the proxy signer can generate the proxy blind signature. In proxy key pair generation phase, only the proxy signer nominated by the original signer can generate the proxy blind signature private key d_p. In proxy blind signature generation phase, the proxy signer Bob uses Zhang and Kim's ID-based blind signature scheme to generate the proxy blind signature. The unforgeability depends on the security of Zhang and Kim's scheme. (2) Only the owner of a proxy blind signature can generate a valid designated verifier signature. If 7-tuple $(m_w, U_1, U_2', \delta, T, m, ID_D)$ generated by Cindy is a designated verifier signature then $\delta = H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D}))$ where $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$. Since $\delta = H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D})) = H_4(e(sU_2' + H_3(m, U_2')d_p, TQ_{ID_D}))$ and Cindy cannot know Dave's private , then Cindy must hold $sU_2' + H_3(m, U_2')d_p$. Let $V_2' = sU_2' + H_3(m, U_2')d_p$, we can verify that $(m_w, U_1, U_2', V_2', m)$ is a legitimate proxy blind signature. We note that for a given message m the proxy signer Bob can also be a holder of the proxy blind signature on message m.

Restrictive verifiability: The designated verifier Dave nominated by the holder of the proxy blind signature, with the secret key d_{ID_D}, can verify the designated verifier signature $(m_w, U_1, U_2', \delta, T, m, ID_D)$ through the equation $\delta = H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D}))$ where $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$. Anyone without the secret key cannot compute $H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D}))$.

Non-transferability: The designated-verifier Dave nominated by Bob cannot prove to a third party that the designated verifier signature $(m_w, U_1, U_2', \delta, T, m)$ is generated by the holder of the proxy blind signature because the receiver Dave can also generates $(m_w, U_1, U_2', \delta, T, m)$ through computing $\delta = H_4(e(U_2' + H_3(m, U_2')Q_p, Td_{ID_D}))$ where $Q_p = U_1 + H_2(m_w, U_1)(Q_{ID_A} + Q_{ID_B})$.

4 Conclusion

Combining universal designated-verifier signatures and proxy blind signatures, we have introduced universal designated-verifier proxy blind signatures for E-commerce. In this paper, we have also proposed an ID-based universal designated-verifier proxy blind signature scheme based on Zhang and Kim's blind signatures.

Acknowledgments

We thank the support of the National Grand Fundamental Research 973 Program of China (No.2004CB318004), the National Natural Science Foundation of China (NSFC90204016, NSFC60373048) and the National High Technology Development Program of China under Grant (863, No.2003AA144030).

References

1. Chaum, D.: Blind signatures for untraceable payments, Advances in Cryptology - CRYPTO' 82, Plenum (1983) 199–203
2. Cheng,X., Liu,J., Wang,X.:An Identity-Based Signature and Its Threshold Version, Proceedings of the 19th International Conference on Advanced Information Networking and Application, vol. 1. (2005)973–977
3. Dong, Z., Zheng,H., Chen, K., Kou, W.:ID-based proxy blind signature, Proceedings of the 18th International Conference on Advanced Information Networking and Application, vol. 2. (2004)380–383
4. Lin, W.D., Jan,J.K.: A security personal learning tools using a proxy blind signature scheme, Proceedings of International Conference on Chinese Language Computing, Illinois, USA. (2000)273–277
5. Mambo,M., Usuda,K., Okamoto, E.: Proxy signature: Delegation of the power to sign messages. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E79-A, No. 9. (1996)1338-1353
6. Mambo,M., Usuda,K., Okamoto, E.: Proxy signatures for delegating signing operation. Proceedings of 3rd ACM Conference on Computer and Communications Security, ACM Press. (1996)48–57
7. Shamir,A.: Identity-based cryptosystems and signature schemes, Advances in Cryptology-Crypto 84, Lecture Notes in Computer Science, Vol. 196, Springer-Verlag. (1984)47–53
8. Steinfeld, R., Bull, L. ,Wang, H., Pieprzyk, J.: Universal Designated Verifier Signatures, Proceedings of Asiacrypt'03, Lecture Notes in Computer Science, Vol. 2894, Springer-Verlag.(2003)523–542
9. Zhang, F., Kim, K: Efficient ID-Based Blind Signature and Proxy Signature from Bilinear Pairings, ACISP'03, Lecture Notes in Computer Science, Vol. 2727, Springer-Verlag. (2003)312–323

An Efficient Control Method for Elevator Group Control System

Ulvi Dagdelen[1], Aytekin Bagis[1], and Dervis Karaboga[2]

[1] Erciyes University, Dept. of Electronic Eng., 38039, Kayseri, Turkey
[2] Erciyes University, Dept. of Computer Eng., 38039, Kayseri, Turkey
{dagdelen, bagis, karaboga}@erciyes.edu.tr

Abstract. This paper presents an efficient control approach for the elevator group control system. The essential of the method used is based on an operation strategy with a talented algorithm. In order to analyze the performance of the presented system, the control method was evaluated by considering different performance characteristics in the elevator group control system. The results of the presented method compared with the results of the area weight algorithm.

1 Introduction

Elevator group control systems (EGCS) are control systems managing to assign an appropriate service car of multiple elevators in a building in order to efficiently transport the passengers waiting in a hall. The main aim of the optimal elevator group control is to obtain the highest handling capacity, the shortest waiting and/or traveling time of passengers, and the minimum power consumption.

Since the controller considers many different cases, the EGCS is a difficult and very complex control problem. In the EGCS, there are many uncertain factors such as number of passengers, hall calls, and car calls in any time. If a group controller manages m elevators and assigns n hall calls to the elevators, the controller considers m^n cases. Hence, the complexity in controlling an elevator group increases exponentially as the number of floor and number of elevator increase in the system. Furthermore, it must be possible for a skilled system operator to change the control strategy since some operators want to operate the system to minimize the passenger waiting time while others do to reduce the power consumption in the elevator control system [1].

Some methods used to achieve the elevator group control have been presented in open literature [1-8]. The most significant drawbacks of these approaches are to exhibit the limited improvement in the overall performance of the group control system and to have various specific parameters to be carefully adjusted in the group control procedure when the system is highly complex and the system state cannot be easily estimated.

In this paper, a more simple and efficient control method than those mentioned in the previous paragraph is presented for the elevator group control with in-traffic mode that differs from our previous work [9] with out-traffic mode. The method is based on an operation strategy with a talented algorithm. In the operation strategy, the area

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1042 – 1047, 2005.
© Springer-Verlag Berlin Heidelberg 2005

weight algorithm is employed by the talented algorithm. The EGCS is introduced in the second section. The proposed control problem is defined in Section 3. In this section, the operation approach and simulation results are also presented. The results are discussed in Section 4.

2 Elevator Group Control System (EGCS)

The structure of an EGCS used for simulation purposes in this paper is shown in Fig. 1. The EGCS consists of the passenger generation, group controller, elevator controller and monitoring units [9].

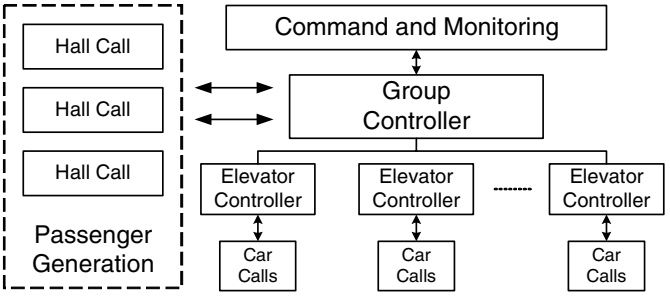


Fig. 1. Architecture of Elevator group controller

Hall and car information of calls are kept in group controller module and the arrangement of the passengers is realized by using the information about the elevator cars. There are two task lists for hall calls and car calls containing the passengers getting on and already existing in the elevator car, respectively. Stop and way information for the cars is determined by using this task list. When the car is charged for a given way, it does not give any response to requests in the reverse direction.

In elevator controller module, the current position and target floor of the elevator car are compared and the car then move to the target floor if there is a difference between the comparison. There is some information in this module such as the passengers in the car, target floor and the number of the passengers carried on a round trip.

The selection of the elevator car is made in order to minimize the average waiting time of passengers, the long wait probability and service time. This selection method of the appropriate elevator is called the hall call assignment method. In this method, an evaluation function is used to achieve the above multiple objectives. The function is evaluated for each elevator car and the elevator car with the smallest function value is selected. Let $\phi(k)$ be the evaluating function for the k[th] elevator, and then this function can be represented with the following formula [10]:

$$\phi(k) = (\sum_{stop} T_{stop}(k) + \sum_{drive} T_{drive}(k)) - \alpha T_{\alpha}(k), \quad \text{k=1, 2, 3..... N} \tag{1}$$

In Equation (1), stop term means floors where hall calls and car calls are assigned, and drive term means floors where there are no calls near the floor. α and T_{α}(k) repre-

sent the area weight and the area value, respectively. Area weight is a weighting factor for area value and it affects the performance of EGCS by the patterns of passenger traffic. The value of the T_{α}(k) is determined by the positions of assigned calls for each elevator.

3 Proposed Talented Control Algorithm and Simulation Results

Many evaluation criteria can be used to estimate the performance of the EGCS. In this paper, four main factors are considered by this control algorithm: (1) *Average waiting time (AWT)* is the time until the service elevator arrives at the floor after a passenger presses a hall call button. AWT is the average of all waiting times in a unit time. (2) *Average journey time (AJT)* is the total time a passenger spends within an elevator system first by waiting and then by riding inside a car. AJT is the average of all journey times in a unit time. (3) *Long waiting percent (LWP)* is the percentage of the passengers who wait more then 60 sec in a unit time. (4) *Run count (RNC)* is the number of the elevator start and stop moves during the running, and it is used to obtain an opinion about the power consumption of a system [1, 11-14].

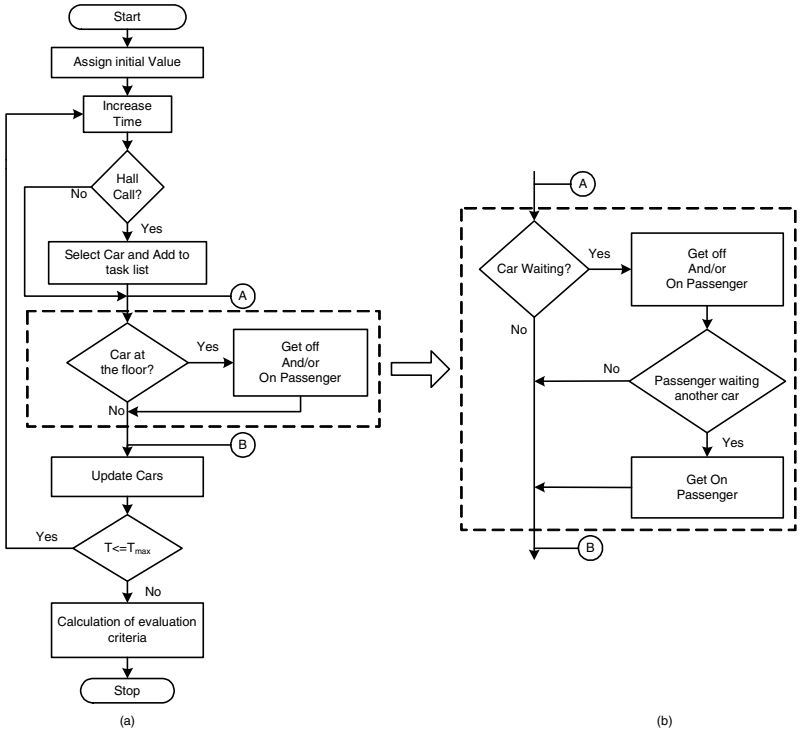


Fig. 2. (a) Flowchart of the AW algorithm, (b) Modification in the algorithm to perform the proposed talented algorithm

The flowchart of the general area weighting (AW) algorithm is given in Fig. 2a. In this flowchart, proposed talented control algorithm is also represented as the dotted area. In the AW algorithm, firstly, the main parameters such as the number of floors in the building, the number of elevator cars, speed and capacity of elevator cars, time that elevator car stops at floor to load/unload users are preserved by the algorithm. The situation of the passengers is randomly generated. For the each elevator car, a task list is obtained, and this list is updated in the predetermined time intervals. The performance of the algorithm is computed by using the evaluation criteria at the end of this process.

Table 1. Simulation conditions

Number of floors-person	16 floors-50 to 200 persons per 5 min.
Number of elevator-capacity	6 cars-20 persons
Elevator speed	2 seconds per floor
Stopping time	14 seconds
In traffic	1st to other floors : 50% Others to 1st floor : 25% Others to other floors : 25%

The main feature of the proposed talented algorithm is with the flexible task lists of the elevator cars. When a car stops at any floor, its task list is compared with the task lists of the other cars at the different floors. If the other cars have a task at this floor, these tasks are undertaken by this car at this floor (Fig. 2b). Thus, the time consumption is significantly decreased by the proposed algorithm. The simulation conditions are given Table 1.

For the evaluation criteria given above, the simulation results are graphically presented in Fig. 3. In this figure, the (AW) algorithm and the proposed talented algorithm are compared to each other for different performance indexes such as average waiting time (AWT), average journey time (AJT), long weighting percent (LWP), and average stopping time (RNC). Since the value of α significantly affects the performance of AW, as mentioned in Section 2, it has a dynamic and a limited range. As can be seen, for example in [14], that, its minimum value is zero while the maximum value is around 40. Therefore, the simulations were repeated for different value of α (0, 15, 30 and 40) [14].

From the figures, it is clearly shown that the presented talented algorithm is a simple and efficient control approach for the elevator group control systems. The average waiting time for the passengers is the most crucial factor in the elevator group control. As seen from Fig. 3a, this factor of the proposed control algorithm is less than that of the AW algorithm. As the number of passenger increases, this factor is significantly decreased. For example, when the number of passenger is 200, the AWT value of the talented algorithm is 47.32 sec. For the same condition, the minimum value of the AWT of area weight algorithm is 77.37 sec ($\alpha = 40$).

From Fig. 3b, it can be clearly seen that the average journey time for the talented algorithm is also shorter than the area weight algorithm. For a numerical example, the differences between the average journey times of the talented algorithm and the area weight algorithm for 200 person are 49.26 sec and 31.89 sec for $\alpha = 0$ and $\alpha = 40$,

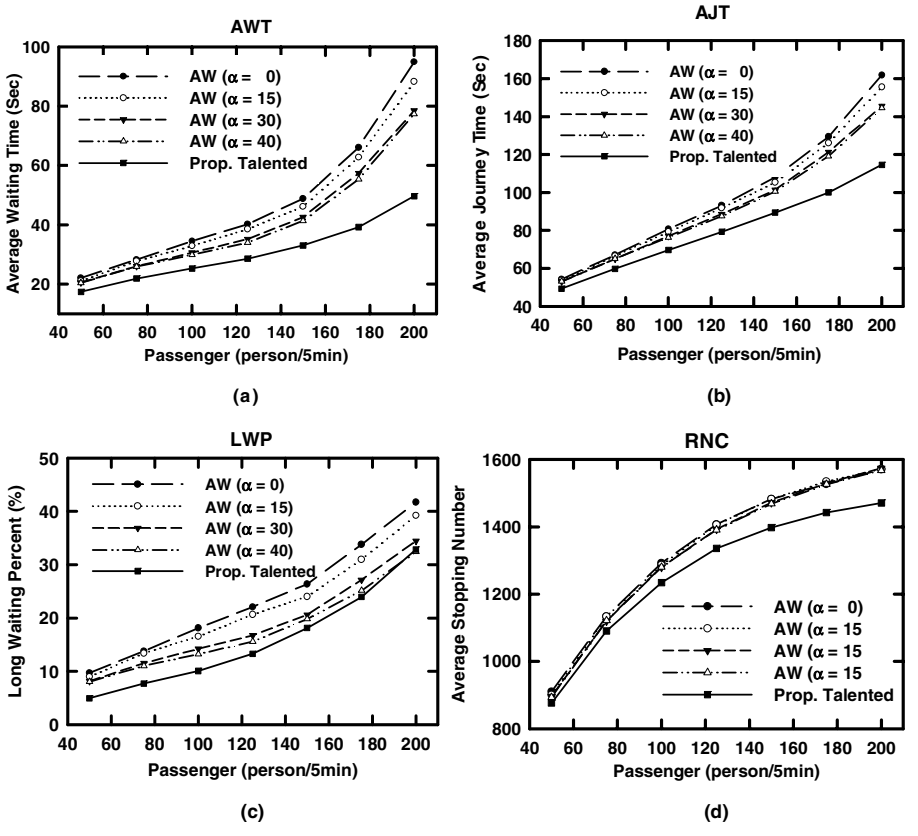


Fig. 3. (a) Average waiting time, (b) Average journey time, (c) Long waiting percent, (d) Average stopping number

respectively. As apparently seen from Figs. 3c-3d, long waiting percent of the passengers and average stopping number for the elevator car is considerably decreased in the proposed algorithm. In an elevator system, the major power consumption comes from the acceleration and deceleration stages [15]. Hence, significant energy savings is expected with the reduction in the number of starts/stops per hour of elevator operation; this can be accomplished with the aid of the proposed talented algorithm.

4 Conclusion

In this paper, an efficient control approach based on a talented algorithm is presented for elevator group control system. When it is compared with the area weighting algorithm, for any condition of the passengers, it is verified that the proposed control algorithm is a very effective and reliable control method with the minimum power consumption requirements of the elevator system as well as time savings of the passengers. The presented algorithm is also capable to adopt for different process parameters such as the number of elevator car, capacities of the cars, floor numbers, different traffic situations, and it is a distinctive advantage of the algorithm.

References

1. Kim, C.B., Seong, K.A., Kwang, H.L., Kim, J.O.: Design and Implementations of a Fuzzy Elevator Group Control System. IEEE Trans. on Systems Man and Cybernetics-Part A 28(3) (1998) 277–287
2. Ishikawa, T., Miyauchi, A., Kaneko, M.: Supervisory Control for Elevator Group by Fuzzy Expert System Which Also Addresses Traveling Time. IEEE International Conference on Industrial Technology Proceeding (2000) 87–94
3. Tobita, T., Fujino, A., Segawa, K., Yoneda, K., Ichikawa, Y.: A Parameter Tuning Method for an Elevator Group Control System Using a Genetic Algorithm. Electrical Engineering in Japan 124(1) (1998) 55–64
4. Fujino, A., Tobita, T., Segawa, K., Yoneda, K., Togawa, A.: An Elevator Group Control System with Floor–Attribute Control Method and System Optimization Using Genetic Algorithm. IEEE Trans. On Industrial Electronics 44(4) (1997) 546–552
5. Cho, Y.C., Gagov, Z., Kwon, W.H.: Elevator Group Control with Accurate Estimation of Hall Call Waiting Times. IEEE Proceeding of Te International Conference on Robotics&Automation (1999) 447–452
6. Ho, Y.W, Fu, L.C.: Dynamic Scheduling Approach The Group Control of Elevator Systems with Learning Ability. IEEE Proceeding of The International Conference on Robotics&Automation (2000) 2410–2415
7. Gudwin, R., Gomide, F., Netto, M.A.: A Fuzzy Group Controller with Linear Context Adaptation. Fuzzy Systems Proceedings (1998) 481–486
8. Crites, R.H., Barto, A.G.: Elevator Group Control Using Multiple Reinforcement Learning Agents. Machine Learning 33 (1998) 235–262
9. Dagdelen, U., Bagis, A., Karaboga, D.: Elevator Group Control by Using Talented Algorithm. The 14th Turkish Symposium on Artificial Intelligence and Neural Networks (2005) 423–430
10. Kim, C.B., Seong, K.A., Kwang, H.L., Kim, J.O., Lim, Y.B.: A Fuzzy Approach to Elevator Group Control System. IEEE Trans. on Systems Man and Cybernetics 25(6) (1995) 985–990
11. Barney, G.: Elevator Traffic Handbook. Spon Press London (2003)
12. Strakosch, G.R.: Vertical transportation: Elevators and escalators. Wiley-Interscience NewYork (1983)
13. Peters, R.D.: The Theory and Practice of General Analysis Lift Calculations. In Elevcon Proceedings (1992) 197–206
14. Kwang, H.L., Kim, C.B.: Techniques and Applications of Fuzzy Theory to an Elevator Group Control System. Fuzzy Theory Systems: Techniques Applications Vol. 1 Academic Press (1999)
15. Law, Y.W., So A.T.P.: Energy–Saving Schemes for Lift and Escalator-A Review. Hong Kong Engineer (1991) 24–28

Next Generation Military Communication Systems Architecture

Qijian Xu[1], Naitong Zhang[1], Jie Zhang[2], and Yu Sun[3]

[1] Harbin Institute of Technology, Harbin, China
xuqijian@126.com, ntzhang@hope.hit.edu.cn
[2] Department of ordnance engineering, Naval Aeronautical Engineering Institute, Yantai, China
zhj_txzh@126.com
[3] Institute No.54th, China electronics technology group corporation, Shijiazhuang, China

Abstract. Next generation military communication system (NGMCS) is described as a kind of typical information communication networks (ICN), and primary requirement and evaluation criterion of NGMCN architecture are summarized briefly. Then a general architecture model for NGMCN, composed of four sections, application, system, technology, and equipment, is set up based on vector space and Fuzzy Sets theory. Further the modeling method is presented with the discussions in detail on fuzzy sets and fuzzy relations associated with the architecture model.

1 Introduction

Along with the abroad applications of high and new technologies in military area, information is becoming a key factor of battle effectiveness for a future army. Making information power already became the restrict post for rivalry both sides in future information warfare. Military communication systems as important part of C^4I will always play a *nerve center* for informatization of arm equipment, and automation of command and control.

Next generation military system (NGMCS) as a kind of typical information communication networks (ICN)[1], should capable to support a wide range of services and information flows, including real time/streaming/non-real time services, point-to-point, multipoint, broadcast and multicast voice, data, video and multi-media applications. The challenges to develop NGMCS man faced now are when and how to build what a kind of architecture so that can direct the technique development, equipment manufacture, networks construction, and services delivery to satisfy various communication demands of diverse customers from different army services, for different operation of war.

This paper tries to find a general method of modeling the architecture of military communication system, based on vector space and Fuzzy Sets theory is presented, and to built up an general architecture for NGMCS.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1048 – 1053, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Requirement and Evaluation Criterion of NGMCS Architecture

2.1 Requirement of NGMCS Architecture

NGMCS, a kind of ICN, as a target network of next generation networks should meet the basic requirements:(1)Provide all kinds of services for information communication; (2)Easily introduce new technologies; (3)Sufficiently utilize existing resources; (4)Has an open and scalable architecture. Those primary requirements should extend to a requirements set for NGMCS, notes *R*, which is commonly a fuzzy set.

2.2 Evaluation Criterion of NGMCS Architecture

In order to evaluate a NGMCS architecture for advance, validity and efficiency, the evaluation criterion set should be established. Two aspects of which, as primary contents can be listed below.

General nature. (1)Systematization— cover all aspects of NGMCS inside and outside; (2)Integrity— all components of the architecture encircle one common target of NGMCS; (3)Pertinency— describe relations such as interoperability among all elements of NGMCS; (4)Hierarchization— divide functions of NGMCS into different layers; (5)Expansibility— make development and evolvement of NGMCS smoothly.

Characteristics. (1)Architectural strain— architecture promoting continuance development of NGMCS; (2)Universal services— supporting all existing and developing services; (3)Network convergence — covering existing kinds of network systems such as military telephone network, computer network and video network; (4) Technology relevancy — compatible all existing and developing network technologies; (5)System flexibility— adapted to all application environment such strategic or tactical, fixed or maneuverable, on land, ocean, air and space.

Above criterion need to be made in more details according with real engineering application thereby performing an evaluating criterion set, notes P, also a fuzzy set.

3 Model of NGMCS Architecture

3.1 Definition of Architecture Model

Definition 1：Model of Modeling for NGMCS Architecture, Ω
Consist of requirement set *R*, architecture set *X*, and their relation set *Y*, commonly are fuzzy sets. That is

$$\Omega=\{R, X, Y\} \tag{1}$$

Where，*X* consist of application architecture *A*, system architecture *S*, technology architecture *T*，and equipment architecture *E*, that is

$$X=\{A, S, T, E\} \tag{2}$$

Y is a set of relations between every two elements in $\{R, A, S, T\}$, that is

$$Y=\{ R_{ij} \},\ i, j\in \{R, A, S, T, E\} \tag{3}$$

Ignore some subordinate relations according with research on a real application; conveniently, Y can be predigested as

$$Y=\{R_{RA}, R_{AS}, R_{ST}, R_{AT}, R_{SE}, R_{TE}\} \tag{4}$$

Based on definition 1, the architecture for NGMCS consists of four parts: application architecture *A*, system architecture *S*, technology architecture *T* and equipment architecture *E*.

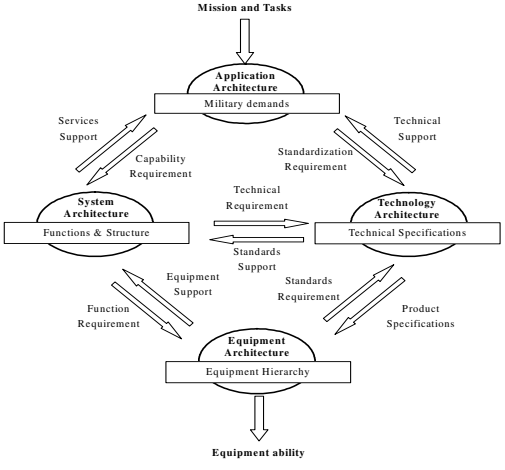


Fig. 1. Components and their relations of NGMCS architecture model

Definition 2: Requirement, *R*

Suppose the total requirements attributes space for information communication is $U=\{U_1, U_2, \ldots\ldots, U_p\}$, where, $U_i \bigcap U_j = \phi$, $i \neq j$, $i, j \leq p$, $U_i=\{u_{1i}, u_{2i}, \ldots\ldots, u_{ni}\}$. Then the requirement for a specific NGMCS, *R*, is a fuzzy sub-set on *U*, i.e. $R \in F(U)$. Its subjection function is $R(r)=(r_1,r_2,\ldots\ldots,r_n)$, $n=\sum_{i=1}^{p} n_i$, $0 \leq r_j \leq 1$, $j=1,2,\ldots,n$, which describes the numbers and correlation levels of requirement elements of the NGMCS.

Said above, the requirements for NGMCS are in various aspects, such as service types, information transmission capability, quality of service (QoS), equipment availability, etc. Each aspect also includes many different demands, for instance, services types include voice, data, fax, multimedia etc. Therefore *R* is a two layer model.

Definition 3: Application architecture, *A*

Suppose the services attributes space of NGMCS providing to customers is $V=\{V_1,V_2,\ldots\ldots,V_q\}$, where, $V_i \bigcap V_j = \phi$, $i \neq j$, $i, j \leq q$, $V_i=\{v_{1i}, v_{2i}, \ldots\ldots, v_{mi}\}$. Then the Application architecture of a specific NGMCS, *A*, is a fuzzy set on *V*, i.e. $A \in F(V)$. Its subjection function is $A(a)=(a_1, a_2, \ldots\ldots, a_m)$, $m=\sum_{i=1}^{q} m_i$, $0 \leq a_j \leq 1$, $j=1,2,\ldots,m$, which describes the numbers and correlation levels of elements in NGMCS application architecture.

The architecture of NGMCS also has a hierarchy model. Application architecture of NGMCS represents the relationship between customers and NGMCS which provide services for them. It can be divided into two planes, tactical plane and technical plane. The former describes tactical communication requirement such as types, scope and grade of communication support ability of NGMCS; and the latter describes the application services NGMCS should support, and defines every service attributes NGMCS providing to subscribers, such as service types, representation media, quality of service(QoS), etc. Each attribute also includes various sub-attributes, for instance, service types including message, retrieving, conversation, distribution, etc. Therefore *A* is a two layer model.

In the same way, we can give definitions based on system space, *V*, technology space, *Q*, and equipment space, *P*, of system architecture, *S*, representing the functional components and logical construction of NGMCS; technology architecture, *T,* giving a set of principles describing configuration, relation and interaction of various components in a NGMCS; and equipment architecture, *E*, gives equipment hierarchy composing the NGMCS and describes application arrangement of the equipment.

3.2 Fuzzy Relations in the Model

Hereinbefore, the architecture of NGMCS can be set up by some operations on fuzzy sets step by step. Obviously, first step is how to get requirement set *R* and its subjection function *R(r)*. Generally, it could only be obtained through consultation on experts who have much practical engineering experiences. Fuzzy relation matrixes given in formula (4), the relations among fuzzy sets *R, A, S, T*, and *E*, might also be gotten, generally, only by experts consultation. Those fuzzy relations are discussed and defined as follows.

Every element in requirement space *U* may be relative to numbers of element in application architecture space *V*, and the relation degrees are various. Whereas, every element in application architecture space *V*, generally, does not support only a specific element in requirement space *U*. Therefore, a fuzzy transform from *U* to *V* exist and can map the requirement R to application architecture *A* for a specific NGMCS.

Definition 4: Requirement-Application Relation R_{RA}
A dualistic fuzzy relation from requirement space *U* to application architecture space *V*, i.e. $R_{RA} \in F(U \times V)$. Space *U* and *V* are all finite space, so the subjection function is a fuzzy matrix.

$$R(r,a)=(r_{ij})_{n\times m}, \quad r_{ij} \in [0, 1] \tag{5}$$

Where, r_{ij} describes depending level of the element number *i* in requirement space *U* on the element number *j* in application architecture space *V*.

In the similar way, we can give definitions of application-system relation, R_{AS}, system-technology relation R_{ST}, application-technology relation R_{AT}, system-equipment relation, R_{SE}, and technology-equipment relation, R_{TE}.

Above relations are enough considered in a practical engineering, although inverse relations of them exist theoretically. Subjections of above fuzzy relations could only be obtained through consultation on experts who have much practical engineering experiences.

3.3 Getting the Architecture

According with definitions and discussions before-mentioned, the fuzzy sets of NGMCS architectures can be gotten through following fuzzy sets synthesis operations, depending on giving requirement *R(r)*, requirement-application relation *R(r,a)*, application-system relation *R(a,s)*, system-technology relation *R(s,t)*, application-technology relation *R(a,t)*, system-equipment relation *R(s,e)* and technology-equipment relation *R(t,e)*.

$$A=R\circ R_{RA} \tag{6}$$

$$S=A\circ R_{AS} \tag{7}$$

$$T=(A\circ R_{AT})\bigcup(S\circ R_{ST}) \tag{8}$$

$$E=(S\circ R_{SE})\bigcup(T\circ R_{TE}) \tag{9}$$

Arithmetic operators in above fuzzy synthesis operation formulas generally are *Zadeh* operators.

$$A(a)=\bigvee_{r\in U}(R(r)\wedge R(r,a)) \tag{10}$$

Denotation in matrix, $R(r)=(r_i)_{1\times n}$, $R(r,a)=(r_{ji})_{n\times m}$, $r_{ij}\in[0,\ 1]$, $A(a)=(a_i)_{1\times m}$, where, $a_i=\bigvee_{j=1}^{n}(r_j\wedge r_{ji})$, $i=1,2,\ldots,m$. In the same way,

$$S(s)=\bigvee_{a\in V}(A(a)\wedge R(a,s)) \tag{11}$$

Denotation in matrix, $A(a)=(a_i)_{1\times m}$, $R(a,s)=(\tau_{ji})_{m\times L}$, $\tau_i\in[0,\ 1]$, $S(s)=(s_i)_{1\times L}$, where, $s_i=\bigvee_{j=1}^{n}(a_j\wedge\tau_{ji})$, $i=1,2,\ldots,L$. Similarly,

$$T(t)=(\bigvee_{a\in V}(A(a)\wedge R(a,t)))\vee(\bigvee_{s\in W}(S(s)\wedge R(s,t))) \tag{12}$$

Denotation in matrix, $A(a)=(a_i)_{1\times m}$, $R(a,t)=(\mu_{ji})_{m\times k}$, $\mu_{ij}\in[0,\ 1]$, $S(s)=(s_i)_{1\times L}$, $R(s,t)=(\gamma_{ji})_{L\times k}$, $\gamma_{ij}\in[0,\ 1]$, $T(t)=(t_i)_{1\times k}$, where, $t_i=(\bigvee_{j=1}^{m}(a_j\wedge\mu_{ji}))\vee(\bigvee_{j=1}^{L}(s_j\wedge\gamma_{ji}))$, $i=1,2,\ldots,k$. As such,

$$E(e)=(\bigvee_{s\in W}(S(s)\wedge R(s,e)))\vee(\bigvee_{t\in Q}(T(t)\wedge R(t,e))) \tag{13}$$

Denotation in matrix, $S(s)=(s_i)_{1\times L}$, $R(s,e)=(v_{ij})_{L\times b}$, $v_{ij}\in[0,\ 1]$, $T(t)=(t_i)_{1\times k}$, $R(t,e)=(w_{ij})_{k\times b}$, $w_{ij}\in[0,\ 1]$, $E(e)=(e_i)_{1\times b}$, where,

$$e_i=(\bigvee_{j=1}^{L}(s\wedge\gamma_{ji}))\vee(\bigvee_{j=1}^{k}(t_j\wedge w_{ji})),\ i=1,2,\ldots,b \tag{14}$$

In practical engineering, other operators might be used in order to get more perfect results. A clear architecture for a specific NGMCS can be conducted by clearing the fuzzy architecture sets and evaluation.

4 Conclusion

Along with the development of information science and technology, military communication system has being evolved from the era of telegram，through telephone and data times, now to a information age. Next generation military communication system (NGMCS) is emerging as a kind of information communication network (ICN). Advanced military communication system architecture should accord with primary requirement and evaluation criterion. Then a general NGMCS architecture model based on vector space and Fuzzy Sets theory is set up with four components: *A, S, T*, and *E*, i.e. application architecture, system architecture, technology architecture, and equipment architecture. Those can be gotten through a series of synthesis operations on fuzzy sets and fuzzy relations associated with them. Of course this paper only presents a primary method of modeling architecture for military communication system, and the results will be just the basis of further study.

References

1. Xu Qijian: Studies on Informaiton Communication Networks' Architecture. High Technology Letters, Vol.12 No.2. ISSN 1002-0470 cn 11-2770/N. (2002) 16-20
2. Abdi R. Modarressi, Seshadri Mohan: Control and management in next-generation networks: Challenges and opportunities. IEEE Communications Magazine, vol. 38, no. 10, October (2000) 94-102
3. J. M. Bennett: Voice over packet reliability issues for next generation networks, ICC 2001-IEEE International Conference on Communications, No.1. (2001) 142-145
4. Y. Inoue et al.: The TINA Consortium. IEEE Commun. Mag.,Vol.36(9).(1998) 130-136
5. ITU-T Recommendation Y.130: Information communication architecture. (2000)
6. Xu Qijian: Comments on IP Framework. ITU-T SG13 delayed contribution D896. ITU-T SG13 Kyoto meeting. Feb/Mar (2000)

Early Warning for Network Worms

Antti Tikkanen and Teemupekka Virtanen

Helsinki University of Technology
{antti.tikkanen, teemupekka.virtanen}@hut.fi

Abstract. Network worms, very similar to viruses, are malicious programs that use vulnerabilities in software to spread between computers that are somehow connected using a computer network.

We remind that a computer should be understood broadly - in the near future, a worm might infect a mobile phone as easily as it now infects a personal desktop computer.

We have proposed a modular architecture for an early warning system. We also implemented a prototype system consisting of several detection and analysis modules. While the prototype was limited in nature, it meets the most of the requirements we have set for such a system.

Keywords: Network worm, early warning, malicious code.

1 Introduction

Network worms depend on vulnerabilities in software. Perhaps not surprisingly, even as awareness on how to write secure software is improving, the number of reported vulnerabilities is on the rise. It should be clear that as the size and complexity of software continues to grow, the number of vulnerabilities will also keep rising. Another issue, noted by [1], is something the software vendors know as the delivery challenge: the ever shortening delivery times for software mean shorter development cycles, which easily leads to lower system quality.

Recent worm outbreaks have caused notable disruptions in computing systems all over the world. However, the potential losses due to a malicious, well-planned network worm are even more devastating: a study by Weaver et al. [2] estimates worldwide losses in excess of $100 billion in a worst-case scenario.

This work attempts to solve the problem of automatically detecting and analyzing new network worms. We will focus on evaluating and constructing an early warning system that attempts to detect and analyze new, rapidly spreading malicious mobile code.

2 Recent Trends with Network Worms

By looking at how worms have evolved through their short time of existence, we hope to gain some insight into what might come in the near future. We will examine a very small subset of worms that have been seen in the wild.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1054–1059, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The Morris worm [3] was the first wide-spread worm. There have been three worms associated with the name Code Red, the first of which caused major disruptions on July 12th, 2001. The Slammer worm was the fastest worm so far, doubling the amount of infected machines every 8.5 seconds [4]. On August 11th, 2003, the Blaster worm, also known as MSBlast and Lovsan, began spreading on the Internet. Data from Microsoft suggests that it was able to infect as many as 8 million hosts, The Witty worm was released on March 19, 2004. The vulnerable population was approximately 12,000 hosts, and it took 45 minutes for the worm to reach full saturation [5]. Shannon et al. discuss some of the distinctive features of Witty. The Sasser worm was released on April 30, 2004. It used a previously published exploit to attack Windows 2000, Windows 2003 and Windows XP operating systems. The worm was self-activated and used second-channel distribution by transferring the worm body via File Transfer Protocol (FTP).

Finding patterns and common behavior in worms from the recent years is important. An early warning system, by definition, is only useful if it can detect new, previously unseen worms. Almost every worm seen in the last decade or so has conformed to a similar release process. The timeline begins after a vendor publishes a patch for a vulnerability that has the suitable characteristics. After this an exploit is published. After a period that seems to be getting shorter each time, a worm that uses this exploit is released.

By looking at the structure of the latest worms, we can see patterns getting more and more common. The target discovery function is usually a simple random scanning algorithm, possibly with local preference. A previously published exploit is copied almost as-is into the worm. The entire worm body is delivered using a second channel server.

3 An Early Warning System for Network Worms

3.1 General Architecture

As input, the system receives network traffic that is either sent or received by the hosts we are monitoring. The internals of the system can be divided into two separate layers. As we shall describe, each layer is modular, so that functionality can be either changed or complemented easily.

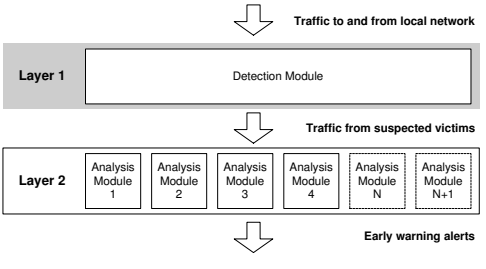


Fig. 1. Early warning system architecture overview

The first layer has only one purpose. It detects suspected infections by looking at the traffic to and from the hosts being monitored. The detection module must be very light-weight, because it will see and analyze all traffic sent and received by the hosts we are monitoring. The detection must be accurate, because all hosts tagged as suspected by this layer will generate an alert.

The second layer consists of many different modules. The purpose of these modules is to generate more information about the hosts we suspect are infected; we call it the analysis layer. These modules receive traffic only from the hosts that have been tagged as suspected by the first layer.

After an infected host has been detected by the first layer and the infection analyzed by the second, the system generates as output an early warning alert.

3.2 Implementation

The early warning system prototype has been implemented as a preprocessor module for Snort, the open source intrusion detection system.

Layer 1, which is responsible for tagging suspected hosts, consists of a single module. The second layer will consist of many modules, responsible for analyzing hosts that have been tagged as suspects by the first layer.

A technique that best fits the requirements set for the system is described in [6]. The mathematic details are discussed in [7]. Threshold Random Walk (TRW) is an algorithm that attempts to detect scanning hosts by looking at the connection attempts they make. [7] note that the key requirement is prompt detection.

[6] describe how TRW should be implemented efficiently both in hardware and in software by approximating certain characteristics. We call their implementation Approximated Threshold Random Walk (ATRW). We closely follow the implementation of ATRW, leaving out the containment functionality we do not need.

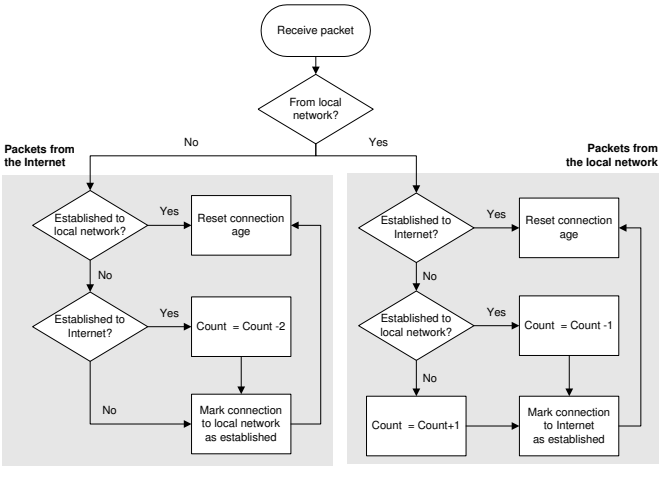


Fig. 2. Flow diagram of ATRW

ATRW, at a very high level, works simply like this. We keep a record of the state of each connection and the state of all the hosts we are monitoring. Each failed connection attempt raises the anomaly count of a host, while each successful connection lowers it. Figure 2 shows the flow diagram for ATRW.

4 Test Results and Analysis

The test data we used consisted of traffic captures from two different source networks:

- A /20 network at a university campus, with approximately 3,000 computers connected to it. The data from this network contains only headers. All local network addresses were sanitized. The test data includes:
  - Capture 1: A capture of 10,000,000 packets, from a time span of 5 minutes and 11 seconds with data from 1,215 hosts in the local network.
  - Capture 2: Another capture of 10,000,000 packets, from a time span of 4 minutes and 8 seconds with data from 1,179 hosts.
- A controlled test bed of three virtual hosts connected to the Internet, with different operating systems and applications installed. The data from this network contained full traffic information. This data includes a single capture, namely:
  - Capture 3: A capture of 47,155 packets, including all the most common networking protocols (SMB/CIFS, HTTP, HTTPS, DNS, SSH, FTP and TELNET).

4.1 Detection Layer Test Results

To test the ATRW detection module, we used both traffic captures from the university campus network (capture 1 and capture 2).

We used a threshold value of T=10, so that the module tagged all hosts exceeding this value. To determine the actual cause of why each host exceeded

Anonymized IP	Maximum count	Comment
217.44.9.201	5821	eDonkey
217.44.91.17	**5202**	**worm or trojan**
217.44.57.219	**3312**	**worm or trojan**
217.44.200.69	1437	eDonkey
217.44.124.81	315	eDonkey, Gnutella
217.44.202.5	152	eDonkey
217.44.164.108	92	eDonkey
217.44.216.74	65	eDonkey, Gnutella
217.44.55.91	62	BitTorrent
217.44.190.21	46	BitTorrent
217.44.14.88	42	eDonkey, Gnutella
217.44.62.179	33	eDonkey, Gnutella

Fig. 3. Outbound hosts tagged by ATRW with threshold 10 (capture 1)

Anonymized IP	Maximum count	Comment
217.44.4.16	**9170**	**worm or trojan**
217.44.93.17	**9126**	**worm or trojan**
217.44.122.67	**1862**	**worm or trojan**
217.44.200.69	1135	eDonkey
217.44.102.201	**1057**	**worm or trojan**
214.44.221.133	333	eDonkey, Gnutella
217.44.14.88	86	eDonkey, Gnutella
214.44.21.209	43	eDonkey
217.44.190.21	33	BitTorrent
217.44.124.81	29	eDonkey, Gnutella
217.44.33.102	13	BitTorrent

Fig. 4. Outbound hosts tagged by ATRW with threshold 10 (capture 2)

this limit, we examined packet information manually. The results can be seen in Figure 3 and Figure 4. The second column indicates the highest anomaly count that the host reached during the capture.

As we see, all entries from the first capture are results of peer-to-peer traffic or true positives in the form of a worm or trojan scanning the network. Both true positives were sending traffic to port 445/tcp, but without payloads, we could not determine which vulnerability they were scanning.

The results for the second capture are very similar. The two captures were taken a week apart, so the infected hosts from the previous capture are no longer online, but we still have four new worm or trojan infections, again scanning port 445/tcp.

From these results, we can reach certain conclusions. As noted in [6] and [8], peer-to-peer traffic can be a source of false positives with any technique that attempts to detect worms by the way they scan for new victims.

We find that ATRW is very suitable to be used as the detection module for our early warning system. The single deficiency is its lack of ability to detect non-scanning worms.

4.2 Analysis Layer Test Results

Overall, we have shown that the architecture and implementation of our system meets the requirements we set reasonably well.

The one shortcoming that needs to be solved, hopefully with future research, is the completeness of early warning systems: there does not exist a reliable method of detecting non-scanning worms. On the other hand, requirements of informational value and transparency are met by the design, implementation and selected platform very well.

During the design of our system, we had to take into account a number of conflicts arising from the number of requirements. These included:

- Transparency vs. informational value. To gather more information, the state of the hosts could be queried directly from them. Requiring software installations, this conflicts with the requirement of transparency.

- Performance vs. informational value. Even though the analysis modules have limited effect on the performance of the system, the number of the modules is limited.
- Transparency vs. performance. If we want the system to be able to handle heavy traffic loads, a partial hardware implementation is a possible solution. Dedicated hardware conflicts with the transparency requirement.

5 Conclusions

The evolution of network worms from the early days of the Morris worm has been interesting and also very disturbing. The continuous change in the field of malicious code requires constant research and improvements to ensure a working infrastructure.

We have proposed a modular architecture for an early warning system. We also implemented a prototype system consisting of several detection and analysis modules. It meets the most of the requirements we have set for such a system.

References

1. Sommerville, I.: Software Engineering, 6th edition. Pearson Education Ltd., UK (2001)
2. Weaver, N., Paxson, V., Staniford, S.: A worst-case worm (2003) Silicon Defense Technical Report.
   Available: http://www.silicondefense.com/papers/worstcase.pdf.
3. Spafford, E.: The internet worm program: An analysis (1998) Purdue Technical Report CSD-TR-823.
4. Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., Weaver, N.: The spread of the sapphire/slammer worm (2003) Available:
   http://www.caida.org/outreach/papers/2003/sapphire/sapphire.html.
5. Shannon, C., Moore, D.: The spread of the witty worm (2004) Available:
   http://www.caida.org/analysis/security/witty/.
6. Weaver, N., Staniford, S., Paxson, V.: Very fast containment of scanning worms. In: Proceedings of the 13th Usenix Security Symposium, USA, Usenix (2004)
7. Jung, J., Paxson, V., Berger, A., Balakrishnan, H.: Fast portscan detection using sequential hypothesis testing. In: 2004 IEEE Symposium on Security and Privacy, USA, IEEE (2004)
8. Schechter, S., Jung, J., Berger, A.: Fast detection of scanning worm infections. In: Recent Advances in Intrusion Detection (RAID), USA (2004)

Skeleton Representation of Character Based on Multiscale Approach

Xinhua You[1], Bin Fang[3], Xinge You[1,2],
Zhenyu He[2], Dan Zhang[1], and Yuan Yan Tang[1,2]

[1] Faculty of Mathematics and Computer Science, Hubei University, China
[2] Department of Computer Science, Hong Kong Baptist University
xyou@comp.hkbu.edu.hk
[3] Chongqing University

Abstract. Character skeleton plays a significant role in character recognition. This paper presents a novel algorithm based on multiscale approach to extract skeletons of printed and hand-written characters. The development of the method is inspired by some desirable characteristics of the modulus minima of wavelet transform. Namely, the local minima of wavelet transform are scale-independent and locate at the medial axis of the symmetrical contours of character stroke. Thus it is particularly suitable for characterizing the inherent skeletons of character strokes. The proposed skeletonization algorithm contains two major steps. First, by thresholding for the modulus minima of wavelet transform, the modulus minima points underlying the character strokes are extracted as the primary skeletons. Based on these primary skeletons, the modulus minima points are being eventually computed as the final skeleton by iteratively performing wavelet transform. The skeleton form the proposed method can be exactly located on the central line of the stroke, and the artifacts and branches of skeletons from traditional methods can be avoided. We tested the algorithm on handwritten and printed character images. Experimental results indicate that the proposed algorithm is applicable to not only binary image but also gray-level image.

1 Introduction

The skeleton is especially suitable for describing character since they have natural and inherent axes [7, 10, 11]. The thinned characters are widely used for recognition [7]. At present, it covers a wide range of applications, such as graphic recognition, handwriting recognition, signature verification [5], etc. From a practical point of view, the skeleton-based representation of characters by sets of thin curves rather than by a raster of pixels may consume less storage space and processing time, and be sufficient to describe their shapes for many applications. It has found that this representation is particularly effective in extracting relevant features of the character for optical character recognition [6].

The objective of skeletonization is to find the medial axis of a character. Skeleton has been defined in various ways in the different literatures. All kinds

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1060–1067, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of definitions result in a variety of skeletionization algorithms as well. However, it is generally agreed that a good skeleton should have some good properties. For instance, a skeleton should be homotopic to the underlying shape, and it should be thin and well-centered, it should have a pleasing visual appearance.

A great deal of skeletonization techniques have been proposed. All these algorithms of skeletonization can be classified into symmetry-based and nonsymmetry-based approach [1, 3, 18].

An essential challenge for computing the primary skeleton of the characters by symmetry technique is that it is difficult to decide the symmetric pairs from the boundary of the character strokes in both continuous and discrete domains [1]. Especially, in the discrete domain, it may not be possible to find their symmetric counterparts on the opposite contour such that they form the local symmetries. Even if a pair of symmetric contour pixels can be exactly found, their corresponding symmetric center may not exist exactly.

The "quasi-symmetric" approach is a sound solution for this problem [6, 14, 18]. Some typical approaches are the regularity-singularity analysis technique, which is based on the Constrained Delaunary Triangulation [18], Principal Curve [6] and cross-sections [3].

Unfortunately, their implementation often suffers from the complicated computation and the resulting skeletons are affected by contour noise. Moreover, these approaches produce the skeleton with strong deviation from the center of the underlying character strokes and rough human perceptions [18]. The resulting skeletons are inaccurate.

Meanwhile, character strokes contain a variety of intersections and junctions, For such cases, another major problem with the traditional symmetry-based skeletonization methods is that their results do not always conform to the human perceptions, since they often contain unwanted artifacts such as noisy spurs and spurious short branch in the junctions of character strokes [3, 4, 7, 11, 17]. It is difficult to detect junctions correctly [2, 11]. To remove all sorts of artifacts from the primary skeleton is still another challenging topic.

Many non-symmetric approches have been proposed to compute skeletons. There are four types of non-symmetric-based methods, which are based on thinning techniques [7], distance transform[11], and Voronoi diagrams [9].

Thinning algorithms produce homotopic skeletons and are easy to implement. However, thinning results are not guaranteed to medial due to the use of discrete data and local operations [7, 9]. Second, skeletons obtained may not preserve geometric properties of the underlying shapes. Third, their computation complexity is high. It usually takes $O(n^3)$ [11] time to compute the skeleton of an $n \times n$ image by thinning.

The recently proposed technique based on maximum modulus symmetry of wavelet transform(MMSWT)[12, 16] improved greatly in these respects. Anyway, it has a poor performance with wide-structure objects. Moreover, the computation cost depends fully on the structure width of object. The wider of shape is, the higher the computation cost is. Therefore, the MMSWT-based algorithm is not suitable for computing skeleton of object with relatively wide structure.

In addition, as we mentioned in our previous discussion [12], it is still puzzled how to choose proper scale for wavelet transform according to the width of shape in practice.

In this paper, a new multiscale approach is proposed to directly extract the skeleton of the character strokes. This new approach is based on the desirable properties of the quadratic spline wavelet function. It shows that these desirable properties are particularly suitable to be used to characterize and compute inherent skeleton of character strokes. The proposed algorithm is also applicable for computing skeletons of character strokes with distinct width structures, but the algorithm based on maximum modulus symmetry analysis fail to deal with them in [12].

2 Characteristics of Modulus Minima of the Quadratic Spline Wavelet

A typical multiscale analysis of image is implemented by smoothing the surface with a convolution kernel $\theta(x, y)$ that is dilated at dyadic scales s. Such an edge detector can be computed with two wavelet functions that are the partial derivative of $\theta(x, y)$. $\psi^1(x, y) := \frac{\partial}{\partial x}\theta(x, y)$ and $\psi^2(x, y) := \frac{\partial}{\partial y}\theta(x, y)$. Let us denote $\theta_s(x, y) := \frac{1}{s^2}\theta(\frac{x}{s}, \frac{y}{s})$. For wavelets $\psi^i(x, y),\ i = 1, 2$, defined above, their scale wavelet transforms can be defined as in [12].

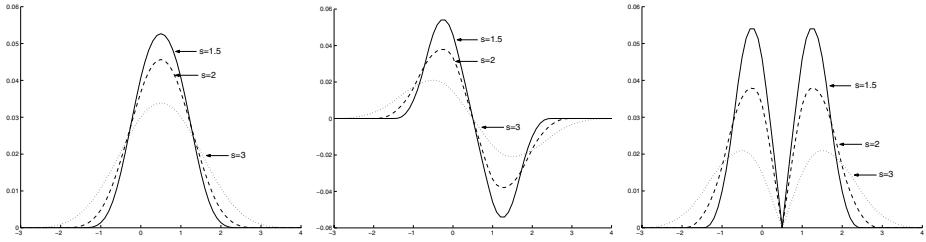
Most of multiscale shape representation or analysis methods are based on Gaussian kernel. However, as pointed out by Mokhtarian and Mackworth [8], the representation should satisfy the efficiency and ease of implementation in order that it be useful for practical shape recognition tasks in computer vision. Obviously, Gaussian kernel does not satisfy such criteria since the computational burden is high at large scales and it is not compact support as well. It has been shown in [13, 15] that that the following quadratic B-spline wavelet perform better than other wavelets for singularities detection applications.

$$\psi(x) := \begin{cases} 24|x|^3/x - 16|x|^2/x, & \text{if } |x| \le 0.5, \\ -8|x|^3/x + 16|x|^2/x - 8|x|/x, & \text{if } 0.5 \le |x| \le 1, \\ 0 & \text{if } |x| \ge 1. \end{cases} \tag{1}$$

Where the corresponding smoothing function $\theta(x)$ is given by

$$\theta(x) := \begin{cases} 8|x|^3 - 8|x|^2 + 4/3, & \text{if } |x| \le 0.5, \\ -(8/3)|x|^3 + 8|x|^2 - 8|x| + 8/3, & \text{if } 0.5 \le |x| \le 1, \\ 0 & \text{if } |x| \ge 1. \end{cases} \tag{2}$$

As well known, B-spline is a good approximation to Guassian kernel. Besides inheriting almost all the good properties of Gaussian scale-space, B-spine derived scale-space exhibits many advantages [12]. Here we proved mathematically the following properties of B-spline wavelet function, which are shown that they are especially suitable to describe the skeletons of character strokes.



(a) (b) (c)

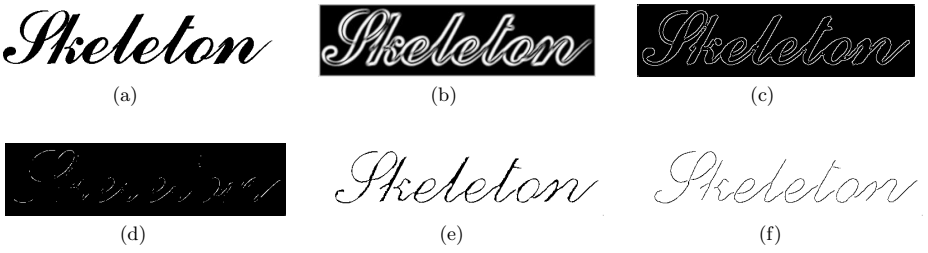
Fig. 1. (a) The convolution $f * \theta_s(x)$ of the original image $f(x)$ and the smoothing function $\theta_s(x)$; (b) The wavelet transform $W_s f(x)$; (c) The moduli of wavelet transform $W_s f(x)$

Theorem. *Let l_d be a straight segment of character stroke with width d. If the scale of the wavelet transform $s \geq d$, the local minima of the WT modulus locate in the mid position of the stroke segment and are fixed at different scales, especially, for some fine scale s, this point exists uniquely between the two consecutive local maximum. Thus the local modulus minima of the WT of the wavelet function locate at the center line of l_d and the location of the modulus minima are independent of the scale of WT. Two corresponding modulus maxima are symmetric with respect to this modulus minimum along the gradient direction.*

In fact, the local minimum locates in the center between the two consecutive local maximum and describes the slowest change points of the underlying character stroke. In order to make the analysis more clear, we consider the following example for one dimension case. It is illustrated by Fig. 1. Where let f be the piecewise function on $(-\infty, \infty)$, i.e., $f(x) := \begin{cases} (x-0.5)^2, & if\ 0 \leq x \leq 1, \\ 0, & otherwise. \end{cases}$ be regarded as analysed signal. Obviously, the local minima of the WT characterize a particular class of the slowest change points of the underlying stroke. In this paper, these desirable properties of minima moduli based on B-spline wavelets will be used to develop multiscale-based algorithm for extracting skeletons of character. Consequently, a set of simple and direct strategies for characterizing and extracting skeleton of the underlying character stroke is strongly motivated by modulus minima analysis of the WT.

3 Multiscale-Based Algorithm and Experiments

Our basic idea for algorithm implementation is as follow: For each image analysed, we randomly choose a scale of wavelet transform and compute its corresponding all local minimum moduli. Thus a primary skeletons which consist of multi-pixels and enclose the inherent central curves of the underlying character stroke can be obtained. They are apparent thinned ribbons from the original stroke with thinned down width structure. For the image contains these primary skeletons, we choose much smaller scale than prior one to perform the



(a) (b) (c) (d) (e) (f)

Fig. 2. (a) Pattern consisting of English character; (b) The raw output of modulus obtained from wavelet transform with scale $s = 6$; (c) Modulus maxima image after performing the wavelet transform ; (d) The raw output of skeletons obtained from the MMSWT [12]; (e) Initial skeletons obtained by the first WT with scale $s = 6$; (f) Final skeletons obtained by the second WT with scale $s = 4$



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

Fig. 3. (a) The original image; (b) The image of the modulus maxima of the first wavelet transform with $s = 6$; (c) The primary skeletons extracted by the proposed algorithm; (d) The final skeletons obtained by performing the second WT.

second wavelet transform on it and compute the second level primary skeletons. Repeating the above computing procedure until the central curves which consist of single pixels are extracted eventually. The corresponding algorithm is designed as follows:

Algorithm 1. *Let $f(x, y)$ be an image which contains the underlying characters.*

(1) *Select the initial scale to perform the first wavelet transform with respect to the quadratic spline wavelet.*
(2) *Calculate all moduli of the wavelet transform* $\{W_s^1 f(x, y),\ W_s^2 f(x, y)\}$.
(3) *Set an initial threshold T, for every point of the underlying stroke, compare its wavelet transform modulus $|\nabla W_s f(x, y)|$ with the initial threshold and preserve all points whose modulus are less than an initial threshold T as the candidate points of initial skeleton.*
(4) *Repeat Step (1) until every skeleton curve are connected by single pixel one by one.*

The recently proposed technique based on maximum modulus symmetry of wavelet transform (MMSWT)[12] improves greatly in traditional methods. Anyway, it has a poor performance with wide-structure objects. Especially, it perform poorly for objects that it happens the some cases beleow: The distance between two ribbon-like structures (e.g. parts of a character) is so close, or structure widths of objects in the same image have the strong difference as shown in Fig. 2(a). The ideal scale of wavelet transform which is suitable for skeletonization objects with various different width structure is not decided easily even does not exists completely. As a result, some natural skeleton points of the character can not be extracted completely as shown in Fig. 2(d). It is shown that the algorithm based on the MMSWT failed to find proper scale to implement it due to contrast change of width structures of the character object as shown in Fig. 2(e). In addition, as we mentioned in our previous discussion [12], it is still puzzled how to choose proper scale for wavelet transform according to the width of shape in practice.

Here an original image with English character "skeleton" is shown in Fig. 2(a). Apparently, the width of the character structures sharply changes in the global object. Without any width structure information being provided beforehand, we applied the proposed multiscale approach to computing its primary skeletons, as shown in Fig. 2(e). As expected, by perform the second WT for the primary image in Fig. 2, the final skeletons, as shown in Fig. 2(c), are extracted successfully and comply well with the geometrical structure of the underlying character.

In Fig 3, we choose an image which consists of Chinese characters. By computing the modulus minima of the first WT with scale $s = 6$, the primary skeletons are obtained in Fig. 3(c). Further, to extract final skeleton, we implement the second detection of minimum moduli of the WT with the scale $s = 4$. The final skeleton is shown in Fig. 3(d).

4 Conclusions

We propose a simple and efficient scheme for thinning characters based on multiscale approach. The new scheme depend on some significant properties of the modulus minima of wavelet transform with the quadratic spline function. Compared with existed traditional method, the proposed scheme has the following features 1) The extracted skeleton is centered inside the underlying stroke in the mathematical sense while the skeleton obtained by most of the existing algorithms slightly deviates the center; 2) The skeletons from the proposed technique retain sufficient information of the original characters and have pleasing visual appearance; 3) The proposed method is also applicable for computing skeleton of the underlying characters with various complicated width structures; 4) The scale of the WT can be selected freely independent of the width of the character strokes over the drawbacks of the modulus-maximum-symmetry-based method; 5) The algorithm improves greatly in the poor performance which most traditional methods are used to computing skeleton in the intersection of the

underlying shape. However, how to choose the denoising threshold are currently under investigation.

Acknowledgments

This research was partially supported by a grant (60403011 and 10371033) from National Natural Science Foundation of China and a grant (2003ABA012) and (20045006071-17) from Science &Technology Department, Hubei province and Wuhan respectively, China. This research was also supported by the grants (RGC and FRG) from Hong Kong Baptist University.

References

1. Blum, H.: "Biological shape and visual science (part1)" *J. Theoret, Eds. Biology.* (1973)
2. brandt, J. W., Algazi, V. R. Continuous skeleton computation by voronoi diagram. *CVGIP: Image Understanding.*, 55 (1992) 329–338
3. Chang ,H. S., Yan, H.: "Analysis of Stroke Structures of Handwritten Chinese Characters". IEEE Trans. Systems, Man, Cybernetics (B), 29 (1999) 47–61
4. Ge, Y., Fitzpatrick, J. M.: "On the Generation of Skeletons from Discrete Euclidean Distance Maps". IEEE Trans. Pattern Anal. Mach. Intell., 18 (1996) 1055–1066
5. Janssen, R. D. T.: "Interpretation of maps: from bottom-up to model-based". *H. Bunke and P. S. P. Wang, Eds., Handbook of Character Recognition and Document Image Analysis.* World Scientific publisher, Singapore (1997)
6. Kgl B., Krzyżak, A.: Piecewise linear skeletonization using principal curves. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(1) (2002) 59–74
7. Lam, L. , Lee, S. W. , Suen, C. Y.: "Thinning Methodologies - a Comprehensive Survey". IEEE Trans. Pattern Anal. Mach. Intell., 14 (1992) 869–885
8. Mokhtarian, F., Mackworth,A. K.: "A theory of multiscale curvature-based shape representation for planar curves". IEEE Trans. Pattern Anal. Mach. Intell., 14(8) (1992) 789–805
9. Ogniewicz, R. L., Kubler, O.: "Hierarchic Voronoi skeletons". Pattern Recognition, 28 (1995) 343–359
10. Rosenfeld,A.: "Axial Representations of Shapes". Comput. Vis. Graph. Image Process., 33 (1986) 156–173
11. Smith,R. W. : "Computer Processing of line images: a survey". Pattern Recognition, 20 (1987) 7–15
12. Tang,Y. Y., You, X. G.: Skeletonization of ribbon-like shapes based on a new wavelet function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9) (2003) 1118–1133
13. Unser, M., Aldroubi, A., Eden, M.: B-spline signal processing: Part ii-efficient design and applications. *IEEE Trans. on SP.*, 41(2) (1993) 834–847
14. Wang, C., Cannon, D. J., Kumara, R. T., Lu,G.: "A Skeleton and Neural Network-Based Approach for Identifying Cosmetic Surface Flaws". IEEE Trans. Neural Networks, 6(5) (1995) 1201–1211
15. Wang,Y.P., Lee,S. L. : "Scale-space derived from B-splines". IEEE Trans. Pattern Anal. Mach. Intell., 20(10) (1998)1040–1050

16. Yang, L. H., You, X.G., Haralick, R. M., Phillips, I. T., Tang,Y.Y. : "Characterization of Dirac Edge with New Wavelet Transform". In Proc. 2th Int. Conf. Wavelets and its Application, 1 (2001) 872–878
17. Zou, J. J. , Yan, H.: Extracting strokes from static line images based on selective searching. Pattern Recognition, 32 (1999) 935–946
18. Zou, J. J. , Yan, H.: "Skeletonization of Ribbon-Like shapes Based on Regularity and Singularity Analyses". IEEE Trans. Systems. Man. Cybernetics (B), 31(3) (2001) 401–407

Channel Equalization Based on Two Weights Neural Network

Wenming Cao, Wanfang Chai, and Shoujue Wang

Collage of Information engineering, Zhejiang University of Technology,
Hangzhou, Zhejiang 310014, China
wmcao@semi.ac.cn

Abstract. In this paper, we discuss the application of a two weights neural network (TWNN) to the channel equalization problem. In particular, the purpose of the paper is to improve the previously developed TWNN equalizer with training using K-means and LMS methods; reducing the TWNN network size by considering a lesser number of TWNN kernels, and developing new techniques for determining channel order which is required to specify the structure of an TWNN equalizer. A linear regression model was used for estimating the channel order. The basic idea of reducing the network size is to select the centers, based on the channel lag. This work includes the comparison of the limits of mean square error (MSE) convergence of both a linear equalizer and an TWNN equalizer. mean square error.

1 Introduction

For several years, there has been attention given to applying neural networks to the channel equalization problem[1]. Recently, the TWNN [2,3,4] has received a great deal of attention from many researchers because of its structural simplicity and more efficient learning. The TWNN network consists of 2 layers (hidden, and output layer). Each hidden unit produces the norm distance (usually Euclidean) between its own direction weights and kernel weights and the network input. The norm distance is then divided by a kernel spread parameter and the result is passed through a nonlinear function. The output response of a TWNN network is a mapping F,

$$Y = F\left[\sum_{j=1}^{n}\left(\frac{W_j}{|W_j|}\right)^{s} \left| \; W_j\left(X_j - W_j^{'}\right) \; \right|^{P} - \theta\right] \tag{1}$$

An application of two weights neural networks to channel equalization has been discussed in some articles [5]. Through the paper, we present a new approach for channel order estimation and a learning algorithm for reducing the number of kernels used for training., and compared with the linear equalizer cast. From the simulation results, we conclude that the training of TWNN equalizer with the proposed method is much faster than that of the previously developed one with the conventional method.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1068 – 1073, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Mathematical Description of Two Weights Neural Network

A mathematical description of two weights neural network *P* of class *A*:

$$P = U_i P_i \tag{2}$$

$$P_i = \left\{ x \mid \rho(x, y) \le k, y \in \mathrm{B}_i, x \in R^n \right\}, \tag{3}$$

$$B_i = \left\{ x \mid x = \alpha S_i + (1-\alpha) S_{i+1}, \alpha = [0,1] \right\} \tag{4}$$

Where S_i is an information data of class A in information space. Let

$$d\left(x, \overline{x_1 x_2}\right) = \min_{\alpha \in [0,1]} d\left(x, \alpha x_1 + (1-\alpha) x_2\right) \tag{5}$$

be the distance of *x* and line segment $\overline{x_1 x_2}$. And the two weights neural networks $\mathrm{S}(x_1, x_2; r)$ is:

$$S(x_1, x_2; r) = \left\{ x \mid d^2\left(x, \overline{x_1 x_2}\right) < r^2 \right\} \tag{6}$$

If x_1 and x_2 are the same information data in information space, then $d\left(x, \overline{x_1 x_2}\right)$ is equivalent to $d(x, x_1)$, and $\mathrm{S}(x_1, x_2; r)$ is equivalent to $S(x_1; r)$ (two weights neural networks becomes a hypersphere). If x_1 and x_2 are different points, then a two weights neural networks $S(x_1, x_2; r)$ is a connection between x_1 and x_2, with compact coverage compared to a hypersphere.

A new neuron model, a two weights neural networks, is defined by the input-output transfer function:

$$f(x; x_1, x_2) = \phi\left(d\left(x, \overline{x_1 x_2}\right)\right) \tag{7}$$

Where $\phi(\cdot)$ is the non-linearity of the neuron, the input vector $\mathrm{x} \in R^n$, and $x_1, x_2 \in R^n$ were two centers. A typical choice of $\phi(\cdot)$ was the Gaussian function used in data fitting, and the threshold function was a variant of the Multiple Weights Neural Network (MWNN)[6,7,8]. The network used in data fitting or system controlling consisted of an input layer of source nodes, a single hidden layer, and an output layer of linear weights. The network implemented the mapping:

$$f_s(x) = \lambda_0 + \sum_{i=1}^{n_s} \lambda_i \varphi\left(d\left(x, \overline{x_{i1} x_{i2}}\right)\right) \tag{8}$$

Where $\lambda_i, 0 \le i \le n_s$ were weights or parameters. The output layer made a decision on the outputs of the hidden layer. One mapped

$$f_s(x) = \max_{i=1}^{n_s} \phi\left(d\left(x, \overline{x_{i1}x_{i2}}\right)\right) \tag{9}$$

Where $\phi(\cdot)$ was a threshold function.

3 Estimation of Channel Order

Linear regression models [1] were used fix estimating the channel order, which is required to specify the structure of the TWNN equalizer. This technique uses the fact that the residual error sum of squares (SSE) of the correctly estimated channel model is smaller than all the SSE of other channel models. The Analysis uses increasing regression model order until the minimum SSE is found. The equations related to this process are:

$$Y = SH + e, \hat{H} = (S^T S)^{-1} S^T Y \tag{10}$$

$$SSE = Y^T Y - \hat{H} S^T Y = \sum_{i=1}^{2^{l+1}} [y_i - E[y_i]] \tag{11}$$

Where *S*, *Y*, $\hat{H}$, *1*, and e denote the channel input matrix, the corresponding noisy channel output vector, the estimated vector of regression model coefficients, the order of regression model, and the vector of deviations of the observed y_i's from their expected value E [y_i], *i* = 1,2, ..., *m*, respectively. The matrices above are demonstrated in what follows:

$$\begin{aligned} S &= [s_1, s_2, ..., s_m]^T \\ Y &= [y_1, y_2, ...y_m]^T \\ H &= [h_{0,} h_{1,} ....., h_l]^T \\ e &= [e_1, e_2, ..., e_m]^T \end{aligned} \tag{12}$$

Where

$$\begin{aligned} s_i &= [s(k_i), s(k_i - 1), ..., s(k_i - l)] \\ y_i &= \sum_{j=0}^{l} s(k_i - l) h_j, i = 1, 2, ..., m \end{aligned} \tag{13}$$

k_i is the time indexes. All the combinations of components in s, result in *m* (*m* = 2*l*+1) row vectors in *S*; thus none of the row vectors are the same (transmitted symbols are assumed to be binary sequences).

4 Reducing Two Number of TWNN Kernels

As shown in fig.1, a digital sequence $s(k)$ is transmitted through a time dispersive channel with transfer function:

$$H(Z)=\sum_{i=0}^{n_c} h_i Z^{-i} \tag{14}$$

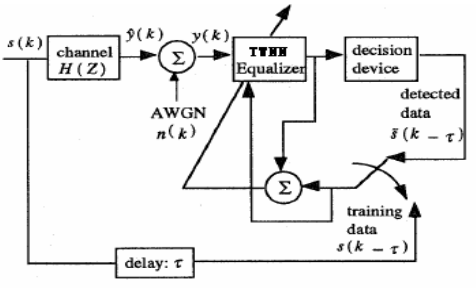


Fig. 1. Block diagram of a TWNN equalizer system

Where *n*, denotes the channel order (the number of delay elements). Then there are $M=2^{n_c+n_e+1}$ combinations of channel input vector with assumption that the transmitted symbols are equiprobable binary sequences

$$s(k)=[s(k),s(k-1),.....,s(k-n_c-n_e)] \tag{15}$$

Where n_e denotes the equalizer order (the number of tap delay elements in the equalizer). This results in *M* desired channel output vector

$$\hat{Y}(k)=[\hat{y}(k),\hat{y}(k-1),.....,\hat{y}(k-n_c)]^T \tag{16}$$

These *M* channel output states can be grouped in pairs by considering the values of $s(k-\tau)$.

$$D^+=\{\hat{Y}(k)\,|\,s(k-\tau)=1\}, D^-=\{\hat{Y}(k)\,|\,s(k-\tau)=-1\} \tag{17}$$

Where τ denotes the channel lag. The noisy channel output vector with additive white Gaussian (AWGN) is described as:

$$Y(k)=[y(k),y(k-1),.....,y(k-n_c)]^T \tag{18}$$

The proposed method of reducing the number of centers

$$s_{sub}(k)=[s(k\ ),s(k-1),.....,s(k-\tau)] \tag{19}$$

This results in selections of 2^{τ} kernels. In order to set up the learning algorithm, we denote the M combinations of $s(k)$ as $s_i, 1 \le i \le M$, and the 2^{τ} combinations of $s_{sub}(k)$ as *subj*, $1 \le i \le 2^{\tau}$ respectively.

5 Simulation Results

A number of equalization algorithms that are well suited for linear channels may perform poorly when the channel becomes nonlinear. To illustrate that the two weight neural network is able to perform equally well in both linear as well as nonlinear channel conditions, the following nonlinear channel examples are used to illustrate its effectiveness.

In this example, the following nonlinear channel [1] was chosen for a 2PAM signaling scheme:

$$y(t) = x(t) + 0.2x^2(t) + e(t), H(z) = X(z)/S(z) = 0.348 + 0.8704z^{-1} + 0.3482z^{-2} \qquad (20)$$

The equalizer order was chosen to be m = 2. The delay was set as τ = 0. 1000 training data bits mixed with noise were used to train the TWNN. The SSE performance of the TWNN was evaluated for a training set of 10 dB SNR, and also for a training set of 20dB SNR. The other parameter values were E = 0.5 and M= 10. The resulting TWNN had 7 and 12 hidden units for 10 dB and 20 dB SNR, respectively, as compared with the 64 desired channel states for a Bayesian equalizer.

The value of the thresholds θ , were set to 0.1. The other parameters were set as ε = *0. 5* and *M*= 10. At the end of training, the network has built up 12 hidden nodes. This is much less than the 18 desired channel states that would be required by the Bayesian equalizer. However, it can be seen that the Bayesian boundary is still network boundary deviates from the Bayesian boundary, at the bottom region of the figure, but this can be seen to be less critical in the equalization task from the BER curves shown in Fig.2. The testing was done with a million test data of various SNR well approximated, at the critical region, which is at the kernels of the figure. The testing was done with a million test data of various SNR to obtain the BER curves. As can be seen, the performance of the TWNN is comparable to that of the Bayesian equalizer.

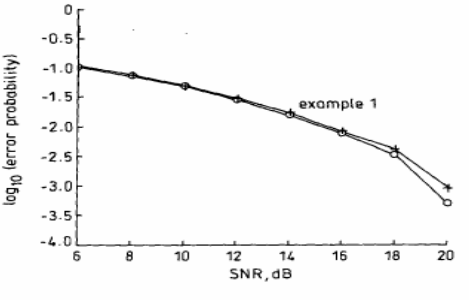


Fig. 2. Error curves for Bayesian and TWNN for nonlinear channel examples

6 Conclusion

From the simulation results, we conclude that the training of TWNN equalizer with the proposed method is much faster than that of the previously developed one with the conventional method. This method of reducing the centers can be also applied for the TWN equalizer with more delay taps. And the error rate performance of this equalizer system with the least number of centers was the same as the case with full number of centers.

References

1. Gibson, G. J., S. Siu, and C. F. N. Cowan: Application of multilayer perceptrons as adaptive channel equalizers, In Roc. IEEE Int. Conf. Acous., Speech, Signal Processing, Glasgow, Scotland, (1989) 1183-1186
2. Wang shoujue: A new development on ANN in China - Biomimetic pattern recognition and multi weight vector neurons, Lecture Notes in Artificial Intelligence, Vol.2639. Springer-Verlag, Berlin Heidelberg New York (2003) 35-43
3. Wenming Cao, Jianhui Hu, Gang Xiao, Shoujue Wang:Application of multi-weighted neuron for iris recognition, Lecture Notes in Computer Science, Vol.3497. Springer-Verlag, Berlin Heidelberg New York (2005) 87-92
4. Wenming Cao, Xiaoxia Pan, Shoujue Wang: Continuous speech research based on two-weight neural network, Lecture Notes in Computer Science, Vol.3497. Springer-Verlag, Berlin Heidelberg New York (2005) 345-350
5. Wenming Cao, Feng Hao, Shoujue Wang: The application of DBF neural networks for object recognition. Information Science 160(1-4): (2004) 153-160
6. Wang SJ, Zhao XT: Biomimetic pattern recognition theory and its applications , Chinese Journal of Electronics 13 (3): (2004) 373-377
7. Wang SJ, Chen YM, Wang XD, et al.: Modeling and optimization of semiconductor manufacturing process with neural networks ,Chinese journal of electronics 9 (1), (2000) 1-5
8. Wang SJ, Lai JL Geometrical learning, descriptive geometry, and biomimetic pattern recognition , Neurocomputing 67: (2005) 9-28

Assessment of Uncertainty in Mineral Prospectivity Prediction Using Interval Neutrosophic Set

Pawalai Kraipeerapun[1], Chun Che Fung[2], and Warick Brown[3]

[1] School of Information Technology, Murdoch University, Australia
[2] Centre for Enterprise Collaboration in Innovative Systems, Australia
{p.kraipeerapun, l.fung}@murdoch.edu.au
[3] Centre for Exploration Targeting, The University of Western Australia, Australia
wbrown@segs.uwa.edu.au

Abstract. Accurate spatial prediction of mineral deposit locations is essential for the mining industry. The integration of Geographic Information System (GIS) data with soft computing techniques can improve the accuracy of the prediction of mineral prospectivity. But uncertainty still exists. Uncertainties always exist in GIS data and in the processing required to make predictions. Quantification of uncertainty in mineral prospectivity prediction can support decision making in regional-scale mineral exploration. This research deals with these uncertainties. In this study, interval neutrosophic sets are combined with existing soft computing techniques to express uncertainty in the prediction of mineral deposit locations.

1 Introduction

In mineral exploration, the most important goal is to find new mineral deposits. After Geographic Information System (GIS) became available, several methods for the prediction of new mineral deposit locations based on GIS were created. In [2], these methods are separated into three categories: knowledge-driven, data-driven, and hybrid methods. Some examples of the knowledge-driven approach are index overlay and fuzzy logic. Neural networks and statistically-based methods such as multiple linear regression, weights of evidence, and Dempster-Shafer Evidential Belief Theory are examples of data-driven approaches. Data-driven methods can only be applied in well-explored areas which contain many known deposits of a particular type. A sufficient number of deposits (> 40) is required for training purposes or statistical analysis. Knowledge-driven methods rely on a deposit expert to make subjective assessments about the relative importance of input data and also on deposit models which are often incomplete or need to be updated in the light of new discoveries [2]. In [1], a neural network method was found to give better prediction results than the existing empirical statistically-based methods. There have been several studies in which GIS and neural networks have been applied to the prediction of mineral prospectivity [7, 5, 6].

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1074–1079, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The accurate spatial prediction of mineral deposit locations is crucial to mineral exploration. However, uncertainty in geoscience data sets is often high due to both the uncertainty in geological mapping and the uncertainty in the spatial locations of the geological features. There is always some degree of imprecision, inaccuracy and vagueness in geoscience data [3]. Mineral prospectivity predictions are calculated from several sources of data which are always imprecise; e.g. geology, geochemistry, or geophysics. Also, these input data are often incomplete because not all the data that is required can be collected from the study area. Inaccuracy occurs in the input data because of measurement or observation error. Mineral potential estimates also possess a degree of vagueness. Few locations have one hundred percent similarity or zero percent similarity to mineral deposits. Most locations have features that correspond to degrees of similarity between these two extremes. The definition of what constitutes a deposit location and what constitutes a barren location is vague. Hence, each location contains uncertain information about the degree of deposit, degree of non-deposit, and degree of indeterminable information. The uncertainty in mineral prospectivity prediction should be considered in a detail because a quantification of uncertainty has the potential to significantly improve the process of decision making in regional mineral exploration. This research proposes an integration of Interval Neutrosophic Set (INS) [10] into the existing mineral prospectivity prediction models. A feed-forward backpropagation neural network is used in our proposed prediction model. The result of the prediction is expressed by three values: truth-membership, indeterminacy-membership, and false-membership values, which all are constituted to interval neutrosophic sets. A multidimensional interpolation method is also applied to estimate the indeterminacy-membership values, which are uncertainties in our model.

The rest of this paper is organized as follows. Section 2 presents the basic theory of interval neutrosophic sets. Section 3 explains the proposed model for the assessment of uncertainties using interval neutrosophic sets and a neural network. The GIS data set used for the experiment and the results are presented in section 4. Conclusions and future work are presented in section 5.

2 Interval Neutrosophic Set

An interval neutrosophic set is an instance of a neutrosophic set [9]. Florentin Smarandache [8] is the researcher who first proposed the concept of neutrosophic sets, neutrosophic logic, neutrosophy, neutrosophic probability and statistics. The neutrosophic set generalizes the concept of a classical set, fuzzy set, interval-valued fuzzy set, intuitionistic fuzzy set, interval-valued intuitionistic fuzzy set, paraconsistent set, dialetheist set, paradoxist set, and tautological set [9]. The membership of an element to the neutorsophic set is expressed by three values: t, i, and f. These values represent truth-membership, indeterminacy-membership, and false-membership, respectively. The three memberships are independent. They can be any real sub-unitary subsets. For example, let A be neutrosophic set. Then, $x((30-40), \{25, 35\}, 20)$ belongs to A means

that x is in A to a degree between $30 - 40\%$, x is unknown to a degree of 25% or 35%, and x is not in A to a degree 20%.

The neutrosophic set was originally developed from a philosophical point of view. In order to apply neutrosophic sets to science or engineering problems, the neutrosophic set and set-theoretic operators need to be specified. Consequently, a special type of neutrosophic set called the interval neutrosophic set (INS) was created to support the scientific approach. This research follows the definition of INS that is defined in [9]. This definition is described below.

Let X be a space of points (objects). An interval neutrosophic set in X is

$$A = \{x(T_A(x), I_A(x), F_A(x)) \mid x \in X \ \wedge \ T_A : X \longrightarrow [0,1] \ \wedge \ I_A : X \longrightarrow [0,1] \wedge F_A : X \longrightarrow [0,1]\}.$$

T_A is the truth-membership function.
I_A is the indeterminacy-membership function.
F_A is the false-membership function.

If X is discrete, an interval neutrosophic set A can be written as

$$A = \sum_{i=1}^{n} \langle T(x_i), I(x_i), F(x_i)\rangle / x_i, \ x_i \in X$$

The operations of INS will also be applied in this research. Details of INS operations can be found in [9].

3 Assessment of Uncertainties Using Interval Neutrosophic Sets and a Neural Network

In order to predict new mineral deposit locations, this paper applies GIS input data to neural networks and utilizes the interval neutrosophic set to express uncertainties in the prediction. The membership of an element to the interval

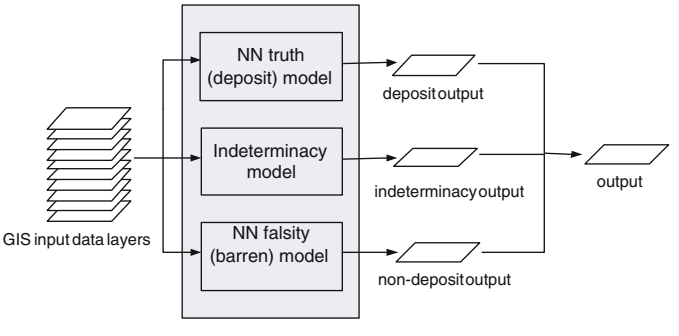


Fig. 1. Uncertainty model based on the integration of interval neutrosophic sets (INS) and neural networks (NN)

neutrosophic set is expressed by three values: t, i, and f. These values represent truth-membership, indeterminacy-membership, and false-membership, respectively. In our proposed model, we consider the set of predicted mineral deposit locations, which is an output in Figure1, as an INS. Let C be an output GIS layer. $C = \{c_1, c_2, ..., c_n\}$ where c_i is a cell at location i. Let interval neutrosophic set A be the set of predicted mineral deposit locations. Hence, A can be written as

$$A = \sum_{i=1}^{n} \langle T(c_i), I(c_i), F(c_i) \rangle / c_i, c_i \in C.$$

T_A is the truth (deposit) membership function.
I_A is the indeterminacy membership function.
F_A is the false (barren) membership function.

The deposit membership values (representing the degree of favourability for deposits) are calculated from the output of the neural network, which is trained to identify deposit locations. The barren membership values are computed from the same type of neural network, with all the same inputs as the deposit network, but this network is trained to identify barren locations using the complement of target outputs used in the deposit training data. For example, if the target output is 1, its complement is 0. In order to calculate the indeterminacy membership values for the training data, the maximum output error of the neural network that calculates the truth (deposit) membership values and the neural network that calculates the false (barren) membership values is selected for each training pattern.

The next step is to use the indeterminacy values for the training data to estimate the uncertainty in the neural network outputs for both truth (deposit) and false (barren) membership values for new data. This is equivalent to estimating the uncertainty in the mineral prospectivity prediction. To represent new data, we use a test data set. The correct outputs for this test data set are known but the values are only used to check the results of our uncertainty estimate. The uncertainty values estimated for the training data are used to estimate the uncertainty of mineral potential values in the test data set. The known errors are plotted in the feature space of the input data layers. Hence, we have to deal with multidimensional space. Two methods are proposed to quantify the uncertainties. First, multidimensional interpolation will be used to calculate interpolated uncertainty values. Second, scaling techniques will be used to reduce high dimensional space into low dimensional space, and then interpolation methods will be used to calculate the interpolated uncertainty values. This paper applies the first technique to the real data set shown in the next section.

4 GIS Data Set

Our study area corresponds to an approximately 100×100 square km region of the Archaean Yilgarn Block, near Kalgoorlie, Western Australia. The GIS data

set used to create input feature vectors in this study consists of ten layers in raster format. This data set is the same as the one described in [1]. Examples of the information in these layers are solid geology, distance to the nearest fault, and magnetic anomaly. Each layer is divided into a grid of square cells of 100 m side. Hence, the map area contains 1,254,000 cells. Each cell stores a single attribute value such as a distance to the nearest fault. All input attribute values are scaled to a range $[0, 1]$. Our training and test data sets consist of only 120 cells representing deposits with total contained gold greater than 1,000 kg, and 148 barren or non-deposit cells. Therefore, the total is 268 cells. We use 85 deposit cells and 102 barren cells for training. The rest is used for testing; i.e. 35 deposit cells and 46 barren cells.

We use a feed-forward backpropagation neural network for the training. This network is the most popular neural network and is suitable for a large variety of applications [4] and was consequently chosen for our experiments with interval neutrosophic sets. In this paper, two neural networks are used for training. The first one represents the truth (deposit) model and is trained on the original target values to yield estimates of the deposit membership values. The second one represents the falsity (barren) model which is trained using the complement of the original target values and provides estimates of the barren membership values. Both networks contain ten input-nodes and a single output node. The same parameter values are applied to the two networks and both networks are initialized with the same random weights. The only difference is that the target values for the second network are equal to the complement of the target values used to train the first network. After obtaining outputs from both deposit and barren networks, the uncertainty estimation in the neural network outputs is calculated from a multidimensional interpolation. We consider errors to represent uncertainties. For each input feature vector or pattern used in the training phase, the maximum value is selected from the deposit and barren membership errors. This training error is then plotted in the feature space of the input data layers. We use multidimensional nearest neighbour interpolation function in Matlab to calculate the uncertainty (indeterminacy membership values) in the neural network outputs corresponding to input feature vectors in the test data set. Only 60 patterns from the input training data are plotted in the input feature space because of memory limitations of the computer used in the experiment. The average maximum error from training is 0.269 while the average maximum error from the testing data is 0.322. The error is calculated from the difference between the actual value and the predicted value. The average error obtained from multidimensional interpolation is 0.263, which is very similar to the training error and much better than the predicted values based on individual cells. This demonstrates the potential of the interpolation approach.

5 Conclusion and Future Work

High quality mineral prospectivity prediction is needed in mineral exploration especially for prediction in poorly-explored areas containing few known deposits.

The data collected from these poorly-explored areas always contains a high level of uncertainty. Existing methods used to predict mineral deposit locations also contain a considerable degree of uncertainty. This paper proposes a method for quantifying uncertainties in mineral potential maps. The INS is integrated with neural networks to express uncertainties in the prediction. Quantification of uncertainty can help support decisions made in regional-scale mineral exploration. Initial work has demonstrated the potential of the INS approach. The interpolation method also gives comparative results similar to the training. In the future, we plan to apply our model to ensembles of neural networks. A cellular automata [11] and a knowledge-based model of mineral deposits will be added to our system to improve the quality of the mineral prospectivity prediction.

References

1. W. M. Brown, T. D. Gedeon, D. I. Groves, and R. G. Barnes. Artificial neural networks: a new method for mineral prospectivity mapping. *Australian journal of earth sciences*, 47:757–770, 2000.
2. Warick Brown. *Artificial Neural Networks: a new method for mineral-prospectivity mapping*. PhD thesis, University of Western Australia, 2002.
3. Matt Duckham. Uncertainty and geographic information: computational and critical convergence. In *Representation in a digital geography*, New York, 2002. John Wiley.
4. C.C. Fung, V. Iyer, W. Brown, and K.W. Wong. Comparing the Performance of Different Neural Networks Architectures for the Prediction of Mineral Prospectivity. In *the Fourth International Conference on Machine Learning and Cybernetics (ICMLC 2005)*, Guangzhou, China, August 2005.
5. V. Iyer, C.C. Fung, W. Brown, and T. Gedeon. A General Regression Neural Network Based Model for Mineral Favourability Prediction. In *Proceedings of the Fifth Postgraduate Electrical Engineering and Computing Symposium (PEECS 2004)*, pages 207–209, Perth, Western Australia, September 2004.
6. V. Iyer, C.C. Fung, W. Brown, and T. Gedeon. Use of Polynomial Neural Network for a mineral prospectivity analysis in a GIS environment. In *Proceedings of the IEEE Region 10 conference (TenCon 2004)*, volume B, pages 411–414, Chiang Mai, Thailand, November 2004.
7. Andrew Skabar. Mineral Potential Mapping using Feed-Forward Neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 1814–1819, July 2003.
8. Florentin Smarandache. *A Unifying Field in Logics: Neutrosophic Logic, Neutrosophy, Neutrosophic Set, Neutrosophic Probability and Statistics*. Xiquan, Phoenix, third edition, 2003.
9. Haibin Wang, Florentin Smarandache, Yan-Qing Zhang, and Rajshekhar Sunderraman. *Interval Neutrosophic Sets and Logic: Theory and Applications in Computing*. Neutrosophic Book Series, No.5. http://arxiv.org/abs/cs/0505014, May 2005.
10. H.B. Wang, D. Madiraju, Y-Q Zhang, and R. Sunderraman. Interval Neutrosophic Sets. *International Journal of Applied Mathematics and Statistics*, 3(M05):1–18, March 2005.
11. Stephen Wolfram. *Theory and Applications of Cellular Automata*. World Scientific, 1986.

Ring-Based Anonymous Fingerprinting Scheme

Qiang Lei, Zhengtao Jiang, and Yumin Wang

Key Laboratory on Computer Networks Information Security of Ministry Education, Xidian University, Xi'an 710071, China
{qlei, ymwang}@mail.xidian.edu.cn

Abstract. Fingerprinting schemes are important techniques for protection of intellectual property. However, the previous schemes have some trust assumption made on the registration center, meanwhile they always consider the collusion between the registration and the merchant, also the buyer. To overcome those disadvantages, an anonymous fingerprinting scheme using a modification of Schnorr ring signature is presented. When a buyer wants to buy digital goods, he can sign the text m that describes the deal on behalf of the ring, which can bring fully anonymity and unlinkability. Once redistributing fingerprinted copies, a merchant still could trace the buyer with a slightly different version.

1 Introduction

With the fast development of information technology and electronic commerce recently, people can deal with digital goods in an easy and cheap way. How to protect intellectual property in digital goods has been a subject of research for many years and led to the development of various techniques. Among them, digital fingerprinting which is a method for the copyright protection of digital data has been proposed. When a buyer wants to purchase digital goods, some information about him will be embedded in the copy, it means every buyer having a slightly different "copy". Obviously, the identifying information, called fingerprint, must be such that a buyer cannot easily detect, remove or change it. Upon finding a distributed copy, the merchant could identify the buyer of the redistributed copy by its fingerprint.

Anonymous fingerprinting [1], introduced by Pfitzmann and Waidner in 1997, is to protect buyer's privacy. Here, the buyer can purchase anonymously and no longer be revealed his identity unless he redistributes the fingerprinted copy. In recent years, many papers have devoted to works for anonymous fingerprinting. [4], [5], [9],[10].

1.1 Our Contribution

Comparing to previous schemes, some of them only have a weak sense of anonymity, some are rather inefficient, and some can not offer one or two properties which make the proposed scheme more perfect. Additionally, there have some trust assumption made on the registration center, meanwhile they always consider the collusion between the registration center and the merchant or the buyer.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1080–1085, 2005.
© Springer-Verlag Berlin Heidelberg 2005

To address these problems, we presents an anonymous fingerprinting scheme using ring signature to give fully anonymity and unlinkability, and even not require a trusted third party as registration center. Furthermore, we do not assume that the registration center does not collude with buyers, or merchant. Because registration center does not take part in fingerprinting operation actually.

1.2 Organization

The paper is organized as follows. In Section 2, we briefly review the ring signature and introduce the other related knowledge. Then a ring-based fingerprinting scheme is described in Section 3. Section 4 is the discussion of security of the proposed scheme. Finally, we conclude in Section 5.

2 Related Knowledge

2.1 Ring Signature Scheme

A new digital signature scheme–Ring Signature Scheme was proposed in Asiacrypt2000 [3], which makes it possible to allow a member of a set to sign a message on the ring's behalf, and makes it impossible to revoke the anonymity of the actual signer. To produce a ring signature, any signer can choose possible ring's signers including himself, and sign any message by using his secret key and the others' public keys without getting their approval. One can verify the signature, but there is no way to revoke the anonymity of the actual signer.

We assume that there are n possible signers $A_1, A_2, ..., A_n$, of which has a public key y_i $(1 \leq n \leq i)$. The actual signer A_s, $s \in 1, 2, ..., n$, wants to sign a message m on behalf of the ring $A_1, A_2, ..., A_n$. Anyone who obtains the signature can assure it generated by some one in the ring, but he gets nothing any more.

2.2 Bit Commitment Scheme

The bit commitment scheme [7], [8] is a basic component of many cryptographic protocol, which is a technique to commit to data yet keep it secret for some period. The signer commits to the value of a bit in a way that prevents the other party from learning it without the signer's help. Later the signer can open the commitment to show which bit he has committed to and he should not be able to cheat by not showing the genuine bit that he chooses to commit.

2.3 Requirements of Anonymous Fingerprinting Scheme

Requirements of anonymous fingerprinting scheme can be listed [2], [10]:

Anonymity: A buyer should able to purchase digital goods anonymously, nothing about the purchase behavior of honest buyers becomes known to any other party. Without obtaining an illegal copy, the merchant cannot identify a buyer.

Unlinkability: Nobody can decide whether any digital goods were purchased by the same buyer.

No Repudiation: Once distributing an illegal copy the buyer should not be able to deny his duty.

No Framing: An honest buyer should not be falsely accused by a malicious merchant or other buyers.

Traceability: The buyer who has distributed copy illegally can be traced.

2.4 Roles of Each Party

There are three parties in our scheme: a registration center, a buyer and a merchant. Define the role of each party as follows:

1.*Registration Center* (***RC***)
We do not even need a trusted third party as registration center actually, only require a conventional CA which will be responsible for managing the registration buyers. Including record every buyer's information, conserve their identities and publish their corresponding public keys.

2.*Buyer* (***B***)
Any buyer has registered in RC before purchase. An honest buyer should not be falsely accused, and a dishonest buyer will be traced.

3.*Merchant*(***M***)
M should embed the buyer's secret information into digital goods without knowing it, but can extract the buyer's information if finding the redistributed copy. He can trace the traitor, and not make wrong accusation.

3 Ring-Based Anonymous Fingerprinting Scheme

3.1 Notations

Let p be a prime with q a large prime factor of p-1, g is a generator of order q in the group Z_p^*. $H(\cdot)$ is a collision resistant one way hash function.Here a set of potential signers $A_1, A_2, ..., A_n$, every potential signer A_i has a secret key $x_i \in Z_q^*$ respectively and publish the public keys $y_i = g^{x_i}(modp)$. The commitment to the bit string a is defined by $com(a)$. $Sig_B(m)$ is the signature generated by B. A buyer chooses secret u, v in Z_q^* satisfying $x_B/u \in Z_q^*$, $x_B/v \in Z_q^*$, then gets two pairs of keys (u, U) and (v, V), where $U = g^{x_u}(modp)$, $V = g^{x_v}(modp)$.

3.2 The Proposed Anonymous Fingerprinting Scheme

Registration
1. One buyer sends his signature on his real identity and his public key to RC.
2. RC verifies the signature using the buyer's public key y_B. If the signature is valid, RC gets a record $(y_B, Sig_B(y_B, ID_B))$ and registration is successful.

Ring Generation
B wants to buy a digital goods P_0, he choose n possible signers in RC randomly, then signs the text m that describes the deal on behalf of the ring $B_1, B_2, ..., B_n$, where the actual signer B_s is the buyer B, $s \in 1, 2, ..., n$. B_i's secret key is x_i and his corresponding public key is y_i, $i = 1, 2, ..., n$. The procession we use Schnorr ring signature [6] is as follows:

1. For all $i \in 1, 2, ..., n$, $i \neq s$, choose a_i randomly in Z_q^*, $a_i \neq a_j$, $i \neq j$. Then compute $R_i = g^{a_i}(modp)$, $S_i = R_i^u(modp)$ and $T_i = R_i^v(modp)$.
2. Choose a random $a_s \in Z_q^*$ and compute $R_s = g^{a_s} \prod_{i \neq s} y_i^{-h_i}(modp)$, where $h_i = H(m, R_i, S_i, T_i)$, $i = 1, 2, ..., n$. If $R_s = 1$ or $R_s = R_i$ for some $i \neq s$, then go to choose another $a_s \in Z_q^*$ and compute R_s. Next generate $S_s = g^{a_s u} \prod_{i \neq s} y_i^{-h_i}(modp)$ and $T_s = g^{a_s v} \prod_{i \neq s} y_i^{-h_i}(modp)$.
3. Compute $\alpha = \sum_{i=1}^{n} a_i + x_s/u \cdot h_s(modq)$; $\beta = \sum_{i=1}^{n} a_i + x_s/v \cdot h_s(modq)$; $\sigma = \sum_{i=1}^{n} a_i + x_s h_s(modq)$, where $h_s = H(m, R_s, S_s, T_s)$, $s \in 1, 2, ..., n$.
4. Produce $Sig = (m, R_1, R_2 ... R_n, S_1, S_2 ... S_n, T_1, T_2 ... T_n, h_1, h_2 ... h_n, \sigma, \alpha, \beta, U, V)$ as the signature of message m made by the ring $B_1, B_2, ..., B_n$.

Fingerprinting

B and M is executed in a secure embedding process, which both parties will not leak their secret information, and upon finding the redistributed fingerprinted copy M could extract the embedded information without any help.

1. B makes commitment $com(u)$ and $com(v)$, then sends $com(u)$, $com(u)$, U and V to M, next he proves in zero-knowledge way that $com(u)$ and $com(v)$ indeed commits to the discrete logarithm of U and V with respect to g.
2. M verifies the signature using the ring buyers' public keys $y_1, y_2, ..., y_n$, which the buyers have registered in RC before. After checking $h_i = H(m, R_i, S_i, T_i)$ for $i = 1, 2, ..., n$, he accepts the signature if $U^\alpha = S_1 S_2 ... S_n y_1^{h_1} y_2^{h_2} ... y_n^{h_n}(modp)$, $V^\beta = T_1 T_2 ... T_n y_1^{h_1} y_2^{h_2} ... y_n^{h_n}(modp)$ and $g^\sigma = R_1 R_2 ... R_n y_1^{h_1} y_2^{h_2} ... y_n^{h_n}(modp)$ hold.
3. If the signature is valid, M and B could work an embedding protocol, where M's input is the original copy P_0, $com(u)$ and $com(v)$, whereas B's input is u, v, $com(u)$ and $com(v)$. The output of B is a fingerprinted copy P_B of P_0.
4. M gets a record (U, V, Sig), which helps open buyer's identity when needed.

Identification

When M finds a redistributed copy $\tilde{\mathrm{P}}$, he could trace an illegal buyer:

1. Using the extraction algorithm M extracts u and v from redistributed copy $\tilde{\mathrm{P}}$, which can compute $g^u(modp)$, $g^v(modp)$. Then by comparing with U and V he searches in database and finds a record (U, V, Sig).
2. From the signature Sig, computing $\alpha - \beta = x_s h_s(1/u - 1/v)(modq)$. M can get $x_s h_s$, then generates $\sum a_i = \sigma - x_s h_s(modq)$, where $h_s = H(m, R_s, S_s, T_s)$. M computes $r_j = g^{\sum a_i} \prod_{i \neq j} y_i^{-H(m, R_i, S_i, T_i)}(modp)$for all $i, j = 1, 2, ..., n$, and compares r_j with $R = R_1 R_2 ... R_n(modp)$. Once r_s is equal to R, $s \in 1, 2, ..., n$, M will obtain y_s. In other words, y_s is B's public key.
3. M gives (u, v, U, V, y_s, Sig) to RC. RC verifies $U = g^u(modp)$, $V = g^v(modp)$ and checks whether Sig is a valid signature. If it holds, he searches in database for possible record $(y_B, Sig_B(y_B, ID_B))$ with $y_s = y_B$.
4. RC sends $Sig_B(y_B, ID_B)$ to M, M checks it by using the public key y_B and is convinced that B is a traitor if it goes through the verification.

4 Security Analysis

Anonymity: As has been discussed above, the merchant only obtains a ring signature on a message that describes the deal, and does not get any secret information unless the merchant finds a redistributed copy and extracts the embedded u, v in it. Thus, the buyer's anonymity is provided if the merchant can not compute discrete logarithms.

M should not have probability greater than 1/n to guess the identity of the real buyer who has computed a ring signature.

Unlinkability: For buyer's unlinkability of his different purchases the ring is generated independently every time, so they could not be linked. Anyway nobody can tell whether any two purchases were made by the same buyer or not. Even a collusion between M and RC could not identify an honest buyer because of the unlinkability between the message-signature and the view of the signer.

What is more, a buyer has been convinced to be a traitor, his new purchases will still be safe. Because the buyer should avoid use of the same value u and v, also choose different ring users in other purchases.

No Repudiation: Since only the buyer keeps secret u and v, besides he has ability to sign the signature $Sig_B(y_B, ID_B)$, the others cannot recreate an illegal copy. Thus the buyer can not deny his duty once finding a redistributed copy.

No Framing: Anyone would not be able frame an honest buyer until he can get u, v at least. The merchant only can get U, V, $com(u)$, $com(v)$ and the zero-knowledge proof on u and v, but knowledge on u and v in secure embedding procedure is not leaked. Since computing the discrete logarithm of U and V is difficult and the cryptographic primitives are assumed to be secure. Only M obtains a illegal copy , can he get u and v by using extraction algorithm.

Traceability: Due to the properties of the watermarking and digital signature techniques, anyone can not change, remove or substitute a fingerprint. In addition, fingerprint detecting can guarantee that the merchant can extract the unique watermarks (fingerprint) that belongs to a traitor.

5 Concluding

In this paper, we propose an anonymous fingerprinting scheme based on a ring signature, which not only can provide fully anonymity and unlinkability, but also could prevent the collusion problem. Even though RC colludes with B or M, the dealing between B and M is still fair. Because RC which is only a conventional CA, does not take part in operation actually. Honest buyers should never be falsely accused by anyone else, and dishonest buyers must be traced if they distribute illegal copies.

But our scheme has the drawback that it may need more computations, however, the construction of the proposed scheme could offer much security. With the development of hardware technique of computer, we believe this problem

will be solved well. Furthermore, the buyer could save data like $g^{a_i}(modp)$ in computer. When dealing with merchant, he directly chooses some results stored in computer which will reduce the amount of buyer's computations.

For adapting to the development of electronic purchase, research work still need to be done on the improvement of the watermarking techniques as well as secure embedding and extracting processes.

Acknowledgment

The authors would like to acknowledge the valuable comments and suggestions of their associates. Special thanks are expressed to Mr. S.B.Lei for many encouragements and discussions. The financial support of the National Natural Science Foundation of China under Grant No. 60473042 is also gratefully acknowledged.

References

1. B.Pfitzman, M.Waidner: Anonymous fingerprinting. Eurocrypt'97, Lecture Notes in Computer Science, Vol. 1233, Springer-Verlog, (1997) 88-102.
2. Hak-Soo Ju, Hyung-Jeong Kim, Dong-Hoon Lee, Jong-In Lim: An Anonymous Buyer-Seller Watermarking Protocol with Anonymity Control. ICISC2002, Lecture Notes in Computer Science, Vol. 2587, Springer-Verlog, (2002) 421-432.
3. Ronald L.Rivest, Adi Shamir, Yael Tauman: How to leak secret. Asiacrypt 2001, Lecture Notes in Computer Science, Vol. 2248, Springer-Verlog, (2001) 552-565.
4. Jam Camenisch: Efficient Anonymous Fingerprinting with Group Signature. Asiacrypt 2000, Lecture Notes in Computer Science, Vol. 1976, Springer-Verlog, (2000) 415-428.
5. Birgit Pfizmann, Ahmad-Reza Sadeghi: Anonymous Fingerprinting with Direct Non-repudiation. Asiacrypt 2000, Lecture Notes in Computer Science, Vol. 1976, Springer-Verlog, (2000) 401-414.
6. Javier Herranz, Germn Sez: Forking Lemmas for Ring signature Schemes. Indocrypt 2003, Lecture Notes in Computer Science, Vol. 2904, Springer-Verlog, (2003) 266-279.
7. Moni Naro: Bit Commitment Using Pseudo-randomness. Crypto'89, Lecture Notes in Computer Science, Vol. 435, Springer-Verlog, (1990) 128-136.
8. G.Brassard, D.L.Chaum, C.Crpeau: Minimum Disclosure Proofs of Knowledge. Journal of Computer and System Science, vol. 37, no. 2, Oct. 1998, 156-189.
9. Yan Wang, Shuwang L, Zhenhua Liu: A Simple Anonymous Fingerprinting Scheme Based on Blind Signature. ICICS 2003, Lecture Notes in Computer Science, Vol. 2836, Springer-Verlog, (2003) 260-268.
10. Jae-Gwi Choi, Kouichi Sakurai, Ji-Hwan Park: Proxy Certificates-Based Digital Finger-priting Scheme for Mobile Communication. Proceedings. IEEE 37th Annual 2003 Interna-tional Carnahan Conference on Security Technology, (2003) 587-594.

Scalable and Robust Fingerprinting Scheme Using Statistically Secure Extension of Anti-collusion Code

Jae-Min Seol and Seong-Whan Kim

Department of Computer Science, University of Seoul, Jeon-Nong-Dong, Seoul, Korea
seoleda@hotmail.com, swkim7@uos.ac.kr

Abstract. Fingerprinting schemes use digital watermarks to determine originators of unauthorized/pirated copies. Multiple users may collude and collectively escape identification by creating an average or median of their individually watermarked copies. Previous fingerprint code design including ACC (anti-collusion code) cannot support large number of users, and we present a practical solution, which defines scalability over existing codebook generation scheme. To increase the robustness over average and median attack, we design a scalable ACC scheme using a Gaussian distributed random variable. We experiment with our scheme using human visual system based watermarking scheme, and the simulation results with standard test images show good collusion detection performance over average and median collusion attacks.

1 Introduction

A digital watermark or watermark is an invisible mark inserted in digital media, and fingerprinting uses digital watermark to determine originators of unauthorized/pirated copies. Multiple users may collude and collectively escape identification by creating an average or median of their individually watermarked copies. An early work on designing collusion-resistant binary fingerprint codes for generic data was based on marking assumption, which states that undetectable marks cannot be arbitrarily changed without rendering the object useless. However, multimedia data have very different characteristics from generic data, and we can embed different marks or fingerprints in overall images, which biased strict marking assumption. Recently, an improvement was to merge the low level code (primitive code) with the direct sequence spread spectrum embedding for multimedia and extend the marking assumption to allow for random jamming [1, 2]. Min Wu presented the design of collusion-resistant fingerprints using code modulation. They proposed a (k-1) collusion-resistant fingerprints scheme, and the (k-1) resilient ACC is derived from (v, k, 1) balanced incomplete block design (BIBD) [3]. The resulting (k-1) resilient ACC code vectors are v-dimensional, and can represent $n = (v^2 - v) / (k^2 - k)$ users with these v basis vectors.

We present a scalable ACC fingerprinting design scheme, which extends ACC for large number of user support. Simply replicating ACC codebook does not work,

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1086 – 1091, 2005.
© Springer-Verlag Berlin Heidelberg 2005

because there are ambiguous cases for determining who colluders are. We extend the ACC (anti-collusion code) scheme using a Gaussian distributed random variable for medium attack robustness. We evaluate our scheme with standard test images, and show good collusion detection performance over two powerful attacks: average and median collusion attacks.

2 Related Works

An early work on designing collusion-resistant binary fingerprint codes was presented by Boneh and Shaw in 1995 [2], which primarily considered the problem of fingerprinting generic data that satisfy an underlying principle referred to as the marking assumption. Figure 1 illustrates a fingerprinting example for Log 5 value in logarithm table. As shown in Figure 1, a fingerprint consists of a collection of marks, each of which is modeled as a position in a digital value (denoted as boxes) and can take a finite number of states. A mark is considered detectable when a coalition of users does not have the same mark in that position. The marking assumption states that undetectable marks cannot be arbitrarily changed without rendering the object useless; however, it is considered possible for the colluding set to change a detectable mark to any state (collusion framework).

Correct value	0.6987000433601880478626110527551
Value for User A	0.6987000433601880478627110427541
Value for User B	0.6987000433601880478627110327531
Value for User C	0.6987000433601880478625110427531

Fig. 1. Fingerprint for log 5 in logarithm table

Min Wu presented the design of collusion-resistant fingerprints using code modulation [3]. The fingerprint signal w_j for the j-th user is constructed using a linear combination of a total of v orthogonal basis signals $\{\mathbf{u}_i\}$ as (1). Here the coefficients $\{b_{ij}\}$, representing the fingerprint codes, are constructed by code vectors with $\{\pm 1\}$.

$$w_j = \sum_{i=1}^{v} b_{ij}\mathbf{u}_i. \tag{1}$$

Anti-collusion codes can be used with code modulation to construct a family of fingerprints with the ability to identify colluders. An anti-collusion code (ACC) is a family of code vectors for which the bits shared between code vectors uniquely identifies groups of colluding users. ACC codes have the property that the composition of any subset of K or fewer code vectors is unique. This property allows for the identification of up to K colluders. A K-resilient ACC is such a code where the composition is an element-wise AND operation. It has been shown that binary-valued ACC can be constructed using balanced incomplete block design (BIBD) [3]. The

definition of (v, k, λ) BIBD code is a set of k-element subsets (blocks) of a v-element set χ, such that each pair of elements of χ occur together in exactly λ blocks. The (v, k, λ) BIBD has a total of n = (v^2 -v)/(k^2 -k) blocks, and we can represent (v, k, λ) BIBD code using an v x n incidence matrix, where M(i, j) is set to 1 when the i-th element belongs to the j-th block, and set to 0 otherwise. The corresponding (k − 1)-resilient ACC code vectors are assigned as the bit complements (finally represented using -1 and 1 for the 0 and 1, respectively) of the columns of the incidence matrix of a (v, k, 1) BIBD. The resulting (k-1) resilient ACC code vectors are v-dimensional, and can represent n = (v^2 -v) / (k^2 -k) users with these v basis vectors.

3 Scalable and Robust Fingerprint Scheme

Min Wu's fingerprinting scheme cannot easily extend to support large number of users because it is based on (v, k, λ) BIBD code design. Because we should have more overhead for bigger BIBD code, we designed a scalable fingerprint scheme, which can make large number of fingerprints from small BIBD code. However, simply extending ACC over multiple image blocks does not work. In our scheme, we construct each user's fingerprint as the composition of ACC (w_i) and a Gaussian distributed random signal λ as shown in Figure 2. The dimension of code vectors (M) can be increased to fit the size of fingerprinting users.

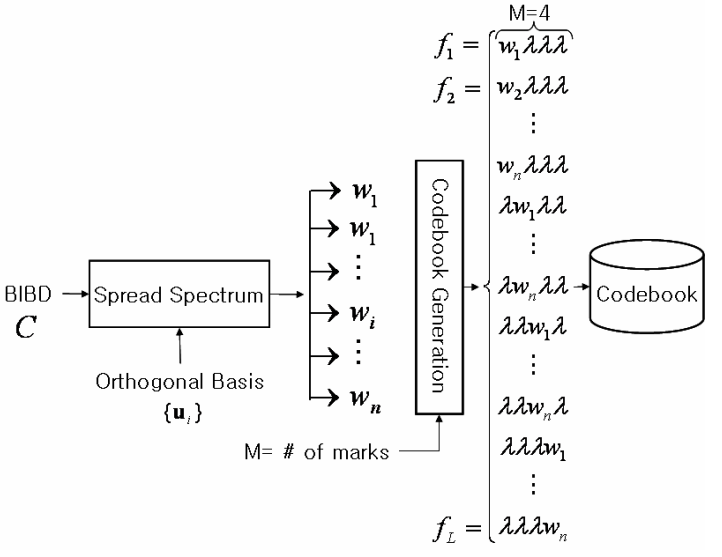


Fig. 2. Scalable fingerprint codebook, extending ACC base code with λ

In our scheme, we used the Hadamard matrix for spread spectrum basis, because it is orthogonal, secure, and widely used in MPEG-4 Part 10 AVC (Advanced Visual Coding: H.264) video compression. We can view our scheme as a two level

spreading: direct sequence spreading and frequency hopping. We used same spreading (direct sequence spreading) as ACC and we spread the ACC over M image blocks (frequency hopped), thereby we can increase the number of fingerprint codes. This scheme has strong advantages that we can control the number of fingerprint codes easily.

Once we generated the code vectors, we embed the fingerprint codes over M x R selected regions as shown in Figure 3. Fingerprinting regions (blocks) are chosen based on the model of NVF (Noise Visibility Function) [4]. Each user l's fingerprint f_l is constructed by repetition and permutation like Dan Boneh's fingerprinting scheme. For example, $w_1\lambda\lambda\lambda$ (M = 4 case) is enlarged 2 (R = 2 case) times ($w_1w_1\lambda\lambda\lambda\lambda\lambda\lambda$) and shuffled ($w_1\lambda\lambda\lambda\lambda\lambda w_1\lambda$). The permutation sequence is unique to all users, but unknown to attackers. Repetition and permutation prevent interleaving collusion attack. $f_l(i)$ is inserted signal into i-th block, each $f_l(i)$ can be ACC or λ signal, and is embedded in the M×R selected image blocks (there are R ACC signals and (M-1)× R λ signals.

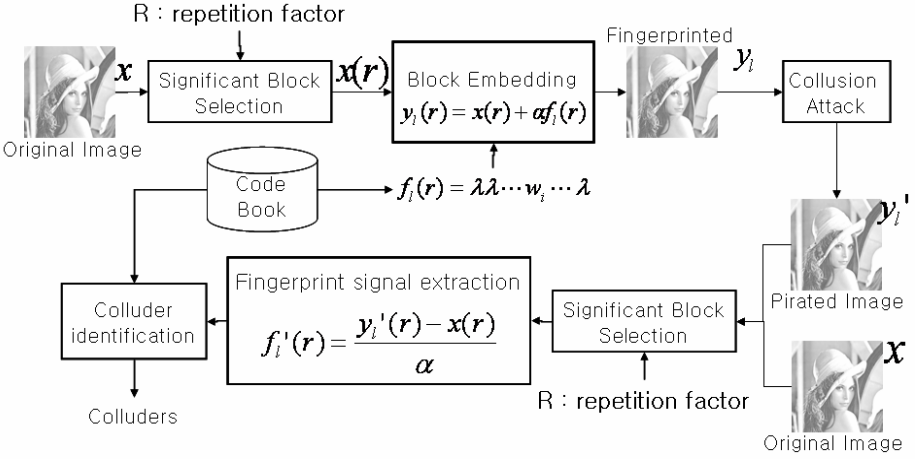


Fig. 3. Scalable fingerprint embedding/ extraction of fingerprint f_l for user l

To embed ACC (w_i) signal in a specified block y_i, we use (2). To increase the fingerprint robustness, we inserted ACC signal over R image blocks. All the ACC (w_i) are the same, however, the resulting watermark will be different, depending on $1-NVF$, which considers the local HVS masking characteristics.

$$y_i = x + (1-NVF)w_i, \qquad w_i = \sum_{j=1}^{v} c_{ij}\mathbf{u}_\mathbf{i}. \tag{2}$$

Likewise, λ signal embedding uses (3). To un-correlate the λ signal, we use the random variable for λ signal, instead of simply choosing zero vectors. μ is determined to optimize decision threshold, and σ is used to increase median attack robustness. If we increase σ, we can increase the median attack robustness because there is little difference between λ signals and w_i; however, we can risk the decrease of detection precision.

$$y_i = x + (1 - NVF)\lambda, \quad \lambda = \sum_{j=1}^{v} n\mathbf{u}_{\mathbf{i}\cdot}, \; n \sim N(\mu, \sigma^2) \tag{3}$$

We used non-blind scheme for fingerprint detection. To detect collusion, we used the collusion detection vector T, which can be computed using the same equation as Min Wu's [3].

4 Analysis and Experimental Results

We used Gaussian distributed random variable N(0.5, 0.04) for λ signal, and we chose σ^2 as 0.04 experimentally to tradeoff between median attack robustness and false positive error rate. We can compute the probability of false positive (there is no collusion, but the λ signal makes similar results as collusion occurs) error as (4). To increase both the detection precision and median attack robustness, we can increase the R in embedding step. If we use the average of λ signal over R blocks, we can correctly differentiate λ signals and w_i with high probability, because the variance of test vector Ti gets smaller (σ^2 / R) than the variance of single λ signal (σ^2).

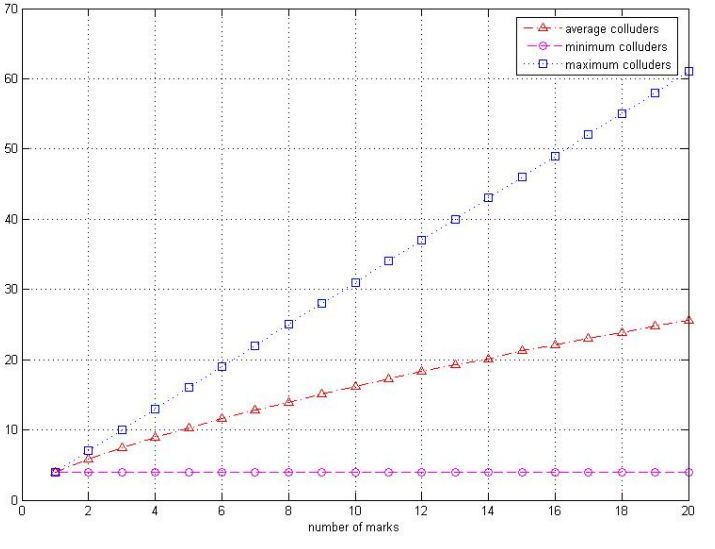


Fig. 4. Number of marks (m) versus average number of colluder for successful collusion

$$P(t_i \mid H_0) = \Phi\left(\frac{t_i - \mu}{\sigma / \sqrt{R}}\right), \quad where\ \Phi(z) = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw \tag{4}$$

We analyzed the average number of fingerprinted images to erase fingerprints (successful collusion using average and median collusion attack [5]). Figure 4 shows that colluders should have 40 fingerprinted images (or 40 colluding members) on average case, to erase the fingerprints for our scalable fingerprinting scheme with (16, 4, 1) BIBD and scalability m = 40. If we use the larger BIBID codes (e.g. (61, 5, 1) BIBD code), we can get much bigger collusion robustness.

5 Conclusions

In this paper, we presented a scalable ACC fingerprinting scheme, which covers large number of fingerprint codes. Previous fingerprint code design including ACC (anti-collusion code) cannot support large number of users. We constructed the scalable fingerprint by spreading BIBD codes over M×R (M: number of marks; R: repetition factor) image blocks. To improve the detection performance, we repeated embedding the same fingerprints over R image blocks. To increase the robustness over average and median attack, we designed a scalable ACC scheme using a Gaussian distributed random variable. We evaluated our fingerprints on standard test images, and showed good collusion detection performance over average and median collusion attacks.

References

1. Yacobi, Y.: Improved Boneh-Shaw content fingerprinting. in Proc. CTRSA2001 (2001) 378–91
2. Boneh, D., Shaw, J.: Collusion-secure fingerprinting for digital data. IEEE Trans. Inform. Theory, vol. 44, Sept. (1998) 1897–1905
3. Trappe, W., Wu, M., Wangm Z. J., Liu, K. J. R.,: Anti-collusion fingerprinting for multimedia. IEEE Trans. Signal Proc. vol. 51, Apr. (2003) 1069-1087
4. Voloshynovskiy, S., Herrige, A., Baumgaertner, N., Pun, T.: A stochastic approach to content adaptive digital image watermarking. Lecture Notes in Computer Science: 3rd Int. Workshop on Information Hiding, vol. 1768, Sept. (1999) 211-236
5. Zhao, H., Wu, M., Wang, J., Ray Liu, K. J.: Nonlinear collusion attacks on independent Fingerprints for multimedia. ICASSP. vol. 5, Apr. (2003) 664-667

Broadcast Encryption Using Identity-Based Public-Key Cryptosystem*

Lv Xixiang[1] and Bo Yang[2]

[1] National key lab of ISN, Xidian University, Xi'an 710071, China
xxlv@mail.xidian.edu.cn
[2] College of Information, South China Agricultural University, Guangzhou 510640, China
byang@center.njtu.edu.cn

Abstract. In this paper, a new public-key broadcast encryption scheme, which is asymmetric and without involvement of trusted third parties, is proposed by using the "subset-cover" framework and the identity-based encryption scheme of Boneh and Franklin. This scheme is the first concrete construction of an asymmetric identity-based public-key broadcast encryption that does not rely on trusted agents and the costly Oblivious Polynomial Evaluation mechanism. Moreover, this novel work contains other desirable features, such as efficient encryption and decryption, low memory requirements, traitor tracing, asymmetry and non-repudiation. The tracing algorithm of this system is more efficient than that of the previous ones.

1 Introduction

The area of Broadcast Encryption deals with methods to efficiently broadcast information to a dynamically changing group of users who are allowed to receive the data. In the network environment, the broadcast encryption has a growing number of applications. Among them are pay-TV applications and secure distribution of copyright-protected material. In such scenarios it is also desirable to have a Tracing Mechanism, in which each subscriber has a decryption key that is associated with his identity (it can be thought of as a fingerprinted decryption key) and the distributor (or the authorities) possesses a "traitor tracing" procedure that, given the pirate decoder, is capable of recovering the identities of at least one of the traitors. The area of Broadcast Encryption was first formally studied by Fiat and Naor in [3]. Since then, there have been a number of schemes in the literature. Some schemes allow traitor tracing. However, none of the above schemes is identity-based and all of them are computationally expensive. In this paper, a very efficient public-key broadcast encryption scheme, which is asymmetric and without involvement of trusted third

* This work is supported by the fund of national nature science: Study on Traitor Tracing (60372046) and the fund of the key lab of the national defence science and technology: Study on the mobile electronic commerce security (51436040204DZ0102).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1092 – 1097, 2005.
© Springer-Verlag Berlin Heidelberg 2005

parties, is proposed by using the subset-cover framework given by Dalit Naor[2] and the identity-based encryption scheme of Boneh-Franklin[1]. Our scheme is identity-based so that the broadcaster can easily broadcast a message or messages to a group in term of the ID of the group. Our scheme contains some desirable and necessary features, such as key-revocation, asymmetry, and black-box traceability for traitor tracing scheme.

2 Subset-Cover Framework

Let N be the set of all users, $|\mathrm{N}| = N$, and $\Gamma \subset \mathrm{N}$ be a group of $|\Gamma| = \mu$ users whose decryption privileges should be revoked. In brief, the basic idea behind the "subset-cover" framework for broadcast encryption (in the symmetric setting) proposed by Naor, Naor, and Lotspiech [2] is to define a family S of subsets of the set N of users, and to assign a key to each subset. In this framework an algorithm defines a collection of subsets $S_1, S_2, \ldots, S_w$, $S_j \subseteq \mathrm{N}$. Each subset S_j is assigned (perhaps implicitly) a long-lived key L_j, each member u of S_j should be able to deduce L_j from its secret information. Given a revoked set Γ, the remaining users $\mathrm{N} \setminus \Gamma$ are partitioned into disjoint subsets $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ so that $\mathrm{N} \setminus \Gamma = \bigcup_{j=1}^{m} S_{i_j}$ and a session key K is encrypted m times with $L_{i_1}, L_{i_2}, \ldots, L_{i_m}$. In [2], two specific methods that realize the above subset-cover framework: the "Complete Subtree (CS)" method and "Subset Difference (SD)" method. Since our scheme is well applicable to the CS method, we review it in detail as follows. The collection of subsets $S_1, S_2, \ldots, S_w$ in the CS scheme corresponds to all complete sub-trees in the full binary tree with N leaves. For any node v_i in the full binary tree (either an internal node or a leaf, $2N-1$ altogether) let the subset S_i be the collection of receivers u that correspond to the leaves of the sub-tree rooted at node v_i. The key assignment method simply assigns an independent and random key L_i to every node v_i in the complete tree. Provide every receiver u with the $\log N + 1$ keys associated with the nodes along the path from the root to leaf u. For a given set Γ of revoked receivers, let $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ be all the subtrees of the original tree whose roots are adjacent to nodes of outdegree 1 in $ST(\Gamma)$, but they are not in $ST(\Gamma)$. It follows immediately that this collection covers all nodes in $\mathrm{N} \setminus \Gamma$ and only them. The cover size is at most $\mu \log(N/\mu)$. This is also the average number of subsets in the cover. At decryption, given a message $<[i_1, i_2, \ldots, i_m, E_{L_{i_1}}(K), \ldots, E_{L_{i_m}}(K)], F_K(M)>$ a receiver u needs to find whether any of its ancestors is among $i_1, i_2, \ldots, i_m$, note that there can be only one such ancestor, so u may belong to at most one subset. This lookup can be facilitated efficiently by using hash-table lookups with perfect hash functions. Refer to [2] for more details.

3 Traitor Tracing Scheme Using Pairing Based System

We will construct our broadcast encryption scheme based on the Boneh and Franklin encryption algorithm. Refer [1] for details. The entities involved in our scheme are the broadcast center C and N receivers u.

(1) *Initialization*: Let G_1 and G_2 denote two groups of prime order q in which the discrete logarithm problem is believed to be hard and for which there exists a computable bilinear map $t : G_1 \times G_1 \to G_2$.

We shall write G_1 with an additive notation and G_2 with a multiplicative notation, since in real life G_1 will be the group of points on an elliptic curve and G_2 will denote a subgroup of the multiplicative group of a finite field. The following cryptographic hash functions are defined: $H_1 : \{0,1\}^* \to G_1$, $H_2 : G_2 \to \{0,1\}^*$.

Let the collection of subsets $S_1, S_2, \ldots, S_w$ correspond to all complete subtrees in the full binary tree with N leaves. The center picks a random $s \in F_q$ as his secret and set $R = sP$ as his public key for some given fixed point $P \in G_1$. For each subset S_i, the center computes sP_i with $P_i = H_1(ID_i)$. Let $s_i \in F_q$, $s_i P_i$ and $k_i = sP_i + s_i P_i$ be the secret key of the subset S_i, and the public key is $Q_i = s_i P$. Provide every receiver u with the $\log N + 1$ keys associated with the nodes along the path from the root to leaf u. Every receiver u is assigned private information I_u. From the conclusion of the literature [2], for all $1 \le i \le w$ such that $u \in S_i$, I_u allows u to deduce the key $k_i = sP_i + s_i P_i$ corresponding to the set S_i.

(2) *Encryption and Broadcast*: Given a set Γ of revoked receivers, the center finds a partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $N \setminus \Gamma$. Now $Q_{i_j} = s_{i_j} P$ and $P_{i_j} = H_1(ID_{i_j})$. Let $A = (a_{xy})$ denote an $m \times m$ non-singular matrix over F_q whose first column is equal to one and $B = (b_{xy}) = A^{-1} (\mathrm{mod}\, q)$. The center precomputes the following matrix relations:

$$\begin{pmatrix} Q_V \\ T_1 \\ \vdots \\ T_{i_m - 1} \end{pmatrix} = B \begin{pmatrix} Q_{i_1} \\ Q_{i_2} \\ \vdots \\ Q_{i_m} \end{pmatrix}, \quad \begin{pmatrix} P_V \\ T_1' \\ \vdots \\ T_{i_m - 1}' \end{pmatrix} = B \begin{pmatrix} P_{i_1} \\ P_{i_2} \\ \vdots \\ P_{i_m} \end{pmatrix}. \tag{3-1}$$

Then

$$Q_{i_j} = Q_V + \sum_{k=2}^{i_m} a_{i_j k} T_{k-1} \; ; \; P_{i_j} = P_V + \sum_{k=2}^{i_m} a_{i_j k} T'_{k-1}, \; j = (1, 2, \ldots, m) \tag{3-2}$$

The center chooses a session encryption key s_m. To encrypts the session key s_m the center picks a random $r \in F_q$ and computes:

$$U = rP, \; U_j^{(1)} = rT_j, \quad U_j^{(2)} = rT_j' \text{ for } 1 \le j \le i_m - 1 \tag{3-3}$$

$$V = s_m \oplus H_2(t(R + Q_V, rP_V)), \tag{3-4}$$

$$h = (U, U_1^{(1)}, U_2^{(1)}, \ldots, U_{i_m - 1}^{(1)}, U_1^{(2)}, U_2^{(2)}, \ldots, U_{i_m - 1}^{(2)}, Q_V, V) \tag{3-5}$$

Here s_m is the session key and h is the enable header. Finally the cipher-text is produced by computing and broadcasting: $C = (h, ENC_{s_m}(m))$.

Note that the center constructs the matrix relations only if the set Γ of revoked receivers alters and it is not necessary to construct the matrix relations at each broadcast.

(3) *Decryption*: At the receiver u, upon receiving a broadcast message $C = (h, ENC_{s_m}(m))$ where $h = (U, U_1^{(1)}, U_2^{(1)}, \ldots, U_{i_m-1}^{(1)}, U_1^{(2)}, U_2^{(2)}, \ldots, U_{i_m-1}^{(2)}, Q_V, V)$, the receiver finds i_j such that $u \in S_{i_j}$ (in case $u \in \Gamma$ the result is null). He then extracts the corresponding key $k_{i_j} = sP_{i_j} + s_{i_j} P_{i_j}$ from I_u, computes the session key s_m as:

$$U^{(1)\prime} = \sum_{k=2}^{i_m} a_{jk} U_{k-1}^{(1)}, \quad U^{(2)\prime} = \sum_{k=2}^{i_m} a_{jk} U_{k-1}^{(2)}, \tag{3-6}$$

$$s_m = V \oplus H_2(t(U, sP_{i_j} + s_{i_j} P_{i_j}) \bullet t(R + Q_V, -U^{(2)\prime}) \bullet t(-U^{(1)\prime}, P_{i_j})). \tag{3-7}$$

Then, the receiver can get the plaintext m by executing the symmetrical cryptographical arithmetic $ENC_{s_m}(\bullet)$. We will skip over the proof for the correctness of the above decryption process because of the paper length limit.

4 Desirable Features

Our broadcast encryption scheme can provide desirable features such as traitor tracing, asymmetry, revocation and high efficiency etc.

(1) *Traitor Tracing*. Traitor tracing is an algorithm for finding illegal users who help build a pirate decoder with their secret keys. Naor et al. [2] presented a subset tracing algorithm to locate the traitors in logarithmic time if the system is secure and satisfies a bifurcation property. In our system, we will modify their algorithm and make it work for our identity-based broadcast scheme.

Subset tracing: An important procedure in our tracing mechanism is one that given a set Γ of revoked receivers, a partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $N \setminus \Gamma$ and an illegal box outputs one and only one of two possible outputs: (a) The box cannot decrypt when the encryption is done with partition S, then Γ contains a traitor. (b) The box can decrypt when the encryption is done with partition S, then finds a subset S_{i_j} such that S_{i_j} contains a traitor. Such a procedure is called subset tracing and is described below.

From the literature [2], the Complete Sub-tree method satisfies the above requirement: in the Complete Sub-tree method each subset, which is a complete sub-tree, can be split into exactly two equal parts, corresponding to the left and right sub-trees.

Traitor tracing algorithm: We now describe the tracing algorithm, assuming that we have a good subset tracing procedure (which is described below). The algorithm maintains a set Γ of revoked receivers and a partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $N \setminus \Gamma$. At each phase one of the subsets is partitioned, and the goal is to partition a subset only if it contains a traitor. Each phase initially applies the subset-tracing procedure with the current partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$. (a) *Tracing initialization:*

At the first phase execute the Subset-tracing procedure **A** with the partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$. The procedure will output a set S_{i_j} which contains the traitor. Let $S_{i_x} \Leftarrow S_{i_j}$. If S_{i_x} contains only one possible candidate - it must be a traitor and we permanently revoke this user; Otherwise we split S_{i_x} into two roughly equal subsets, that is $S_{i_x} = S_{i_{x1}} \bigcup S_{i_{x2}}$, and continue with $\Gamma \Leftarrow \Gamma \bigcup S_{i_{x1}}$ and the new partitioning $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $\mathrm{N} \setminus \Gamma$. (b) *Tracing algorithm:* Do the following phase again and again until the algorithm find the traitor. Apply the Subset-tracing Procedure **B** with the new partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ that the previous phase outputs. If the procedure outputs that the box cannot decrypt with S, then $S_{i_{x1}}$ must contain the traitor. Let $S_{i_x} = S_{i_{x1}}$; Otherwise, the procedure outputs that the box can decrypt with S, then $S_{i_{x2}}$ must contain the traitor. Let $S_{i_x} = S_{i_{x2}}$. If S_{i_x} contains only one possible candidate - it must be a traitor and we permanently revoke this user, this doesn't hurt a legitimate user. Otherwise we split S_{i_x} into two roughly equal subsets, that is $S_{i_x} = S_{i_{x1}} \bigcup S_{i_{x2}}$, and continue with $\Gamma \Leftarrow \Gamma \bigcup S_{i_{x1}}$ and the new partitioning $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $\mathrm{N} \setminus \Gamma$. The existence of such a split is assured by the bifurcation property.

Subset-tracing procedure **A:**

Input: the partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $\mathrm{N} \setminus \Gamma$.

Output: the subset S_{i_j} that contains the traitor.

The Subset Tracing procedure needs to find a subset S_{i_j} that contains a traitor. The tracer encrypts a message with a set Γ of revoked receivers. The tracer inputs the cipher-text to the pirate box and observes whether it decrypts the cipher-text correctly or not. If its output is (a) correct on the input where a set of revoked receivers is Γ and (b) incorrect on the input where a set of revoked receivers is $\Gamma \bigcup S_{i_j}$, then the tracer decides S_{i_j} is the subset that contains the traitor. The following is the detailed tracing algorithms.

Step 1*:* Let Γ be the set of revoked receivers. Set $Lo \leftarrow 0$, $Hi \leftarrow i_m$ and $l \leftarrow 1$. Then, repeat the following producers for $1 \le l \le \lfloor \log_2 i_m \rfloor + 1$:

(1.1) Set $Mid \leftarrow \lfloor (Lo + Hi)/2 \rfloor$ and $\Gamma \leftarrow \{S_{i_1}, S_{i_2}, \ldots, S_{i_{Mid}}\}$. Then, encrypt a message with the partition $S = S_{i_{Mid+1}}, S_{i_{Mid+2}}, \ldots, S_{i_m}$ .Input the cipher-text to the pirate box and observe the output: If D outputs the correct plaintext, then set $Lo \leftarrow Mid$; Otherwise (D outputs incorrectly), set $Hi \leftarrow Mid$.

(1.2) Increase l by one and go to (1.1).

Step 2: After the $\lfloor \log_2 i_m \rfloor + 1$ tests, there exists one subset $S_{i_j} \in \{S_{i_1}, S_{i_2}, \ldots, S_{i_m}\}$ which satisfies the following two conditions for some $l \in \{1, 2, \ldots, \lfloor \log_2 i_m \rfloor + 1\}$: Φ

the pirate box outputs the correct plaintext with Γ; ② the pirate outputs the incorrect plaintext with $\Gamma \cup S_{i_j}$. Then Output S_{i_j} as the subset that contains the traitor.

Subset-tracing Procedure **B:**
*Input***:** the partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $N \setminus \Gamma$.

Output: "correct" or "incorrect".

Encrypt a message as the section 3.2 with the partition $S_{i_1}, S_{i_2}, \ldots, S_{i_m}$ covering all users in $N \setminus \Gamma$ and input the cipher-text to the pirate box D and observe the output: If D outputs the correct plaintext, on the input, then output "correct"; Otherwise (D outputs incorrectly), output "incorrect".

(2) *Asymmetry*. The secret key $k_i = sP_i + s_i P_i$ is composed of two parts. The supplier and the users have one half each. The users' corresponding half $s_i P_i$ (or s_i) is not leaked to the broadcast center. Thus, our scheme is asymmetric and supports non-repudiation for traitor tracing.

(3) *Revocation.* With the subset cover mechanism, our scheme can provide key revocation without redistribute new keys to the other receivers. To revoke a receiver, the center does nothing but place the receiver into the group Γ of which users' decryption privileges is revoked.

(4) *Efficiency.* Note that in the above scheme, the length of the enable header is only $2(i_m\text{-}1)$, where $i_m = \mu \log(N/\mu)$ and $\mu = |\Gamma|$. The main advantage of our scheme over the previous ones is that it is computationally much more efficient as we just need to compute $2(i_m\text{-}1)$ scalar multiplications instead of large numbers of pairings in the encryption process. Note also that in the decryption process, we need three pairing computations, $2(i_m\text{-}1)$ scalar multiplications and $2(i_m\text{-}1)$ additions in the group G_1.

5 Conclusion

We have shown how pairing based systems can allow a simple way to construct a broadcast encryption mechanism for the digital contents copyright protection.

References

1. D. Boneh and M. Franklin: Identity based encryption from the Weil pairing. In Advances in Cryptology – CRYPTO 2001, Springer-Verlag LNCS 2139(2001) 213–229,.
2. D. Naor, M. Naor and J. Lotspiech: Revocation and Tracing Schemes for Stateless Receivers, Advances in Cryptology-Crypto'01, LNCS 2139, Springer-verlag(2001) 41-62.
3. A. Fiat and M. Naor. Broadcast encryption. In Advances in Cryptology, Crypto'93, Lecture Notes in Computer Science 773, Springer Verlag (1994)480–491.

Multimedia Digital Right Management Using Selective Scrambling for Mobile Handset

Goo-Rak Kwon, Tea-Young Lee, Kyoung-Ho Kim, Jae-Do Jin, and Sung-Jea Ko

Department of Electronics Engineering, Korea University,
5-1 Anam-Dong, Sungbuk-ku, Seoul 136-701, Korea
Tel: +82-2-3290-3228
{grkwon, spirou79, tongsin, jiny218, sjko}@dali.korea.ac.kr

Abstract. In this paper, we propose a novel solution called *joint encryption*, in which audio and video data are scrambled efficiently by using modified phase scrambling, modified motion vector, and wavelet tree shuffling. Experimental results indicate that the proposed joint encryption technique is very simple to implement, has no adverse impact on error resiliency and coding efficiency, and produces level of security.

1 Introduction

In many systems for scrambling digital A/V [1, 2, 3], A/V data is first subject to compression, and then the compressed A/V data is treated as ordinary data and is encrypted/decrtpted using traditional cryptographic techniques such as digital encryption standard (DES) [4], advanced encryption standard (AES) [5], and RSA [6]. Due to the high bit rate of A/V (even compressed A/V), they usually add a large amount of processing overhead to meet the real-time A/V delivery requirement [3, 7]. To address the complexity issue, [7] is proposed to shuffled the DCT coefficients in each block. This technique, although very simple to implement, may increase the bit rate of the compressed video. To reduce the amount of processing overhead, selective encryption of the MPEG compressed video data [3, 8] has been proposed. Although it may achieve real-time encryption/decryption for low bit rate sequence, it will have real-time implementation problem for high bit rate sequence.

In this paper, we propose a novel solution called *joint encryption*, in which audio and video data are scrambled efficiently by using phase scrambling, modified motion vector, and wavelet tree shuffling. In this approach, A/V compressed data is encrypted and scrambled without affecting the compressing efficient significantly. The proposed has the main feature that the scrambling process in very simple and efficient. And it also provides level of security, has no adverse impact on error resiliency and coding efficiency, and allows more flexible selective encryption.

The rest of the paper is organized as follows. The architecture of the proposed DRM system is presented in Section 2. Section 3 and 4, respectively, describe

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1098–1103, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the proposed joint A/V encryption methods in detail. Experimental results are shown in Section 5 and Section 6 gives some concluding remarks.

2 Proposed DRM Architecture

Fig. 1 shows the architecture of the proposed DRM system composed of two entities: Contents Provider and User. Each of the two entities has a common Public Mode Set (PMS). The encrypted A/V file and metadata are created by Contents Provider using the PMS and then we adopted DES for encryption of metadata. In user part, every mobile device has unique Device-Key which distinguishes individual mobile devices. This Key has a role in access authentication and Metadata encryption. The process of the proposed DRM system is described as follows:

1. User accesses Server of Contents Provider and sends ID, Password, and a Device-Key. Then, User can be authenticated with ID and Password by Server.
2. Server encrypts metadata with Device Key.
3. Server transmits the encrypted A/V file and metadata to User's mobile device.
4. Mobile device decrypts received files with PMS and Device Key.
5. User can watch and listen to the played A/V data.

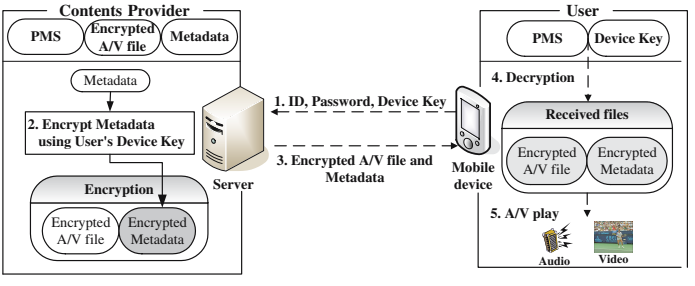


Fig. 1. The architecture of the proposed DRM system

3 Proposed Joint Audio Encryption Method

3.1 Modified Phase Coding

The proposed copyright management consists of the scrambling processes. The former based on modified phase coding embeds the secret key imperceptibly in the music signal. The Secret Key is used for authorizing the contents. And the latter scrambles the original music signal for degrading the tone quality. In the process of MP3 compression, after transform, both the Secret Key and

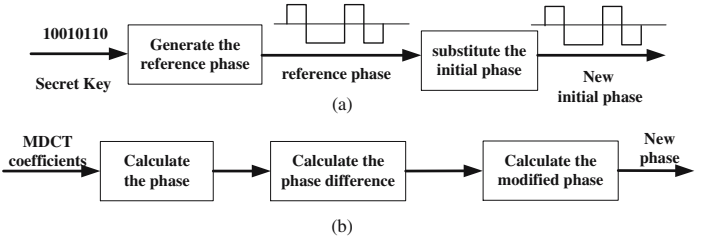


Fig. 2. The procedures of modified phase coding. (a) The procedure for embedding a Secret Key in the first frame. (b) The procedure for preserving the relative phase between frames.

the Scrambling Key are embedded by using the phase of coefficients in the frequency domain. The scrambled data is quantized to achieve the target bitrate. The psycho-acoustic model is used for imperceptibility. Finally, the quantized coefficients are converted into an encoded bitstream through Huffman coding. Fig. 2 represents the procedure of the modified phase coding. The procedure consists of two main parts. The first part inserts the Secret Key in the first frame. As shown in Fig. 2 (a), the Secret Key generates the reference phase. The new initial phase, $\phi'_0(\omega_{\mathrm{k}})$ is equal to the reference phase, $\phi_{\mathrm{ref}}(\omega_{\mathrm{k}})$ where ω_k is the k^{th} frequency coefficient. The second part preserves the relative phase between frames. In Fig. 2 (b), we calculate the phase of the current frame and the phase differences between adjacent frames to keep relative phase differences, that is given by

$$\Delta\phi_n(\omega_k) \;=\; \phi_n(\omega_k) - \phi_{n-1}(\omega_k), \quad (1 \le n < N), \tag{1}$$

where $\Delta\phi_n(\omega_k)$ is the phase differences between the n^{th}frame and $n-1^{th}$frame, $\phi_n(\omega_k)$ is the phase of the n^{th}frame, and N is the number of the phase-coded frames. The new phase,$\phi'_n(\omega_k)$, is obtained by adding the phase difference of the current frame and the modified phase of the previous frame as follows:

$$\phi'_n(\omega_k) \;=\; \phi'_{n-1}(\omega_k) + \Delta\phi_n(\omega_k). \tag{2}$$

Fig. 3 shows the scrambling procedure. The proposed method needs to use the Device Key, which is shared with the encoder and the decoder, for perfect recovery of the quality of music. The scrambling process is to select the bands to be converted by using the Device Key and to reverse the phase of conversion bands. The proposed method does not change the total number of bits of MP3 files since there is no magnitude change and additional information.

In the MP3 decoding process, we calculate the reference phase by using reconstructed MDCT coefficients after dequantization. If the Secret Key is detected correctly, the recovering procedure for phases in the conversion bands is executed.

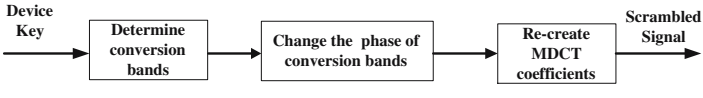


Fig. 3. The scrambling process

4 Proposed Joint Video Encryption Method

4.1 Modified Motion Vector

Fig. 4 shows the detail method of encryption and decryption. In Fig. 4 (a), the encryption method modifies the MV of the MB that is extracted from video encoder. The modified MV is obtained by

$$(MV'_x,\ MV'_y) = (MV_x + \Delta x_i,\ MV_y + \Delta y_i), (0 \le i \le n) \tag{3}$$

where MV_x and MV_y of MB are modified by Δx_i and Δy_i of i^{th} mode's weighted parameter in the PMS. Metadata consists of selected modes for the modified MVs. These modes are selected from the PMS. Metadata is also encrypted with a Device Key received from User. In the decryption method shown in Fig. 4(b), the encrypted metadata is decrypted with a Device Key. Then, MVs are reconstructed by using the decrypted metadata and the PMS. Therefore, the decrypted video is finally obtained.

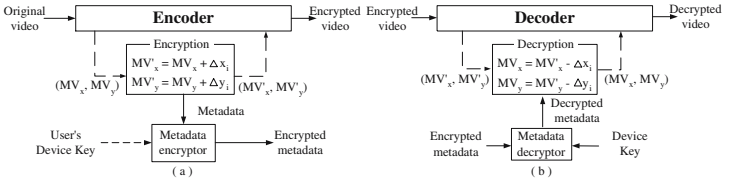


Fig. 4. The proposed Encryption/Decryption process: (a) Encryption process. (b) Decryption process.

4.2 Wavelet Tree Shuffling

To increase the level of security, traditional technique is block shuffling and rotation because it is very simple to implement. However, it is difficult to decide the block size due to it generally needs to be selected to trade security for statistical coding impact, with larger and fewer blocks producing less security but less impact on statistical coding. To solve these problems, wavelet tree shuffling in wavelet transform is proposed without considering the block size decision.

We use 4-level wavelet transform. With 4-level decomposition, there are 13 sub-band. The wavelet coefficients are grouped according to wavelet trees except the coefficients of high frequency bands ($L1^1$, $L1^2$, and $L1^3$) which contain little energy. We use the coefficients in sub-band $L4^1$, $L4^2$, and $L4^3$ as roots to form wavelets trees [10]. Wavelet trees will be shuffled according to a shuffling table generated using PMS and a Device Key.

5 Experimental Results

Through simulation, we have compared the performance of the proposed method with the existing methods. In audio scrambling, we use three test music and the LAME MP3 codec.



(a) (b)

Fig. 5. Simulation results (30 fps , 64 kbit/s): (a) Decoded picture without scramble. (b) Decoded picture with scramble.

In video scrambling, Fig. 5 shows the displays of the encrypted video on two differential authorized and unauthorized mobile devices. And experimental testing is performed on "foreman" sequence. The encryption is encoded with MPEG-4 Simple Profile@Level 3 (SP@L3). Fig. 5 shows the decoded picture with authorized device [Fig. 5(a)] and the reconstructed picture with unauthorized mobile handset [Fig. 5(b)].

Table 1. Comparisons of different scrambling techniques for "foreman" sequence, all at the same PSNR

Algorithm	Bit rate (kbit/s)	Bit Overhead	Complexity	Security
No scrambling	47.23	0	N/A	No
Sign+Slice+MV(sign) [9]	53.38	23.6	20(%), +slice memory	high
Shuffle within block [7]	73.21	55	very simple	low
Proposed (A/V scrambling) WTS + MMV + (MPC)	53.23	12.7	+slice memory (shuffling table)	high+

Table 1 summarizes the estimated complexity, security level for different scrambling techniques. The estimated complexity for sign encryption is only about 15-20% of the wholesale encryption. In [9], several scrambling algorithms are mixed. Therefore, Security is high but Scrambling overhead increases the bit rate by about 23.6%. In [7], implementation is very simple but overhead increment is about 55%. In our experiments, we found that for all scrambling schemes tested, if motion vector information was not encrypted, then we could have the perception that foreman head was moving in the scene, although the detail may not be visible. Therefore, we focused on the level of security and the overhead reduction, as shown in Table 1.

6 Conclusions

We proposed a novel solution called *joint encryption*, in which audio and video data are scrambled efficiently by using phase scrambling, modified motion vector, and wavelet tree shuffling. This approach showed the minimal cost encryption

scheme for securing the copyrighted MP3 and MPEG-4 A/V data using the DES encryption technique. The proposed scrambling techniques appear to achieve a very good compromise between several desirable properties such as speed, security, and file size, therefore it is very suitable for network A/V applications.

References

1. MPEG4 IPMP FPDAM. ISO/IEC 14496-1:2001/AD3, ISO/IEC JTC1/SC 29/WG11 N4701, **3** (1996)
2. Agi, I. and Gong, L.: An empirical study of secure MPEG video transmissions. in The Internet Society Symposium on Network and Distributed System Security (1996)
3. Meyer, J. and Gadegast, F.: Security Mechanisms for Multimedia Data With the Example MPEG-1 Video. http://www.cs.tuberlin.de/phade/phade/secmpeg.html (1995)
4. Data Encryption Standard. FIPS PUB 46. **1** (1977)
5. Advanced Encryption Standard (AES). FIPS PUB 197. (2001)
6. Rivest, R., Shamir, A. and Adleman, L.:A method for obtaining digital signaures and public key cryptosystems. Commun. ACM. **21** (1978) 120–126
7. Tang, L.:Methods for encrypting and decrypting MPEG video data efficiently. Proc. Fourth ACM (1996) 219–229
8. Maples, T. and Spanos, G.:Performance study of a selective encryption scheme for the security of networked, real-time video. 4th Inter. Conf. Computer Communications and Networks (1995)
9. Zeng, W. and Lei, S.:Efficient Frequency Domain Selective Scrambling of Digital Video. IEEE Trans. Multimedia. Vol. 5 (2003) 118–129
10. Shapiro, M.:Embedded Image Coding using Zerotrees of Wavelet Coefficients. IEEE Trans. Signal Processing. Vol. 41 (1993) 3445–3462

Design and Implementation of Crypto Co-processor and Its Application to Security Systems

HoWon Kim[1], Mun-Kyu Lee[2], Dong-Kyue Kim[3], Sang-Kyoon Chung[4], and Kyoil Chung[1]

[1] Department of Information Security Basic, Electronics and Telecommunications Research Institute, South Korea
howonkim@gmail.com, kyoil@etri.re.kr

[2] School of Computer Science and Engineering, Inha University, South Korea
mklee@inha.ac.kr

[3] School of Electrical and Computer Engineering, Pusan National University, South Korea
ddkim1@pusan.ac.kr

[4] Department Department of Computer Science, Inje University, South Korea
skkim@cs.inje.ac.kr

Abstract. In this paper, we will present the design and implementation of the crypto coprocessors which can be used for providing the RFID security. By analyzing the power consumption characteristics of our crypto coprocessors, we can see which crypto algorithms can be used for highly resource constrained applications such as RFID. To provide the security mechanism such as integrity and authentication between the RFID tag and the RFID reader, crypto algorithm or security protocol should be implemented at the RFID tag and reader. Moreover, the security mechanism in the RFID tag should be implemented with low power consumption because the resources of RFID tag are highly constrained, that is, the RFID tag has low computing capability and the security mechanism which operates in the RFID tag should be low power consumed. After investigating the power consumption characteristics of our designed crypto processors, we will conclude which crypto algorithm is suitable for providing the security of RFID systems.

Keywords: Crypto Coprocessor, Crypto Algorithm, Power Consumption.

1 Introduction

Though the Radio frequency identification (RFID) is not newly developed technology, the term RFID has been used for only a couple of decades. In the 1980s and 1990s, long after the early use of this kind of technology in World War II, the evolutions of semiconductor technology enabled RFID tags to be used in many kinds of applications such as supply chain management, library and rental sectors, baggage tagging, access control, animal identification, document tracking,

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1104–1109, 2005.
© Springer-Verlag Berlin Heidelberg 2005

security, manufacturing system, electronic appliances, and healthcare systems. By using RFID, companies can achieve significant cost savings and increase the range of service they can provide. Surely, the people can get many benefits from this technology.

As RFID technology becomes commonplace, new problems such as the security problems may arise. These problems may include espionage, monitoring, counterfeiting, cloning, and privacy violation. Without proper protection, RFID tags could violate consumer privacy by revealing sensitive information and even destroy the benefits of RFID technology.

In these days, drug counterfeiting has become a global problem. So we can use RFID technology to prevent the counterfeiting of drug and check the supply chain history. However, what happened if malicious people cloned the RFID tag which is attached to drug and then deceived innocent consumers and retailers? It may destroy the benefits of RFID technology and the consequences of compromised pharmaceutical can be catastrophic.

There are many research works and studies to address the RFID vulnerability problems. Cryptographic and non-cryptographic solutions have been proposed and are now underway [WSRE03, VA03, FDW04]. Among these solutions, cryptographic approaches can provide strong security and privacy to RFID systems. However, the resource limitations in RFID tags are big barrier to adopt the crypto algorithms in RFID tags. The resource limitation of RFID is more serious than that of smart card, which is well known to the public as a resource limited application [KKB+03]. Therefore, there are necessities to investigate the characteristics of crypto algorithms in the viewpoint of power consumption and then develop the technology to make crypto algorithms operable with low power consumed. Until recently, crypto research community mainly focused on the performance and the level of security issues in crypto algorithms [LCHR04, LL03, KPR+03]. However, the advent of the RFID will increase interests on the design of light weight crypto algorithm and its low power consumed implementation like a main purpose of our paper.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of the background on power consumption characteristics of a circuit and the methodology for power consumption estimation. Section 3 describes about the power consumption characteristics of crypto coprocessors that we made. Our results and some suggestions on RFID security are also given in Section 3. Finally, we end this paper with some conclusions.

2 The Methodology for Power Consumption Estimation at the Gate Level

In general, the power dissipated in a circuit falls into two categories such as dynamic power and static power. Dynamic power is the power dissipated when the circuit is active. When a circuit is in active, the static power is far less than the dynamic power. So we can disregard this factor when the target design is most in active mode. However, for designs that are usually idle, we should consider the factor of static power.

To estimate the power consumption of a certain crypto algorithm, we have used the following methodology. (1) First, we have extracted the input and output information of our HDL described crypto coprocessor by using HDL compiler tool. More technically speaking, we call the extracted output file as "forward SAIF (Switching Activity Inter-change Format) annotation file." This file provides the input and output port information and hierarchical information of our crypto coprocessor. (2) Second, we have compiled and simulated crypto coprocessor with inputs of forward annotation files and test vectors. In this step, we can get the most important information for logic level power estimation, that is, the switching activity of the circuit. After this step, we can get the "SAIF back annotation file" which has toggling information of all nodes of our design. (3) Lastly, we have simulated and estimated the power consumption of our design after synthesis with the back annotation information. Now we get the estimated value of power consumption of a crypto coprocessor.

For more accurate measurement of power consumption, we have used well-modeled technology library that a fabrication vendor carefully prepared. We have used the 0.25 μm CMOS technology library and we have verified that the power estimation technique of a fully annotated design can be within 20% of power analysis result with SPICE.

3 Security Mechanism for RFID Systems

3.1 Security Mechanism for Resource Constrained RFID Tag

Many research works have done to deal with security problem of RFID systems. Blocker tags mechanism is proposed to protect consumers from tracking [JRS03]. This method is suitable for low cost RFID tag which cost less that US$ 0.05. However, the Block tag may act as a DOS (Denial of Service) attacking device. Also, many authentication protocols targeted to low-cost RFID tags have been proposed. The simplest protocol is using exclusive OR operation for RFID tag to reader and read to RFID tag communication as follows [VA03]: $R \rightarrow T \ : \ x \oplus k_1, \ \ T \rightarrow R \ : \ x \oplus k_2$. This protocol would be provably secure if the keys k_1 and k_2 are selected randomly in each run of the protocol. However, this means the key establishment problem is crucial factor for security of this protocol.

There are also many research works to provide strong security to RFID systems by using crypto algorithms. For example, Hash Lock mechanism is proposed as a reader authentication by a RFID tag [WSRE03]. This mechanism can be used to control the access by reader by authenticating the reader. In this mechanism, a tag may be locked so that it refuses to reveal its ID until it is unlocked. The locked tag has a value of meta-ID y, and it is only unlocked by presentation of a key or PIN value x such that $y = h(x)$ for a standard one-way hash function h.

RFID security protocols based on the symmetric key crypto algorithm such as AES are also proposed to authenticate the RFID reader by RFID tag [FDW04]. The AES can be used for realizing the challenge response protocol according to ISO/IEC 9798-2. This protocol is represented as following equation. $R \rightarrow T \ :$

r_B, $T \rightarrow R : E_k(r_B)$. where r_B is a random value as a challenge. If the RFID reader is legal one, the RFID reader can decrypt the response value from the tag, $E_k(r_B)$, with a correct k key value. It means only legal RFID reader can access the RFID tag.

3.2 Power Consumption Characteristics of Crypto Algorithms

Now we present the power consumption estimation results of crypto algorithms when they are made to hardware block. And we investigate if we can use these algorithms at the highly resource constrained environment such as RFID tag.

Hash Function: Hash function is the core crypto algorithm which can be used to authenticate the reader by a tag. Though it is reported that there exist some security weaknesses in SHA-160 (also called as SHA-1), most people still believe that SHA-1 can provide enough security to RFID systems.

The SHA-1 takes as input a message with a maximum length of less that 264 bits and produces as output a 160-bit message digest. The input is processed in 512-bit blocks. It requires 4 rounds of 20 operations each, resulting in a total of 80 operations, to generate the message digest. The 4 rounds have a similar structure, but each uses different primitive non-linear function logics such as f_1, f_2, f_3, and f_4. Details can be found in [Sta03].

To measure the power consumption of SHA-1, we have implemented the SHA-1 block as it is, which means we have not applied special hardware design techniques (that is, gated clock, parallelism, operand isolation, state machine optimization, glitch free technique, etc.) to reduce the power consumption of SHA-1. We have only applied the pipeline technique to implement our prototyped SHA-1 hardware block. This pipeline technique exploits the characteristics of SHA-1 that requires a different non-linear function every 20 clock cycles [Sta03]. After implementing SHA-1 with HDL (Hardware Description Language), we have measured consumed power according to the methodology shown in this paper. Results tell us that the amount of power consumption of SHA-1 is about $1,320\mu W$ at 10MHz operating frequency, which means SHA-1 consumes more power than AES. Actually, because SHA-1 hash function is made for the target of software applications, hardware implemented SHA-1 consumes more power than might be thought.

Symmetric Key Crypto Algorithm: Symmetric key crypto algorithm can be used to authenticate the RFID reader and some efforts have done to apply AES crypto algorithm to a RFID tag. The AES crypto operates on blocks of data, so called State, which a fixed size of 128 bits. The state is organized as a matrix of 4 rows and 4 columns of bytes. The key lengths are variable with 128 bits, 192 bits, or 256 bits. For our prototyped AES hardware module, we have implemented the AES with key size of 128 bits.

The AES crypto algorithm has room to reduce the power consumption, that is, the power consumption of S-box can be reduced when it is made with combinational logics instead of using memory such as ROM. It is possible from the fact that the non-linear function of S-box of the AES has relatively a good structure

to make it with combinational logics. Also, we can further reduce the power consumption by applying the gated clock power reduction technique to the registers in the AES crypto module. The main operation logic of the AES crypto module can be made with a power optimization technique and a glitch free design technique. Glitch is generally known that it consumes much power consumption. The SubByte, MixCulumn, AddRoundKey operation can be executed with the size of 8 bits and then this can save the power consumption. Our power measured results show that AES consumes $810\mu W$ at 10MHz operating frequency. This is less than that of Hash function.

The power consumption characteristics of crypto algorithms which we have implemented with hardware logics are summarized at Table 1. From these data, we can see that the Hash function consumes more power than AES. For reference, the contactless smart card requires the power consumption of less than $30mW$. For the case of RFID tag, the passive RFID tag requires around less than $1mW$ power consumption. Actually, the power consumption specification of RFID depends on the distance between RFID reader and tag and the application of RFID tag. For these reasons, we can not say the exact specification of power consumption of RFID tag, but, we can estimate the reasonable power consumption amount of crypto algorithms with considering the source's power capability and the reading range.

Table 1. Power consumption characteristics of crypto algorithms which can be applied to RFID security

	Hash Function (SHA-1)	Symmetric Key Crypto Algorithm (AES)
Power Consumption (μW)	1,320 (μW) (@ 10MHz)	810 (μW)(@ 10MHz)

Regarding to the Hash function, many people believe that the Hash lock mechanism (or variants of Hash lock mechanism) is a good solution to be commercially available. So we think that we have to make techniques to implement the conventional Hash function such as SHA-1 and MD5 to low power consumed or develop new Hash function which is appropriate to high resource constrained environments such as RFID tags. It will be our future research topics. The AES consumes reasonable power and it has many architectural features to further reduce the power consumption and we can apply conventional low power techniques to make a hardware based AES crypto module.

4 Concluding Remarks

We have analyzed the power consumption characteristics of crypto algorithms which can be applied to highly resource constrained environment, the RFID systems. Two kinds of crypto algorithms such as hash function (SHA-1) and symmetric key crypto algorithm (AES) were modeled with HDL language and then

implemented with hardware with verifications. We have analyzed the power consumption characteristics according to the power estimation methodology which we have described in this paper. The Hash function, SHA-1, and symmetric key crypto algorithm, AES, consumes $1,320\mu W$ and $810\mu W$, respectively. In implementing the SHA-1 hardware crypto module, we have applied the pipeline technique to reduce the power consumption. And for implementing the AES hardware crypto module, we have applied various low power design techniques to reduce the power consumption. AES crypto algorithm has good features from the viewpoint of power consumptions. From the power consumption estimation result, we can say that the AES crypto algorithm can be a good crypto algorithm which can be implemented at the RFID tag and reader. And regarding to the Hash function, more research works should be done to make new Hash algorithm which is appropriate to highly resource constrained applications.

References

[FDW04] Martin Feldhofer, Sandra Dominikus, and Johannes Wolkerstorfer. Strong authentication for RFID systems using the AES algorithm. In *CHES*, pages 357–370, 2004.

[JRS03] Ari Juels, Ronald L. Rivest, and Michael Szydlo. The blocker tag: selective blocking of RFID tags for consumer privacy. In Vijay Atluri and Peng Liu, editors, *Proceedings of the 10th ACM Conference on Computer and Communication Security (CCS-03)*, pages 103–111, New York, October 27–30 2003. ACM Press.

[KKB+03] Wonjong Kim, Seungchul Kim, Younghwan Bae, Sungik Jun, Youngsoo Park, and Hanjin Cho. A platform-based soc design of a 32-bit smart card. *ETRI Journal*, 25(6):510–516, December 2003.

[KPR+03] Ju-Sung Kang, Bart Preneel, Heuisu Ryu, Kyo Il Chung, and Chee Hang Park. Pseudorandomness of basic structures in the block cipher KASUMI. *ETRI Journal*, 25(2):89–100, April 2003.

[LCHR04] Dong Hoon Lee, Seongtaek Chee, Sang Cheol Hwang, and Jae-Cheol Ryou. Improved scalar multiplication on elliptic curves defined over $F_{2^{mn}}$. *ETRI Journal*, 26(3):241–251, June 2004.

[LL03] Chanho Lee and Jeongho Lee. A scalable structure for a multiplier and an inversion unit in $GF(2^m)$. *ETRI Journal*, 25(5):315–320, October 2003.

[Sta03] W. Stallings. *Cryptography and Network Security : Principles and Practice.* Prentice Hall, Upper Saddle River, New Jersey, USA, third edition, 2003.

[VA03] Istv An Vajda and Levente Butty An. Lightweight authentication protocols for low-cost RFID tags, August 06 2003.

[WSRE03] Stephen A. Weis, Sanjay E. Sarma, Ronald L. Rivest, and Daniel W. Engels. Security and privacy aspects of low-cost radio frequency identification systems. In *SPC*, pages 201–212, 2003.

Continuous Speech Research Based on HyperSausage Neuron

Wenming Cao[1,2], Jianqing Li[1], and Shoujue Wang[2]

[1] Institute of Intelligent Information System, Information College,
Zhejiang University of Technology, Hangzhou 310032, China
[2] Institute of Semiconductors, Chinese Academy of Science, Beijing 100083, China
wmcao@semi.ac.cn

Abstract. In this paper, we presents HyperSausage Neuron based on the High-Dimension Space(HDS), and proposes a new algorithm for speaker independent continuous digit speech recognition. At last, compared to HMM-based method, the recognition rate of HyperSausage Neuron method is higher than that of in HMM-based method.

1 Introduction

In the recent years, Wang Soujue[2,3,6,7] proposed a HyperSausage Neuron Learning theory in high dimensional space (HDS), which uses the HDS descriptive geometry as tools, maps multidimensional digital signals to be points in HDS, and discuss relationship among multidimensional digital signals via analysis of point distribution and geometric object characteristics in HDS. Since this method is simple and feasible, and is able to accelerate computation by parallel operation, it has get remarkable achievements in many application fields[8,9].

In the light of descriptive geometry and notions in set theory, this paper defines the HyperSausage Neuron, presents some property with respect to the Geometrical Learning based on the HyperSausage Neuron Learning theory, finally, this paper also proposes a new algorithm for speaker independent continuous digit speech recognition on the basis of the HDS point dynamic searching theory without endpoints detection and segmentation. First from the variant digit syllables in real continuous digit speech, we establish the covering area in feature space for continuous speech. During recognition, we make use of the HyperSausage Neuron Learning dynamic searching theory in HDS[4,5] to do recognition, and then get the satisfying recognition results. At last, compared to HMM-based method[1], from the development trend of the comparing results, as sample amount increases, the difference of recognition rate between two methods will decrease slowly, while sample amount approaching to be large enough, two rates all close to 100% little by little. As seen from the results, the recognition rate of HDS Geometrical Learning method is higher than that of in HMM-based method, because, the Geometrical Learning describes the morphological distribution of speech in HDS, whereas HMM-based method is only a probability distribution, whose accuracy is certainly inferior to HyperSausage Neuron.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1110 – 1115, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 HyperSausage Neuron

A mathematical description of HyperSausage Neuron P of class A:

$$\mathrm{P}=U_i \mathrm{P}_i,$$

$$P_i = \left\{ x \mid \rho(x, y) \le k, y \in \mathrm{B}_i, x \in R^n \right\}, \tag{1}$$

$$\mathrm{B}_i = \left\{ x \mid x = \alpha S_i + (1-\alpha) S_{i+1}, \alpha = [0,1] \right\}. \tag{2}$$

Where s_i is a sample of a class A in feature space. Let

$$d\left(x, \overline{x_1 x_2}\right) = \min_{\alpha \in [0,1]} d\left(x, \alpha x_1 + (1-\alpha) x_2\right) \tag{3}$$

be the distance of x and line segment $\overline{x_1 x_2}$. Then

$$d^2\left(x, \overline{x_1 x_2}\right) = \begin{cases} \|x - x_1\|^2, q(x, x_1, x_2) < 0, \\ \|x - x_2\|^2, q(x, x_1, x_2) > \|x_2 - x_1\|, \\ \|x - x_1\|^2 - q^2(x, x_1, x_2), otherwise \end{cases} \tag{4}$$

$$q(x, x_1, x_2) = < (x - x_1), \left(\frac{x_2 - x_1}{\|x_2 - x_1\|} \right) > . \tag{5}$$

And the HyperSausage Neurons $\mathrm{S}(x_1, x_2; r)$ is:

$$S(x_1, x_2; r) = \left\{ x \mid d^2\left(x, \overline{x_1 x_2}\right) < r^2 \right\} \tag{6}$$

If x^1 and x^2 are the same point in feature space, then $\mathrm{d}\left(x, \overline{x_1 x_2}\right)$ is equivalent to $d(x, x_1)$, and $\mathrm{S}(x_1, x_2; r)$ is equivalent to $S(x_1; r)$ (HyperSausage Neurons becomes a hypersphere). If x^1 and x^2 are different points, then a HyperSausage Neurons $\mathrm{S}(x_1, x_2; r)$ is a connection between x^1 and x^2, with compact coverage compared to a hypersphere.

A new neuron model, a HyperSausage Neurons, is defined by the input-output transfer function:

$$f(x; x_1, x_2) = \phi\left(d\left(x, \overline{x_1 x_2}\right)\right) \tag{7}$$

Where $\phi(\cdot)$ is the non-linearity of the neuron, the input vector $\mathrm{x} \in R^n$, and $x_1, x_2 \in R^n$ were two centers. A typical choice of $\phi(\cdot)$ was the Gaussian function

used in data fitting, and the threshold function was a variant of the Multiple Weights Neural Network(MWNN) [9,10]. The network used in data fitting or system controlling consisted of an input layer of source nodes, a single hidden layer, and an output layer of linear weights. The network implemented the mapping

$$f_s(x) = \lambda_0 + \sum_{i=1}^{n_s} \lambda_i \varphi\left(d\left(x, \overline{x_{i1} x_{i2}}\right)\right) \tag{8}$$

where $\lambda_i, 0 \le i \le n_s$ were weights or parameters. The output layer made a decision on the outputs of the hidden layer. One mapped

$$f_s(x) = \max_{i=1}^{n_s} \phi\left(d\left(x, \overline{x_{i1} x_{i2}}\right)\right) \tag{9}$$

where $\phi(\cdot)$ was a threshold function. In this paper , $n_i = 11$.

Algorithm:

Let S denote the filtered set that contains the expression pattern which determine the network and X denote the original set that contains all the expression pattern in the order.

Begin

1. Put the first expression pattern into the result set S and let it be the fiducially expression pattern s_b , and the distance between the others and it will be compared. Set $S = \{ s_b \}$. $s_{\max} = s_b$ and $d_{\max} = 0$
2. If no expression pattern in the original set X ,stop filtering. Otherwise , check the next expression pattern in X , then compute its distance to s_b ,i.e., $d = \|s - s_b\|$.
3. If $d > d_{\max}$,goto step 6. Otherwise continue to step 4.
4. If $d < \varepsilon$,set $s_{\max} = s$, $d_{\max} = d$, goto step 2. Otherwise continue to step 5.
5. Put s into the result set: $S = S \cup \{s\}$,and let $s_b = s$, $s_{\max} = s$, and $d_{\max} = d$. Then go to step 2.
6. If $| d_{\max} - d | > \varepsilon_2$, go to step 2. Otherwise put $s_{\max}$ into the result set: $S = S \cup \{s_{\max}\}$,and let $s_b = s_{\max}$, $d_{\max} = \|s - s_{\max}\|$ go to step2.

3 Experiment and Analysis

The continuous speech sample used by built network is from 28 people. Select randomly 10 sample points in each digit class for each person, 280 sample points in each class, which builds up the digital neural network covering area in this class, experimental result is given in Tab.1.

Table 1. Recognition result of randomly selected sample network

	Set A	Set B	Sum
Recognizable digit amount	225	32	256
digit amount of accuracy recognition	191	20	211
digit amount of misrecognition	18	8	26
Accuracy recognition rate	85.27%	62.5%	82.42%

The continuous speech sample used by built network is from 24 people, everyone picks up 10 sample points, but which is only by simple screening，these sample points by direction tracing to the source, the original speech segment is the digit read which can be recognized by ears. Applying 240 sample points to created network for the digit of each class, Tab. 2 is the experimental result.

Table 2. Recognition result of built sample network by screening

	Set A	Set B	Sum
Recognizable digit amount	200	64	264
Digit amount of accuracy recognition	177	52	229
Digit amount of misrecognition	14	6	20
Accuracy recognition rate	88.5%	81.25%	86.74%

During experiments, train the same training sample (according to the number of training sample in each class, there are five groups 48,96,144,192,240 to attend comparison) via hypersausage neuron learning and HMM model method, then recognize the same recognized sample by these two methods respectively. Experimental results demonstrate the method present in this paper is superior to the traditional HMM model method.

Table 3. Threshold used by HDS neural network model building(TH) and neuron number(NN)

Training sample amount	\	Ling	Yi	Er	San	Si	Wu	Liu	Qi	Ba	Jiu	Yao
48	TH	120	110	100	120	110	110	120	100	100	110	110
	NN	45	43	43	45	42	43	44	41	43	44	42
96	TH	120	110	110	120	110	110	120	100	100	100	100
	NN	88	82	79	89	86	79	83	80	84	90	83
144	TH	120	110	110	120	110	110	110	100	100	100	100
	NN	125	112	114	123	124	120	129	115	112	130	129
192	TH	120	110	110	120	110	110	120	100	100	100	100
	NN	165	150	149	163	127	158	156	151	150	172	164
240	TH	120	100	100	120	120	120	120	110	100	120	100
	NN	202	199	194	201	198	162	188	178	175	183	200

Table 3 gives the method adopting hypersausage neuron learning theory, threshold(TH) in the neural network model constructing and the used neuron number (NN)at last. Table 3 shows the state number and Gauss density function amount with the highest recognition rate after adjustment many times based on HMM model method to create an acoustic model. Table 3-6 is the number statistics about recognized sample.

Table 4. Optimal state amount used by HMM model building and Gauss density function amount

Training sample amount	48	96	144	192	240
State amount	4	5	5	5	5
Gauss density function amount	4	3	6	4	6

Table 5. The recognition sample amount in hypersausage neuron and HMM-based method

Class	Ling	Yi	Er	San	Si	Wu	Liu	Qi	Ba	Jiu	Yao
amount	463	534	555	650	950	719	471	1087	315	888	676

Table 6. Comparison of overall recognition rate under the variant training sample amount in each class

Training sample amount in each class	Geometrical Learning	HMM
48	0.927203	0.8441
96	0.955939	0.8859
144	0.959907	0.9221
192	0.970033	0.9245
240	0.974001	0.9357

Finally，from the development trend of comparing result，as the sample amount increases, the distance between two methods gets to be reduced slowly, when the sample amount approaches infinity, two recognition rates will all approach to100%. In the case of enough neuron number, speech method based on hypersausage neuron learning has a higher recognition rate than HMM-based method. The reason is the former depicts the morphological distribution in HDS, whereas the latter is only a kind of probability distribution, whose accuracy of course is less than the former method.

4 Conclusion

Because dividing continuous speech is difficult, and it also directly influence the speech recognition rate. Some experiments that have been researched indicate that if we correct the previous division mistakes, the error rate of system's words would decrease at least 11.7%. So we can say that, to a great extent, the other systems' low recognition rates to continuous speech closely relate to the low accurate rates of the detection of continuous speech endpoint. So this paper changes the traditional speech

recognition's pattern that is to be syncopated firstly then to be recognized. It uses the algorithm with dynamic searching. It achieves the continuous speech recognition without syncopating. The paper is from the complicated hypersausage neuron. It gives a new algorithm in speech with hypersausage neuron learning algorithm. We hope it can be using in continuous speech of large vocabulary.

References

1. W.-J. Yang: Hidden Markov Model for Mandarin lexical tone recognition, IEEE Trans. Acoust. Speech Signal Process. 36 (1988) 988–992.
2. Wang Shou-jue: Biomimetic (Topological) Pattern Recognition —— A New Model of Pattern Recognition Theory and Its Applications (in Chinese), Acta Electronica Sinica, Vol.30, No.10, (2002) 1417-1420,
3. Wang SJ, Zhao XT: Biomimetic pattern recognition theory and its applications , Chinese Journal of Electronics 13 (3): JUL (2004) 373-377
4. Wenming Cao, Feng Hao, Shoujue Wang: The application of DBF neural networks for object recognition. Information Science 160(1-4): (2004) 153-160
5. Cao Wenming: The application of Direction basis function neural networks to the prediction of chaotic time series ,Chinese Journal of Electronics 13 (3): JUL (2004) 395-398
6. Cao WM, Feng H, Zhang DM, et al: An adaptive controller for a class of nonlinear system using direction basis function. Chinese journal of electronics 11 (3): JUL (2002) 303-306
7. Wang shoujue: A new development on ANN in China - Biomimetic pattern recognition and multi weight vector neurons, Lecture Notes in Artificial Intelligence, Vol.2639. Springer-Verlag, Berlin Heidelberg New York (2003) 35-43
8. Wenming Cao, Jianhui Hu, Gang Xiao, Shoujue Wang:Application of multi-weighted neuron for iris recognition, Lecture Notes in Computer Science, Vol.3497. Springer-Verlag, Berlin Heidelberg New York (2005) 87-92
9. Wenming Cao, Xiaoxia Pan, Shoujue Wang: Continuous speech research based on two-weight neural network, Lecture Notes in Computer Science, Vol.3497. Springer-Verlag, Berlin Heidelberg New York (2005) 345-350
10. Wang SJ, Chen YM, Wang XD, et al.: Modeling and optimization of semiconductor manufacturing process with neural networks , Chinese Journal of Electronics 9 (1): (2000) 1-5

Variable-Rate Channel Coding for Space-Time Coded MIMO System*

Changcai Han and Dongfeng Yuan

School of Information Science and Engineering,
Shandong University, Jinan 250100, China
cchan03@sdu.edu.cn

Abstract. A new scheme integrating variable-rate channel coding with the multilayered space-time trellis codes (ML-STTC), is proposed to compensate for the performance loss due to the limited diversity gain of the firstly-detected layer in the group interference suppression and cancellation algorithm at the receiver. In the proposed scheme, the code rate of the channel coding in each layer is adaptive to the diversity gain of this layer, with intent to guarantee the performance of the first layer which suffers from diversity loss. In this way, the error propagation is reduced and the overall performance is improved.

1 Introduction

Recent research reveals the capacity of various channels can be significantly increased by deploying multiple antennas at both the transmitter and receiver [1]-[3]. The goal of space-time signal design [4]-[6] for the multiple-input and multiple-output (MIMO) systems is to exploit this capacity in practice. Space-time trellis code (STTC) [4], based on the joint design of coding, modulation and diversity, is a bandwidth/power-efficient solution. However, for large number of transmit antennas, it may become too complex to be implemented. Thus, combined group array processing, the scheme of multilayered Space-time Trellis codes (ML-STTC) has been proposed in [7], which generalizes and improves upon layered space-time architecture [8].

In ML-STTC, the transmit antennas are partitioned into small layers and individual STTC is used to transmit signals from each layer of antennas. At the receiver, group interference suppression and cancellation (G-ISC) algorithm and Viterbi decoder are applied to decode the signals of each layer. In this way, the diversity gain and high data rate can be achieved within and among the layers, respectively. However, using G-ISC, the firstly-detected layer suffers from serious diversity loss compared with the layers detected later, which aggravates the error propagation and decreases the overall performance of ML-STTC.

In this paper, the performance of ML-STTC concatenated channel coding is evaluated. Based on this, the variable-rate coding scheme is proposed to compensate for the diversity loss, in which the code rate of each layer is adaptive to the diversity gain

* Supported by National Scientific Foundation of China (60372030), China Education Ministry Foundation for Visiting Scholar ([2003]406), Key Project of Provincial Scientific Foundation of Shandong (Z2003G02).

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1116 – 1121, 2005.
© Springer-Verlag Berlin Heidelberg 2005

while keeping the average code rate unchanged. The intuition is that, with high probabilities, the signals in the first layer can be correctly detected and their interference to other layers can be cancelled perfectly, which in turn facilitates the detection of the signals in other layers. Simulation results reveal the proposed scheme can improve the overall performance of the ML-STTC upon the fixed-rate channel coding scheme.

2 Multilayered Space-Time Trellis Codes

The MIMO system with n transmit and m receive antennas is considered. The antennas at the transmitter are partitioned into K layers and the k-th, $1 \le k \le K$ layer is comprised of n_k antennas with $\sum_{k=1}^{K} n_k = n$. The input bits are divided into K blocks. The k-th, $1 \le k \le K$ block is encoded by STTC encoder C_k called component code, and transmitted simultaneously from antennas in the k-th layer. Let $c_t^{i,k}$ be the signals transmitted from antenna i in the k-th layer at time slot t. The coefficient E_k denotes average energy of each antenna in the k-th layer. The signals transmitted from different antennas undergo independent Rayleigh fading. Then the received signals of antenna j, $1 \le j \le m$ at time slot t is given by

$$r_t^j = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sqrt{E_k}\, \alpha_{i,j}^k c_t^{i,k} + \eta_t^j \tag{1}$$

where $\alpha_{i,j}^k$ is the path gain from transmit antenna i in the k-th layer to receive antenna j and modeled as the sample of independent quasi-static complex Gaussian random variable (RV) with zero mean and variance 0.5 per dimension. The noise η_t^j is the sample of independent complex Gaussian RV with zero mean and variance $N_0/2$ per dimension. The vector form of (1) is given by

$$\mathbf{r}_t = \sum_{k=1}^{K} \sqrt{E_k}\, \mathbf{\Omega}_k \mathbf{c}_t^k + \boldsymbol{\eta}_t \tag{2}$$

where $\mathbf{r}_t = (r_t^1, r_t^2, \ldots, r_t^m)^T$, $\mathbf{c}_t^k = (c_t^{1,k}, c_t^{2,k}, \ldots, c_t^{n_k,k})^T$, $\boldsymbol{\eta}_t = (\eta_t^1, \eta_t^2, \ldots, \eta_t^m)^T$ and $\mathbf{\Omega}_k$ is a matrix, in which the (j, i)th element is $\alpha_{i,j}^k, 1 \le i \le n_k$, $1 \le j \le m$.

Assume that the ideal channel state information (CSI) is available at the receiver. We further assume $n_k \ge n - \sum_{i=1}^{k-1} n_i - m + 1$, $1 \le k \le K$. Then group interference suppression and cancellation algorithm [7] is employed to detect signals transmitted from different layers. Without loss of generality, we take the first layer as an example. Let $\mathbf{\Theta}_1 = (v_1^1, v_2^1, \ldots, v_{m-n+n_1}^1)^T$, where v_i^1, $1 \le i \le m - n + n_1$ are orthonormal vectors in the null space spanned by $\mathbf{\Lambda}_1 = [\mathbf{\Omega}_2, \ldots, \mathbf{\Omega}_K]$. Multiply both sides of (2) by $\mathbf{\Theta}_1$ to arrive at

$$\tilde{\mathbf{r}}_t^1 = \sqrt{E_1}\tilde{\boldsymbol{\Omega}}_1\mathbf{c}_t^1+\tilde{\boldsymbol{\eta}}_t^1 \tag{3}$$

where $\tilde{\mathbf{r}}_t^1 = \boldsymbol{\Theta}_1\mathbf{r}_t$, $\tilde{\boldsymbol{\Omega}}_1 = \boldsymbol{\Theta}_1\boldsymbol{\Omega}_1$ and $\tilde{\boldsymbol{\eta}}_t^1 = \boldsymbol{\Theta}_1\boldsymbol{\eta}_t$ .Proved in [7], $\tilde{\boldsymbol{\Omega}}_1$ and $\tilde{\boldsymbol{\eta}}_t$ are independent complex Gaussian RVs.

In (3), the interferences from other layers are suppressed, then Viterbi algorithm is applied to decode C_1. After decoding it, the contribution of these signals in the first layer is subtracted from different receive antennas. This gives a system with $n-n_1$ transmit and *m* receive antennas. Proceed in this manner to decode signals transmitted from antennas of other layers.

For the *k*-th layer, after using the group detection, we get

$$\tilde{\mathbf{r}}_t^k = \sqrt{E_k}\tilde{\boldsymbol{\Omega}}_k\mathbf{c}_t^k+\tilde{\boldsymbol{\eta}}_t^k + \boldsymbol{\Theta}_k\sum_{i=1}^{k-1}\sqrt{E_i}\boldsymbol{\Omega}_i(\mathbf{c}_t^i-\hat{\mathbf{c}}_t^i) \tag{4}$$

where $\tilde{\mathbf{r}}_t^k = \boldsymbol{\Theta}_k(\mathbf{r}_t - \sum_{i=1}^{k-1}\sqrt{E_i}\boldsymbol{\Omega}_i\hat{\mathbf{c}}_t^i)$, $\tilde{\boldsymbol{\Omega}}_k = \boldsymbol{\Theta}_k\boldsymbol{\Omega}_k$, $\tilde{\boldsymbol{\eta}}_t^k = \boldsymbol{\Theta}_k\boldsymbol{\eta}_t$, $\boldsymbol{\Theta}_k$ is an orthonormal basis of the null space spanned by $\boldsymbol{\Lambda}_k = [\boldsymbol{\Omega}_{k+1},\ldots,\boldsymbol{\Omega}_K]$ and $\hat{\mathbf{c}}_t^i$ is the vector of the decoded signals of the *i*-th layer. In the right-hand side of (4), the first and the second term denote the received signals of the *k*-th layer and Gaussian noise, respectively. The third term is the interference due to the imperfect cancellation of the *k*-1 layers detected before the *k*-th one, and given $\mathbf{c}_t^i$ and $\hat{\mathbf{c}}_t^i$, it becomes a zero mean complex Gaussian RV [9]. Therefore, in order to improve the overall performance of the system, the error propagation should be decreased.

The diversity gain of the *k*-th detected layer is given by

$$d_k = n_k(m-n+\sum_{i=1}^{k}n_i) \tag{5}$$

Assuming $n_1 = n_2 =,\ldots,n_K$, the sequence $d_K, d_{K-1},\ldots,d_1$ is a decreasing sequence [7]. The diversity loss of those layers detected at earlier stage aggravates the error propagation. Thus, in this paper, a new scheme integrating variable-rate channel coding with ML-STTC is developed to compensate for the performance loss.

3 Variable-Rate Channel Coding for ML-STTC

The block diagram of the proposed communication system integrating variable-rate channel coding with ML-STTC, is given in Fig. 1. The input data stream *L* is divided into *K* sub data flows of length $L_1, L_2,..., L_K$,with $\sum_{k=1}^{K}L_k = L$. Each data flow L_k , $1 \le k \le K$ is encoded by channel code with code rate R_k and Space-time Trellis code C_k . The output of C_k composed of n_k sequences of constellation symbols, are simultaneously transmitted from the antennas of the *k*-th layer. At the receiver, the group detection algorithm for ML-STTC is employed to detect signals of each layer. The sequences of symbols allocated to each antenna in different layers should have

the same length, so the length of the uncoded sub data flow L_k , $1 \le k \le K$ is subjected to

$$L_k = \frac{LR_k}{KR} \tag{6}$$

where R is the average code rate given by

$$R = \frac{1}{K}(R_1 + R_2 + \cdots + R_K)\,. \tag{7}$$

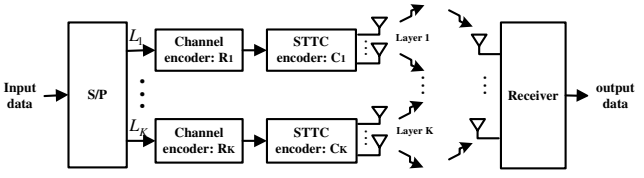


Fig. 1. Block diagram of the proposed system

First, let us analyze the performance of each layer in ML-STTC. For the *k*-th layer, the bit error rate (BER) P_b^k can be expressed as

$$P_b^k = f\left(C_k, d_k, N_0, Var(e^{k-1}), R_k\right) \tag{8}$$

where C_k denotes the STTC applied to the *k*-th layer, d_k is the diversity gain, N_0 is the variance of the noise, $Var(e^{k-1})$ is the variance of the interference due to the imperfect cancellation of those *k*-1 layers detected before the *k*-th one, and R_k is the code rate of the channel coding applied to the *k*-th layer. Thus the average BER of the system can be given by

$$P_b = \frac{1}{L}\sum_{k=1}^{K} L_k P_b^k = \frac{1}{KR}\sum_{k=1}^{K} R_k P_b^k \tag{9}$$

Then, let us analyze how to allocate the code rate of channel coding in each layer so as to combat the diversity loss and improve the performance. As stated above, the firstly-detected layer suffers from the most serious diversity loss and the error can be propagated to the next layer. Thus in our scheme, the layer with lower diversity gain d_k is protected by channel coding with the lower code rate R_k, while keeping the average code rate R unchanged. That is $R_1 < R_2 < \cdots < R_K$, with $R_1+R_2+\cdots+R_K=KR$. In this way, with high probabilities, the signals in the *k*-th layer can be correctly detected and their interference to the other layers can be cancelled perfectly, which in turn facilitates the detection of the signals in the (*k*+1)th layer. That is why the variable-rate coding scheme can improve the overall performance of the Space-time coded system compared with the fixed-rate scheme in which $R_1=R_2=\cdots=R_K=R$.

4 Simulation Results

In this section, simulation results are given to illustrate the validity of the proposed scheme. Consider the MIMO system with four transmit and four receive antennas. The signals transmitted from different antennas undergo independent quasi-static Rayleigh fading. The four transmit antennas are equally divided into two layers and the 32-state 4-PSK modulated Space-time Trellis codes [4] are used as the component codes for the two layers. Based on this, the variable-rate or fixed-rate convolutional codes are combined. In the variable-rate coding, different rate allocation schemes are tested and compared while keeping the average code rate R=1/2.

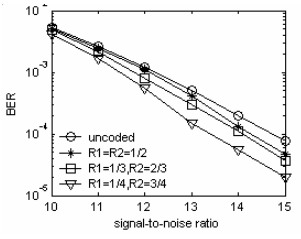


Fig. 2. Overall performances

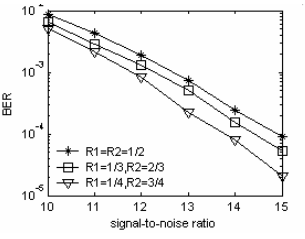


Fig. 3. First layer performance

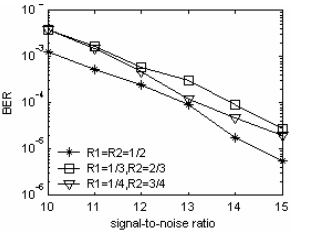


Fig. 4. Second layer performance

The system performances are compared in Fig. 2 when different code rate allocation schemes are employed. Seen from Fig.2, both of the variable-rate scheme R_1=1/3, R_2=2/3 and R_1=1/4, R_2=3/4 can improve the system performance upon the fixed-rate scheme R_1=R_2=1/2, while keeping average code rate R=1/2 unchanged. Further more, the scheme R_1=1/4, R_2=3/4 is more effective than scheme R_1=1/3, R_2=2/3, and then we draw a conclusion that the performance can be improved further by lowering the code rate of the first layer.

In order to compare those different rate allocation schemes in details, the performances of the first and second layer are given in Fig. 3 and Fig. 4, respectively. In Fig. 3, as the code rate R_1 is lowered, the performance of the first layer can be improved, which reduces the error propagation $Var(e^1)$ in (8). For the second layer, as R_1 decreases and R_2 increases, the reduced $Var(e^1)$ improves its performance while the increased R_2 depresses it, which makes it more complicated. In Fig. 4, the second layer performance of scheme R_1=1/4, R_2=3/4 is improved compared with the scheme R_1=1/3, R_2=2/3, and this confirms that when R_1 is reduced to some extent, the reduced error propagation can combat the increased R_2 and guarantee the performance of the second layer. That is why the variable-rate coding scheme can combat the diversity loss for the firstly-detected layer and improve the overall performance of the system.

5 Conclusions

Considering the error propagation effect, the performance of the multilayered Space-time Trellis codes concatenated channel coding is analyzed. Based on this, a new scheme integrating variable-rate channel coding with ML-STTC is proposed to com-

pensate for the performance loss due to the limited the diversity gain of the firstly-detected layer. Proved by simulations, the proposed scheme can improve the system performance upon the fixed-rate channel coding scheme at the same average code rate. Further more, for the variable-rate coding scheme, a better performance can be achieved by lowering the code rate of the firstly-detected layer further, while keeping the average code rate unchanged.

References

1. Teletar E.: Capacity of multi-antenna Gaussian channels. AT&T Bell Labs, Tech. Rep., June (1995)
2. Marzetta Thomas L., Hochwald B.: Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. IEEE Trans. Inform. Theory, **45** (1999) 139–158
3. Biglieri E., Caire G., Taricco G.: Limiting performance for blockfading channels with multiple antennas. IEEE Trans. Inform. Theory, **47** (2001) 1273–1289
4. Tarokh V., Seshadri N., Calderbank A. R.: Space-time codes for high data rate wireless communication: Performance criterion and code construction. IEEE Trans. Inform. Theory, **44** (1998) 774–765
5. Tarokh V., Jafarkhani H., Calderbank A. R.: Space–time block codes from orthogonal designs. IEEE Trans. Inform. Theory, **45** (1999) 1456–1467
6. Gamal H. E., Hammons A. R., Jr.: On the design and performance of algebraic space–time codes for BPSK and QPSK modulation. IEEE Trans. Commun., **50** (2002) 907–913
7. Tarokh V., Naguib A., Seshadri N., Calderbank A. R.: Combined array processing and space–time coding. IEEE Trans. Inform. Theory, **45** (1999) 1121-1128
8. Foschini G. J.: Layered space–time architecture for wireless communication in a fading environment when using multielement antennas. Bell Labs Tech. J., **1** (1996) 41–59
9. Kim Il-Min: Space–time power optimization of variable-rate space–time block codes based on successive interference cancellation. IEEE Trans. Commun., **52** (2004) 1204–1213

A New Watermarking Method Based on DWT

Xiang-chu Feng and Yongdong Yang

School of Science, Xidian University, 710071 Xi'an, China
xcfeng@mail.xidian.edu.cn

Abstract. A new hiding scheme of digital image is introduced in this paper, which is robust to image processing such as JPEG compression, median filtering, adding Gauss noise, cropping, and histogram equalization. In this paper, we embed watermark into DWT domain, using integer modulo, retrieval of embedded watermark does not need original image.

1 Introduction

The rapid growth of digital imagery, the increasingly easy access to the digital medium, and the increasingly powerful tools available for manipulating digital media has made. The copyright's protect of digital media (digital image, digital video, digital audio, etc.) has already become the problem that urgently demands to solve. Conventional cryptographic systems permit only valid key holders access to encrypted data, but once such data is decrypted, there is no way to track its reproduction or retransmission. Over the last decade, digital watermarking has been presented to complement cryptographic processes. A hiding scheme of digital image is most among these researches, it is a process of embedding information (or signature) that can confirm the ownership or integrality of the image directly into the media data by making small modifications to them.

The watermark is signal or modulo which is embedded the digital image into. It can be robust watermark and fragile (or semi-fragile) watermarks by apply aim or perceptually visible and invisible by vision. We focus on robust and invisible watermarks in this paper. In general, there are three most common requirements of invisible watermarks: 1) Imperceptibility: the watermark shouldn't be found in vision; 2) Robustness: the watermark should be resistant to intentional attacks that attempt to remove the watermarks; 3) Confidentiality: it has been claimed that digital watermarks can be used to identify the rightful owner.

Recently much work has been devoted to designing robust watermarking schemes. In general, the watermarking schemes can be classified into two categories: spatial domain approach or spectrum domain approach. People seem to attach importance to spectrum domain, because this method is invisibile. It can't satisfy the need of the copyright protection that uses pseudorandom sequence as embedding watermarks; One of the crucial requirements might be that the watermarks should be detected without using a second image (usually the original image) [1]. In this paper, we substitute copyright image for 1-D pseudorandom sequence and embedded in spectrum domain by modular arithmetic, the watermarks should be detected without using the original image.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1122 – 1126, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Embedding Algorithm

It shows that human's retina can divide image into several frequency bands, each one has an octave, and independent each other. Similarly, image can be decomposed into different subbands in logarithms scales in MRA, so we hope to handle each subband composition independently by DWT. This paper expatiated a new method of embedded watermarks which is based on discrete wavelet transform. Algorithm is as follows: Let I of size $M_1 \times N_1$ be original image, W of size $M_2 \times N_2$ be a permutation of watermark by key K, where $M_1/4 \geq M_2$, $N_1/4 \geq N_2$. Let f_k^i be the k-th level of the DWT sub-bands coefficient.

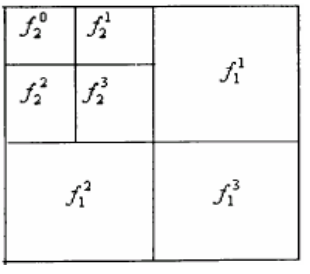


Fig. 1. Subband of DWT

To $f_2^i(x,y), 0 \leq x \leq M_2 - 1,\ 0 \leq y \leq N_2 - 1,\ i = 1,2,3$, we modulate f_k^i as follow:

$$S_i = \mathrm{mod}(f_2^i(x,y), 4 \times G) \tag{1}$$

Then let

$$f1_2^i(x,y) = \begin{cases} f_2^i(x,y) + a \times G - S & \text{W(x,y)=0} \\ f_2^i(x,y) + b \times G - S & \text{W(x,y)=1} \\ f_2^i(x,y) & \text{otherwise} \end{cases} \tag{2}$$

To $M_2 \leq x \leq M_1/4 - 1,\ N_2 \leq y \leq N_1/4 - 1,\ f1_2^i(x,y) = f_2^i(x,y)$, i=1, 2, 3. Where *a, b, G* are embedding intensity factors, and *a, b, G* are positive integers, the choice of whose values is determined according to the actual conditions. For the invisible of watermark, the change of coefficients in spatial domain or spectrum domain should be small. Meanwhile the absolute values of *a* and *b* are at least more than 2 for the need of detection *a, b, G* satisfied

$$\sum_{x=0}^{M_2-1} \sum_{y=0}^{N_2-1} S_i(x,y)/(M_2 \times N_2) \cong (a+b) \times G \tag{3}$$

On the basis of the experimental results, visual effect of the images hidden watermark can be assured through select lesser G□Whereas, in the images hidden watermark, if choosing biggish G, the ability to the anti-noise and anti-filter has been made stronger. Therefore, we may choose appropriate G according to the practice. Set a = 1, b = 3 and G = 8 in our paper. At the same time, it is not strict to the size of watermark image.

The embedded image can be achieved by high pass band of embedded watermark $f1_2^i$, f_1^i, f_2^0, i = 1, 2, 3 through wavelet reconstruction. This paper utilizes the method of repetitious embedding which is to embed the same watermark in three detail subimages f_2^i, i=1,2,3 of wavelet decomposition.

3 Detecting Algorithm

The detection of annotation watermark takes a few steps which are similar to its embedding process. Firstly, the border of the watermarked image is decomposed into its DWT sub-bands: high pass band coefficients ff_k^i, k = 1, 2. i=1, 2, 3 and low pass band coefficients ff_2^0 .Then, modulo coefficients of embedded position:

$$SS_i = \mathrm{mod}(ff_2^i(x, y), 4\times G) \tag{4}$$

G is same value as embedded. Then let :

$$\mathrm{w(x,y)}=\begin{cases} 1 & \sum_{i=1}^{3} SS_i(x, y) \geq |b-a|\times G \\ 0 & else \end{cases} \tag{5}$$

Finally, the detected watermark is obtained by reversal permutation of w according to key K.

We can define similarity NC:

$$NC = \sum_{i=0}^{M_1-1}\sum_{j=0}^{M_2-1}[W(i, j)W1(i, j)] / \sum_{i=0}^{M_1-1}\sum_{j=0}^{M_2-1} W(i, j)^2 \tag{6}$$

where W is original watermark ,W1 is detected watermark.

We modify W and W1 as follow:

$$W(i, j) = \begin{cases} -1 & w(i, j) = 0 \\ 1 & w(i, j) = 1 \end{cases} \tag{7}$$

$$W1(i, j) = \begin{cases} -1 & w(i, j) = 0 \\ 1 & w(i, j) = 1 \end{cases} \tag{8}$$

it vails to obtain the accurate similarity between original and detected watermark.

The algorithm has several advantages: (1) We can permute watermark before embedding watermark (permutation algorithm is shown in paper [3]) to make embedded watermark have the ability to anti-cutting and strengthened the confidentiality; (2)The watermark permuted is embedded in mid-frequency domain, which ensured the image embedded watermark algorithm is effective on against destroy of compressive code; (3)This algorithm is easier implement than the embedded algorithm in DCT domain paper [4]; (4)The size of watermark is smaller than original image, thus the embedded position can be freely chosen, and the algorithm strengthens robustness to many kinds of attacks; (5)The method of repetitious embedding make detecting algorithm obtain the ability of error-Correcting; (6) The watermarks should be detected without using the original image; (7)The experiment shows that the algorithm can been realized more easily and it is much stable comparing with paper [5].

4 Experimental Results

We test on the 512×512 peppers image. The watermark is 64×64 Binary Image of"西电科大".

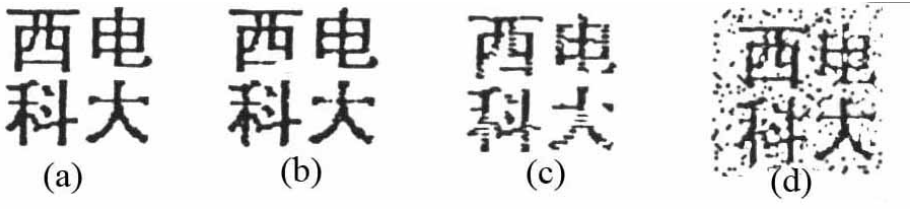


Fig. 2. The detecting results comparision: (a) original watermark; (b) cut 1/8 image; (c) cut 1/4 ; (d) adding Gauss noise

Table 1. Experimental results

Image Process	Detected in [5]	Detected in paper	Cut 1/4	Add Gauss noise(M=8)	Histogram Balance	JPEG (35%)	Median filtering
PSNR	32.072	36.663	12.11	28.015	29.899	28.972	16.147
Similarity (NC)	0.59668	1	0.52566	0.55444	0.58573	0.44305	0.51752

The experimental results are as follows: this algorithm is superior to paper [5]. This resulting watermarked image is much more robust to additive noise, histogram Balance, median Filter and MPEG compression. Since the watermark algorithm is based on DWT partitioning, it can't be resilient to geometrical attacks. This disadvantage can be solved by integrating a rotation-invariant log-polar mapping (LPM) with the wavelet transform [6]. For color image, watermark can be embedded into arbitrary color component, then we synthesize new color image.

5 Conclusions

In this paper, a novel, yet simple, wavelet domain watermarking scheme for image authentication is presented. Experimental results show that these algorithms are robust to many kinds of attacks, including compression, filtering, blurring. The watermark detection is also efficient and blind. Since the watermark embedded by K-permutation and repetitious embedded the malicious detection of the watermark would not be possible without knowing the key K. The scheme also achieves higher embedding rates than other methods.

References

1. Zeng W., Liu B.: A Statistical Watermark Detection Technique without Using Images for Resovling Rightful Ownerships of Digital Images, *IEEE Transaction on images processing*, Vol. 8 (1999) 1534-1548
2. Zhou Yaxun, Ye Qinwei, Xu Tifeng: A Text Watermark Scheme Based on Wavelet Multire solution Data Combination, *ACTA Electronica Sinica,* Vol .28 (2000) 142-144
3. Ding Wei, Qi Dong-Xu: Digtal Image Transformation and Information Hideing and Disguising Technology, *Chinese J. Computers,* Vol. 21 (1998) 838-843
4. Xia GuangSheng, Chen MingQi, Yang YiXian, Hu ZhengMing: Watermark Algor ithm Based on Modulo, *Chinese J. Computers,* Vol. 23 (2000) 1146-1150
5. Tsai. M-J, Yu K-Y, Chen Y-Z: Jiont Wavelet and Spatial Transformation for Digtal Watermarking, *IEEE Transactions on Consumers Electronics,* Vol. 46 (2000) 241-245
6. Gui Xie, Hong Shen: Robust Wavelet-Based Blind Image Watermarking Against Geometrical Attacks, *2004 IEEE Internatianai Conference on Multimedia and Expo* (ICME) (2004) 2051-2054

Efficient Point Rendering Method Using Sequential Level-of-Detail

Daniel Kang and Byeong-Seok Shin

Inha University, Department of Computer Science and Information Engineering,
253 Yonghyeon-Dong, Nam-Gu, Inchon 402-751, Rep. of Korea
rich5264@medialab.inha.ac.kr, bsshin@inha.ac.kr

Abstract. We propose an extension of sequential point trees, a data structure that allows adaptive rendering of point clouds completely on the graphics processor (GPU). Using a sequential level-of-detail selection, we exploited the programmable graphics pipeline for rendering of large data sets. By adding position and radius of parent node to all points in the sequence, hole and overdraw problems of the sequential point trees technique were resolved, and as a result better image quality and rendering performance were achieved.

1 Introduction

In recent years, for its simplicity and intuitive approach, point-based rendering techniques became a popular research area, and a variety of methods were introduced. The general idea of point-based rendering is to represent a surface using a large set of points and render these points directly.

Rusinkiweicz and Levoy proposed a multi-resolution point rendering system called *QSplat* using a point-hierarchy along with an effective LOD mechanism [1]. Later they extended their work to view-dependent progressive transmission of data, which is suitable for streaming environment in network [6].

Zwicker *et al*. proposed a framework for high quality splatting based on elliptical Gaussian kernels [3]. Their method avoids aliasing artifacts, and additionally features anisotropic texture filtering for point-based rendering. Later, Ren *et al*. extended the system to be accelerated by the graphics hardware [7].

Coconu and Hege used the octree-based data structure to sort the points back to front, and as a result correct blending was possible without using an expensive multi-pass rendering and the z-buffer [8]. Koo and Shin improved the rendering performance by using octree-based visibility computation and precomputed splat kernels [9].

Pajarola *et al.* proposed a technique using object-space point interpolation of densely sampled surfaces [6]. With transformation-invariant covariance matrix of a set of points, they efficiently determined splat sizes in a multi-resolution hierarchy.

Sequential point trees, proposed by Dachsbacher *et al*. [2] exploited the modern graphics hardware with a sequential LOD technique. They first build the bounding hierarchy from the input point set, then sequentialize and rearrange the points to send them to the GPU. Programmable pipeline of the GPU selects the points to be rendered and shades them independently. This allows the CPU to do other non-rendering com-

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1127 – 1133, 2005.
© Springer-Verlag Berlin Heidelberg 2005

putations. However sequential point trees technique has potential problems. When the viewing distance varies from point to point, the GPU fails to determine the upper-bound, and then results in holes or overdraws. To solve this problem, they added an interval overlap as big as the distance between points. However, the number of overdraws increased and extra computation is required.

In this paper, we propose an extension of sequential point trees to resolve inherent problems of original method. Schematically, holes and overdraws are caused by incorrect calculations of the upper bounds while determining the detail level of the model. Hence, for each point, we added spatial information of the upper-level point. With this simple extension, we could determine an exact viewing distance of the upper level points, and avoid holes as well as overdraws.

In Sect. 2 and 3, we describe the preprocessing step and rendering pipeline in detail. Sect. 4 discusses our implementation and results, and we summarize our work.

2 Preprocessing

To maximize rendering performance we used a sequential LOD technique. In our algorithm, projected area of each point is a single criterion in determining the detail level of the model. Bounding hierarchy derived from the input point set is an appropriate structure for hierarchical LOD. However dependencies between parent node and its children of hierarchical structure have to be eliminated to exploit the programmable pipeline of the graphics hardware. Consequently, to de-couple these dependencies between nodes, each node stores additional data derived from its parent node, namely, parent node's point size.

However, as in an original algorithm, projected area of parent node cannot be computed correctly because of the following reasons: Viewing distance of a point required to compute projected area can varies according to several viewing conditions in a perspective projection. And the viewing distances of points is not constant through the entire model. For these reasons, holes on the surface of the model and overdrawing problems aroused. So we extended an original algorithm by adding parent node's position data to determine an exact viewing distance of the parent node.

Preprocessing includes building a bounding hierarchy, sequentialization and quantization steps, as seen in Fig. 1.

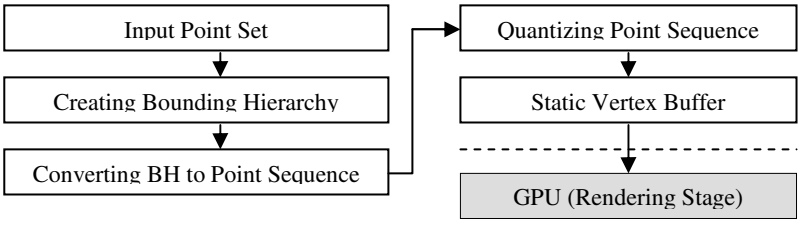


Fig. 1. Preprocessing steps of the proposed method

We assume that every input point includes a center position $\boldsymbol{p}$, a surface normal $\boldsymbol{n}$, a radius of point $\boldsymbol{r}$ and optionally a color value $\boldsymbol{c}$. In the first step of preprocessing, a bounding hierarchy is created from the input point set. Although a variety of volumes

such as axis-aligned bounding box and *k*-DOP are possible, we used the bounding sphere as a bounding volume for simplicity. Based on an octree structure, all the points are rearranged as leaf nodes, and internal node of the tree is created by combining its children. While combining the child nodes, point-related values are averaged and re-normalized as necessary, and bounding volume is extended large enough to cover all of its children.

The hierarchical data created in the first step cannot be processed directly on the graphics hardware due to its recursive characteristics. Because the core of our algorithm is utilizing the graphics hardware to render models, this hierarchical structure has to be sequentialized.

Since the graphics hardware has a parallel processing unit, every point in the sequence has to be selected and shaded independently. To remove a recursive test, each point stores additional values derived from the parent node in the hierarchy: parent point's radius $\boldsymbol{r}_p$ and center position $\boldsymbol{p}_p$. After sequentialization process, we sort the points in the sequence by $\boldsymbol{r}_p$.

For the maximum rendering performance and memory efficiency, points in the sequence are compressed. Currently many graphics hardware prefer vertex format of multiple of 32 bytes size. Hence, in our algorithm, points are compressed to 32 bytes format (originally about 47 bytes).

3 Rendering

All rendering steps are processed on the graphics hardware except the first stage, determination of rendering range. This simple step done on the CPU does not exceed 10% of workload of the CPU in most cases.

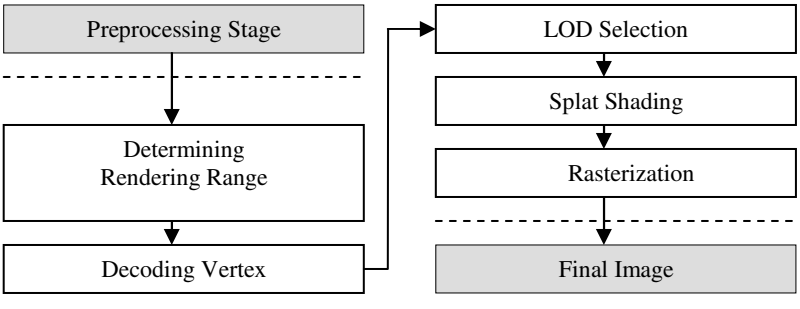


Fig. 2. Rendering pipeline of our method

After quantization process, the quantized point sequence for the model is used to render the model. Because points in this sequence do not store view-dependent values, we can use the static vertex buffer in DirectX (or vertex array in OpenGL). Locking a vertex buffer can have a significant performance penalty. So our strategy of using a static buffer results in an improved rendering performance. In our experiments, when using a static vertex buffer, rendering performance was increased about 20%.

Although we use a static buffer, we can limit the number of points to be streamed to the GPU within the buffer. As illustrated in Fig. 3, model in the distance requires a certain amount of points in the front section of the buffer for rendering. Hence, by calculating the minimum point size to render the whole object, we can reduce the bandwidth of streaming data.

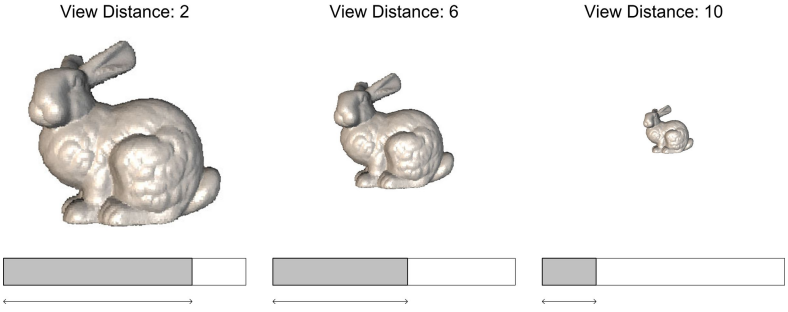


Fig. 3. Rendering range required to render model varies according to the viewing distance

To determine the minimal-sized point in the sequence, we assume that all the points in the sequence are positioned at the front-most position depicted as $\boldsymbol{p}_n$ in Fig. 4 within the model's bounding volume. This assumption makes the determination of rendering range a simple and fast operation. First, a lower bound is computed at the front-most position. Then we search the first point which is smaller than a lower bound. Since the point sequence is sorted, we can find this minimal point with binary search within a short time. After determining the rendering range, limited amount of data are passed to the GPU.

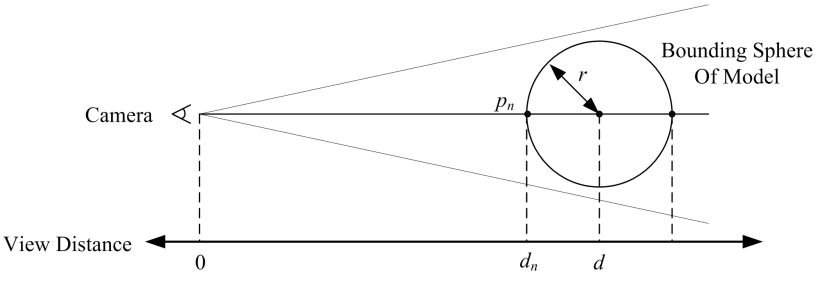


Fig. 4. Determining a lower bound. At position $\boldsymbol{p}_n$, all points' pixel length are maximized in perspective projection.

After all the points are streamed to the graphics hardware, each point is processed on the GPU independently. Because points are compressed, vertex shader decodes a given vertex data.

To determine a point's projected area on the screen, we divide point's size by its viewing distance in a perspective projection. $\boldsymbol{e}$ is a user defined accuracy parameter to control the detail level of the model. When $\boldsymbol{e}$ equals one pixel, all the points of which projected area is under one pixel are culled out.

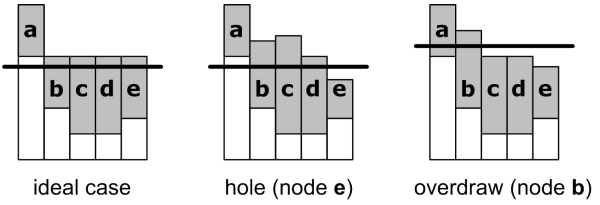


Fig. 5. The maximum values of every child nodes are equal logically (left). When the maximum values are under-estimated, holes may occur (center). When the maximum values are over-estimated, overdraws may occur (right).

In an original algorithm (see Fig. 5), hole and overdraw problems may occur when the $\boldsymbol{r}_p$'s pixel length is calculated at incorrect viewing distance. In a perspective projection projected area of a point varies according to its viewing distance (see Fig. 6). Hence we added position of parent node's point to calculate correct maximum value.

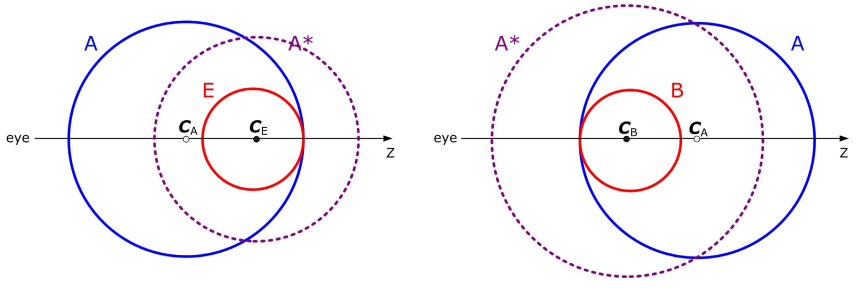


Fig. 6. When $\boldsymbol{C}_A < \boldsymbol{C}_E$, node E's maximum value is calculated at the distance of $\boldsymbol{C}_E$, possibly resulting in a hole because $\boldsymbol{A} > \boldsymbol{A}^*$ (left). When $\boldsymbol{C}_B < \boldsymbol{C}_A$, node B's maximum value is calculated at the distance of $\boldsymbol{C}_B$, possibly resulting in an overdraw because $\boldsymbol{A}^* > \boldsymbol{A}$ (right).

4 Experimental Results

We implemented our algorithm under Managed DirectX. Vertex shader and pixel shader were used to perform the sequential processing. With the hardware-accelerated point sprites and alpha texturing technique, we tried several splat kernels including hardware-accelerated rectangular shape, opaque circle and fuzzy circle. All experiments are carried on ATI RADEON 9550 graphics hardware and screen resolution of 1280 by 1024.

Table 1 shows rendering performance of our algorithm. Performance was measured in two rendering condition: using a static vertex buffer and using a dynamic vertex buffer in our method. As the results show, when using a static vertex buffer, rendering performance was increased by more than 7% than using a dynamic buffer.

Table 2 shows the number of holes and overdraws detected in the original algorithm. The number of holes and overdraws are maximized when the bounding volume of the model is longest along the sight of view.

Table 3 shows the number of holes and overdraws when using various pixel cutoff values. As seen in the results, larger pixel cutoff caused more holes and overdraws. In our method, however, no hole and no overdraw presented at all viewing conditions.

Fig. 7 shows rendering image using our method. It is very hard to find over-blurred regions and holes on those images.

Table 1. Rendering speed of various models using our algorithm. Rendering speed (fps) was measured in three different conditions: using a dynamic vertex buffer, using a static vertex buffer and using an original sequential point trees algorithm.

model	# of input points	fps (original)	fps (static)	fps (dynamic)
Bunny	35,130	333.5	330.4	290.0
Buddha	1,052,607	30.0	30.4	27.0
Dragon	1,267,577	27.5	27.5	25.7

Table 2. Number of holes and overdraws while rendering various models using original algorithm

model	max. # of holes	max. # of overdraws
Bunny	21	23
Buddha	162	158
Dragon	183	203

Table 3. Number of holes and overdraws while rendering dragon model (1,267,577 points) using original algorithm according to pixel cutoff

pixel cutoff	max. # of holes	max. # of overdraws
1.0	4	4
2.0	59	62
5.0	93	99
10.0	130	131

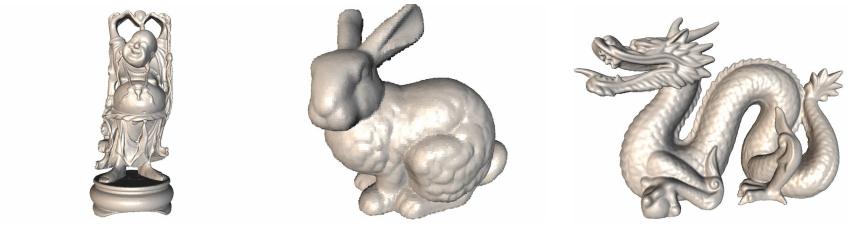


Fig. 7. Rendering result images. Buddha (left), bunny (middle) and dragon (right)

5 Conclusions

We proposed an extension of sequential point trees, avoiding hole and overdraw problems. While making the most of sequential LOD's high rendering performance, we achieved a hole-free surface reconstruction and eliminated overdraws as well. As seen in the experiment results, our algorithm shows practically the same rendering performance as an original sequential point trees.

Acknowledgement

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (D00600).

References

1. Rusinkiweicz, S., Levoy, M.: QSplat: A Multiresolution Point Rendering System for Large Meshes, Proceedings ACM SIGGRAPH (2000) 343-352
2. Dachsbacher, C., Vogelgsang, C., Stamminger, M.: Sequential Point Trees, Proceedings ACM SIGGRAPH (2003) 657-663
3. Zwicker, M., Pfister, H., Van Barr, J., Gross, M.: EWA Splatting, IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 3 (2002) 223-238
4. Pajarola, R., Sainz, M., Guidotti, P.: Confetti: Object-Space Point Blending and Splatting, IEEE Transactions on Visualization and Computer Graphics, Vol. 10, No. 5 (2004) 598-608
5. Saintz, M., Pajarola, R., Lario, R.: Points Reloaded: Point-Based Rendering Revisited, Proceedings Eurographics Symposium on Point-Based Graphics (2004) 121-128
6. Rusinkiewicz, S., Levoy, M.: Streaming QSplat: A viewer for networked visualization of large, dense models, Symposium for Interactive 3D Graphics Proceedings (2001) 63-68
7. Ren, L., Pfister, H., Zwicker, M.: Object Space EWA Surface Splatting: A Hardware Accelerated Approach to High Quality Point Rendering, Proceedings Eurographics (2002) 461-470
8. Coconu, L., Hege, H-C.: Hardware-Oriented Point-Based Rendering of Complex Scene, Proceedings Eurographics Workshop on Rendering (2002) 43-52
9. Koo, Y., Shin, B.: An Efficient Point Rendering Using Octree and Texture Lookup, Lecture Notes in Computer Science, Vol. 3482 (2005) 1187-1196
10. Botsch, M., Kobbelt, L.: High-Quality Point-Based Rendering on Modern GPUs, Proceedings Pacific Graphics (2003) 335-343
11. Chen, B., Nguyen, M.X.: Pop: A Hybrid Point and Polygon Rendering System for Large Data, Proceedings IEEE Visualization (2001) 45-52
12. Pfister, H., Zwicker, M., van Baar, J., Gross, M.: Surfels: Surface Elements as Rendering Primitives, Proceedings SIGGRAPH (2000) 335-342

Construction of a Class of Compactly Supported Biorthogonal Multiple Vector-Valued Wavelets

Tongqi Zhang[1] and Qingjiang Chen[2]

[1] Institute of Math. and Appli.Math., Weinan Teachers College, Weinan 714000, China
wnzhtq@wntc.edu.cn
[2] School of Science, Xi'an Jiaotong University, Xi'an 710049, China
chenxj6684@mail.xjtu.edu.cn

Abstract. In this paper, we introduce the notion of vector-valued multiresolution analysis. We discuss the existence of biorthogonal multiple vector-valued wavelets. An algorithm for constructing a class of compactly supported biorthogonal multiple vector-valued wavelets associated with the biorthogonal multiple vector-valued scaling functions is presented by using multiresolution analysis and matrix theory.

1 Introduction

Multiwavelets[1,2], due to their good characteristics, have widely been applied to many aspects in technology and science recently, such as, image compress, signal processing and so on. It is noticed that multiwavelets can be generated from the component functions in vector-valued wavelets. Studying vector-valued wavelets is useful in multiwavelet theory. Xia and Suter[3] introduced the notion for vector-valued wavelets and studied the existence and construction of orthogonal vector-valued wavelets. Fowler and Hua[4] implemented biorthogonal vector-valued wavelet transforms to study fluid flows in oceanography and aerodynamics. However, multiwavelets and vector-valued wavelets are different in the following sense. Prefiltering is usually required for discrete multiwavelet transforms[5] but not necessary for discrete vector-valued wavelet transforms. Therefore, it is necessary to study vector-valued wavelets. However, as yet there has not been a general method to obtain biorthogonal vector-valued wavelets. The main aim of the paper is to study the construction of biorthogonal vector-valued wavelets.

Notations: Let R and C be the set of real and complex numbers, respectively. Z denotes all integers. Let $s \in Z$ be a constant and $s \geq 2$. $\mathbf{I}_s$ and $\mathbf{O}$ denote the $s \times s$ identity matrix and zero matrix, respectively. The space of multiple vector-valued functions $L^2(R, C^{s\times s})$ is defined as

$$L^2(R\,,\,C^{s\times s}) := \left\{ \mathcal{F}(t) = \begin{pmatrix} f_{11}(t)\ f_{12}(t)\cdots f_{1\,s}(t) \\ f_{21}(t)\ f_{22}(t)\cdots f_{2\,s}(t) \\ \cdots\quad\cdots\quad\cdots\quad\cdots \\ f_{s1}(t)\ f_{s2}(t)\cdots f_{s\,s}(t) \end{pmatrix} : \begin{matrix} f_{kl}(t) \in L^2(R) \\ k,l = 1,2,\cdots,s \end{matrix} \right\}$$

For $\mathcal{F} \in L^2(R, C^{s\times s})$, the integration of vector-valued function $\mathcal{F}(t)$ is defined as follow $\int_R \mathcal{F}(t)dt = (\int_R f_{k,\,l}(t)dt\,)_{s\times s}$, and $||\mathcal{F}||$ represents the norm

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1134–1139, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of operator $\mathcal{F}$, i.e., $||\mathcal{F}|| := (\sum_{k,\,l=1}^{s} \int_{\mathcal{R}} |f_{k,\,l}(t)|^2\, dt\,)^{1/2}$. The Fourier transform of $\mathcal{F} \in L^2(R, C^{s\times s})$ is defined to be $\widehat{\mathcal{F}}(\alpha) := \int_R \mathcal{F}(t) \cdot \exp\{-i\alpha t\}\, dt$. For two functions $\mathcal{F}t)$, $\mathcal{G}(t) \in L^2(R, C^{s\times s})$, their *symbol inner product* is defined as

$$[\,\mathcal{F}, \mathcal{G}\,] := \int_R \mathcal{F}(t)\, \mathcal{G}(t)^*\; dt \tag{1}$$

where * means the transposition and the complex conjugation.

A sequence of vector-valued functions $\{\,\mathcal{F}_k(t)\}_{k\in Z} \subset \mathbf{U} \subset L^2(R, C^{s\times s})$ is called a *Riesz* basis of $\mathbf{U}$ if (1) For any $\Upsilon(t) \in \mathbf{U} \subset L^2(R, C^{s\times s})$, there exists a unique sequence of $s \times s$ matrix $\{P_k\}_{k\in Z}$ such that

$$\Upsilon(t) = \sum\nolimits_{k\in Z} P_k \mathcal{F}_k(t). \tag{2}$$

(2) there exist constants $0 < C_1 \le C_2 < \infty$ such that, for any $s \times s$ constant matrix sequence $\{P_k\}_{k\in Z}$,

$$C_1\, ||\{P_k\}||_\dagger \;\le\; ||\textstyle\sum_{k\in Z} P_k \mathcal{F}_k(t)||_{L_2} \;\le\; C_2\, ||\{P_k\}||_\dagger$$

where $||\{P_k\}||_\dagger$ denotes the norm of the matrix sequence $\{P_k\}_{k\in Z}$.

2 Vector-Valued Multiresolution Analysis

First, we introduce vector-valued multiresolution analysis and give the definition for biorthogonal multiple vector-valued wavelets. Next, we study the existence of the biorthogonal multiple vector-valued wavelets.

Definition 2.1. A vector-valued multiresolution analysis of $L^2(R, C^{s\times s})$ is a nested sequence of closed subspaces $\mathbf{S}_j$, $j \in Z$ of $L^2(R, C^{s\times s})$ such that
(*a*) $\mathbf{S}_j \subset \mathbf{S}_{j+1}, \quad \forall\, j \in Z$;
(b) $\mathcal{F}(\cdot) \in \mathbf{S}_0 \Longleftrightarrow \mathcal{F}(2^j \cdot) \in \mathbf{S}_j,\ \forall\, j \in Z$;
(*c*) $\bigcap_{j\in Z} \mathbf{S}_j = \{\mathbf{O}\}\,;\ \bigcup_{\mathbf{j}\in\mathbf{Z}} \mathbf{S_j}$ is dense in $L^2(\mathcal{R}, \mathcal{C}^{s\times s})$.
(d) There is a $\mathbf{\Phi}(t) \in \mathbf{S}_0$ such that its translates $\mathbf{\Phi}_k(t) := \mathbf{\Phi}(t-k), k \in Z$, form a *Riesz* basis for $\mathbf{S}_0$.

Since $\mathbf{\Phi}(t) \in \mathbf{S}_0 \subset \mathbf{S}_1$, by definition 2.1 and (2) there exists a $s \times s$ matrice sequence $\{A_k\}_{k\in Z}$ such that

$$\mathbf{\Phi}(t) = 2 \cdot \textstyle\sum_{k\in Z} A_k\, \mathbf{\Phi}(2t-k). \tag{3}$$

Equation (3) is called a refinement equation and $\mathbf{\Phi}(t)$ is called a multiple vector-valued scaling functions. Let

$$\mathcal{A}(\alpha) = \textstyle\sum_{k\in Z} A_k \cdot \exp\{-ik\alpha\}, \quad \alpha \in \mathcal{R}. \tag{4}$$

Then
$$\widehat{\mathbf{\Phi}}(\alpha) = \mathcal{A}(\frac{\alpha}{2})\, \widehat{\mathbf{\Phi}}(\frac{\alpha}{2}), \quad \alpha \in \mathcal{R}. \tag{5}$$

Let W_j, $j \in Z$, stands for the complementary subspace of $\mathbf{S}_j$ in $\mathbf{S}_{j+1}$ and there exists a vector-valued function $\mathbf{\Psi}(t) \in L^2(R, C^{s\times s})$ such that $\mathbf{\Psi}_{j,\,k}(t) = 2^{j/2}\mathbf{\Psi}(2^j\, t - k)$, j, $k \in Z$ forms a *Riesz* basis of W_j.

It is clear that $\mathbf{\Psi}(t) \in W_0 \subset \mathbf{S}_1$. Hence there exists a $s \times s$ sequence of matrices $\{B_k\}_{k\in Z}$ such that

$$\mathbf{\Psi}(t) = 2 \cdot \textstyle\sum_{k\in Z} B_k \, \mathbf{\Phi}(2t-k). \tag{6}$$

By taking *Fourier* transform for both sides of (6), we have

$$\widehat{\mathbf{\Psi}}(\alpha) = \mathcal{B}(\frac{\alpha}{2})\, \widehat{\mathbf{\Phi}}(\frac{\alpha}{2}) \quad \alpha \in \mathcal{R}. \tag{7}$$

where
$$\mathcal{B}(\alpha) = \textstyle\sum_{k\in Z} B_k \cdot \exp\{-ik\alpha\} \quad \alpha \in \mathcal{R}. \tag{8}$$

We call $\mathbf{\Phi}(t)$, $\widetilde{\mathbf{\Phi}}(t) \in L^2(\mathcal{R}, \mathcal{C}^{s\times s})$ a pair of biorthogonal vector-valued scaling functions, if

$$[\, \mathbf{\Phi}(\cdot), \widetilde{\mathbf{\Phi}}(\cdot - n)\,] = \delta_{0,\,n}\mathbf{I}_s \qquad n \in Z. \tag{9}$$

We say that two vector-valued functions $\mathbf{\Psi}(t)$, $\widetilde{\mathbf{\Psi}}(t) \in L^2(\mathcal{R}, \mathcal{C}^{s\times s})$ are a pair of biorthogonal vector-valued wavelets associated with vector-valued scaling functions $\mathbf{\Phi}(t)$, $\widetilde{\mathbf{\Phi}}(t)$, if

$$[\, \mathbf{\Phi}(\cdot)\,, \widetilde{\mathbf{\Psi}}(\cdot - n)\,] = [\, \widetilde{\mathbf{\Phi}}(\cdot)\,, \mathbf{\Psi}(\cdot - n)\,] = \mathbf{O} \quad n \in Z; \tag{10}$$

$$[\, \mathbf{\Psi}(\cdot)\,, \widetilde{\mathbf{\Psi}}(\cdot - n)\,] = \delta_{0,\,n}\mathbf{I}_s \quad n \in Z, \tag{11}$$

and the sequence of functions $\{\mathbf{\Psi}(t-k)\}_{k\in Z}$ constitutes a *Riesz* basis of W_0.

Similar to (3) and (6), $\widetilde{\mathbf{\Phi}}(t)$ and $\widetilde{\mathbf{\Psi}}(t)$ also satisfy the following equations:

$$\widetilde{\mathbf{\Phi}}(t) = 2 \cdot \textstyle\sum_{k\in Z} \widetilde{A}_k \widetilde{\mathbf{\Phi}}(2t-k), \tag{12}$$

$$\widetilde{\mathbf{\Psi}}(t) = 2 \cdot \textstyle\sum_{k\in Z} \widetilde{B}_k \widetilde{\mathbf{\Phi}}(2t-k). \tag{13}$$

By taking Fourier transform for both sides of (12) and (13), resp., we have

$$\widehat{\widetilde{\mathbf{\Phi}}}(\alpha) = \widetilde{\mathcal{A}}(\tfrac{\alpha}{2})\, \widetilde{\mathbf{\Phi}}(\tfrac{\alpha}{2}) \quad \alpha \in \mathcal{R} \tag{14}$$

$$\widehat{\widetilde{\mathbf{\Psi}}}(\alpha) = \widetilde{\mathcal{B}}(\tfrac{\alpha}{2})\, \widetilde{\mathbf{\Phi}}(\tfrac{\alpha}{2}) \quad \alpha \in \mathcal{R} \tag{15}$$

where
$$\widetilde{\mathcal{A}}(\alpha) = \textstyle\sum_{k\in Z} \widetilde{A}_k \cdot \exp\{-ik\alpha\} \quad \alpha \in \mathcal{R}; \tag{16}$$

$$\widetilde{\mathcal{B}}(\alpha) = \textstyle\sum_{k\in Z} \widetilde{B}_k \cdot \exp\{-ik\alpha\} \quad \alpha \in \mathcal{R}. \tag{17}$$

Lemma 2.1[2]**.** Let $\mathcal{F}(t)\,, \widetilde{\mathcal{F}}(t) \in L^2(\mathcal{R}, \mathcal{C}^{s\times s})$. Then $\mathcal{F}(t)$ and $\widetilde{\mathcal{F}}(t)$ are a pair of biorthogonal functions if and only if

$$\textstyle\sum_{k\in Z} \widehat{\mathcal{F}}(\alpha + 2k\pi)\widehat{\widetilde{\mathcal{F}}}(\alpha + 2k\pi)^* = \mathbf{I}_s, \ \alpha \in \mathcal{R}. \tag{18}$$

Lemma 2.2[2]**.** Let $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$ defined in (3) and (12), respectively, be a pair of biorthogonal vector-valued scaling functions. $\mathcal{A}(\alpha)$ and $\widetilde{\mathcal{A}}(\alpha)$ is defined in (4) and (16). Then

$$\mathcal{A}(\alpha)\widetilde{\mathcal{A}}(\alpha)^* + \mathcal{A}(\alpha+\pi)\widetilde{\mathcal{A}}(\alpha+\pi)^* = \mathbf{I}_s, \quad \alpha \in \mathcal{R}. \tag{19}$$

Formula (19) is equivalent to $\quad \textstyle\sum_{l\in Z} A_l\, \widetilde{A}^*_{l+2k} = \frac{1}{2}\delta_{0,\,k}\mathbf{I}_s. \tag{20}$

Theorem 2.3. Let $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$ be a pair of biorthogonal vector-valued scaling functions. Assume $\mathbf{\Psi}(t)$ and $\widetilde{\mathbf{\Psi}}(t)$ defined in (6) and (13) are vector-valued functions, and $\mathbf{\Psi}(t)$ and $\widetilde{\mathbf{\Psi}}(t)$ are a pair of biorthogonal vector-valued wavelet functions associated with $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$. Then

$$\mathcal{A}(\alpha)\widetilde{\mathcal{B}}(\alpha)^* + \mathcal{A}(\alpha+\pi)\widetilde{\mathcal{B}}(\alpha+\pi)^* = \mathbf{O}, \ \ \alpha \in \mathcal{R}. \tag{21}$$

$$\widetilde{\mathcal{A}}(\alpha)\mathcal{B}(\alpha)^* + \widetilde{\mathcal{A}}(\alpha+\pi)\mathcal{B}(\alpha+\pi)^* = \mathbf{O}, \ \ \alpha \in \mathcal{R}. \tag{22}$$

$$\mathcal{B}(\alpha)\widetilde{\mathcal{B}}(\alpha)^* + \mathcal{B}(\alpha+\pi)\widetilde{\mathcal{B}}(\alpha+\pi)^* = \mathbf{I}_s, \ \ \alpha \in \mathcal{R}. \tag{23}$$

Proof. If $\mathbf{\Psi}(t)$ and $\widetilde{\mathbf{\Psi}}(t)$ are biorthogonal associated with $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$, then (10) and (11) hold. Similar to Lemma 2.1, we derive from (10) that

$\sum_{k\in Z} \widehat{\mathbf{\Phi}}(2\alpha+2k\pi)\,\widehat{\widetilde{\mathbf{\Psi}}}(2\alpha+2k\pi)^* = \mathbf{O} \ \ \alpha \in \mathcal{R}$. Thus

$$\begin{aligned}\mathbf{O} &= \textstyle\sum_{k\in Z} \mathcal{A}(\alpha+k\pi)\widehat{\mathbf{\Phi}}(\alpha+k\pi)\widehat{\widetilde{\mathbf{\Phi}}}(\alpha+k\pi)^*\widetilde{\mathcal{B}}(\alpha+k\pi)^* \\ &= \sum_{\mu=0}^{1} \mathcal{A}(\alpha+\mu\pi) \sum_{\kappa\in Z} \widehat{\mathbf{\Phi}}(\alpha+\mu\pi+2\kappa\pi)\,\widehat{\widetilde{\mathbf{\Phi}}}(\alpha+\mu\pi+2\kappa\pi)^*\,\widetilde{\mathcal{B}}(\alpha+\mu\pi)^* \\ &= \mathcal{A}(\alpha)\widetilde{\mathcal{B}}(\alpha)^* + \mathcal{A}(\alpha+\pi)\widetilde{\mathcal{B}}(\alpha+\pi)^*.\end{aligned}$$

Hence, (21) holds and so does (22). The following is the proof of (23). According to (11), we have $\sum_{l\in Z} \widehat{\mathbf{\Psi}}(2\alpha+2l\pi)\,\widehat{\widetilde{\mathbf{\Psi}}}(2\alpha+2l\pi)^* = \mathbf{I}_s, \ \ \alpha \in \mathcal{R}$. we thus get that

$$\begin{aligned}\mathbf{I}_s &= \textstyle\sum_{l\in Z} \mathcal{B}(\alpha+l\pi)\widehat{\mathbf{\Phi}}(\alpha+l\pi)\widehat{\widetilde{\mathbf{\Phi}}}(\alpha+l\pi)^*\widetilde{\mathcal{B}}(\alpha+l\pi)^* \\ &= \sum_{\nu=0}^{1} \mathcal{B}(\alpha+\nu\pi) \sum_{\sigma\in Z} \widehat{\mathbf{\Phi}}(\alpha+\nu\pi+2\sigma\pi)\,\widehat{\widetilde{\mathbf{\Phi}}}(\alpha+\nu\pi+2\sigma\pi)^*\,\widetilde{\mathcal{B}}(\alpha+\nu\pi)^* \\ &= \mathcal{B}(\alpha)\widetilde{\mathcal{B}}(\alpha)^* + \mathcal{B}(\alpha+\pi)\widetilde{\mathcal{B}}(\alpha+\pi)^*.\end{aligned}$$

Formulas (21)-(23) are equivalent to

$$\textstyle\sum_{l\in Z} A_l\,\widetilde{B}^*_{l+2k} = \mathbf{O}; \quad \sum_{l\in Z} \widetilde{A}_l\,B^*_{l+2k} = \mathbf{O}; \quad \sum_{l\in Z} B_l\,\widetilde{B}^*_{l+2k} = \frac{1}{2}\delta_{0,\,k}\mathbf{I}_s. \tag{24}$$

Thus, both Theorem 2.3 and formulas (20), (24) provide an approach to construct compactly supported biorthogonal vector-valued wavelets.

3 Construction of Biorthogonal Vector-Valued Wavelets

We will proceed to study the construction of biorthogonal vector-valued wavelets and present an algorithm for constructing them.

Lemma 3.1. Let $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$ be a pair of compactly supported biorthogonal vector-valued scaling functions in $L^2(\mathcal{R}, \mathcal{C}^{s\times s})$ satisfying the equations:

$$\mathbf{\Phi}(t) = 2\cdot\textstyle\sum_{k=0}^{\mathcal{L}} A_k\,\mathbf{\Phi}(2\,t-k), \qquad \widetilde{\mathbf{\Phi}}(t) = 2\cdot\sum_{k=0}^{\mathcal{L}} \widetilde{A}_k\,\widetilde{\mathbf{\Phi}}(2\,t-k),$$

and set $\mathbf{\Phi}_\dagger(t) = \mathbf{\Phi}(2t) + \mathbf{\Phi}(2t-1)$, $\widetilde{\mathbf{\Phi}}_\dagger(t) = \widetilde{\mathbf{\Phi}}(2t) + \widetilde{\mathbf{\Phi}}(2t-1)$. Suppose $A_{2n} + A_{2n-2} = A_{2n+1} + A_{2n-1}$, $\widetilde{A}_{2n} + \widetilde{A}_{2n-2} = \widetilde{A}_{2n+1} + \widetilde{A}_{2n-1}$, $n = 0, 1, \cdots, \lceil \frac{\mathcal{L}}{2} \rceil$ where $\mathcal{L}$ is a positive integer and $\lceil x \rceil = \inf\{n : n \geq x, n \in Z\}$. Then

(1) $\mathbf{\Phi}_\dagger(t)$ and $\widetilde{\mathbf{\Phi}}_\dagger(t)$ are a pair of compactly supported biorthogonal vector-valued scaling functions, and $supp\,\mathbf{\Phi}_\dagger(t) \subset [\,0, \lceil \frac{\mathcal{L}}{2} \rceil\,]$; $supp\,\widetilde{\mathbf{\Phi}}_\dagger(t) \subset [\,0, \lceil \frac{\mathcal{L}}{2} \rceil\,]$.

(2) $\mathbf{\Phi}_\dagger(t)$ and $\widetilde{\mathbf{\Phi}}_\dagger(t)$ satisfy the following matrix dilation equations, respectively,

$$\mathbf{\Phi}_\dagger(t) = 2 \cdot \textstyle\sum_{n=0}^{\lceil \mathcal{L}/2 \rceil} (A_{2n} + A_{2n-2}) \mathbf{\Phi}_\dagger(2t-n), \tag{25}$$

$$\widetilde{\mathbf{\Phi}}_\dagger(t) = 2 \cdot \textstyle\sum_{n=0}^{\lceil \mathcal{L}/2 \rceil} (\widetilde{A}_{2n} + \widetilde{A}_{2n-2}) \widetilde{\mathbf{\Phi}}_\dagger(2t-n). \tag{26}$$

According to Lemma 3.1, without loss of generality, we only discuss the problems about construction of vector-valued wavelets with 3-coefficient.

Theorem 3.2. Let $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$ be a pair of 3-coefficient compactly supported biorthogonal vector-valued scaling functions satisfying:

$$\mathbf{\Phi}(t) = 2A_0\mathbf{\Phi}(2t) + 2A_1\mathbf{\Phi}(2t-1) + 2A_2\mathbf{\Phi}(2t-2). \tag{27}$$

$$\widetilde{\mathbf{\Phi}}(t) = 2\widetilde{A}_0\widetilde{\mathbf{\Phi}}(2t) + 2\widetilde{A}_1\widetilde{\mathbf{\Phi}}(2t-1) + 2\widetilde{A}_2\widetilde{\mathbf{\Phi}}(2t-2). \tag{28}$$

Assume that there is an integer $\ell,\ 0 \le \ell \le 2$, such that the matrix Q ,defined in the following equation, is invertible matrix:

$$Q^2 = (\tfrac{1}{2} I_s - A_\ell\, \widetilde{A}_\ell^*)^{-1} A_\ell\, \widetilde{A}_\ell^*. \tag{29}$$

Define

$$\begin{cases} B_j = Q\, A_j, & j \ne \ell, \\ B_j = -Q^{-1} A_j, & j = \ell, \\ \widetilde{B}_j = Q^*\, \widetilde{A}_j, & j \ne \ell, \\ \widetilde{B}_j = -(Q^*)^{-1}\, \widetilde{A}_j, & j = \ell. \end{cases} \qquad j\,,\, \ell \in \{0,1,2\} \tag{30}$$

Then $\mathbf{\Psi}(t) = 2B_0\mathbf{\Phi}(2t) + 2B_1\mathbf{\Phi}(2t-1) + 2B_2\mathbf{\Phi}(2t-2).$

$$\widetilde{\mathbf{\Psi}}(t) = 2\widetilde{B}_0\widetilde{\mathbf{\Phi}}(2t) + 2\widetilde{B}_1\widetilde{\mathbf{\Phi}}(2t-1) + 2\widetilde{B}_2\widetilde{\mathbf{\Phi}}(2t-2).$$

are a pair of biorthogonal vector-valued wavelets associated with $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$.

Proof. For convenience, let $\ell = 1$. By Theorem 2.3 and formula (24), it suffices to show that $\{B_0, B_1, B_2, \widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2\}$ satisfy the following equations:

$$A_0\, \widetilde{B}_2^* = A_2\, \widetilde{B}_0^* = \widetilde{A}_0\, B_2^* = \widetilde{A}_2\, B_0^* = \mathbf{O}, \tag{31}$$

$$A_0\, \widetilde{B}_0^* + A_1\, \widetilde{B}_1^* + A_2\, \widetilde{B}_2^* = \mathbf{O}, \tag{32}$$

$$\widetilde{A}_0\, B_0^* + \widetilde{A}_1\, B_1^* + \widetilde{A}_2\, B_2^* = \mathbf{O}, \tag{33}$$

$$B_0\, \widetilde{B}_2^* = \widetilde{B}_0\, B_2^* = \mathbf{O} \tag{34}$$

$$B_0\, \widetilde{B}_0^* + B_1\, \widetilde{B}_1^* + B_2\, \widetilde{B}_2^* = \tfrac{1}{2}\, \mathbf{I}_s. \tag{35}$$

If $\{B_0, B_1, B_2, \widetilde{B}_0, \widetilde{B}_1, \widetilde{B}_2\}$ are given by (30), then equations (31) and (34) hold from (20). For (32), we obtain from (20) and (29) that

$$\begin{aligned} & A_0\, \widetilde{B}_0^* \,+\, A_1\, \widetilde{B}_1^* \,+\, A_2\, \widetilde{B}_2^* = A_0\, \widetilde{A}_0^*\, Q \,-\, A_1\, \widetilde{A}_1^* Q^{-1} \,+\, A_2\, \widetilde{A}_2^*\, Q \\ &= (\, A_0\, \widetilde{A}_0^* + A_2\, \widetilde{A}_2^*\,)\, Q - A_1\, \widetilde{A}_1^* Q^{-1} = (\, \tfrac{1}{2}\, \mathbf{I}_s - A_1\, \widetilde{A}_1^*\,)\, Q - A_1\, \widetilde{A}_1^* Q^{-1} \\ &= [\,(\, \tfrac{1}{2}\, \mathbf{I}_s - A_1\, \widetilde{A}_1^*\,)\, Q^2 - \, A_1\, \widetilde{A}_1^*\,] Q^{-1} = \mathbf{O}. \end{aligned}$$

Similarly, (33) can be obtained. Finally, we will prove (35) holds.

$$B_0 \widetilde{B}_0^* + B_1 \widetilde{B}_1^* + B_2 \widetilde{B}_2^* = Q\, A_0\, \widetilde{A}_0^*\, Q + Q\, A_2\, \widetilde{A}_2^*\, Q + Q^{-1}\, A_1\, \widetilde{A}_1^*\, Q^{-1}$$

$$
\begin{aligned}
&= Q(A_0\,\widetilde{A}_0^* + A_2\,\widetilde{A}_2^*)Q + Q^{-1}\,A_1\,\widetilde{A}_1^*\,Q^{-1}\\
&= Q^{-1}[\,Q^2(\tfrac{1}{2}\,\mathbf{I}_s - A_1\,\widetilde{A}_1^*)Q^2 + A_1\,\widetilde{A}_1^*\,]Q^{-1}\\
&= Q^{-1}(Q^2\,A_1\,\widetilde{A}_1^* + A_1\,\widetilde{A}_1^*)Q^{-1} = Q\,(\,A_1\,\widetilde{A}_1^* + Q^{-2}A_1\,\widetilde{A}_1^*)Q^{-1}\\
&= Q\,(\,A_1\,\widetilde{A}_1^* + \tfrac{1}{2}\,\mathbf{I}_s - A_1\,\widetilde{A}_1^*)Q^{-1} = \tfrac{1}{2}\,\mathbf{I}_s.
\end{aligned}
$$

Example. Let $\mathbf{\Phi}(t)$, $\widetilde{\mathbf{\Phi}}(t) \in L^2(\mathcal{R}, \mathcal{C}^{2\times 2})$ and $supp\,\mathbf{\Phi}(t) = supp\,\widetilde{\mathbf{\Phi}}(t) = [-1, 1]$, be a pair of 3-coefficient biorthogonal vector-valued scaling functions satisfying the following equations[9]:

$$\mathbf{\Phi}(t) = 2\sum_{k=-1}^{1} A_k\,\mathbf{\Phi}(2t-k), \qquad \widetilde{\mathbf{\Phi}}(t) = 2\sum_{k=-1}^{1} \widetilde{A}_k\,\widetilde{\mathbf{\Phi}}(2t-k).$$

where

$$A_{-1} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ -\frac{1}{2} & -\frac{1}{5} \end{pmatrix}, \qquad A_0 = \begin{pmatrix} \frac{\sqrt{2}(1+i)}{4} & 0 \\ 0 & \frac{1+i\sqrt{3}}{8} \end{pmatrix}, \qquad A_1 = \begin{pmatrix} \frac{1}{4} & -\frac{1}{10} \\ \frac{1}{2} & -\frac{1}{5} \end{pmatrix}.$$

$$\widetilde{A}_{-1} = \begin{pmatrix} \frac{1}{4} & \frac{5}{8} \\ -\frac{7}{32} & -\frac{35}{64} \end{pmatrix}, \qquad \widetilde{A}_0 = \begin{pmatrix} \frac{\sqrt{2}(1+i)}{4} & 0 \\ 0 & \frac{1+i\sqrt{3}}{8} \end{pmatrix}, \qquad \widetilde{A}_1 = \begin{pmatrix} \frac{1}{4} & -\frac{5}{8} \\ \frac{7}{32} & -\frac{35}{64} \end{pmatrix}.$$

Let $\ell = 0$. By using (29) and (30), we get

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\sqrt{7}}{7} \end{pmatrix}, \qquad Q^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{7} \end{pmatrix}.$$

$$B_{-1} = \begin{pmatrix} \frac{1}{4} & \frac{1}{10} \\ -\frac{\sqrt{7}}{14} & -\frac{\sqrt{7}}{35} \end{pmatrix}, \qquad B_0 = \begin{pmatrix} -\frac{\sqrt{2}(1+i)}{4} & 0 \\ 0 & -\frac{\sqrt{7}(1+i\sqrt{3})}{8} \end{pmatrix}, \qquad B_1 = \begin{pmatrix} \frac{1}{4} & -\frac{1}{10} \\ \frac{\sqrt{7}}{14} & -\frac{\sqrt{7}}{35} \end{pmatrix}.$$

$$\widetilde{B}_{-1} = \begin{pmatrix} \frac{1}{4} & \frac{5}{8} \\ -\frac{\sqrt{7}}{32} & -\frac{5\sqrt{7}}{64} \end{pmatrix}, \qquad \widetilde{B}_0 = \begin{pmatrix} -\frac{\sqrt{2}(1+i)}{4} & 0 \\ 0 & -\frac{\sqrt{7}(1+i\sqrt{3})}{8} \end{pmatrix}, \qquad \widetilde{B}_1 = \begin{pmatrix} \frac{1}{4} & -\frac{5}{8} \\ \frac{\sqrt{7}}{32} & -\frac{5\sqrt{7}}{64} \end{pmatrix}.$$

From Theorem 3.2, we conclude that

$$\mathbf{\Psi}(t) = 2\sum_{k=-1}^{1} B_k\,\mathbf{\Phi}(2t-k), \qquad \widetilde{\mathbf{\Psi}}(t) = 2\sum_{k=-1}^{1} \widetilde{B}_k\,\widetilde{\mathbf{\Phi}}(2t-k).$$

are a pair of biorthogonal vector-valued wavelets associated with $\mathbf{\Phi}(t)$ and $\widetilde{\mathbf{\Phi}}(t)$.

References

1. Chui C K. and Lian J., *A study on orthonormal multiwavelets*, J. Appli. Numer. Math., 20(1996): 273-298.
2. Yang S, Cheng Z and Wang H, *Construction of biorthogonal multiwavelets*, J. Math. Anal. Appl. 276(2002) 1-12.
3. Xia X. G., Suter B. W., *Vector-valued wavelets and vector filter banks*. IEEE Trans. Signal Processing, 44(1996) 508-518.
4. Fowler J. E., Hua L., *Wavelet Transforms for Vector Fields Using Omnidirectionally Balanced Multiwavelets*, IEEE Trans. Signal Processing, 50(2002) 3018-3027.
5. Xia X. G., Geronimo J. S., Hardin D. P., Suter B. W., *Design of prefilters for discrete multiwavelet transforms*, IEEE Trans. Signal Processing, 44(1996) 25-35.

Metabolic Visualization and Intelligent Shape Analysis of the Hippocampus

Yoo-Joo Choi[1,5], Jeong-Sik Kim[4], Min-Jeong Kim[2], Soo-Mi Choi[4], and Myoung-Hee Kim[2,3]

[1] Institute for Graphic Interfaces, Ewha-SK telecom building, 11-1 Daehyun-Dong, Seodaemun-Ku, Seoul, Korea
choirina@ewhain.net
[2] Department of Computer Science and Engineering, Ewha Womans University, Daehyun-Dong, Seodaemun-Ku, Seoul, Korea
kimmj@ewhain.net
[3] Center for Computer Graphics and Virtual Reality, Ewha Womans University, Daehyun-Dong, Seodaemun-Ku, Seoul, Korea
mhkim@ewha.ac.kr
[4] School of Computer Engineering, Sejong University, Seoul, Korea
jskim@sju.ac.kr, smchoi@sejong.ac.kr
[5] Department of Computer Application Technology, Seoul University of Venture and Information, Seoul, Korea

Abstract. This paper suggests a prototype system for visualization and analysis of anatomic shape and functional features of the hippocampus. Based on the result of MR-SPECT multi-modality image registration, anatomical and functional features of hippocampus are extracted from MR and registered SPECT images, respectively. The hippocampus is visualized in 3D by applying volume rendering to hippocampus volume data extracted from the MR image with color coded by registered SPECT image. In order to offer the objective and quantitative data concerning to the anatomic shape and functional features of the hippocampus, the geometric volume and the SPECT intensity histogram of hippocampus regions are automatically measured based on the MR and the registered SPECT image, respectively. We also propose a new method for the analysis of hippocampal shape using an integrated Octree-based representation, consisting of meshes, voxels, and skeletons.

1 Introduction

The correlations between the anatomical shape of brain subsystems and brain diseases have been widely researched in order to diagnose and prevent the diseases. In particular, significant attention has been paid to the analysis of the hippocampus in MR images because of its intimate connection to memory, emotion and learning [1-4]. MRI-based hippocampal volumetric measurements are useful in the diagnosis of patients with mesial temporal lobe epilepsy(TLE). Past studies have mainly focused on the analysis of hippocampal volumetric values and geometric shapes using the MR image. In the functional images such as SPECT and PET, it is impossible to directly extract the hippocampus regions because of the low image resolution of the functional

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1140 – 1148, 2005.
© Springer-Verlag Berlin Heidelberg 2005

images. But if the multi-modality image registration can precede, not only the geometrical features but also the functional features of brain subsystem of interest can be extracted. Furthermore, 3-D shape analysis technologies can be effectively applied to compare geometric features of brain subsystem such as the hippocampus between healthy person group and patient group, or to analyze the shape deformation of brain subsystem according to the time.

In this paper, we present a prototype system to analyze the hippocampus in SPECT image as well as in MR image based on a stable and accurate multi-modality surface-based registration using surface distance and curvature optimization that is independent of initial position and orientation of objects to be registered. Functional features of hippocampus are represented on the geometrical surface of hippocampus by color coded based on the registered SPECT image. The volume and SPECT intensity histogram of the hippocampus measured from the MR and the registered SPECT images are furnished to objectively investigate geometrical and functional differences between subjects. Furthermore, in order to analyze local shape deformation efficiently, we propose a new method for the analysis of hippocampal shape using an integrated octree-based representation, consisting of meshes, voxels, and skeletons. The proposed method supports a hierarchical level-of-detail (LOD) representation consisting of hybrid information such as boundary surfaces, internal skeletons, and original medical images.

The rest of the paper is organized as follows: We present the visualization of shape and metabolic function in Section 2. Section 3 describes the quantitative analysis of the hippocampal structure. Section 4 shows the experimental results using the MR and SPECT images of healthy persons and patients with epilepsy. We conclude the paper in Section 5.

2 Visualization of Shape and Metabolic Function

For the MR/SPECT multi-modality brain registration, cerebral cortex areas are preferentially extracted from MR and SPECT images. Hippocampus areas also should be segmented from MR image in pre-processing phase for the analysis of the geometric and functional features of hippocampus by applying the region-growing method and morphological operations. The final step of segmentation is to build a 3D distance map with respect to the surface of cerebral cortex. In the 3D distance map, each non-surface voxel is given a value that is a measure of the distance to the nearest surface voxel. The 3D distance map is used to effectively measure the distance of two brain surfaces to be represented by MR and SPECT images.

In order to guarantee a stable registration result that is independent of the initial position or direction of the test object, moment-based initial transformation is processed before the fine registration process. Moment-based transformation means the translation and rotation of the object-centered coordinate system for a test object in order to overlap object spaces for reference and test objects. We calculated a centroid and 3D principal axes as the moment information based on extracted surface voxels for each volume image. To find the 3D principal axes, the covariance matrix is defined. The eigen values and eigen vectors of the covariance matrix are calculated. The eigen vector with the maximum eigen value is matched to the long principal axis of

the object and the eigen vector with the minimum eigen value represents the short principal axis of the object.

The origin of a coordinate system before initial registration is the first voxel position of the first 2-D image slice in each volume image. *X*, *Y* and *Z* axes mean width, height and depth direction of volume data, respectively. The object-centered coordinate system extracted using the moment information is represented with respect to the initial coordinate system. In initial registration, the transformation matrix for overlapping two object spaces is calculated as shown in Eq. 1:

$$\begin{aligned} M &= T(-C_{x2},-C_{y2},-C_{z2}) \cdot S(Vr_x, Vr_y, Vr_z) \cdot R_Y(\rho_2) \cdot R_Z(\theta_2) \\ &\cdot R_X(\varphi_2) \cdot R_X(-\varphi_1) \cdot R_Z(-\theta_1) \cdot R_Y(-\rho_1) \cdot T(C_{x1}, C_{y1}, C_{z1}) \end{aligned} \tag{1}$$

where, T is the translation transformation and R_X, R_Y and R_Z are rotation transformations with respect to x-, y-, and z-axes, respectively. S is the scaling transformation. (C_{x2}, C_{y2}, C_{z2}) are the coordinates of object centroid in the test image and (C_{x1}, C_{y1}, C_{z1}) are the coordinates of object centroid in the reference image. Vr_x, Vr_y,Vr_z are a ratio of voxel size of the test image to that of the reference image with respect to x-, y-, and z-directions, respectively. ρ_1 and ρ_2 are, respectively, the angles between the projected long axis to XZ plane and x-axis in the reference and test images. θ_1 and θ_2 are angles between the projected long axis to XY plane and x-axis in the reference and test images. The short axis of the test image is transformed into YZ plane by applying $R_Y(\rho_2) \cdot R_Z(\theta_2)$. φ_1 and φ_2 are angles between the transformed short axis to YZ plane and y-axis in the reference and test images.

Surface matching is done by optimizing a cost function that is usually defined by the generalized distance between the surface voxels of two volume images. In this study, we define surface matching cost function by surface distance and surface curvature difference. We use a pre-computed 3D distance map in order to get the distance between one sampling surface voxel of the test image and the extracted surface of the reference image. All surface voxels of reference and test images have their own surface curvature value that is extracted using the Freeman & Davis algorithm. To evaluate the cost function at each iteration, the surface curvature difference between the test and reference images is computed. Eq. 2 shows the proposed cost function:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|Dmap[s_i]\right|^2 \times \left|Curv_{test}[s_i] - Curv_{ref}[s_i]\right|^2\right)} \tag{2}$$

where, n is the number of sampling surface voxels and s is the coordinates of a sampling surface voxel in the test image. *Dmap* is the distance between a sampling surface voxel of the test image and an extracted surface of the reference image. $Curv_{test}$ is the surface curvature of a voxel in the test image and $Curv_{ref}$ is the surface curvature of a voxel in the reference image. Non-surface voxels have a large curvature valueThe geometric shape of the hippocampus is reconstructed based on the MR images and its metabolic function is extracted from the rearranged SPECT images. The intensities of regions in the rearranged SPECT image corresponding to the hippocampus regions of the MR image are extracted based on the registration result. The extracted intensities, that is, metabolic features of the hippocampus are represented on the hippocampus surface by colors coded according to the color palette. The hippocampus is divided into two regions according to the center plane of object-oriented bounding box. For

the quantitative analysis, the volumes of two regions are respectively measured. The SPECT intensity histogram of the hippocampus regions is also automatically extracted from the registered SPECT image.

3 Quantitative Analysis of Hippocampal Structure

In this section, we describe how to analyze the shape of hippocampal structure. Our method is composed of three main steps. First, we construct Octree-based shape representation. Three different types of shape information (i.e. meshes, voxels, and skeletons) are integrated into a single Octree data structure. Then, we normalize this representation into canonical coordinate frame using Iterative Closest Points (ICP) algorithm. Finally, we extract shape features from hippocampus models using L_2 Norm metric. And then we adopt a neural network-based classifier to discriminate between normal controls and epilepsy patients.

The Octree is suitable for generating a local LOD representation. Multiresolutional binary voxel representation can be generated by using depth buffer of OpenGL from the meshes model [5]. The computation time of voxelization is independent of the shape complexity and is proportional to the buffer resolution. The skeletal representation is extracted from the binary voxel representation. Here, the skeletal points are computed by the center of the object in each slice image or by 3-D distance transformation in more general case. The gaps between neighboring skeletal points are linearly interpolated. Figure 1 shows how to integrate three types of shape information into the Octree. To normalize the position and rotation, we adapt the Iterative Closest Point (ICP) algorithm proposed by Zhang *et al.* [6] in which they apply a free-form curve representation to the registration process.

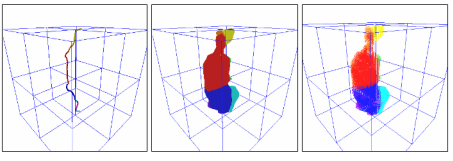


Fig. 1. An integrated Octree-based representation of the right hippocampus: Left: Labeled skeleton, Middle: Labeled meshes, Right: Labeled voxels

After normalizing the shape, we estimate the shape difference by computing the distance for the sample meshes extracted from the deformable meshes using the L_2 Norm metric. The L_2 norm is a metric to compute the distance between two 3D points by Eq. 3, where x and y represents the centers of corresponding sample meshes.

$$\mathrm{L}_2(x, y) = \left(\sum_{i=0}^{k} |x_i - y_i|^2 \right)^{1/2} \tag{3}$$

An artificial neural networks is the mathematical model which imitates the ability of the human brain. Today a great deal of effort is focused on the development of neural networks for applications such as pattern recognition and classification, data compression and optimization [7]. As we adopt a supervised back-propagation neural

network to the implementation of a classifier, we can confirm the feasibility for discriminating between normal controls and epilepsy patients. We use skeletons and sample meshes representation to compute the input of a classifier. The input layer's values are composed of the distances from skeleton points to the sample meshes. The output layer has two kinds of values - 0 or 1 (where '0' means "epilepsy patients" and '1' means "normal controls").

4 Experimental Results

The overall procedures proposed in this paper were implemented using Visual C++ on a Pentium IV. We applied the proposed registration method to MR and PET volume data acquired from five healthy persons and five patients with epilepsy. The hippocampus is reconstructed based on the segmented MR image in 3D and is displayed with MR images of three orthogonal planes. This enables the shape and location features of hippocampus with respect to the overall brain to be clearly analyzed. Figure 2 shows the 3D hippocampus model of one healthy person (Data 6) displayed in MR images of orthogonal planes.

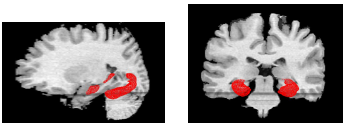


Fig. 2. 3D hippocampus model displayed with MR images of orthogonal planes. Left: sagittal view, Right: coronal view.

Table 1 compares five healthy persons with five epilepsy patients in the volume of the hippocampus. In case of healthy persons, the left and right hippocampi are very symmetric in the volume and shape sides. However, the volumes and shapes of hippocampus of patients with epilepsy are very different as shown in |A-B| field of Table 1 and Figure 3. Figure 3 shows the visualization results of hippocampi in 3D with color coded by the SPECT intensity. Table 2 shows the comparison of SPECT intensity distribution of healthy persons and patients. In case of health persons, the SPECT intensities are normally distributed, whereas the intensities of hippocampus of patients are irregularly distributed and the mean intensity is lower than that of the healthy group. We can easily and intuitively analyze that the overall SPECT intensities of hippocampus range of a patient are lower that those of a healthy person by the surface color to be extracted from the rearranged SPECT image.

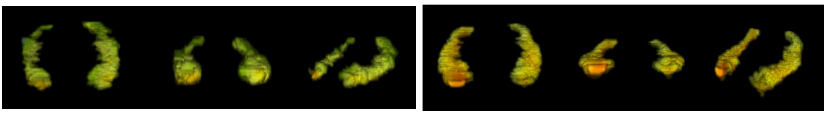


Fig. 3. Color-coded 3D visualization of Hippocampus using MR and SPECT images. Left: The hippocampi of patient with epilepsy, Right: The hippocampi of healthy person.

We also tested the ability of the proposed normalization algorithm for the hippocampal models and of the back-propagation neural network based classifier. In order to estimate the capacity of our normalization method, we tested 270 skeleton pairs from the deformed models. We tested five kinds of the skeletons to check the efficiency of our method (6 points skeleton, 11 points skeleton, 31 points skeleton, 51 points skeleton, 101 points skeleton). We were able to acquire 0.051 average sec for normalizing two mesh models (each model has 5,000 meshes).

Table 1. Comparison of hippocampus volume. (unit : mm^3)

		Left Hippocampus			Right Hippocampus			|A-B|
		First Half	Second Half	Whole (A)	First Half	Second Half	Whole (B)	
Patients with epilepsy	Data 1	1875	1342	3217	2103	1614	3717	500
	Data 2	1735	1431	3166	964	1385	2349	817
	Data 3	1331	851	2182	2154	1981	4135	1953
	Data 4	1328	1134	2462	1837	1454	3291	829
	Data 5	1135	870	2005	1339	1125	2464	459
	Mean	**1480.8**	**1125.6**	**2606.4**	**1679.4**	**1511.8**	**3191.2**	**911.6**
	Std Dev.	**310.39**	**265.00**	**558.71**	**514.09**	**316.08**	**777.06**	**607.15**
Healthy Persons	Data 6	1654	1180	2834	1198	1233	2431	403
	Data 7	1582	862	2444	1745	1112	2857	413
	Data 8	1759	1219	2978	1410	1263	2664	314
	Data 9	1572	1086	2658	1101	1458	2559	99
	Data10	1933	1043	2976	1895	1190	3085	109
	Mean	**1700**	**1078**	**2778**	**1469.8**	**1251.2**	**2719.2**	**267.6**
	Std Dev.	**150.16**	**139.84**	**228.19**	**342.78**	**128.80**	**257.18**	**154.27**

Table 2. Comparison of SPECT intensity distribution

		Mean	Standard Deviation
Patients with Epilepsy	Data 1	128.64	18.11
	Data 2	116.07	19.36
	Data 3	109.12	32.76
	Data 4	141.75	29.59
	Data 5	116.95	41.41
Healthy Persons	Data 6	147.42	19.85
	Data 7	123.38	26.22
	Data 8	161.13	19.82
	Data 9	142.85	17.17
	Data 10	170.61	24.50

Figure 4(left) shows the results of global analysis between the hippocampal structures of a normal subject and a patient with epilepsy. Figure 4(middle) and (right) show how to compare two hippocampal shapes based on the proposed Octree scheme. It is possible to reduce the computation time in comparing two 3-D shapes by

picking a certain skeletal point (Figure 4(middle)) or by localizing an Octree node (Figure 4(right)) from the remaining parts. It is also possible to analyze the more detail region by expanding the resolution of the Octree, since it has a hierarchical structure. The result of shape comparison is displayed on the surface of the target object using color-coding.

Table 3 gives the results of global shape difference between the normal left hippocampus (N_L) and three deformed targets (T1, T2, and T3) in the upper area. From Table 1, it shows that T1 and T3 are 5.7% and 11.2%, respectively, which is smaller than N_L, whereas T2 is 9.3% larger than N_L. Table 4 summarizes the result of local shape differences by comparing the 3-D shapes between the reference models (P_L and N-R) and deformed targets (T4~T7), respectively. P_L is an abnormal left hippocampus in epilepsy and N_R is a normal right hippocampus. T4~T7 are deformed targets at specific region (i.e. upper-front-right, bottom-front-left, upper-back-left, and the bottom region, respectively). In Table 4, we observe that the similarity error at deformed region is higher than at other regions. As shown in Table 3 and 4, our method is able to discriminate the global shape difference and is also able to distinguish a certain shape difference at a specific local region in a hierarchical fashion.

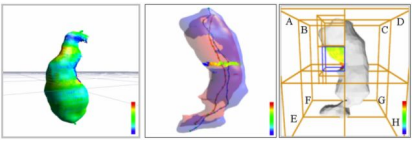


Fig. 4. Global and hierarchical local shape analysis. Left: global analysis result, Middle: skeletal point picking based local shape analysis, Right: Octree-based hierarchical shape analysis.

Table 3. The result of global shape analysis

	L_2 Norm	Volume difference	Rank
N_L:T1	1.220	94.3%	1
N_L:T2	1.554	109.3%	2
N_L:T3	2.420	88.8%	3

Table 4. The result of local shape analysis based on the Octree structure

	A	B	C	D	E	F	G	H
P_L:T4	0.15	0.77	0.84	3.15	0.00	0.00	0.00	0.15
P_L:T5	1.20	0.00	0.00	0.00	3.12	2.00	1.00	1.44
N_R:T6	0.06	1.02	0.06	0.00	0.00	0.12	0.00	0.00
N_R:T7	0.00	0.00	0.00	0.00	1.54	1.31	1.313	1.54

To confirm the capacity of our classifier, we used 80 hippocampal mesh models and 400 skeletons extracted from the experimental models for the learning and the test. And, through the cross-validation technique, we organized more learning and test set. Figure 5 shows the result of the classification based on the back-propagation neural network.

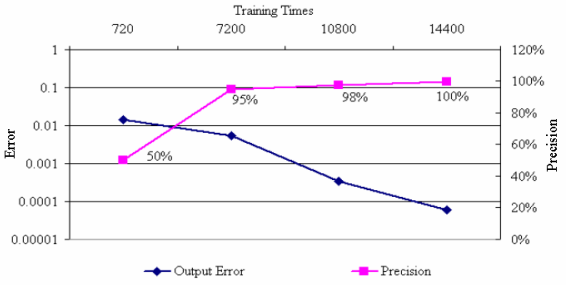


Fig. 5. The result of the neural network based classification for the hippocampal models

5 Conclusion and Future Work

In this paper, we compared the hippocampus features of two subject groups based on MR and SPECT volume data acquired from five healthy persons and five patients with epilepsy using the proposed prototype system. In order to compare the anatomical and functional features of hippocampus of two groups, the MR-SPECT multi-modality brain registration was preferentially performed. In this paper, the surface curvature-based registration method was proposed. Our proposed registration method improved error rate by about 40.65 % with respect to the commercial surface-based registration tool [8]. This improvement is the result of reducing the local minimum risk caused by neglecting the 3D shape properties of objects. It was also reported that the proposed registration method has a stable error rate, without regard to the position and orientation of subjects, due to the initial registration based on moment information.

In the proposed system, the volume and SPECT intensity histogram of the hippocampus measured from the MR and the registered SPECT images are furnished and the color-coded hippocampus model is visualized with orthogonal image planes. These functions can enable diagnosticians to analyze the geometric and metabolic features of the hippocampus objectively and effectively. Furthermore, the proposed shape analysis method can not only discriminate the global shape difference, but also distinguish a certain shape difference at a specific local region by applying a hierarchical level-of-detail (LOD) representation consisting of hybrid information such as boundary surfaces, internal skeletons, and original medical images.

Acknowledgements

This work was partially supported by the Korean Ministry of Science and Technology under the NRL Program and in part by the Korean Ministry of Information and Communication under the ITRC Program.

References

1. S. Bouix, J. C. Pruessner, D. L. Collins, K. Siddiqi : Hippocampal Shape Analysis Using Medial Surfaces. MICCAI (2001) 33 - 40.
2. Lei Wang, Sarang C. Joshi, Michael I. Miller, John G. Csernansky: Statistical Analysis of Hippocampal Asymmetry in Schizophrenia, NeuroImage 14 (2001) 531 – 545.

3. Martin Stin, Jeffrey A. Lieberman, Guido Gerig : Boundary and Medial Shape Analysis of the Hippocampus in Schizophrenia. MICCAI (2003)
4. R. Edward Hogan, Richard D. Bucholz, Sarang Joshi: Hippocampal Deformation-based Shape Analysis in Epilepsy and Unilateral Mesial Temporal Sclerosis. Epilepsia, 44(6) (2001) 800-806
5. Karabassi, E.A., Papaioannou, G., Theoharis, T.: A Fast Depth-Buffer-Based Voxelization Algorithm. Journal of Graphics Tools, ACM, Vol. 4. No.4, (1999) 5-10.
6. Zhang Z.: Iterative point matching for registration of freeform curves and surfaces. International Journal of Computer Vision, Vol. 13, No. 2, (1994) 119-152.
7. Sordo, M.: Introduction to Neural Networks in Healthcare, Open Clinical Document, October, (2002).
8. Analyze 5.0 Tutorials(copyright 1999 - 2003 AnalyzeDirect, copyright 1986 -2003 BIR, Mayo Clinic).

Characteristic Classification and Correlation Analysis of Source-Level Vulnerabilities in the Linux Kernel*

Kwangsun Ko[1], Insook Jang[2], Yong-hyeog Kang[3], Jinseok Lee[2], and Young Ik Eom[1,**]

[1] School of Information and Communication Eng., Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea
{rilla91, yieom}@ece.skku.ac.kr
[2] National Security Research Institute, 161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Korea
{jis, jinslee}@etri.re.kr
[3] School of Business Administration, Far East University, 5 San Wangjang, Gamgok, Eumseong, Chungbuk 369-851, Korea
yhkang@mail.kdu.ac.kr

Abstract. Although studies regarding the classification and analysis of source-level vulnerabilities in operating systems are not direct and practical solutions to the exploits with which computer systems are attacked, it is important that these studies supply the elementary technology for the development of effective security mechanisms. Linux systems are widely used on the Internet and in intra-net environments. However, researches regarding the fundamental vulnerabilities in the Linux kernel have not been satisfactorily conducted. In this paper, characteristic classification and correlation analysis of source-level vulnerabilities in the Linux kernel, open to the public and listed on the SecurityFocus site for the 6 years from 1999 to 2004, are presented. This study will enable Linux kernel maintenance groups to understand the wide array of vulnerabilities, to analyze the characteristics of the attack abusing vulnerabilities, and to prioritize their development effort according to the impact of these vulnerabilities on the Linux systems.

1 Introduction

There have been ongoing studies conducted by academics and institutes on the classification and analysis of hardware and software vulnerabilities in computer systems since the 1970s. Although these these studies are not direct and practical solutions to the exploits with which computer systems are attacked, it is

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).
** Corresponding author.

Y. Hao et al. (Eds.): CIS 2005, Part II, LNAI 3802, pp. 1149–1156, 2005.
© Springer-Verlag Berlin Heidelberg 2005

important that these studies supply the elementary technology for the development of effective security mechanisms. Software containing a minor security flaw can expose a secure environment and make a system vulnerable to attacks [1][2].

There are a variety of reasons in the existence of security vulnerabilities in computer systems: incorrect implementation, improper configuration and initialization, design errors, no verification of parameters, abusive system calls, and so on. Surely as the social concerns of security vulnerabilities increases, computer systems must be designed to be increasingly secure. In 1991, after the first version of the Linux kernel distributed out by Linus Torvalds, Linux systems are widely used on Internet and in intra-net environments. However, researches regarding the fundamental vulnerabilities inherent in the Linux kernel have not been thoroughly conducted. In this paper, characteristic classification and correlation analysis of 124 source-level vulnerabilities in the Linux kernel, open to the public and listed in the SecurityFocus site for the 6 years from 1999 to 2004, are presented according to Linux kernel versions.

The subsequent sections of this paper are organized as follows. Section 2 presents related studies that have been conducted, in order to classify and analyze vulnerabilities in computer systems. In Section 3 and 4, characteristic classification and correlation analysis of source-level vulnerabilities in the Linux kernel are detailed, respectively. Section 5 concludes this paper.

2 Related Works

Several studies exist regarding vulnerabilities in computer systems: the Research In Secured Operating Systems (RISOS) [2], Security taxonomy [3], Chillarege's Orthogonal Defect Classification [4][5], Spafford's taxonomy [6], Landwehr's taxonomy [6], Bishop's taxonomy [7], Du and Mathur's taxonomy [8], and Jiwnani's taxonomy [2]. In addition, there are many Internet sites that have classified or analyzed the vulnerabilities of computer systems: the SecurityFocus site [9], the Common Vulnerabilities and Exposures (CVE) site [10], the LinuxSecurity site [11], and the iSEC Security Research site [12].

In this paper, necessary information regarding characteristic classification and correlation analysis is obtained from the SecurityFocus site for the 6 years from 1999 to 2004, where source-level vulnerabilities in the Linux kernel are very well defined and categorized among others.

3 Characteristic Classification

The characteristic classification of source-level vulnerabilities in the Linux kernel are done as follows:

- ***Location in the Linux kernel.*** This consists of 8 criteria: 'network', 'device', 'memory', 'system call', 'file', 'process', 'hardware', and 'etc.' according to where vulnerabilities are detected. (Software functionalities of the Linux kernel, 'location' of [6], and 'location of flaws in the system' of [2] are referred to in this paper.)

- ***Impact on the system.*** This consists of 6 criteria: 'privilege escalation', 'memory disclosure', 'denial of service', 'system crash', 'information leakage', and 'etc.' according to how vulnerabilities impact on the system. ('effect domain' of [7] and 'impact of flaws on the system' of [2] are referred to in this paper.)
- ***Existence of exploits.*** This consists of 4 criteria: 'O (there are exploits)', 'X (there is not a exploit)', 'proof of concept (just showing how an exploit code processes)', and 'no exploit is required (e.g. sending abnormal network packets is enough to exploit a vulnerability)' according to the existence of exploits. ([9] is referred to in this paper.)

There are 132 vulnerabilities from the SecurityFocus site for the 6 years from 1999 to 2004: 18 vulnerabilities in 1999; 5 in 2000; 12 in 2001; 16 in 2002; 12 in 2003; and 69 in 2004. Some are, however, also excluded for accuracy. Vulnerabilities like a 'Multiple Local Linux Kernel Vulnerabilities (Aug. 27, 2004)' can not be accepted as raw data of our classification and analysis because of ambiguousness. There are 3 and 5 vulnerabilities in Linux 2.4 kernel and 2.6, respectively. Therefore, 124 vulnerabilities are used for characteristic classification in this paper. For the reference, the reason that the number of vulnerabilities in 1999 are higher than all other years except 2004 is because 11 pre-existing vulnerabilities were simultaneously made public on Jun. 1, 1999.

Based on the SecurityFocus' taxonomy, the numbers and percentages of vulnerabilities are presented in Table 1.

Table 1. The numbers and percentages of vulnerabilities (SecurityFocus' taxonomy)

Criteria	Number	Percentage
Race condition error	6	4.8
Boundary condition error	18	14.5
Access validation error	9	7.3
Serialization error	2	1.6
Failure to handle exceptional conditions	25	20.2
Environment error	2	1.6
Input validation error	3	2.4
Origin validation error	1	0.8
Design error	40	32.3
Unknown	18	14.5
Total	124	100

Table 1 shows that almost 80% of total vulnerabilities are related to 4 criteria: 'Design error' (32.3%), 'Failure to handle exceptional conditions' (20.2%), 'Boundary condition error' (14.5%), and 'Unknown' (14.5%). More than 50% are related to the criteria of 'Design error' and 'Failure to handle exceptional condition'. The reason that the criteria of 'Design error' and 'Unknown' occupy relative much portion of total vulnerabilities is because many functions and mechanisms are incorrectly implemented, supporting a variety of hardware

products and faculties. This fact means that as Linux systems play important roles in IT, Linux kernel maintenance groups intensively fix any 'Design error' and 'Unknown' issues as quickly as possible.

Additionally, there are two considerations. One is the division of vulnerabilities according to Linux kernel versions. Vulnerabilities found prior to the public date of a specific Linux X kernel are associated with the prior version of X. For example, vulnerabilities from Jan. 4, 2000 to Dec. 17, 2003 are associated with Linux 2.4 kernel; Linux 2.6 kernel was released, Dec. 18, 2003. The other is that the sums of vulnerabilities according to Linux kernel versions are accumulated because a number of vulnerabilities do not belong to a only specific Linux kernel version.

3.1 Location in the Linux Kernel

Based on the 'location in the Linux kernel', the numbers of vulnerabilities are presented in Table 2. Vulnerabilities related to inner kernel functions are included in 'System call' criterion, and those related to 'Signal', 'Capability', and 'Structure' are included in 'Process' criterion.

Table 2. The numbers of vulnerabilities (location in the Linux kernel)

Version	Network	Device	Memory	System call	File	Process	Hardware	Etc.
2.2	9	-	2	-	3	4	-	-
2.4	27	4	4	8	8	6	-	2
2.6	39	14	10	22	15	16	4	4

Table 2 show that most vulnerabilities are included under 4 criteria in Linux 2.2 kernel: 'Network', 'Process', 'File', and 'Memory'. However in the posterior Linux kernel versions, those are spread over the entire Linux kernel. It is indirectly confirmed that many vulnerabilities related to the main functionalities in the Linux kernel have been fixed and that Linux systems are becoming increasingly stable in all major faculties. The 'Hardware' criterion firstly appears in Linux 2.6 kernel and will increase continuously because a number of hardware-dependent kernel codes supporting various hardware platforms may be incorrectly implemented. The vulnerabilities of 'System call' criterion which is 13.6% appear in Linux 2.4 kernel and increase more than 17.7% in 2.6. This fact means that no checking the arguments of system calls whether the arguments are correct until the completion of translation from linear to physical address [14] has problems. Additionally, we predict that the vulnerable probability of system calls becomes higher.

3.2 Impact on the System

Based on the 'impact on the system', the numbers of vulnerabilities are presented in Table 3. The criterion of 'etc.' include the 4 criteria: 'buffer overflow', 'hardware access', 'providing the wrong information', and 'implicit effects' criterion.

Table 3. The numbers of vulnerabilities (impact on the system)

Version	Privilege escalation	Memory disclosure	Denial of service	System crash	Information leakage	Etc.
2.2	3	1	7	6	-	5
2.4	12	1	24	10	7	12
2.6	27	4	48	18	18	21

Table 3 show that 'privilege escalation' and 'denial of service' are less than 20% and more than 30%, respectively. These criteria are mainly used when attackers exploit Linux systems. The percentage of 'memory disclosure' is low in all versions, and this may prove that the virtual memory management mechanism in the Linux kernel is very effective. The reason that the number of vulnerabilities in the 'information leakage' criterion increases in posterior to Linux 2.4 kernel is because tracing system calls (e.g. *ptrace()* system call) or the */proc* filesystem are frequently used in order to support many new mechanisms.

3.3 Existence of Exploits

Based on the 'existence of exploits', the numbers of vulnerabilities are presented in Table 4.

Table 4. The number of vulnerabilities (existence of exploits)

Version	O	X	Proof of concept	No exploit required
2.2	16	2	-	-
2.4	28	23	4	4
2.6	39	68	11	6

Table 4 show that the ratio of public vulnerabilities sharply increases while that of the corresponding exploit codes slowly increase. This fact presents that the exploit codes related to vulnerabilities take considerable time to be open. The advent of exploit codes also intentionally decreases in order to prevent unskilled people from easily abusing exploit codes. This is confirmed by the result of the 'proof of concept' criterion.

4 Correlation Analysis

In this section, the results of a correlation analysis are presented based on the characteristic classification above. There are two criteria: (location in the Linux kernel, impact on the system) and (existence of exploits, location in the Linux kernel), are briefly (location, impact), (existence, location), respectively.

4.1 Location and Impact

Based on the (location, impact), the numbers of vulnerabilities are presented in Table 5.

Table 5. The numbers of vulnerabilities in high-ranked criteria of (location, impact)

Version	Location	Impact	Number (duplication)
2.2	Network	Denial of service	3
	Network	System crash	4 (1)
	Process	Denial of service	2 (1)
2.4	Network	Denial of service	6
	Network	Information leakage	4
	Device	Privilege escalation	5 (3)
	System call	Privilege escalation	4
2.6	Network	Denial of service	7
	Device	Privilege escalation	4 (1)
	System call	Denial of service	7

Table 5 shows that the number of (network, denial of service) and (device, privilege escalation) criteria are 16/46 (34.8%) and 9/46 (19.6%) in all Linux kernel versions, respectively. This shows that the impact of vulnerabilities corresponding to the location of the 'network' relates to 'denial of service' and that 'device' relates to 'privilege escalation'. That is, when attackers aim to make the services of a target system helpless against an attack, attackers usually abuse the vulnerabilities related to 'network' faculties. When attackers aim to be an abnormal administrator, attackers usually abuse vulnerabilities relating to 'device'. Needless to say, device drivers are not as susceptible to exploits as other parts of OS and that there seems to be a large lag between discovery of vulnerabilities and actual exploit code [15][16]. No regarding of the vulnerabilities corresponding to device drivers is not useful because device drivers in Linux run as kernel modules and the source code size of the */dev* directory is very larger than others.

4.2 Existence and Location

Based on the (existence, location), the numbers of vulnerabilities are presented in Table 6.

Table 6. The number of vulnerabilities in high-ranked criteria of (existence, location)

Version	Existence	Location	Number
2.2	O	Network	8
	O	File	3
	O	Process	3
2.4	O	System call	5
	X	Network	11
	X	Device	4
2.6	X	Device	10
	X	System call	10
	X	File	7
	X	Process	7

Table 6 shows that the ratio of 'O' clearly decreases from Linux 2.2 kernel to 2.6. The frequency of 'O' is 14/14 (100%) in Linux 2.2 kernel; 5/20 (25%) in 2.4; and 0/34 (0%) in 2.6. This means that the advent of exploit codes to public vulnerabilities take much time or the number of public exploit codes intentionally decreases in order to prevent unskilled people from easily abusing exploit codes. In all Linux kernel versions, about 14/49 (28.6%) of 'X' are in the 'device' criterion of location. This fact is somewhat different from the result of the characteristic classification of 'location in the Linux kernel'; the ratio of vulnerabilities is low. That is, there are a number of source-level vulnerabilities related to 'device' criterion, however, the exploitable probability of vulnerabilities is low. It is indirectly proved that when an attacker wants to exploit vulnerabilities of a target system, he or she has a problem with which source codes of 'device' criterion are chosen because the target system has many source codes in order to support a variety of hundreds of thousands of hardware products.

5 Conclusions

In this paper, characteristic classification and correlation analysis of source-level vulnerabilities in the Linux kernel, open to the public and listed on the SecurityFocus site for the 6 years from 1999 to 2004, are presented. According to characteristic classification, most vulnerabilities in Linux 2.2 kernel are detected in the criteria of 'Network', 'Process', 'File', and 'Memory', however, vulnerabilities in posterior versions are spread over the entire criteria of the Linux kernel. Based on 'impact on the system', the criteria of 'privilege escalation' and 'denial of service' occupy less than 20% and more than 30% of total vulnerabilities, respectively. Additionally, the ratio of public vulnerabilities sharply increases in whole Linux kernel versions, however that of exploit codes related to public vulnerabilities slowly increases. According to correlation analysis, when attackers aim to render target systems denial of service, they mainly exploit vulnerabilities related to 'network' faculties while when attackers want to be abnormal administrators, they mainly abuse vulnerabilities related to 'device'. There are a number of source-level vulnerabilities related to 'device' criterion, however, the exploitable probability of vulnerabilities is low.

This study will enable Linux kernel maintenance groups to understand the wide array of vulnerabilities, to analyze the characteristics of the attack abusing vulnerabilities, and to prioritize their development effort according to the impact of these vulnerabilities on the Linux systems.

References

1. B. Marick: A survey of software fault surveys, Technical Report UIUCDCS-R90-1651, University of Illinois at Urbana-Chamaign, Dec. (1990)
2. K. Jiwnani and M. Zelkowitz: Maintaining Software with a Security Perspective, International Conference on Software Maintenance (ICSM'02), Montreal, Quebec, Canada (2002)

3. Security Taxonomy, http://www.garlic.com/ lynn/secure.htm.
4. R. Chillarege: ODC for Process Measurement, Analysis and Control, Proc. of the Foruth International Conference on Software Quality, ASQC Software Division, McLean, VA, (1994)3-5
5. R. Chillarege, I. S. Bhandari, J. K. Chaar, M. J. Halliday, D. S. Moebus, B. K. Ray, and Man-Yuen Wong: Orthogonal Defect Classification - A Concept for In-Process Measurements, IEEE Transactions on Software Engineering, 18(1992)
6. C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi: A Taxonomy of Computer Program Security Flaws, ACM Computing Surveys, 26(1994)
7. M. Bishop: A Taxonomy of UNIX System and Network Vulnerabilities, Technical Report CSE-95-10, Purdue University (1995)
8. W. Du and A. P. Mathur: Categorization of Software Errors that led to Security Breaches, Proc. of the 21st National Information Systems Security Conference (NISSC' 98), Crystal City, VA (1998)
9. SecurityFocus, http://www.securityfocus.com.
10. Common Vulnerabilities and Exposures, the Standard for Information Security Vulnerability Names, http://www.cve.mitre.org.
11. Guardian Digital: Inc., http://www.linuxsecurity.com.
12. iSEC Security Research, http://www.isec.pl.
13. A. Rubini and J. Corbet, Linux Device Drivers 2nd Ed., O'REILLY (2001)
14. D. P. Bovet and M. Cesati, Understanding the Linux Kernel 2nd Ed., O'REILLY (2003)
15. E. Rescorla: Security Holes Who cares?, Proc. Of the 12the USENIX Security Symposium, Washington D.C. (2003)
16. H. Browne, W. Arbuagh, J. McHugh, and W. Fiothen: A Trend Analysis of Exploitations, In IEEE Symposium on Security and Privacy (2001)

Author Index

Ahn, Dosung II-635
Aine, Sandip I-57
Akhtar, Saqlain I-528
Alagar, Vasu S. I-303
Allende, Héctor I-49
An, Bo I-375
Araki, Kenji I-434
Årnes, André II-388
Aritsugi, Masayoshi II-548
Austin, Francis R. I-902, II-649

Bae, Duhyun II-439
Baek, Joong-Hwan I-729, II-899
Baek, Joon-sik I-729
Baek, Kyunghwan II-285
Bagis, Aytekin II-1042
Bagula, Antoine B. I-224
Bai, Ji II-826
Baik, Doo-Kwon II-725
Bao, Guanjun II-977
Bao, Shiqiang II-989
Bau, Y.T. I-657
Berg, Henrik I-17
Bhattacharya, Bhargab B. I-1057
Bhowmick, Partha I-1057
Bian, Shun I-809
Bin, Liu I-616
Bin, Shen I-704
Biswas, Arindam I-1057
Bo, Liefeng I-909
Bo, Yuan I-704
Bougaev, Anton I-991
Brekne, Tønnes II-388
Brown, Christopher L. II-80
Brown, Warick II-1074
Burkhardt, Hans II-820
Byun, Jin Wook II-143

Cai, Wandong II-457
Cai, Yuanli I-410
Cai, Zhiping II-703
Cao, Chunhong I-145
Cao, Jian I-121, I-267, II-679
Cao, Lei II-679
Cao, Tianjie II-1036
Cao, Wenming II-1068, II-1110
Cao, Yijia I-1106
Cao, Zhenfu II-104, II-167, II-445
Cao, Zhenyu I-267
Castiello, Ciro I-321
Chae, Kyu-Yeol II-947
Chai, Wanfang II-1068
Chakrabarti, P.P. I-57
Chau, K.W. I-645
Chang, Chinchen I-464, II-538, II-554, II-567
Chang, Elizabeth II-55, II-273
Chang, Hye-Young II-260
Chang, Nam Su II-1
Chang, Yilin I-864, I-1051, II-597
Chen, Chun I-941
Chen, Chunlin I-393
Chen, Enhong I-73
Chen, Gen-Cai I-941
Chen, Guimin I-663
Chen, Haibo I-758
Chen, Haixia I-145
Chen, Hua I-101
Chen, Jing II-838
Chen, Junying I-957
Chen, Licong II-629
Chen, Ling I-261, I-941
Chen, Liping II-629
Chen, Mingjun I-638
Chen, Mingyu II-322
Chen, Qingjiang II-1134
Chen, Qingliang I-349
Chen, SiBao I-1003
Chen, Tengbo I-651
Chen, Weihong II-74
Chen, Wensheng I-933
Chen, Xiaofeng II-117
Chen, Xuyang II-851
Chen, Yongqiang II-1017
Chen, Zhiguo II-433
Chen, Ziyi I-200
Chen, Zonghai I-393
Cheng, ChunTian II-421

Cheng, Daijie I-375
Cheng, Xiangguo II-49
Cheng, Xiaochun I-381, I-422
Cheung, Yiu-Ming I-184, I-965
Chi, Yue II-415
Cho, A-Young II-623
Cho, Byeong Heon I-1064
Cho, Ik-Hwan II-623
Cho, SeongJe II-125, II-260
Cho, Tae Ho II-340
Cho, Yookun II-125
Cho, Youn-Ho II-1025
Choi, Jae Ho II-869
Choi, Jongmoo II-125
Choi, Jun I-313
Choi, Jun Yong II-303, II-451
Choi, Kyunghee II-297
Choi, Soo-Mi II-1140
Choi, Sung Jin II-494
Choi, Won-Hyuck II-669
Choi, WoongChul II-260
Choi, Yoo-Joo II-1140
Choudary Gorantla, M. II-110
Chu, Xiuqin I-1009
Chung, Kyoil II-1104
Chung, Sang-Kyoon II-1104
Chung, Yongwha II-635
Coello Coello, Carlos A. I-208
Cruz-Cortés, Nareli I-208
Cui, Can I-864
Cui, Shuning II-731

Dagdelen, Ulvi II-1042
Dai, Guozhong I-355, I-1100
Dai, Qionghai II-863
Dai, Xianhua I-927
Dan, Hongwei I-752
Dash, P.K. I-163
de la Iglesia, B. II-1011
Deng, Naiyang I-580, I-586
Deng, Shengchun I-157, I-735
Deng, Yafeng II-772
Deng, Zhihong I-458
Dhurandhar, Amit I-1021
Ding, Guiguang II-863
Ding, Jianli I-230
Doh, Kyu B. II-780
Doi, Hiroshi II-33
Dong, Daoyi I-393
Dong, Jinxiang II-965
Dong, Lihua II-68
Dong, Liyan I-145
Dong, Ping II-27
Dong, Won-Pyo II-947
Dongliang, Wang I-616
Du, Haifeng II-923
Du, Ruizhong II-244
Du, Tao I-151
Duan, Haixin II-482, II-518
Duan, Peijun I-381
Duan, Shanshan II-445
Duc, Duong Anh II-605
Duthen, Yves I-329

Ekinci, Murat I-973
Engelbrecht, Andries P. I-192
Eom, Young Ik II-1149
Eun, Jongmin I-574
Ewe, Hong Tat I-622, I-657

Fan, Liangzhong II-935
Fan, Wenbing II-838
Fanelli, Anna Maria I-321
Fang, Bin I-833, II-1060
Fang, Binxing II-212, II-415
Fang, Wei I-1015
Fang, Yong I-1015
Feng, Boqin I-747, II-731
Feng, Dengguo II-995
Feng, Huamin I-1015
Feng, Xiang-chu II-1122
Feng, Zhilin II-965
Feng, Zhiyong I-878
Fu, Guojiang I-638
Fu, Peng II-1017
Fu, Zetian I-598
Fung, Chun Che II-1074

Gan, John Q. I-495
Gan, Xiaobing I-902
Gao, Dengpan I-915
Gao, Xiaojuan II-398
Gao, Xinbo I-696, I-997
Gao, Yuelin I-675
Ge, Fei I-915
Gedikli, Eyup I-973
Georgiadis, Panagiotis II-589
Gerardo, Bobby D. I-337, I-1082
Ghose, Supratip I-470
Gong, Maoguo I-793, I-846, I-858

Gonzalez, Jesus A. I-41
Gopalan, Srividya II-236
Gu, Chunxiang II-9
Gu, Lixu I-884, I-890
Gu, Ming I-476
Guan, Chun I-387
Guo, Jun I-979
Guo, Ping II-43
Gustafsson, Lennart I-81

Hahn, Hernsoo II-285
Han, Bing I-997
Han, Changcai II-1116
Han, Dongfeng I-483
Han, Jiuqiang I-133
Han, Lixia I-297
Han, Qi I-663
Han, Song II-55
Han, Sunyoung II-737
Han, Youngjun II-285
Hao, Yue II-796
Haslum, Kjetil II-388
He, Caisheng I-927
He, Hui II-560
He, Xiaoliang I-1088
He, Xin I-381
He, Yan II-463
He, Zengyou I-157, I-735
He, Zhaoyang II-965
He, Zhenyu II-1060
Hippisley, Andrew I-489
Ho, Chang-Hoi I-503
Ho, Anthony T.S. II-661
Ho, Chin Kuan I-622, I-657
Hong, Sungjune II-737
Hong, Wei-Chiang I-512
House, Terry C. II-719
Hou, Yunshu II-875
Hsu, Che-Chang I-550
Hu, Fangming I-1009
Hu, Heping I-442
Hu, Hesuan I-1094
Hu, Jun I-387
Hu, Mingzeng II-560
Hu, Ruijuan II-989
Hu, S.X. II-917
Hu, Xiaofeng II-804
Hu, Yupu II-68
Huang, Dezhi II-812
Huang, Fenggang I-1039
Huang, Houkuan I-9, I-33, I-741
Huang, Huifeng II-554
Huang, Jian I-933
Huang, Jing I-604
Huang, Jiwu II-573
Huang, Mei-juan II-532
Huang, Peiwei II-161
Huang, T.Z. II-917
Huh, Sung-Hoe I-503
Huh, Sung-Hoe I-544
Hui, Li II-180
Huo, Hongwei I-821
Hwang, Jae-jeong I-1082

Imai, Sayaka II-548

Jang, Heejun II-285
Jang, Insook II-1149
Jang, Jong Whan II-869
Jang, MinSeok I-95
Jawale, Rakesh I-1021
Jeong, Chang-Sung II-845, II-947
Jeong, Dong-Seok II-623
Jeong, Dongwon II-725
Ji, Dongmin I-127
Ji, Hongbing I-556, I-997
Ji, Ping I-405
Ji, Shiming II-977
Jia, Jianyuan I-663
Jiang, Gangyi II-935
Jiang, Quanyuan I-1106
Jiang, Zhengtao II-1080
Jiao, Licheng I-89, I-238, I-273, I-696, I-793, I-839, I-846, I-858, I-909
Jiao, Yong-Chang I-247, I-651
Jin, Andrew Teoh Beng II-788
Jin, Hong I-1100
Jin, Jae-Do II-1098
Jin, Long I-721
Jin, Yi-hui I-630
Jing, Wenfeng I-139
Jing, Yinan II-1017
Jo, Geun-Sik I-470
Jones, David Llewellyn I-1074
Joo, Moon G. I-127
Juárez-Morales, Raúl I-208
Juneja, Dimple I-367
Jung, Gihyun II-297
Jung, In-Sung I-107

Jung, Je Kyo I-216
Jung, SeungHwan II-635
Jung, Seung Wook II-86
Jung, Sung Hoon I-1064
Jung, Young-Chul II-743

Kajava, Jorma II-508
Kanamori, Yoshinari II-548
Kang, Daesung I-65
Kang, Daniel II-1127
Kang, Geuntaek I-127
Kang, Hyun-Ho II-581
Kang, Hyun-Soo II-857
Kang, Lishan I-200
Kang, Sin-Jae I-361
Kang, Yong-hyeog II-1149
Karaboga, Dervis II-1042
Karim, Asim I-170
Kato, Nei II-252
Kellinis, Emmanouel II-589
Kim, Joo-Hong I-503
Kim, Byung Ki II-303, II-451
Kim, Chang Han II-1
Kim, Cheol-Ki I-985
Kim, Dong-Kyue II-1104
Kim, Doo-Hyun II-669
Kim, Gwanyeon II-439
Kim, HongGeun II-260
Kim, HoWon II-469, II-1104
Kim, Hyun-Mi II-623
Kim, Jangbok II-297
Kim, Jeong Hyun II-749
Kim, Jeong-Sik II-1140
Kim, Jiho II-439
Kim, Jin-Geol I-176
Kim, Jinhyung II-725
Kim, Jongho I-65
Kim, Jong-Wan I-361
Kim, Jung-Eun I-1082
Kim, Jung-Sun II-669
Kim, Kwang-Baek I-785, I-985
Kim, Kyoung-Ho II-1098
Kim, Min-Jeong II-1140
Kim, Myoung-Hee II-1140
Kim, Nam Chul I-799
Kim, Sang Hyun I-799
Kim, Sang-Kyoon II-635
Kim, Seong-Whan II-643, II-1086
Kim, Seong-Woo II-743
Kim, Shin Hyoung II-869
Kim, Sosun II-1
Kim, Sungshin I-785
Kim, Taehae II-635
Kim, Tae-Yong II-857
Kim, Weon-Goo I-95
Kim, Wookhyun I-25
Kim, Yong-Deak II-935
Kim, Young Dae II-303, II-451
Kim, Young-Tak II-743
Knapskog, Svein Johan II-388
Ko, Kwangsun II-1149
Ko, Sung-Jea II-1098
Kong, Jung-Shik I-176
Kong, Min I-682, I-1003
Koo, Han-Suh II-845
Kraipeerapun, Pawalai II-1074
Kudo, Daisuke II-252
Kumar, Rajeev I-57
Kwak, Jong In I-799
Kwon, Goo-Rak II-1098
Kwon, Ki-Ryong II-581
Kwon, Taeck-Geun II-220
Kwon, Taekyoung II-427

Lai, Jianhuang I-933
Lee, Bo-Hee I-176
Lee, Byungil II-469
Lee, Chong Ho I-216
Lee, Deok-Gyu I-313
Lee, Dong Hoon II-143
Lee, Dong Wook II-749
Lee, Gun-Woo II-911
Lee, Jae-Wan I-337, I-1082
Lee, Jeongjun I-127
Lee, Jinseok II-1149
Lee, JungSan II-538
Lee, Kang-Woong II-899
Lee, Kuhn-Il II-911
Lee, Kwnag-Jae II-669
Lee, Kyoung-Mi II-832
Lee, Mun-Kyu II-1104
Lee, Sang-Gul I-985
Lee, Sang-Jin I-313
Lee, Sang Jun II-303, II-451
Lee, Seon-Gu I-544
Lee, Seung Phil II-869
Lee, Seungwon II-125
Lee, Songwook I-574
Lee, Soon-tak I-729
Lee, Tai Sik II-749

Lee, Tea-Young II-1098
Lee, Wonchang I-127
Lee, Woobeom I-25
Lee, Yong-min I-216
Lee, Yoon Cheol II-33
Lee, Youngseok II-220
Lei, Qiang II-1080
Leung, Ho-fung II-1030
Li, Bicheng II-764
Li, Bo II-291
Li, Changjun II-655
Li, Chunping I-713
Li, Chunzhong I-592
Li, Chuyu I-721
Li, Congcong II-772
Li, Fengxiang I-1045
Li, Fuhai II-812
Li, Gang I-405
Li, Haijun I-145
Li, Haiquan I-405
Li, Hong I-247
Li, Hongwei II-941
Li, HuiXian II-421
Li, Jianhua II-526
Li, Jianqing II-1110
Li, Jie I-696
Li, Jingtao II-1017
Li, Kenli II-463
Li, Lingjuan II-334
Li, Liping II-381
Li, Ming II-267
Li, Minglu II-679
Li, Ning I-638
Li, Qianmu II-542
Li, Qunxia I-979
Li, Renfa II-463
Li, Rui II-941
Li, Wei II-291
Li, Wenhui I-483
Li, Wenjie I-878
Li, Xiang II-173
Li, Xiangxue II-104
Li, Xing II-518
Li, Xinghua II-356
Li, Yajing II-74
Li, Yao II-309
Li, Yushan I-1009
Li, Zhanchun II-309
Li, Zhitang II-309, II-611
Li, Zhiwu I-1094
Liang, Xiaoping II-573
Liao, Jian II-161
Liao, Z.W. II-917
Lim, Chan II-857
Lim, Jongin II-1, II-143
Lim, Sukhyun I-827
Lin, Yongmin I-741
Lin, Chiachen I-464, II-567
Lin, Chih-Sheng I-512
Lin, Dongdai II-131, II-375, II-1036
Lin, Ling I-489
Lin, Mengquan II-526
Lin, Shukuan I-520
Lin, Xuegang II-407
Ling, David Ngo Chek II-788
Linshu, He I-528
Liotta, Antonio I-489
Liu, Bin II-309, II-611
Liu, Bo I-630
Liu, Chi II-703
Liu, Dalian I-184
Liu, Fang I-793, I-852
Liu, Fenglei I-410
Liu, Fengyu II-542
Liu, Gang I-416, I-979
Liu, Jinfeng I-688
Liu, Jing I-238
Liu, Jingmei II-49
Liu, Keqin II-711
Liu, Liangjiang I-609
Liu, Pengfei I-133
Liu, Qin II-695
Liu, Sen I-1015
Liu, Shengli II-117
Liu, Shizhu I-442
Liu, Weiping II-457
Liu, Wu II-482, II-518
Liu, Ying I-902
Liu, Yu I-957
Liu, Zaiqiang II-995
Lu, Bin I-846
Lu, Jiang I-957
Lu, Jiangang I-921
Lu, Zhe-Ming II-820
Luan, Shangmin I-355
Luo, Bin I-1003
Luo, Xiaohua II-687
Luo, Yunlun II-43
Lv, Kewei II-96
Lv, Lintao I-747

Lv, Shaohe II-703
Lv, Tianyang I-771

Ma, Jianfeng II-228, II-356, II-476, II-488
Ma, Jinwen I-915, II-812
Ma, Jun I-101
Ma, Linru II-328
Ma, Lizhuang II-826
Ma, Long II-655
Ma, Rui I-1033
Ma, Wenping I-793
Ma, Xiaoqi I-381, I-422
Maik, Vivek II-929
Marias, Giannis F. II-589
Mateo, Romeo Mark A. I-337
Meng, Deyu I-139
Meng, Hongyun I-839
Meng, Kai II-772
Meng, Zhiqing I-267, I-568
Miao, Chunyan I-375
Miao, Yuan I-375
Min, Yao I-704
Ming, Hua I-821
Mitchell, Chris J. II-149, II-192
Moe, Marie Elisabeth Gaup II-388
Mogami, Yoshio I-115
Moon, Daesung II-635
Moon, Kiyoung II-635
Moon, SangJae II-356, II-488
Moraga, Claudio I-49
Morii, Masakatu II-17
Mu, Chengpo I-9

Nah, Won II-899
Ñanculef, Ricardo I-49
Naqvi, Syed II-348
Nemoto, Yoshiaki II-252
Ni, Zhonghua I-870
Nin, Jordi I-1
Ning, Yufu I-230
Nishiura, Yasumasa I-669

Oh, Sang-Hun II-1025
Oh, Ha Ryoung I-1064
Ohigashi, Toshihiro II-17
Ohn, Jungho II-780
Okamoto, Eiji II-198
Okamoto, Takeshi II-198
Olmos, Ivan I-41
Olsson, Roland I-17
Omran, Mahamed G.H. I-192
Osorio, Mauricio I-41

Pai, Ping-Feng I-512
Paik, Joonki II-929
Pan, Rihong II-629
Pan, Xiaoying I-852
Pan, Yun I-285
Pan, Yunhe I-752, II-687
Pancake, Cherri M. II-548
Panda, G. I-163
Pang, LiaoJun II-421
Pang, Sulin I-1027
Pang, Ying-Han II-788
Panigrahi, B.K. I-163
Papapanagiotou, Konstantinos II-589
Papliński, Andrew P. I-81
Park, Jong-Hyuk I-313
Park, Seon-Ki I-503
Park, Chang Won II-494
Park, Hye-Ung I-313
Park, Ji-Hyung I-450, II-204
Park, Jooyoung I-65
Park, Kyu-Sik II-1025
Park, Sang-ho II-427
Park, Sehyun II-439
Park, Young-Ho II-1
Park, Young-Ran II-581
Patnaik, Lalit M. I-562
Pei, Hui II-820
Peng, Changgen II-173
Peng, Jing I-669
Peng, Lifang I-568
Peng, Qinke II-315
Peters, Terry I-884
Poon, Ting-C II-780
Potdar, Vidyasagar II-273
Puhan, Niladri B. II-661
Purevjii, Bat-Odon II-548

Qi, Yaxuan I-1033
Qi, Yinghao II-161
Qi, Zhiquan I-580
Qian, Haifeng II-104
Qian, Yuntao II-757
Qin, Zheng I-957
Qing, Sihan II-381
Qu, Youli I-741

Rayward-Smith, V.J. II-1011
Riguidel, Michel II-348
Rodríguez-Henríquez, Francisco I-208
Rong, Gang I-688
Rong, Mentian II-161
Rong, Xiaofeng II-398
Ruland, Christoph II-86
Ryu, Jeha I-949
Ryu, Keun Ho I-721

Sachidananda, Saraswathi II-236
Sallhammar, Karin II-388
Salman, Ayed I-192
Samantaray, S.R. I-163
Sanza, Cédric I-329
Savola, Reijo II-508
Saxena, Ashutosh II-110
Sbert, Mateu II-989
Schmucker, Martin II-80
Seo, Hee Suk II-340
Seo, Jungyun I-574
Seo, Kang-hee I-827
Seo, Seong Chae II-303, II-451
Seo, Sungbo I-721
Seol, Jae-Min II-1086
Seong, Yeong Rak I-1064
Shang, Ronghua I-846
Shang, Wenqian I-741
Shankarnarayanan, Kartik I-1021
Sharath, S. I-562
Sharma, A.K. I-367
She, Yanjie I-483
Shen, Haifeng I-979
Shen, Jianjun II-381
Shen, Lincheng I-399
Shen, Qingni II-381
Shen, Zhiqi I-375
Shen, Zhong I-864
Shi, Feng I-604
Shi, Guangming II-851
Shi, Xiangdong II-27
Shi, Yi II-364
Shim, Jaehong II-297
Shim, Kwang-Hyun I-1064
Shin, Byeong-Seok I-827, II-1127
Shin, Jeongho II-929
Shin, Sang-Uk II-581
Shiraishi, Yoshiaki II-17
Sim, Kwee-Bo I-428
Skinner, Geoff II-55
Song, Hee-Jun I-503, I-544
Song, Il-Seop II-220
Song, Li II-983
Song, Ohyoung II-439
Song, Shaoxu I-713
Song, Shuni II-655
Song, Wenchao I-663
Srinivasa, K.G. I-562
Su, Guangda II-772
Su, Han I-1039
Su, Kaile I-349
Su, Ruidan II-398
Su, Yong I-416
Suk, Jung-Youp II-911
Sun, Baolin I-101
Sun, Fan I-536
Sun, Jiaguang I-476, II-959
Sun, Linyan I-405
Sun, Maosong I-536
Sun, Ninghui II-322
Sun, Suqin I-965
Sun, Xi II-500
Sun, Xiaojuan II-322
Sun, Youxian I-921
Sun, Yu II-1048
Sun, Yuxiang II-155
Sun, Zengqi I-255
Sung, Hyun-Sung II-643

Takagi, Tsuyoshi II-198
Taleb-Bendiab, A. I-1074
Tang, Qiang II-149, II-192
Tang, Shaohua II-186
Tang, Shugang I-878
Tang, Wansheng I-230
Tang, Wenyu II-334
Tang, Yuan Yan I-833, II-1060
Tao, Xiaoyan I-556
Thapa, Devinder I-107
Tian, Fengzhan I-33
Tian, Haibo II-500
Tian, Jing I-399
Tian, Junfeng II-244
Tian, Peng I-682
Tian, Shengfeng I-9, I-33
Tian, Yingjie I-580, I-586
Tikkanen, Antti II-1054
Ting, Shen II-180
Torra, Vicenç I-1
Tran, Trung Hau I-329

Triet, Tran Minh II-605
Tu, Li I-261

Urmanov, Aleksey I-991

Valle, Carlos I-49
Varadarajan, Sridhar II-236
Varonen, Rauno II-508
Venugopal, K.R. I-562
Virtanen, Teemupekka II-1054

Waizumi, Yuji II-252
Wan, Shuai I-1051, II-597
Wan, Yuehua II-977
Wang, Zhihai I-741
Wang, BaoBao I-343
Wang, Ben I-495
Wang, Changguang II-476
Wang, Changjie II-1030
Wang, Dayin II-375
Wang, Fangwei II-476
Wang, Fasong II-941
Wang, Feng I-73
Wang, Gi-Nam I-107
Wang, Guoyuo II-893
Wang, Hai I-765
Wang, Hongan I-1100
Wang, Hong F. I-224
Wang, Hui I-1100
Wang, Jianxin II-328
Wang, Jiaxin I-1033
Wang, Jin I-216
Wang, Jinlong I-752
Wang, Junping II-796
Wang, Lei II-959
Wang, Licheng I-285, II-104
Wang, Lihua II-198
Wang, Limin I-145
Wang, Ling I-630, I-909
Wang, Long I-422
Wang, Ping II-212
Wang, Qiangmin II-526
Wang, Qin II-167
Wang, Rangding II-935
Wang, Ruchuan II-334
Wang, Shaomei I-638
Wang, Shoujue II-1068, II-1110
Wang, Sun'an II-923
Wang, Wei I-765
Wang, Weidong I-1094
Wang, Weixing II-887
Wang, Wenjia I-809
Wang, Wentao II-893
Wang, Xiangyang II-617
Wang, Xiaomo I-483
Wang, Xinmei II-49
Wang, Yaonan I-609
Wang, Ye I-343
Wang, Yingluo I-291
Wang, Youqing II-1003
Wang, Yuexuan I-870
Wang, Yumin II-500, II-1080
Wang, Yuping I-184, I-247, I-297
Wang, Zhanquan I-758
Wang, Zhengxuan I-771
Wang, Zhiling II-137
Wang, Zhiquan II-881
Wang, Zhurong I-747
Wang, Zongjiang I-151
Wei, Jinwu II-971
Wei, Ping I-815
Wei, Xiaopeng I-896
Wei, Zhiqiang I-279
Wen, Fengtong II-62
Wen, Qiaoyan II-62
Wenjun, Zhang II-983
Woo, Woontack I-949
Wu, Chanle II-695
Wu, Chen II-273
Wu, Jiangxing II-971
Wu, Jianping II-482, II-518
Wu, Lijun I-349
Wu, Mingni I-464
Wu, Wenling II-62, II-375
Wu, Xiaoyun II-573
Wu, Yanling I-921
Wu, Zhaohui II-687

Xiang, Youjun II-953
Xie, Bo I-941
Xie, Shengli II-953
Xie, Shu-cui II-532
Xie, Xuemei II-851
Xie, Yi II-977
Xin, Zhiyun I-476
Xiong, Hongkai II-983
Xixiang, Lv II-1092
Xu, Benlian II-881
Xu, Chen I-139, II-649
Xu, Chengxian I-675, I-815, I-1088

Xu, Congfu I-752
Xu, Jianfeng I-884, I-890
Xu, Jin An I-434
Xu, Manwu II-542
Xu, Qijian II-1048
Xu, Qing II-989
Xu, Rongsheng II-407
Xu, Wenli II-863
Xu, Xiaofei I-157, I-735
Xu, Yong II-155
Xu, Zhiliang II-953
Xu, Zongben I-777
Xue, Rui II-1036

Yan, Wei II-279
Yang, Bo II-137, II-1092
Yang, Chan-Yun I-550
Yang, Fuzheng I-1051, II-597
Yang, Jiehui II-27
Yang, Jr-Syu I-550
Yang, Lin II-328
Yang, Ping I-598
Yang, Qiu I-101
Yang, Tao II-463
Yang, Xinyu II-364
Yang, Yongdong II-1122
Yang, Yueting I-675
Yang, Zhi II-560
Yao, Li II-826
Yao, Tingting II-826
Yao, Zhiqiang II-629
Ye, Feng I-121
Ye, Qingtai II-804
Ye, Xien II-935
Yeh, Junbin II-567
Yeom, Ki-Won I-450, II-204
Yi, Hong I-870
Yi, Hong Yan II-1011
Yin, Bo I-279
Yin, Hujun II-764
Yin, Jianping II-703
Yin, Jianwei II-965
Yin, Qian II-43
Yoon, Han-Ul I-428
You, Jin Ho II-303, II-451
You, Xinge I-833, II-1060
You, Xinhua II-1060
Youn, Hee Yong II-494
Yu, Hang I-858
Yu, Jian I-33
Yu, Mei II-935
Yu, Mengjie I-1074
Yu, Peifei I-921
Yu, Shaoquan II-941
Yu, Shengsheng I-416
Yu, Xinbo I-1045
Yu, Zhenhua I-410
Yu, Zhenwei I-285
Yuan, Dongfeng II-1116
Yuan, Qiaoling II-977
Yuan, Xiaohui I-669
Yuan, Yubo I-592
Yuen, Pong C. I-933
Yun, Jaeseok I-949
Yun, Xiaochun II-212, II-415

Zeng, Yong II-68, II-228
Zengke, Zhang I-616
Zhan, Daqi I-965
Zhanchun, Li II-611
Zhang, Bin I-858
Zhang, Changshui I-1033
Zhang, Chenbin I-393
Zhang, Cheng II-315
Zhang, Dan II-1060
Zhang, Dongdong II-983
Zhang, Fan II-433, II-488
Zhang, Fangguo II-117
Zhang, Fushun I-651
Zhang, Gendu II-1017
Zhang, Hong II-482, II-542
Zhang, Hongli II-560
Zhang, Ji I-765
Zhang, Jian I-713
Zhang, Jian-zhong II-532
Zhang, Jie II-1048
Zhang, Jinhua II-923
Zhang, Li II-977
Zhang, Ling II-905
Zhang, Ming I-458
Zhang, Naitong II-1048
Zhang, Pengxiang I-1106
Zhang, Qiang I-896
Zhang, Rui I-777, II-989
Zhang, Shaomin I-520
Zhang, S.S. I-151
Zhang, Tongqi II-1134
Zhang, Weiguo I-291
Zhang, Weiqiang II-649
Zhang, Weizhe II-560

Zhang, Xian-Da II-905
Zhang, Xiaohua I-839
Zhang, Xiaoshuan I-598
Zhang, Xin I-864
Zhang, Xinhong II-433
Zhang, Xizhe I-771
Zhang, Yajuan II-9
Zhang, Yanning II-875
Zhang, Ying II-826
Zhang, Yongzheng II-415
Zhang, Yunkai II-476
Zhang, Yuqing II-137
Zhao, Dexin I-878
Zhao, Fengji I-815
Zhao, Hong II-617
Zhao, Jizhong I-476
Zhao, Jun I-273
Zhao, Liyuan II-711
Zhao, Rongchun II-875
Zhao, Shuguang I-273
Zhao, Wei II-267
Zhao, Weidong II-244
Zhao, Wencang I-1045
Zhen, Jina II-838
Zheng, Jin II-291
Zheng, Jun II-711
Zheng, Kougen II-687
Zheng, Mao I-303
Zheng, Qiuhua II-757
Zheng, Yanxing I-399
Zhi, Lijia I-520
Zhong, Hua I-89
Zhong, Shaochun I-381, I-422
Zhong, Weicai I-238
Zhou, Bai-Tao II-899
Zhou, Bing II-291
Zhou, Dongdai I-381
Zhou, Donghua II-1003
Zhou, Dongsheng I-896
Zhou, Gengui I-121, I-267, I-568
Zhou, Hongfang I-747
Zhou, Jingli I-416
Zhou, Jun II-772
Zhou, Lihua II-398
Zhou, Shude I-255
Zhou, Sujing II-131
Zhou, Xinliang II-322
Zhou, Xuebing II-80
Zhou, Yong II-804
Zhou, Yuan II-445
Zhu, Chuanchao I-1088
Zhu, Haibin I-741
Zhu, Huayong I-399
Zhu, Miaoliang II-407
Zhu, Yihua I-568
Zhu, Yuefei II-9
Zhuang, Chuanli I-598
Zhuang, Jian II-923
Zhuang, Xiahai I-884, I-890
Zhuang, Xiaodong I-279
Zou, Beiji II-959
Zou, Boxian II-629
Zou, Zhenyu I-1106
Zuo, Wanli I-771
Zuo, Wuheng II-965